New models and algorithms for the identification of sncRNA-(snc)RNA interactions intra and across-species/kingdoms

Nicolas Homberg

N° d'ordre NNT :

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON
Opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale ED341
Evolution, Ecosystèmes, Microbiologie, Modélisation

Spécialité de doctorat : Bioinformatique

Soutenue publiquement le 15/06/2023, par :
Nicolas Homberg

New models and algorithms for the identification of sncRNA-(snc)RNA interactions intra and across-species/kingdoms

---

Devant le jury composé de :

Etienne Birmelé, Professeur, Université de Strasbourg	Rapporteur
Alain Denise, Professeur, Université Paris-Saclay	Rapporteur
Christine Gaspin, Directrice de recherche, INRAe	Co-directrice de thèse
Thierry Lecroq, Professeur, Université de Rouen	Examinateur
Yves Quentin, Chargé de recherche CNRS, Université Paul Sabatier Toulouse III	Examinateur
Marie-France Sagot, Directrice de recherche, Inria	Co-directrice de thèse
Cristina Vieira-Heddi, Professeure, Université Lyon 1	Examinatrice

Laboratoire de Biométrie et de
Biologie Évolutive
Université Claude Bernard Lyon 1
Bâtiment Grégor Mendel
43 boulevard du 11 novembre 1918
69622 VILLEURBANNE

École doctorale Évolution Écosystèmes
Microbiologie Modélisation
43 boulevard du 11 novembre 1918
69622 VILLEURBANNE

Resumé en français

Préambule

Ce document résume le manuscrit qui décrit le travail réalisé pendant la thèse. Le manuscrit est organisé en 7 chapitres (A à F). Une introduction présente la question adressée dans ce travail de thèse (Chapitre A) : Nouveaux modèles et algorithmes pour l'identification d'interactions ARN-ARN intra/inter espèces. Un état de l'art est ensuite présenté concernant les approches expérimentales et *in silico* qui ont servi de base pour la suite (Chapitre B). Nous avons suite à cela identifié deux critères présents dans une majorité d'approches, l'accessibilité et la « seed », considérés comme essentiels pour prédire des interactions miARN-mARN. Nous avons aussi sélectionné une approche expérimentale qui nous ont permis de constituer plusieurs jeux de données pour deux espèces différentes (chapitre C). Trois approches à l'état de l'art ont été utilisées pour évaluer l'importance des critères « accessibilité » et « seed » (Chapitre D). L'étude de la conservation de motifs intra-espèces en tant que nouveau critère pour améliorer la prédiction *in silico* des interactions miARN-ARNm est étudiée dans le Chapitre E. Enfin, une dernière contribution de ce travail (chapitre F) est le développement d'une interface web permettant de faciliter l'utilisation de l'outil Smile. Le travail réalisé a donné lieu à un premier article accepté en 2023 dans le journal *Genes*.

Chapitre A : Introduction

Dans ce premier chapitre introductif, nous rappelons quelques éléments de vocabulaire permettant de définir les « objets » et les propriétés qui seront manipulés tout au long de la thèse. Nous présentons ensuite la problématique initialement adressée dans ce travail qui concernait le développement de nouveaux modèles et algorithmes pour l'identification d'interactions ARN-ARN intra/inter espèces.

Les ARN non codants (ARNnc) sont des ARN issus d'unités de transcription qui leur sont propres, mais qui peuvent aussi être générés par un processus de maturation à partir de transcrits primaires. Ils ont la capacité à se replier sur eux-mêmes en formant des interactions qui permettent à la molécule d'adopter une structure stable et caractéristique de sa fonction. Il existe aujourd'hui une classification de ces molécules en deux grandes familles distinguant les petits ARNnc de taille inférieure à 200nt, et les longs ARNnc de taille supérieure. De nombreux ARNnc, dits régulateurs, ont la capacité d'interagir avec

d'autres molécules d'ARN, notamment en formant des appariements entre nucléotides (G avec C, A avec U) que nous désignerons par la suite comme des interactions ARN-ARN.

Chez les eucaryotes, l'importance du rôle du transcriptome non codant est reconnue dans un grand nombre de processus essentiels tels que la différentiation ou le développement. Il a aussi été démontré que les ARNnc peuvent être transportés vers des cellules ou tissus distants et y exercer leur fonction. Les relations inter-espèces sont basées sur des inter-connexions complexes entre les cellules des organismes impliqués. La littérature récente fait état du rôle « clé » joué par les ARNnc dans ces relations complexes. Parmi les ARNnc, les miARN représentent une famille des plus étudiées. Ils sont de très petite taille (21nt) et issus de transcrits primaires de taille pouvant atteindre plusieurs kb. Dans la plupart des organismes, ils sont essentiellement étudiés dans un contexte de régulation négative de l'expression des gènes, par la mise en place d'interactions entre miARN et ARN messagers (ARNm). Leur rôle dans les relations hôtes-pathogènes a pu être mis en évidence, par exemple dans les mécanismes de défense des plantes contre l'infection par des champignons et dans la protection de la barrière intestinale via leur ingestion dans le microbiote de l'hôte.

Les nature(s) et mécanisme(s) d'interactions mis en place dans les relations inter-espèces médiées par les miARNs étant peu caractérisés, nous avons investi dans nos travaux de thèse la question d'une meilleure caractérisation de ces interactions en adressant les interactions entre miARN et ARN messager (miARN-ARNm) intra-espèce sous l'hypothèse que les interactions miARN-mARN inter-espèces sont de même nature et que les miARN y jouent un rôle essentiel.

Chapitre B : Etat de l'art

Nous réalisons dans ce chapitre une étude bibliographique présentant quelques approches expérimentales et *in silico* à l'état de l'art pour l'identification des interactions miARN-ARNm. Quatre sections constituent ce chapitre. La première couvre, à travers quelques outils de référence, les approches *in silico* d'intérêt pour la suite de ce travail. Elle précise notamment, pour chaque outil, les principaux critères implémentés (énergie de l'interaction, accessibilité de l'ARNm, appariements entre nucléotides, conservation inter-espèces de l'ARNm et du miARN,...). La deuxième section couvre plusieurs approches expérimentales et met en lumière les difficultés, les inconvénients et/ou intérêts que chacune présente pour l'identification d'interactions miARN-ARNm. Elle met en lumière les apports et faiblesses de la technique CLASH, dont les données issues de deux des études publiées ont fait l'objet de notre choix pour nos propres analyses. Dans une troisième section, nous passons en revue quelques bases de données de référence rassemblant les données de la littérature concernant les interactions miARN-ARNm. Enfin, la quatrième section présente deux approches d'inférence de motifs qui seront mobilisées pour explorer l'intérêt d'intégrer la sur-représentation de motifs comme évidence supplémentaire dans un contexte d'identification de cibles *in silico*.

Chapitre C : Elaboration de jeux de données pour l'exploration des interactions miARN-ARNm

Ce chapitre constitue une première contribution de la thèse. Il était essentiel, dès le début de ce travail de thèse, de disposer de jeux de données nous permettant d'avoir accès à des données « valides » donnant les positions précises des régions des miARN et ARNm en interaction. Sur la base de ressources de référence (miRBase pour les miARN, Ensembl pour les transcrits codants et génomes) et de résultats d'études s'appuyant sur la technique CLASH (Cross-linking Ligation And Sequencing of Hybrids), nous avons développé les scripts nous permettant d'extraire de manière automatique les séquences d'intérêt des régions des ARNm en interaction avec les miARN.

Deux études utilisant la technique CLASH ont été sélectionnées, l'une chez l'humain et l'autre chez la souris. Chacune d'entre elles a permis de sélectionner plusieurs milliers de régions en interaction mobilisant quelques centaines de miARN et quelques milliers d'ARNm. Afin d'évaluer le « bruit » induit par la longueur des ARNm lors d'une prédiction d'interaction *in silico*, 3 sous-ensembles d'ARNm ont été construits : le premier, appelé SeqE, contenant uniquement la région en interaction, le deuxième, appelé SeqW, contenant, pour chaque ARNm, 200 nt en amont et en aval de l'interaction, le troisième, appelé SeqC, intégrant l'ARNm complet.

L'observation inattendue que, dans chacune de ces études, plusieurs modalités d'interactions co-existaient avec le modèle canonique d'interaction (« seed » du miARN avec région 3'UTR de l'ARNm), nous a conduit à générer plusieurs jeux de données permettant d'analyser plus finement l'ensemble des modalités d'interactions en tenant compte des régions mobilisées sur le miARN (internes ou externes à la « seed », 5 à 6 classes) et sur l'ARNm (régions 5'UTR, codantes, 3'UTR, 3 localisations sur l'ARNm).

Chapitre D : Exploration des critères d'accessibilité et de « graine » dans le cadre des interactions miARN-ARNm intra-espèces

La contribution principale de ce chapitre porte sur l'évaluation de l'importance de deux critères présents dans une majorité d'approches de prédiction *in silico* des interactions miARN-ARNm. Le premier critère vise à évaluer l'accessibilité d'une région, c'est-à-dire sa capacité à interagir avec un autre ARN. Le second critère, appelé « seed », définit la sous-séquence du miARN composée des nucléotides 2 à 7. Cette région est considérée comme devant être appariée et comme essentielle dans les interactions miARN-mARN. Le chapitre est organisé en 3 sections présentant respectivement la stratégie utilisée, les résultats et une discussion.

L'importance du caractère accessible du site d'interaction dans l'ARNm est un critère qui reste encore aujourd'hui débattu dans la communauté scientifique. Dans ce chapitre, nous proposons une analyse détaillée de ce critère à la lumière de données fournies par les deux études expérimentales présentées dans le chapitre précédent. Trois méthodes implémentées dans les outils Pita, miRanda et IntaRNA ont servi de support à cette étude. Deux de ces méthodes, miRanda et Pita ont été développées pour identifier

de manière spécifique les interactions miARN-ARNm. Elles encodent la contrainte de l'existence d'une « seed » dans le miARN, aux positions 2 à 6, pouvant former un duplex avec l'ARNm. L'une, Pita, encode le critère d'accessibilité alors que l'autre, miRanda, ne considère pas ce critère. La troisième méthode, IntaRNA, est plus générique et présente l'avantage de pouvoir considérer ou non ces deux critères (optionnels) dans la prédiction des interactions. Chacune de ces méthodes a été mise en œuvre sur les jeux de données obtenus de CLASH en considérant deux types de vrais positifs selon que les interactions prédites recouvrent complètement (Tpstrong) ou pas (TPweak) les régions miARN et ARNm en interaction identifiées dans les études CLASH, un type de faux positifs (FP) intégrant les prédictions non recouvrantes sur l'un au moins des miARN ou ARNm, deux types de faux négatifs (FN) selon que les interactions prédites excluent complètement (FNstrong) ou pas (FNweak) les régions miARN et ARNm en interaction identifiées dans les études CLASH.

Nous avons utilisé pour cette étude des métriques d'évaluation classiques calculées à partir des nombres de vrais positifs, faux positifs, vrais négatifs et faux négatifs : « recall » ou rappel, précision et F-score. Les méthodes intégrant le critère d'accessibilité produisant de meilleurs résultats, nous avons décidé d'explorer les deux critères en les relâchant dans l'outil IntaRNA. De manière inattendue, nous avons pu mettre en évidence une amélioration des résultats lorsque le critère de la « seed » était relâché. Observant une dégradation globale des résultats lorsque le critère d'accessibilité était relâché, nous avons réalisé une étude ciblant cet unique critère en utilisant les outils RNAfold et RNAplfold. Les détails par classe et par localisation sont détaillés dans le manuscrit. Un des résultats de cette dernière étude met en évidence le caractère plus accessible des régions 3'UTR.

Chapitre E : La conservation intra-espèce comme critère d'amélioration de l'identification de cibles de miARN

Partant de l'observation qu'un miARN peut interagir avec de nombreux ARNm et qu'un ARNm peut être ciblé par plusieurs miARN, cette contribution a porté sur l'évaluation de la conservation de motifs intra-espèce en tant que critère pouvant améliorer la capacité prédictive des logiciels de prédiction d'interactions miARN-ARNm. Cette hypothèse n'est pas complètement nouvelle mais, jusque là, la conservation de motifs intra-espèce utilisée comme critère permettant d'améliorer l'identification de cibles de miARN a essentiellement été étudiée en exploitant la conservation de motifs localisés au sein des régions 2 à 7 des miARN. Dans le cadre de ce chapitre, sur la base des données apportées par les études utilisant la technique CLASH, nous avons exploré le cadre plus large de la conservation de motifs au sein des ARNm.

Nous avons d'abord exploré la sur-représentation de motifs de longueur 7 (nombre d'occurrences supérieur ou égal à 10). Le Z-score a été utilisé pour caractériser le caractère aléatoire ou non des observations réalisées. De manière intéressante, deux signaux opposés (un pic négatif suivi d'un pic positif) apparaissent à la localisation de l'interaction sur l'ARNm. L'étude de la composition en nucléotides des régions par classe et par localisation permet notamment de mettre en évidence un biais général en di-nucléotides qui diffère entre données humaines et de souris au début des sites d'interaction mais elle confirme

un biais en dinucléotides AT aux extrémités des sites en interaction de l'ARNm lorsque ces motifs sont localisés dans les régions 3'UTR. Cependant, une analyse de conservation de motifs à l'échelle seule de l'ensemble des ARNm ne permet pas de mettre en évidence des motifs sur-représentés correspondant à des sites d'interactions tels qu'identifiés dans les données CLASH.

Dans une deuxième section, nous avons étudié la conservation de motifs intra-espèce avec les outils de la suite Meme en considérant cette fois l'ensemble des ARNm par miARN. Pour chaque miARN, nous avons créé des jeux de données composés par l'ensemble des sites d'interaction des ARNm ciblés par le miARN i. Comme attendu, les motifs conservés trouvés par Meme sont ceux des classes conservatrices de la « seed ».

Dans une troisième section, nous évaluons ce critère comme pouvant apporter un signal d'évidence supplémentaire permettant d'améliorer la prédiction *in silico* des interactions miARN-ARNm. L'hypothèse sous-jacente est que cette conservation à l'échelle de l'ARNm peut apporter une évidence pour la découverte de nouvelles cibles et de nouveaux miARNs. Les mêmes métriques que celles du chapitre précédent ont été utilisées (rappel, précision et F-score), pour chaque classe et localisation des sous-ensembles de séquences que nous avons établis. Bien qu'encourageants, les résultats obtenus ne permettent pas d'envisager l'utilisation seule de ce critère dans un contexte de prédiction *in silico* des interactions miARN-ARNm, même dans le cas où le miARN est connu.

Chapitre F : Web-service

La dernière contribution de cette thèse porte sur la réalisation d'une interface web permettant au plus grand nombre d'utiliser l'outil d'inférence de motifs, Smile, développé par le passé dans mon équipe d'accueil. Smile présente en effet les intérêts d'une approche exacte et, de manière originale, de pouvoir poser des contraintes permettant de préciser par exemple la composition souhaitée du motif recherché, que ce soit en termes de nucléotides ou de sous-motifs présents. Une première version est disponible à l'URL http ://134.214.213.44.

Conclusion et perspectives

Une première contribution de ce travail de thèse est la constitution de jeux de données permettant d'analyser l'importance de l'usage des critères pris en compte dans les outils de prédiction *in silico* des interactions miARN-ARNm. Ces jeux de données ont été sélectionnés pour la précision qu'ils apportent sur les régions en interaction, que ce soit sur le miARN ou bien sur l'ARNm. Notre objectif était d'évaluer l'importance de l'usage de deux d'entre eux, l'accessibilité et la « seed ». A travers la mise en œuvre de 3 méthodes à l'état de l'art, nous avons pu mettre en évidence l'importance de l'accessibilité pour chacun des organismes étudiés et, globalement, pour l'ensemble des classes et localisations. Nous avons aussi étudié la conservation de motifs comme nouveau critère à considérer. A ce stade de l'étude, il reste difficile d'évaluer sa pertinence. Sa seule utilisation ne permet pas de discriminer les régions d'intérêt mais il apparaît intéressant d'évaluer sa pertinence en le combinant avec les critères d'accessibilité et de conservation inter-espèces.

Contents

List of Figures

List of Tables

Glossary

Argonaute The Argonaute or AGO is a protein that links with the miRNA and forms the miRNA-induced silencing complex (miRISC). Its role is to cleave mRNAs, hence disabling them. 7, 12

complementary DNA Complementary DNA is DNA synthesised from a single-stranded RNA. 7, 43

consensus motif Consensus motifs or consensus logo is a sequence logo simplified to correspond to a text format. A consensus motif only displays the best letter for each position. 90

deoxyribonucleic acid DNA stands for deoxyribonucleic acid and is present in the cells. It serves the purpose of generating proteins. It contains the information that allows all living creatures to survive and reproduce. It is composed of 4 different nucleotides: Adenine (A), Thymine (T), Guanine (G), Cytosine (C). Adenine pairs with Thymine and Guanine with Cytosine. This pairing is the reason for the double helix shape of DNA. The same DNA is shared among all our cells, but thanks to gene transcription, each cell will produce different proteins that are appropriate to their biological context. 5, 7

locus A locus (plural loci) is a precise and fixed position on a chromosome. 9, 12

nucleotide Nucleotides consist of a nucleoside and a phosphate. They serve as monomeric units of the nucleic acid polymers — deoxyribonucleic acid (DNA) and Ribonucleic acid (RNA) – both of which are essential biomolecules within all life-forms on Earth. The diet provides nucleotides which are also synthesised by the liver from common nutrients. 9

promoter A promoter on DNA is the region where transcription is initiated. 12

Ribonucleic acid Ribonucleic acid (RNA) is a polymeric molecule that plays a vital role in many biological processes, including coding, decoding, regulation, and expression of genes. Along with DNA, RNA is one of the two types of nucleic acids, which are essential macromolecules required for all known forms of life. Additionally, nucleic acids, along with lipids, proteins, and carbohydrates, are one of the four primary macromolecules essential for the functioning of living organisms. 5, 7

ribosome The translational apparatus, composed of ribosomes and associated molecules, is responsible for biological protein synthesis or mRNA translation. Ribosomes, which are macromolecular machines found in all living cells, play a crucial role in this process by linking amino acids together in the order specified by the codons of mRNA molecules, ultimately forming polypeptide chains. These ribosomes consist of two main components: the small and large ribosomal subunits, each of which contains one or more ribosomal RNA (rRNA) molecules and many ribosomal proteins. 11

transcript A transcript is an RNA sequence originated from transcribed DNA. 13

Acronyms

AGO Argonaute. 12, 13

cDNA complementary DNA. 43

DNA deoxyribonucleic acid. 5, 9

KSHV Kaposi's sarcoma-associated herpesvirus. 42

MHV68 Murine gammaherpesvirus-68. 42

miRISC miRNA-induced silencing complex. 5, 12, 13

miRNA micro RNA. 11, 12, 19

mRNA messenger RNA. 9, 11, 12, 19

RNA Ribonucleic acid. 5, 9

sncRNA small non coding RNA. 11, 19

sRNA small RNA. 37

tRNA transfer RNA. 9

UTR untranslated region. 12

Chapter A

Introduction

All self-reproducing cellular organisms have DNA which is contained in most cells. The function of DNA is to store genetic information for every type of cell in every organ. In order to be used, the information has to be transcribed into RNA; this step is called transcription, see Figure A.1. DNA is made of nucleotides, each one corresponding to a different type of nitrogenous base (also called nucleobases) which together with deoxyribose and at least one phosphate group compose a nucleotide. There are four types of nitrogenous bases, and thus of nucleotides. The bases are Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). Adenine is complementary to Thymine and Guanine is complementary to Cytosine. During transcription, the RNA sequences are composed of the complementary nucleotides of the DNA. However, thymine is replaced by the nucleotide Uracil (U).

A.1 The RNA's hidden orchestra

Once transcribed into RNA, the information can be used directly or it can be translated into proteins which are composed of amino acids. The information in the RNA sequences is coded with a group of 3 nucleotides called codons. Each codon can either code for one amino acid or indicate the start or end of the translation process.

Every protein is composed of a sequence of amino acids. Proteins are the main actors for every function in a cell. RNA sequences that will be translated and that code for proteins are called messenger RNAs (messenger RNAs). They are transcribed from DNA loci called protein-coding genes.

The translation of an mRNA proceeds in three steps:

— A macromolecular complex called ribosome interacts with the mature mRNA.

— The ribosome decodes the information of the mRNA and successively catches the transfer RNAs that carry an amino acid corresponding to the mRNA's genetic code.

— The ribosome links the amino acid to the growing polypeptide chain until a stop codon is read on the mRNA.

Although the very same DNA is present in all cells, each cell can have diverse roles depending on the developmental time, the location in the organism, and the environment. This process is called differentiation. For instance, cells located in the brain might have

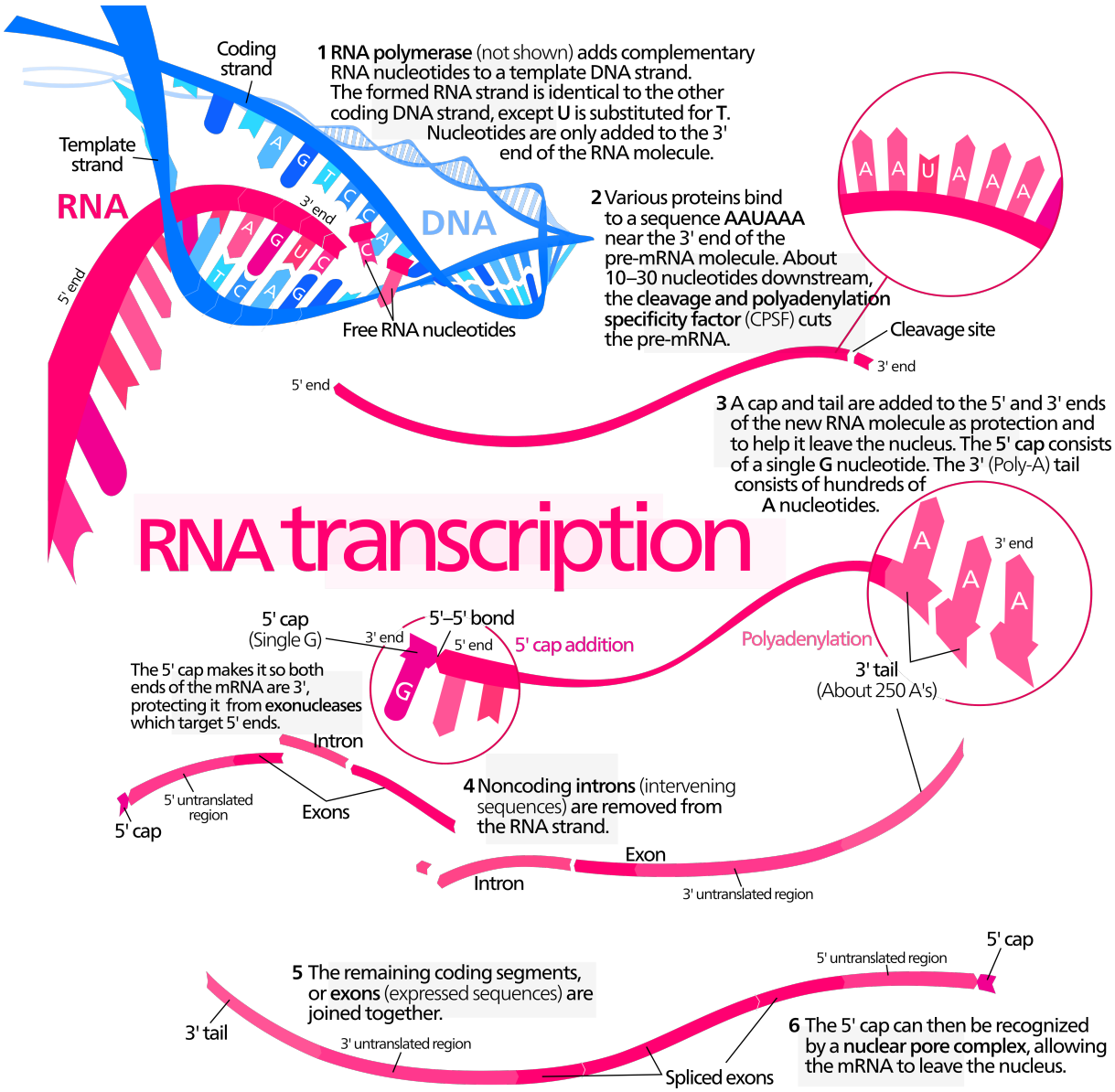


Figure A.1 – Diagram of the transcription and splicing processes. Source[By Kelvinsong - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=23086203]

a significantly different behaviour than cells located in the heart. This specificity is the result of several biological mechanisms. First, the DNA stored in a double helix shape has to be open in order to allow an enzyme called RNA polymerase to perform the transcription. The precise location of the opening varies depending on various mechanisms which remain not well understood but that can involve specific factors in different cell types.

After transcription, the RNA sequences go through several processes of maturation that rely also on the specific proteins present in the cell. One such process is called splicing. This removes the non-coding sequences called introns and leaves the coding sequences called exons. One gene can produce several mature mRNA molecules through what is called alternative splicing in which the exons may be joined in different combinations.

In order to better fit the need in protein of an organism, a last process of differentiation of mRNAs is the up- or down-regulation of their associated protein. One mechanism that triggers this regulation is the interaction between mRNA and RNA sequences called non coding RNAs.

The down-regulation of mRNAs triggered by ncRNAs results in a reduction of the amount of protein created from an mRNA. It involves different mechanisms such as preventing interaction with the ribosome or expediting the degradation of mRNAs [Bartel, 2009].

The up-regulation of mRNA triggered by ncRNAs is a relatively recent discovery [Vasudevan, 2012]. This results in an increase in the amount of protein translated from mRNAs. The current knowledge about the various complex mechanisms involved in the ability to activate the expression of mRNAs by miRNAs is that this is conducted by the same mechanisms involved in down-regulation but through different biological signals [Wilhelm and Smibert, 2005, Kedde and Agami, 2008].

A special class of ncRNAs is composed of small non coding RNAs (sncRNAs) which range in length from 18 to 200 nucleotides and which were suspected several decades ago to communicate between cells [Benner, 1988]. It was later discovered that they are able to move through body fluids such as blood, tears and maternal milk [Chen, 2012, Hoy and Buck, 2012, Taylor and Gercel-Taylor, 2013]. This is specially the case of one type of sncRNAs which is microRNA, micro RNA (miRNA) for short. Recent studies even found that sncRNAs may participate in across-species RNA-RNA interactions [Weiberg et al., 2013]. This is true of sncRNAs in fungi as indicated in [Weiberg et al., 2013] that may thus target eukaryotes, in the case tomato plant and *Arabidopsis thaliana*. Such RNA-RNA interactions have however also been more recently hypothesised in the other direction, that is from a eukaryotic host to fungi or to bacteria with which the host is in close contact [Cai et al., 2018, Liu et al., 2016].

Additionally, it is worth mentioning that miRNAs individually can also be regulated and there are multiple biological processes that regulate the biogenesis pathway of miRNAs [Ha and Kim, 2014].

A.2 MicroRNA

As mentioned above, miRNAs is a subclass of sncRNAs with an average length of 22 nucleotides, although they can range from 17 to 26 nucleotides. They are extensively studied due to their ability to regulate the expression of mRNAs. They are for instance

hypothesised to play a significant role in the treatment of diseases such as cancer. miRNAs exist only in eukaryotes, however something similar is found in prokaryotes [Bloch et al., 2017].

i History of miRNAs

The first miRNA, *lin-4*, was discovered in 1993 by the teams of Ambros and Ruvkun in the species *Caenorhabditis elegans* [Lee et al., 1993, Wightman et al., 1993]. *lin-4* was first thought, in 1980, to be a protein coding gene that regulates the development of the larvae [Horvitz and Sulston, 1980, Chalfie, 1981]. A few years later, in 1987, a gene called *lin-14*, was discovered which has a function opposed to the one of the gene *lin-4* [Ambros and Horvitz, 1987, Ferguson et al., 1987]. *lin-4* was later suggested to be a small non-coding RNA with a complementary sequence to *lin-14*, and that *lin-14* was down-regulated post-transcriptionally by an interaction in its 3'untranslated region (UTR) [Lee et al., 1993, Wightman et al., 1993].

A few years after the discovery of *lin-4*, other miRNAs such as *let-7* were discovered to regulate mRNAs in many different species [Pasquinelli et al., 2000]. Since this pioneer paper, the number of miRNAs discovered never stopped growing and the first database about miRNAs (which will be discussed in more detail in Section B.3) now contains 48885 unique miRNAs from 271 organisms with 1917 miRNAs in human.

Thanks to this ever growing research on miRNAs, a lot of knowledge has been gathered around the biological genesis of miRNAs, mRNA regulation and miRNA-mRNA interactions. Overall, the discovery of the miRNAs has provided a new and important perspective on gene regulation, and opened up new approaches for the development of therapies for a variety of diseases.

ii Biological genesis of miRNAs

Half of the miRNAs originate from gene loci and are mostly from introns with only a few from exons [Kim and Kim, 2007]. The other half comes from intergenic loci and have their own promoters [Kim and Kim, 2007, de Rie et al., 2017].

There are multiple possible biogenesis pathways for miRNAs. The canonical one is the most frequent (depicted in Figure A.2) and starts with the transcription into a hairpin-shaped RNA called primary-miRNA, or pri-miRNA for short.

The pri-miRNA is cleaved by a Drosha-DGCR8 complex, called microprocessor, at the base of the hairpin and forms a precursor-miRNA, pre-miRNA for short, which is exported from the nucleus by the Exportin-5/RanGDP [Guo, 2012]. The pre-miRNA, still with a hairpin shape, is composed of two complementary strands and a bulge which is cleaved by a Dicer protein leaving two bound small RNA sequences as a mature duplex [Zhang et al., 2004]. These two bound small RNAs are called 3p or 5p depending on which strand, 3' or 5', of the pre-miRNA they originated from, and are called mature miRNA. One of the 3p or 5p small RNAs will be loaded into an Argonaute (AGO)2 protein. The mature miRNA loaded onto the AGO protein are the main components of the miRISC, which initiate the regulation process through various mechanisms, such as mRNA cleavage by AGO2 or inhibition of translation by blocking the ribosome. The mechanism of regulation is determined by the degree of complementarity between the miRNA and mRNA. [Jo et al.,

2015]

The various biological pathways of miRNA influence their nomenclature. Therefore miRNAs generated from a longer transcript and miRNAs which share a similar seed or the same characteristics can be grouped in a same family. For instance, the *let-7* family contains 12 different miRNAs.

iii Potential roles of miRNAs in disease treatment

miRNAs have a key role in many processes and are recognised as important actors in diseases.

Some mutations or deletions of miRNAs can trigger diseases such as hereditary hearing loss [Mencía et al., 2009], hereditary cataract [Hughes et al., 2011] and skeletal and growth defects [de Pontual et al., 2011]. There is current research on using miRNAs in the pathogenesis, diagnostic and treatment of cancer [Kong et al., 2012, Uzuner et al., 2022]: low or high levels of some miRNAs can be a good indicator of the evolution of a disease, for instance either a high level of *mir-185* or a low level of *mir-133b* can indicate colorectal cancer metastasis [Akçakaya et al., 2011] or the presence of *miR-205* and *miR-373* can better characterise the type of cancer [Eyking et al., 2016] .

miRNAs are expected to play an important role in treating cancer efficiently without damaging other cells by silencing genes critical of tumor cell growth. However, due to the miRNAs short lifespan, the current biggest challenge is to efficiently deliver miRNAs to diseased tissues [Wang et al., 2011, Fu et al., 2019, Reda El Sayed et al., 2021].

miRNAs have been found to be also linked to kidney diseases [Phua et al., 2015], some functions in the nervous system [Maes et al., 2009], addiction and especially alcoholism [Lewohl et al., 2011], obesity [Romao et al., 2011], and the regulation of genes that benefit viruses [Qureshi et al., 2014].

A.3 Biological features of miRNA-mRNA interactions

A miRNA-mRNA interaction happens when a miRNA and an mRNA are paired to each other (see Figure A.3) via mostly the complementarity between the miRNA and the mRNA. We enumerate in this section the biological characteristics of miRNA-mRNA interactions. These features are also used in the miRNA-mRNA prediction tools. The various algorithms are described in Chapter B.

All mechanisms involved in the selection of the mRNA target by a miRISC are not yet fully discovered or understood. However, there is already quite some work on the subject in the literature. Which mRNA is targeted largely depends on the miRNA and more precisely on its sequence. Indeed, a good complementarity between mRNA and miRNA is mandatory and the degree of such complementarity determines if AGO2 cleaves the mRNA (near-perfect complementarity) or if the mRNA is degraded by the miRISC action [Jo et al., 2015]. However, several other characteristics have to be considered. For instance, *TargetScan* [Agarwal et al., 2015, Agarwal et al., 2018], a method that predicts miRNA targets, uses 26 features.

The main and most important features considered in the literature for identifying targets of a miRNA are the stability of the RNA-miRNA duplex, the seed matching,

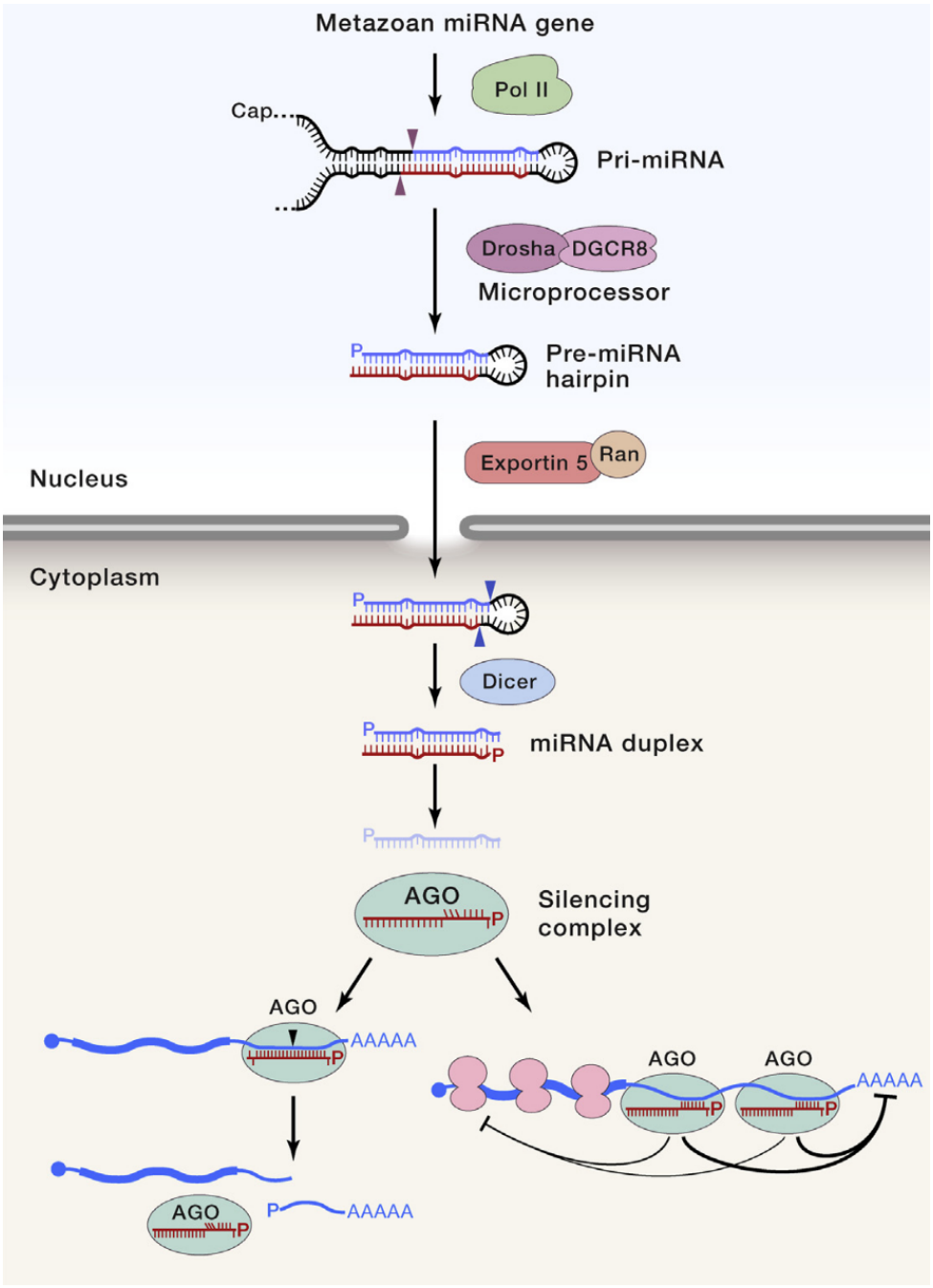


Figure A.2 – The canonical biological pathway of miRNAs. Source [Bartel, 2018]

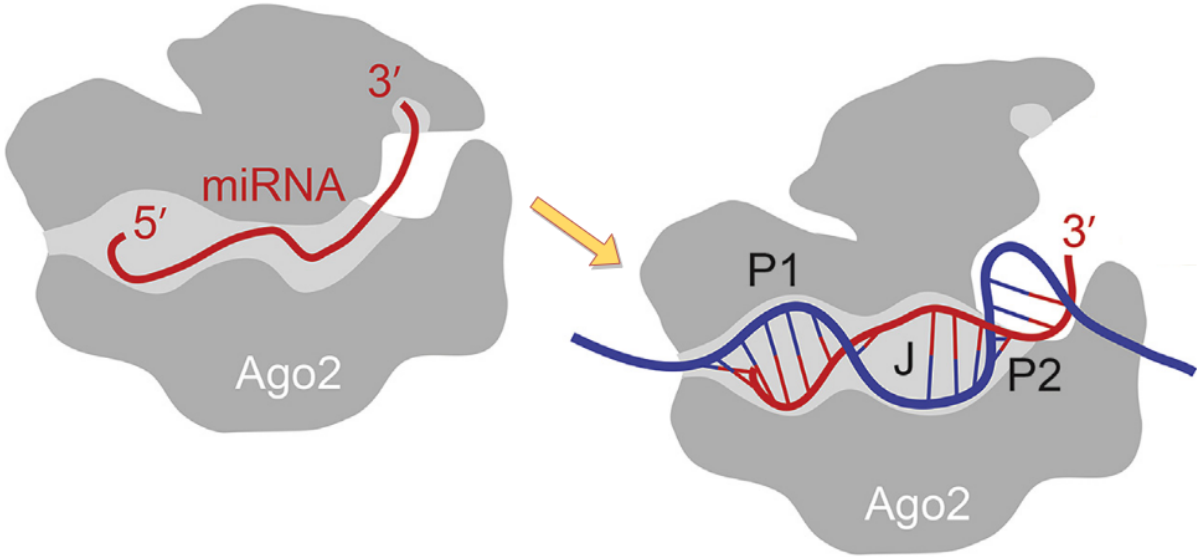


Figure A.3 – Interaction between one miRNA in red and one mRNA in blue. The interaction begins in the 5'end of the miRNA. Adapted from [Sheu-Gruttadauria et al., 2019]

the location of the target on the mRNA, the accessibility of one or both RNAs, and the conservation inter-species.

i Stability of the mRNA-miRNA duplex

The complementarity, and therefore the stability of an interaction between miRNA and mRNA can be measured with a thermodynamic model: the free energy (Gibbs free energy) value of the formed duplex. The free energy, historically called affinity, is a thermodynamic state function used to quantify by its sign the likelihood of whether a process is thermodynamically spontaneous or favourable (negative), or non-spontaneous (positive) (page 201 of the book [Atkins and Paula, 2010]). A non-spontaneous process, also called endergonic, requires external energy injected to this process and it will, therefore, never happen on its own. On the contrary, a spontaneous process, also called exorgonic, can occur on its own and will produce energy.

In the case of an RNA-RNA interaction, both RNA strands are stabilised together via hydrogen bonds induced by the complementarity between nucleotides. The free energy of RNA-RNA interactions, therefore thermodynamically favourable, is always negative. Moreover, the higher is such complementarity and stability, the smaller is the free energy resulting in a higher chance of interaction.

The free energy of the duplex is however not enough to understand all possible interactions: for instance, a perfectly complementary miRNA-mRNA couple may not interact because the mRNA is not present in sufficient quantity and may thus never encounter the miRNA, or else molecules with a lower complementarity may already be blocking the mRNA interaction site [Salmena et al., 2011].

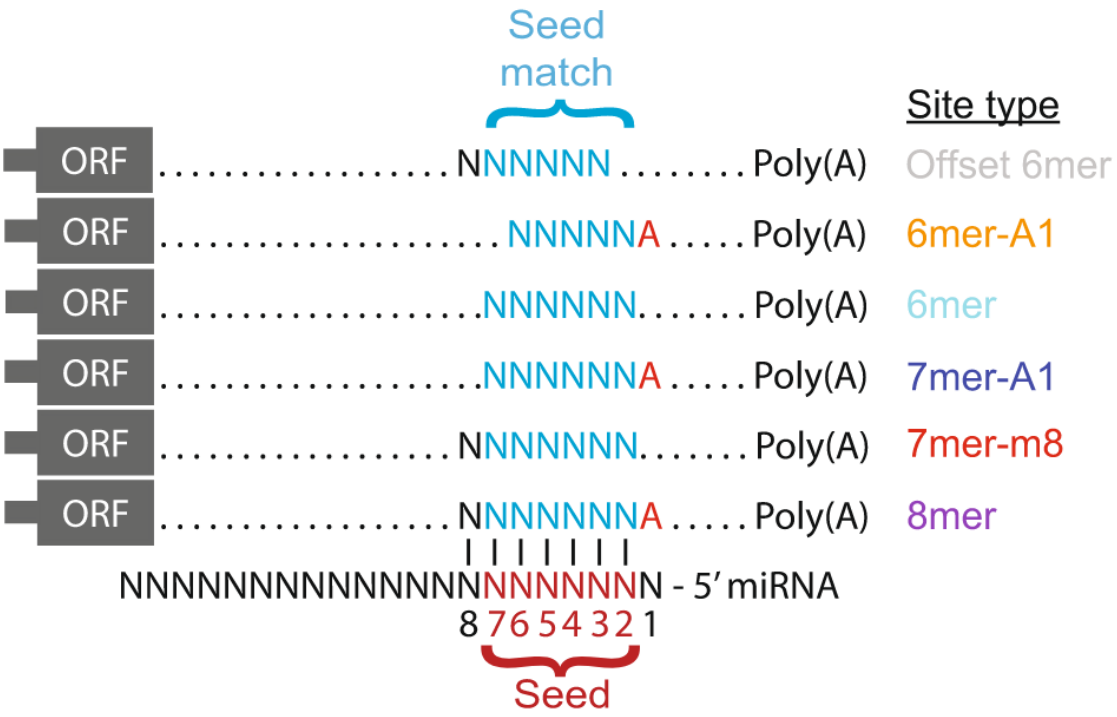


Figure A.4 – The six canonical site types of the seed in Mammalians and Drosophila. Adapted from [Agarwal et al., 2018]

ii Seed

The seed refers to a small portion of the miRNAs and is located within the nucleotides 2 to 8 from the 5' end.

The seed is found to serve as an anchor in 50% [Helwak et al., 2013] to 80% [Grosswendt, 2014] of the miRNA-mRNA interactions in mammalians [Lewis et al., 2005] and in insects [Stark et al., 2007]. It is also in many cases sufficient to significantly induce repression [Doench, 2004] and most of the known miRNA-mRNA interactions which lead to regulation involve the seed [Helwak et al., 2013, Moore et al., 2015].

The seed on the miRNA can interact with mRNAs according to different patterns (see Figure A.4. The nomenclature for these patterns is based on the length of the region of perfect complementarity, which ranges from 6 to 7 nucleotides, and is commonly referred to as canonical interactions. Additionally, it is worth noting that the interaction can start on the miRNA with the nucleotide A resulting in a pattern of length 8 [Grimson et al., 2007].

iii Location

Another important feature to consider in the miRNA-mRNA interactions is the location of such interactions on both the miRNA and the mRNA. According to studies presented in [Sheu-Gruttadauria et al., 2019], the interaction between miRNA and mRNA can extend beyond the seed region, with the seed often interacting with a supplementary region located between nucleotides 13 and 16. The authors of [Sheu-Gruttadauria et al.,

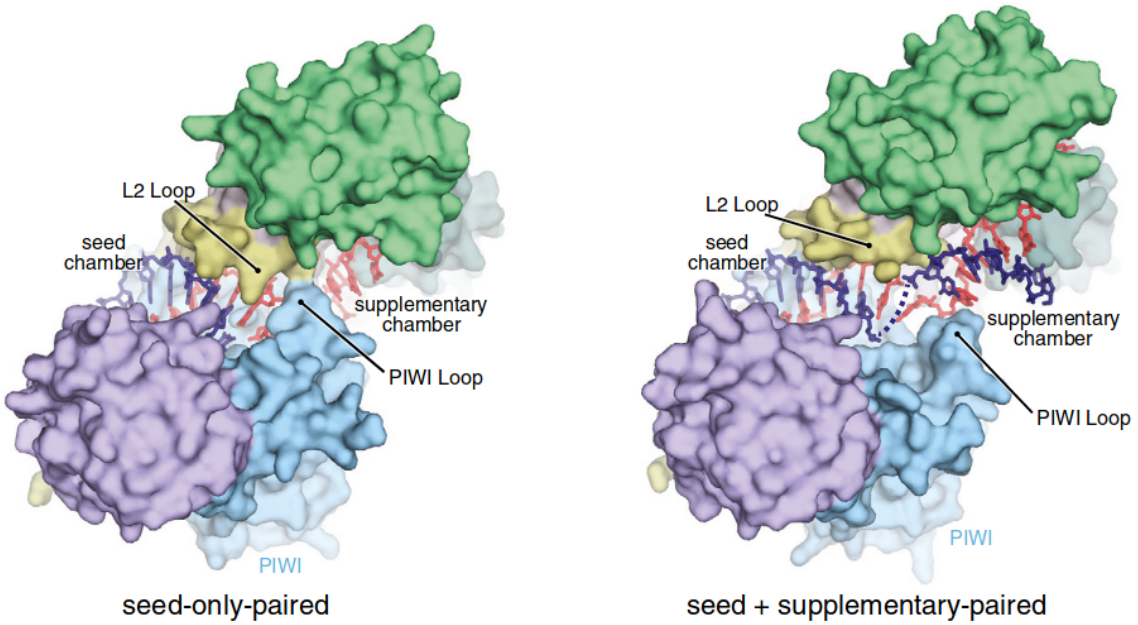


Figure A.5 – AGO structure opening leads to the base pairing of the supplementary region. In red the miRNA and in blue the mRNA. Adapted from [Sheu-Gruttadauria et al., 2019]

2019] suggest that a secondary opening of the AGO molecule may occur, leading to the base pairing of the supplementary region, as shown in Figure A.5. Other studies have also observed this bipartite base pairing pattern, as well as other types of bi- or tripartite base-pairings [Helwak et al., 2013, Moore et al., 2015].

The precise location of the interaction on the mRNA should also be considered. These can be in all regions of the mRNA: 5'UTR, CDS, and 3'UTR. Since historically, the first animal miRNA was discovered in the 3'UTR of *Caenorhabditis elegans* and *Drosophila melanogaster* (see Section i) with the addition that it was shown that the interaction sites present in the 3'UTRs are more efficiently down-regulated than those present in the CDS regions [Grimson et al., 2007, Hafner et al., 2010], the majority of the prediction tools [Akhtar et al., 2016, Bartel, 2009, Alexiou et al., 2009] are focused on searching the 3'UTRs. There are nonetheless studies which show that miRNAs targeting the CDS region may lead to an improvement of the regulation efficiency of interactions with the 3'UTR [Fang and Rajewsky, 2011]. Therefore, there is an attempt by some methods [Marin et al., 2013] to search for interaction sites within the CDS. On the other hand, while interactions in the 5'UTR have been reported in previous studies [Helwak et al., 2013, Moore et al., 2015], they have received relatively less attention and investigation compared to other regions.

Moreover, it is worth noting that miRNA-mRNA interactions may exhibit different characteristics across different kingdoms. For instance, in plants, the miRNA targets typically display a greater degree of complementarity with the mRNA compared to the miRNA targets in human [Millar and Waterhouse, 2005]

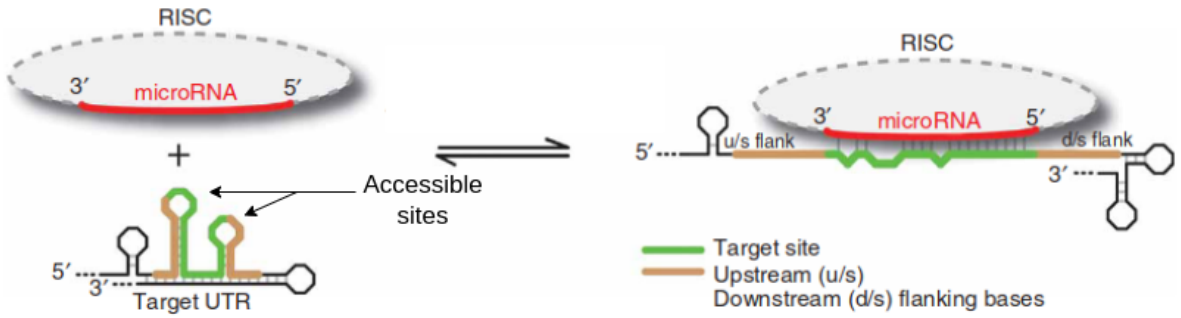


Figure A.6 – An example of accessible site on the mRNA which can induce the interaction with a miRNA. Adapted from [Kertesz et al., 2007]

iv Accessibility

Once transcribed, mRNAs are stabilised through the creation of secondary or tertiary structures. In order to interact, both RNA sequences must be accessible, which means that the interacting region should be unfolded, inside a loop, or folded weakly enough that it can unfold to allow base-pairing. There are four structure levels which are all important for the functionality of an RNA:

— The primary structure refers to the sequence of the RNA itself.

— The secondary structure refers to the folding of the RNA on itself in the plan.

— The third structure refers to the folding of the sequence of itself considering a 3D space.

— The quaternary structure corresponds to the structure formed from the interaction with other RNAs and molecules.

The notion of site accessibility refers to the ease with which miRNA-mRNA interactions can occur at a structural level. In support of this concept, a study by Long et al. [Long et al., 2007] suggests that miRNAs tend to initially bind to a highly accessible region, which then induces the unfolding of the mRNA structure and allows for the rest of the miRNA to bind, as illustrated in Figure A.6.

v Conservation

Most of the times conservation refers to inter-species conservation, and describes the degree of similarity or even identity of a sequence in different species. Conservation is an important additional feature to consider, and can focus on the sequence of the miRNA, of the mRNA or of a combination of the two. With the increasing number of available genomes and transcriptomes, conservation can now be more rigorously assessed. The fact that miRNAs exist in nearly all eukaryotic organisms argues for the fact that when miRNAs are conserved and given that their interaction relies on sequence complementarity some of their targets should also be conserved. This feature was introduced for example in [Friedman et al., 2008]. Furthermore, seed regions in miRNAs are generally more conserved than the rest of the miRNA [Lewis et al., 2003, Lewis et al., 2005].

A.4 Overview and objectives of this thesis

Before my PhD journey, the Team Erable from Inria and the LBBE (LABORATOIRE DE BIOMÉTRIE ET BIOLOGIE ÉVOLUTIVE) of the University of Lyon, already had a fruitful development of multiple computationnal biology tools that covered various fields including the identification of miRNAs. Thus, such teams provided a fertile ground for conducting further research on miRNA interactions and even for exploring a novel field that involves cross-species interactions.

The initial objective of my PhD, whose title is "New models and algorithms for the identification of sncRNA-(snc)RNA interactions intra and across-species/kingdoms", was to make progress on the identification, and thus validation of the existence of such across-cells and specially across-species or even across-kingdoms sncRNA-(snc)RNA interactions. However, during my first year, I concentrated my efforts on the problem of sncRNA target identification ***intra-species***, and furthermore, considered only one type of sncRNA, namely miRNAs.

There are various reasons for this. The first is that I come from computer science. I had close to no previous knowledge of biology except from what I learned in school. Moreover, I had also only little experience in computational biology. Indeed, previous to starting my PhD, I did only one relatively short – less than one year – internship in this area, however on another topic which concerned developing a DNA scaffolder. I therefore felt that I needed to start by focusing on a topic that is better known and thus less challenging to master.

The second, equally important reason is that, together with my supervisors, we realised that, although much more is known about miRNA-mRNA interactions intra-species, there remain many open questions, and even some disagreements on the criteria to be used to infer the targets of miRNAs, and also on the role of such miRNAs. These were initially believed to bind to mRNAs for repression only, but recent research has shown that they may also have a role as up-regulators [Vasudevan, 2012]. All the above is true of eukaryotes where miRNAs are commonly found, and even more so of prokaryotes which were not believed to have any miRNAs but where miRNA-like molecules [Gu et al., 2017] have started to be identified.

Despite the many issues that remain unsolved, the fact that miRNAs have been studied since longer means that we can have access to more literature, and also to more data, including to large datasets which have been at least partially experimentally validated. As such experiments, and notably those focused on validating specific miRNA-mRNA bindings, are both financially and time expensive, the information provided by the existing datasets remain nevertheless quite partial. The information that is provided thus often concerns which pairs of miRNA-mRNA interact but not the precise loci, that is the precise sites on both the miRNA and the mRNA, where such interactions happen. *In silico* approaches allow for a much faster and cheaper analysis of miRNAs and mRNAs to try to identify first which miRNA interacts with which mRNA, and then the precise location of the binding sites on both. However, the currently existing algorithms and methods suffer from a low accuracy and infer too many false positives [Fridrich et al., 2019, Mockly and Seitz, 2019].

We present in this thesis the resulting work, organised in five chapters. The current chapter provided the fundamental ideas necessary to understand the remaining ones. It

is followed by Chapter B that describes the current state of the art with the primary algorithms and how some methods make use of those algorithms.

Chapter C presents the creation and curation of datasets on *Arabidopsis thaliana*, human and on mice where the two latter are used in all the following chapters. Indeed, since we did not have experimental data at the same level of detail and depth as concerns *Arabidopsis thaliana*, we did not have the opportunity to use this dataset anymore. It was however useful and interesting at the beginning of this thesis, which is the reason why its curation is still presented.

Chapter D focuses on the usage of the accessibility and the seed by either a comparison the results of miRNA-mRNA predictors that use or not accessibility and the seed for their prediction or by a direct approach.

Chapter E presents the idea of intra-species conservation. While the concept of over-represented motifs has been previously described in the literature [Miranda et al., 2006], these motifs or patterns were typically inferred from miRNA sequences rather than from mRNA sequences.

The last chapter F describes the implementation of a web service that is currently running the motif inference tool Smile together with some other ones that allow to further analyse the results obtained by Smile.

Part of the work presented in this PhD manuscript was published on March 7, 2023, in a special edition on microRNAs of the journal *Genes*. The content of the corresponding paper and the Supplementary Figures are included in some of the next sections. The paper [Homberg et al., 2023] itself is available in open-access at the URL: `https://www.mdpi.com/2073-4425/14/3/664`.

Chapter B

State-of-the-art

In this chapter, we present the main existing tools, data, and algorithms that were used during this thesis. It is divided in four different sections.

The first one focuses on the computational tools, sometimes several ones, used to take into account each of the features for identifying miRNA–mRNA interaction sites (see Chapter A.3), and describes how each incorporates such features. We then present in the second section some of the current experimental procedures to capture miRNA–mRNA interaction sites. The third section provides a general overview of the databases of miRNA–mRNA interaction sites that were either computationally or experimentally inferred. Finally, we formally describe the two motif inference algorithms, namely Smile and Meme, that we used in this thesis.

B.1 *In silico*-computational methods

Since the beginning of the research on miRNA–mRNA interactions more than 25 years ago, a plethora of interaction site predictors was developed. These were surveyed and compared in a number of studies. The largest one, to our knowledge, did a comparison of 88 different methods [Kern et al., 2021]. These certainly did not include all those that have been developed over the years. Most if not all methods however share a combination of core features used for the inference of such interaction sites such as thermodynamics Gibbs free energy, accessibility, complementarity and/or conservation for inferring the interaction sites.

We present in this section the fundamental tools for computing such features which are used in the miRNA–mRNA interaction predictors. Despite the progress made, there is still room for improvement as the currently best such predictors yield too many false positives as stated in the paper [Fridrich et al., 2019].

i Computation of thermodynamics and accessibility

Thermodynamics "affinity" is one major feature that enables to infer the secondary structure of an RNA sequence or of the duplex between two RNA sequences. The free energy is computed in a similar way in both cases. Some tools that compute the folding structure of an RNA strand can be used to compute a duplex RNA by adding special

ligation characters between the two RNA sequences. In this section, the term free energy always refers to the Gibbs free energy.

We start by explaining formally the RNA structure and continue by presenting the Nearest Neighbour model which is used to compute the free energy of different possible RNA structures. We describe next three different approaches to compute the free energy:

— Minimum Free Energy refers to the free energy of the optimal structure.

— Partition Function refers to the free energy of all possible structures which is used in the computation of the accessibility.

— Maximum Expected Accuracy refers to the free energy of the optimal structure with some highly frequent suboptimal structures.

A more detailed review of the computation of RNA structure thermodynamics can be found in [Raden et al., 2018].

A current state-of-the-art fundamental method that uses all three free energy approaches is RNAFold [Lorenz et al., 2011]. As mentioned above, methods that compute the structure of one RNA can with a simple modification compute the structure of a duplex. This is the case for instance of RNACofold [Hofacker et al., 1994] with a time complexity in $O((n+m)^3$. There is though some limitation with the former approach due to a restriction to nested structures (explained below) which means that some structures such as kissing hairpins can not be predicted. There are some methods specialised on duplex structures, such as for instance RNAHybrid [Kruger and Rehmsmeier, 2006, Rehmsmeier et al., 2004], RNAduplex [Lorenz et al., 2011] and RNAplex [Tafer and Hofacker, 2008]. Note that RNAplex is a new version of RNAduplex which trades a better time complexity against a simplified energy model in the computation of some expensive substructures such as large bulge loops and large interior loops. The time complexity of RNAduplex and RNAHybrid is $O((nm)^2$ whereas the time complexity of RANplex is $O(nm)$.

RNA structure

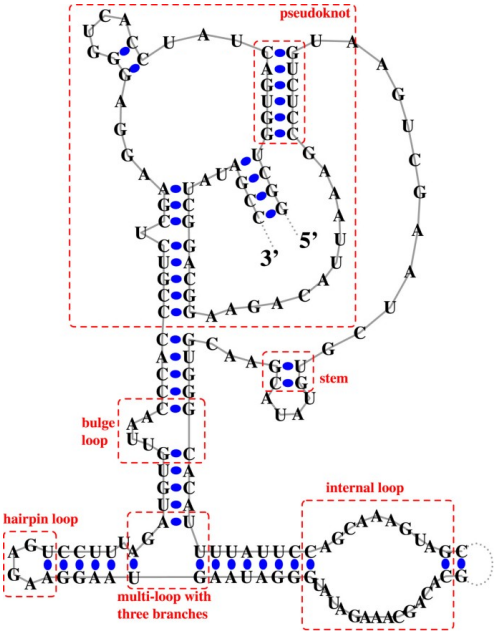
We denote an RNA sequence of length n by S where $S \in A, C, G, U^n$. A possible folding of S is the structure P which is defined as an ordered series of base-pairs. Each of the nucleotides of S can only bond once with a complementary nucleotide, with the exception of the wobble pair GU.

In reality, a structure can include various other features leading to complex structures such as pseudoknots. An example is presented in Figure B.1a. The complexity of a secondary structure is however often simplified [Lorenz et al., 2011] to exclude pseudoknots. This is done by removing possible crossed base-pairings and allowing only for nested base-pairings (see Figure B.1b). This simplification allows for a faster computation time when computing the possible structure of an RNA [Condon et al., 2004] .

A sequence S can have several different structures; we denote the set of all such structures by $\mathcal{P}$.

Nearest Neighbour model to compute the free energy of a structure

We recall that a structure may be formed by one RNA or involve two RNAs. In both cases, we call this structure P which has n base-pairings.



(a) Example of RNA secondary structure of the RNase P RNA molecule of *Methanococcus marapaludis* from the RNase P database. Adapted from [Andronescu et al., 2008]

(b) Arcs diagram of nested and crossing structures. Each arc symbolises a nucleotide bonding.

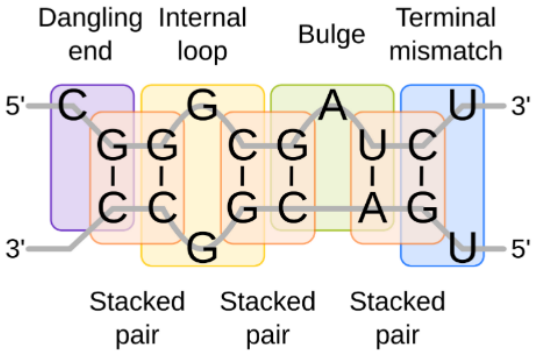
Although the global free energy of P is simply the sum of the free energy of each individual base-pairing, it is not trivial to compute it because, in addition to being sensitive to temperature, the free energy of a base pair depends on its nearest neighbours. For instance, the free energy of the base pair G–C is different when its neighbour is a A–U base pair (energy of the duplex: 3.8 kcal/mol) or a U–A base pair (energy of the duplex: 3.6 kcal/mol), hence the Nearest Neighbour free energy computation model [Borer et al., 1974]. It consists in separating the structure P into substructures such as stacked pairs, bulges and loops. The free energy of these substructures are added by iterating over each base pair substructure.

An example of this model can be seen in Figure B.2

The free energy computed using the Nearest-Neighbour model and the free energy of all possible substructures are experimentally validated and defined in the data table of Tuner and Mathews [Turner and Mathews, 2010] which refined and extended the combination of cases presented in the initial publication of Mathews [Mathews et al., 1999].

Minimum Free Energy (mfe)

The minimum free energy structure P_{mfe} is the energy of the optimal structure which has the lowest free energy among all structures of $\mathcal{P}$. P_{mfe} is the most stable structure because the base-pairing is proven to be the most effective force that maintains the RNA structure [DeVoe and Tinoco, 1962]. The optimal structure P_{mfe} is computed using the Zuker and Stiegler algorithm [Zuker and Stiegler, 1981]. It has a time complexity of $O(n^4)$ and a space complexity of $O(n^2)$. The Zuker and Stiegler algorithm uses dynamic programming and is very close in the way it works to the Nussinov algorithm that identifies



Structural element	Nucleotides	Free energy [kcal/mol]
Dangling end	(C,GC)	−0.3
Stacked pair	(GC,GC)	−3.3
Internal loop	(GC,G|G,CG)	−2.1
Stacked pair	(CG,GC)	−2.4
Bulge	(GC,A,UA)	+1.6
Stacked pair	(UA,CG)	−2.4
Terminal mismatch	(CG,UU)	−1.2
Intermolecular initiation		+4.1
Overall free energy	$\sum$	−6.0

Figure B.2 – Example of free energy computation using the Nearest-Neighbour model with the parameters of Turner and Mathews [Mathews et al., 1999]. Source [Gelhausen and Raden, 2018]

the structure with the most base pairs (Equation B.1. The maximum number of base pairs in the subsequence S_iS_j is stored in the cell $N_{i,j}$ of the dynamic programming matrix N. At the beginning, all the cells of N are initialised with 0. The final number of maximum base pairs can be read in $N_{1,n}$ and the structure is obtained by backtracking on N.

$$N_{i,j} = \begin{cases} N_{i,j-1} & \text{if } S_j \text{ unpaired} \\ \max_{i \leq k < (j-l)}(N_{i,k-1} + N_{k+1,j-1} + 1) & \text{if } S_k, S_j \text{ pair} \end{cases} \tag{B.1}$$

A more detailed implementation of this algorithm supporting the Nearest Neighbour model can be found in the paper [Tafer and Hofacker, 2008] or [Rehmsmeier et al., 2004].

Partition Function (pf)

The previous structure considers only the optimal case of minimum free energy and thus the most probable one. However, it is essential to be able to compute the probability of suboptimal structures. This is mainly modeled with a Boltzmann distribution. Note that for all the following equations, the Boltzmann factor k_B has to be set to a gas constant R when using energy models with units "per mole" which is the case when the free energy of each base pair is computed with the Nearest Neighbour method.

With the Boltzmann distribution, we have the probability of a structure P in function of its free energy $E(P)$ defined by:

$$Pr(P) = \frac{exp(-E(P)/k_BT)}{\sum_{P' \in \mathcal{P}} exp(-E(P')/k_BT)}$$

. The numerator $exp(-E(P)/k_BT)$ is the Boltzmann weight of a structure P where k_B is the Boltzmann factor and T is the temperature of the system which is usually set to the body temperature, namely 37° Celsius. The denominator $Z = \sum_{P' \in \mathcal{P}} exp(-E(P')/k_BT)$

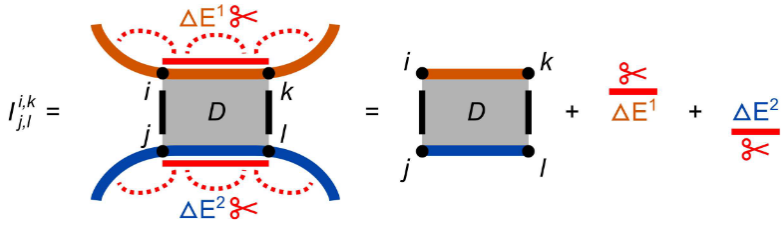


Figure B.3 – Accessibility of the interaction $I_{i..k}^{j..l}$ on two RNAs (blue and orange) by computing the sum of the energy gain D from the hybridisation of two RNAs and the loss of energy that is required to open both RNAs. Source [Raden et al., 2018]

is called the canonical partition function Z. The partition function Z represents all possible structures and the conversion of the free energy of all possible structures is given by $-RT * log(Z)$. Computationally-wise, since the number of possible structures is exponential with n, a direct exhaustive computation is not viable. An algorithm in $O(n^3)$ time that applies the main principles of the Smith–Waterman counting algorithm [Waterman and Smith, 1978] with the Nearest Neighbour model was proposed by McCaskil [McCaskill, 1990].

From the probability of structure $Pr(P)$, we can go further and compute the probability that a region (i, j) on all $P \in \mathcal{P}$ is either paired $Pr^{bp}(i, j))$ or unpaired/single-stranded $(Pr^{ss}(i, j))$, thus accessible. This is given by:

$$Pr^{bp}(i,j) = \frac{\sum_{P \in \mathcal{P}_{i..j}^{bp}} exp(-E(P)/k_B T)}{Z} \tag{B.2}$$

$$Pr^{ss}(i,j) = \frac{\sum_{P \in \mathcal{P}_{i..j}^{ss}} exp(-E(P)/k_B T)}{Z} \tag{B.3}$$

$Pr^{ss}(i, j)$ is computed for instance with RNAplFold [Lorenz et al., 2011].

An alternative way to compute the accessibility (see Figure B.3) of an interaction $I_{i..k}^{j..l}$ with $i..k$ the region of the first and $j..l$ the region of the second RNA is to consider the sum of the energy gain D from the hybridisation of two RNAs and the loss of energy that is required to open both RNAs: $\Delta E^{i..k}$ and $\Delta E^{j..l}$. The opening energy $\Delta E^{i..j}$ is calculated by the difference of energy of all accessible structures ($E(Z_{i..j}^{ss})$ with the whole energy of all structures ($E(Z)$). We therefore have:

$$\begin{aligned} \Delta E^{i..j} &= E(Z_{i..j}^{ss}) - E(Z) \\ &= -(RTln(Z_{i..j}^{ss}) - RTln(Z)) \\ &= -RTln(\frac{Z_{i..j}^{ss}}{Z}) \\ &= -RTln(Pr^{ss}(i,j)) \end{aligned} \tag{B.4}$$

Pita [Kertesz et al., 2007] for instance, computes the accessibility with this method. However, only the opening energy $\Delta E^{i..k}$ of the mRNA is considered (called ΔG_{open}). The authors use the free energy D of the optimal structure for the hybridisation that they call ΔG_{duplex}. Since $\Delta G_{open}^{i..j}$ is equivalent to $Pr^{ss}(i, j)$, they computed it with RNAplFold.

Maximum Expected Accuracy (mae)

The minimum free energy and the partition function are either the optimal or all possible structures from $\mathcal{P}$. Do *et al.* developed the concept of Maximum Expected Accuracy (mae) in the paper [Do et al., 2006]. The aim of the Maximum Expected Accuracy is to consider the optimal structure with some highly frequent suboptimal structures. The weighted factor γ serves to strengthen or loosen the selection of the suboptimal structures. The maximum expected accuracy can be computed with a variation of the Nussinov algorithm B.1 :

$$M_{i,j} = \begin{cases} M_{i,j-1} + Pr^{ss}(j,j) & \text{if } S_j \text{ unpaired} \\ \max_{i \leq k < (j-l)}(M_{i,k-1} + M_{k+1,j-1} + 2\gamma Pr^{bp}(k,j)) & \text{if } S_k, S_j \text{ pair} \end{cases} \tag{B.5}$$

Similarly to N in the Nussinov algorithm, M is set initially equal to 0, the maximum expected accuracy of the structure is found in $M_{i..n}$, and the corresponding structure can be retrieved with a backtrack.

ii Computation of alignment

In this section, we detail the fastest and easier than thermodynamic-based, yet effective, method to compute the interaction between a miRNA and an mRNA by means of an alignment. Indeed, a basic interaction RNA scheme can be related to the alignment of two RNAs by converting one of them to its reverse complement. Furthermore, this method considers only inter-RNA base-pairing, excluding the folding of the mRNA.

There are several alignment methods. We describe here the most used one developed by Needleman and Wunsch [Needleman and Wunsch, 1970]. It is based on a scoring system that penalises gap opening, called score_p, and rewards matches, called score_m. Moreover, it works by storing in a matrix F the best score of the one or the several optimal alignments according to the scoring schemes. The alignment can then be retrieved by backtracking. The matrix F is filled with:

$$F_{i,j} = max \begin{cases} F_{i-1,j-1} + \text{score}_m & \text{if } i \text{ and } j \text{ match} \\ F_{i-1,j} + \text{score}_p & \text{if } i \text{ and } j \text{ unmatch (deletion)} \\ F_{i,j-1} + \text{score}_p & \text{if } i \text{ and } j \text{ unmatch (insertion)} \end{cases} \tag{B.6}$$

Some variation improves this method by splitting the penalty gap to a penalty of starting the gap and a penalty of extending the gap. A different score of matches can also be applied, for instance to take into account wobble-base pairing.

iii Computation of conservation

Another major feature used in various methods [Agarwal et al., 2015, Marín and Vaníček, 2012, Miranda et al., 2006] is the conservation inter-species. Indeed, there is evidence supporting the fact that the majority of the miRNA interaction sites in mammalian [Friedman et al., 2008, Lewis et al., 2005] and insect [Schnall-Levin et al., 2010] mRNAs are conserved. For this feature, there is not a fundamental tool such as RNAfold in thermodynamic computation that is used in most of the methods; instead all methods have their own implementations which are, however, globally similar.

The basic idea behind the computation of the level of conservation is to make use of an alignment file which contains the mRNA sequences of several species aligned. Hence, for a potential miRNA target site on an mRNA (identified from other features), the conservation level is deduced from the depth and quality of the alignments. Additionally, some methods [Gumienny and Zavolan, 2015] inject the evolutionary distance between the targeted species with other species from a phylogenetic tree into the assertion of the level of conservation, and therefore favor alignments based on selection pressure instead of by chance. The conservation of the seed of miRNAs across different species is in the case of some methods [Agarwal et al., 2015, Agarwal et al., 2018] also taken into account in the final conservation score.

Contrary to conservation inter-species, the conservation intra-species is less commonly used among methods. It has different names such as motifs over-represented in PACCMIT-CDS [Marin et al., 2013] or pattern-based in RNA22 [Miranda et al., 2006] and there is actually no consensus of implementation. RNA22 infers variable length patterns with at least 30% of their positions specified from miRNA sequences. For instance, a motif of length 12 must have at least 4 positions that have a specific nucleotide, while the other ones are non specific. These patterns are searched on mRNAs and a putative target site (called Target Island) is suggested when more than 30 different patterns match this mRNA region. This putative target is then validated or rejected based on a thermodynamic affinity. PACCMIT-CDS on the other hand searches for over-represented seed matches in the CDS region and computes the probability that a seed can be found more in mRNA sequences than in random sequences which preserve the same proportion of codons.

However, this concept of intra-species conservation has to be handled with caution when the performance of interactions is considered. Indeed, intra-species conservation has to be linked with target abundance which is proven to reduce the regulation power of miRNAs [Garcia et al., 2011, Arvey et al., 2010].

B.2 *In vivo*-experimental methods

In this section, we will briefly explain some wet-experimental methods used to predict miRNA–RNA interactions. We separated this section into Low and High Throughput experimental methods.

These exploit one or several properties of the miRNA–mRNA interactions in order to infer the level of confidence of the existence or functionality of an interaction:

— Level of expression of an mRNA in the presence or absence of a miRNA.

— Change of phenotype in the presence or absence of a miRNA.

— The interaction between mIRNA and mRNA is dependent on an AGO complex.

— The miRNA and mRNA have a sequence complementarity.

— The miRNA and mRNA interactions are conserved inter-species due to biological advantages.

A more complete explanation can be found for instance in the paper [Mockly and Seitz, 2019].

i Low-throughput experimental methods

Low-throughput methods use a technical approach that produces a low number of results. The results from such methods emphasise generally quality over quantity.

In vivo genetics

This technique led to the first discovery of a miRNA interaction, namely the one involving the *lin-4* miRNA. It is based on mRNA mutants that lack the putative interaction site of miRNA binding; these mRNA mutants are then placed in a cell and the expected behaviour is that these mRNAs will not be regulated by the miRNAs which results in a gain of function induced by the mRNA mutants. This experimentation has the advantage of testing directly the function of putative sites. Furthermore, this test occurs in a normal biological cell, instead of laboratory cells. The disadvantage is the technical challenge to create a viable mRNA mutant without the putative interaction site. This technique is expected to be heavily used in the future with the current sequence alteration technologies.

Quantification of artificial reporter expression in cultured cells

Reporter Expression in Cultured Cells is the most common technique. The idea is to clone the putative interaction downstream of a reporter coding sequence and then to transfect the sequence obtained in a cell containing (naturally or artificially) the putative interacting miRNA. The effectiveness of the regulation is then measured with the expression level of the reporter gene.

Some of the disadvantages of miRNA transfection are: (1) Too many miRNAs can disrupt the natural biological context of the cell and lead to mRNAs not repressed whereas they should have been; (2) Over-repression which can also happen due to the miRNA being in large excess relatively to the natural amount and which will target mRNAs that would normally be unregulated.

An alternative but similar method is to decrease the number of miRNAs in the cell with inhibitory oligonucleotides and measure the potential gain of expression of the mRNAs. This method leads to fewer biases of overexpression, yet inhibitory oligonucleotides have also side effects.

Detection of the reporter expression is also faulty at times. The most common reporter genes are *Luciferase* proteins from the *Firefly*. The *Firefly* mRNA has an intracellular half-life of 3 to 4.5 hours which is the same time needed for this kind of experiment and it may result in an imprecise outcome.

ii High-throughput experimental methods

High-throughput methods use a technical approach that produces a vast number of results. These methods thus generally emphasise quantity over quality which results in the frequent need of a heavy in silico post-processing in order to clean the high-throughput in vivo results.

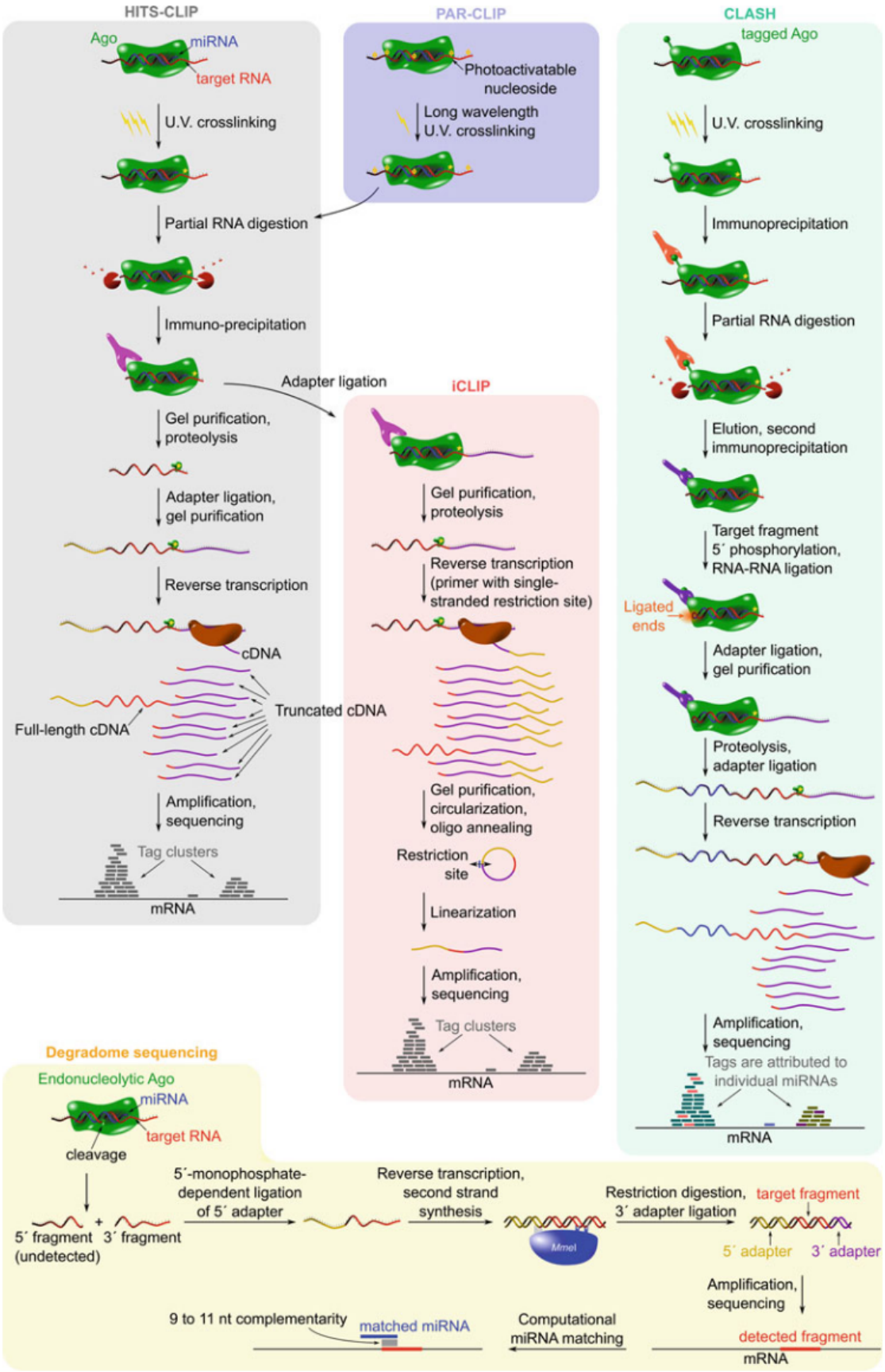


Figure B.4 – Overview of Degradome sequencing, CLASH and CLIP based experimental methods to discover miRNA-mRNA interactions. (source [Mockly and Seitz, 2019])

HITS-CLIP for HIgh-Throughput Sequencing by Cross-Linking and Immuno-Precipitation

The core of the HITS-CLIP methods [Ule et al., 2003, Ule et al., 2005] is to create a bond between protein and RNA, hence a cross-link. This cross-link is obtainable with the property of UV irradiation [Greenberg, 1979]. The resulting protein-RNA complex can then be isolated with immunoprecipitation and then sequenced. The AGO protein, which helps the miRNA and mRNA binding (see A.2), is fortunately a good candidate for such UV-induced cross-linking and can also be isolated with immunoprecipitation [Schirle et al., 2014]. Both mRNA and miRNA are then sequenced independently once the AGO is removed. The final outcome provides targeted mRNAs and miRNAs but a final computational step is mandatory to link each couple of miRNA and mRNA. Some negative critics of this technique are that UV cannot penetrate thick cell skins [Sugimoto et al., 2012]. Moreover, interactions involving uracil nucleotides are preferred leading to potential false negatives [Chi et al., 2009].

PAR-CLIP for Photoactivatable Ribonucleoside-enhanced by Cross-Linking and ImmunoPrecipitation

PAR-CLIP aims to improve the cross-link reaction with a photoactivable nucleotide-like 4-thiouridine (4-SU). This will bind with the mRNA and make it more sensible to long-wave UV, which results in a higher number of cross-links made (100-1000 fold improvement [Hafner et al., 2010]). Thanks to the 4-SU, it is easier to differentiate cross-linked RNA from non cross-linked RNA. The drawback of this method is a technical limitation of incorporating the 4-SU by mRNAs. This tends to make the PAR-CLIP experiment very time consuming with 12 to 16h of exposition to 4-SU. A typical PAR-CLIP experiment needs 60 to 100 Petri-dishes.

iCLIP for individual-nucleotide resolution CLIP

When using the crossed-linked and immunprecipitated methods, some protein fragments stay bonded to the mRNAs thus obstructing the PCR amplification and sequencing. The method used by iCLIP diverges from the one used by HITS-CLIP after the immunoprecipitation process as iCLIP adds a complementary 3' primer adapter to the mRNA. The mRNA is then circularised and re-linearised which helps the cleansing and thus the efficiency of the PCR amplification and of the sequencing (see Figure B.4). The libraries produced by iCLIP thus have 80% more mRNA sequences than those produced by HITS-CLIP and provide a more precise localisation of the interaction on the mRNAs [Sugimoto et al., 2012]). This method is however technically challenging.

CLASH

CLASH stands for Cross-linking, Ligation And Sequencing of Hybrids. All the methods described above do not provide the miRNA and interacting miRNAs which have to be computed post-experimentation whereas in CLASH, the miRNAs are sequenced with their mRNA targets. Indeed, in a CLASH experiment, the miRNA and mRNA are linked with the AGO protein triggered by UV-light as with CLIP, however a special treatment involving a tagged AGO and a two-step immunoprecipitation allows the ligation of the

miRNA with the mRNA. This ligation makes possible the sequencing of hybrid sequences composed of the miRNA and mRNA. A final computational step separates the miRNA and mRNA sequences and aligns them with a reference model. A disadvantage of this technique is that the efficiency of the ligation is low, and hybrids or chimeric reads constitute only about 2% of the total reads.

Degradome sequencing

The method of degradome sequencing makes use of a particular AGO protein which will cut both ends around a miRNA–mRNA interaction. This cleavage only happens when there is a perfect complementarity at the positions 10-11 of the miRNA. It also leaves a characteristic 3' and 5' end which can ligate to a 5' adapter. This 5' adapter with the help of the *MmeI* enzyme enables the ligation of a 3' adapter to the other end of the mRNA site targeted. The mRNA interaction site is hence between two adapters which enables PCR amplification and then a deep sequencing. The property that the positions 10-11 have to be a perfect match with the miRNA is a good help to computationally link the miRNA with this region of an mRNA.

B.3 Overview of the existing databases

In bioinformatics, data are the core of the field; indeed all algorithms are validated or trained on them. With biological data, there are mainly two concerns: quality and quantity. The data can be obtained from *in silico* predictions or through experiments and this will impact the quality. Quantity, on the other hand, is equally important to be able to reliably perform statistical tests. Fortunately, the technological breakthroughs in the last 20 years allowed for a cost reduction coupled with an increasing reliability of the data generation process. With this diversity and quantity of data, the need for common places that keep track of all these data is a necessity. This section is dedicated to briefly present some of the major databases which I used during my thesis.

i NCBI

NCBI [Pruitt et al., 2007] stands for National Center for Biotechnology Information. It is hosted by the United States government and therefore linked with several American services such as the National Institute of Health (NIH) and the National Library of medicine (NLM). *NCBI* contains in fact several different databases and tools that provide a variety of information such as:

— Pubmed– which gives access to biomedical and life science publications.
— GenBank– which is a database that stores DNA, RNA and protein data.
— Blast– which is a set of tools to align and search sequences.
— Entrez– which is a search engine to browse the databases of *NCBI*.

ii Ensembl

Ensembl [Kinsella et al., 2011] is with *NCBI* the leader on bioinformatics data. It was created from a collaboration between the European Bioinformatics Institute (EBI)

and the Wellcome Trust Sanger Institute. *Ensembl* provides comparable information as *NCBI* with, for instance, over 50,000 genomes of different species and the data-mining tool *biomart* which is a precious help to generate datasets from various sequence IDs and various genome versions.

iii miRBase

miRBase [Griffiths-Jones et al., 2007] is the biggest and leader database specialised on miRNAs. It provides annotation and sequence data of miRNAs as well as evidence, in the form of publications supporting the existence of this miRNA. There is nonetheless in *miRBase* little information concerning the mRNA or other targets of a miRNA. The current version 22.1 contains 48,885 miRNAs spread across 206 species. Depending on the version, new miRNAs can be added following a new discovery, or on the contrary, removed due to an erratum in a publication or to a change in the name, for instance to add $3p$ or $5p$ depending on the strand or to link several families of miRNAs that have a similar function or sequence.

iv miRTarBase

miRTarBase [Hsu et al., 2011], which stands for miRNA Target Base, stores experimentally validated and hence manually curated mRNA sites targeted by miRNAs. *miRTarBase* is complementary to *MiRBase* in order to have information on the mRNA targets as well as the miRNAs. Furthermore, *miRTarBase* ranks the experimental validation with two levels of evidence, namely strong and weak. The experimental validations Reporter assay, Western blot or qPCR are considered as strong evidence whereas microarray, NGS, pSilac and CLIP-seq are considered as weak. *miRTarBase* in its version 9.0 contains 2,200,449 targets of 4,630 miRNAs that are spread over 37 different species.

B.4 Two motif finding methods

In this section, we give an overview of two methods of motif inference, namely Smile and Meme, that were used in this thesis. Both methods work by taking a set of sequences as input together with some parameters such as the minimum and maximum length of the motifs to be inferred, and provide as output one or several motifs that were found in the sequences. Each of these two approaches has its strengths and weaknesses.

Meme with the motif distribution mode of one occurrence per sequence is better suited for a motif that is expected to be found in all sequences whereas Smile works well with motifs that are present in only a fraction of the sequences. In addition Smile is an exact algorithm which provides motifs in the form of k-mers, while Meme is a heuristic which provides motifs in the form of a matrix which indicates for each position of the motif a probability associated with each letter of the alphabet.

i Smile

Smile [Sagot, 1998] is a deterministic algorithm that was developed in our team. Its purpose is to find motifs that are common to several biological sequences which can be

written on various alphabets, such as the amino acid or nucleotide alphabet, or any other kind. It is based on a generalised suffix tree which will find all the motifs verifying some parameters given as input. These motifs can then be filtered.

To properly describe the input of SMILE, we first define the parameters quorum q, substitution e and the minimum and maximum length. Formally, a motif m is defined on an alphabet $\mathcal{A} = \{a_1, a_2, ..., a_L\}$. We define $u \in \mathcal{A}^+$ as a word of a sequence $s \in \mathcal{A}^+$ if $s = xuy$ with $x, y \in \mathcal{A}^*$.
We say that the motif $m \in \mathcal{A}^+$ has an *e-occurrence* in s where $e \in N$ if $\exists u$ in s such that the Hamming distance (number of substitutions) between u and m is less than or equal to e. Let n be the number of sequences $(s1, ..., s_n) \in \mathcal{A}^*$ and let $q \in N$ with $1 \leq q \leq n$. We say that a motif m is valid if it possesses at least one $e - occurrence$ in at least q sequences. The parameter q is called the *quorum* that a motif must verify to be considered as valid.

For instance, when we are looking for a motif of length 5 with 1 substitution and with a quorum of 20, this motif must be found in at least 20 sequences where each occurrence found should be identical or with only one letter of difference.

A suffix tree is a structure organised in a tree which represents all suffixes of a string. For a string of size l, the suffix tree has l leaves representing the suffixes and each edge is a substring of the input string. SMILE uses a generalised suffix tree which represents several strings concatenated and separated by different special characters not in $\mathcal{A}$ (depicted in Figure B.5).

In order to efficiently infer the motifs from the generalised suffix tree (action called "spelling" in [Sagot, 1998]), each node of the suffix tree has to store an additional information (not shown in Figure B.5) that is the number of different sequences that those nodes refer to. This information, denoted by CSS_x, corresponds to the Color Size Problem [Hui, 1992]. In the case of zero substitution ($e = 0$), this number is sufficient. If $e > 0$, an additional information of the enumeration of each sequence that refers to this node has to be stored.

Once the motifs are found, we evaluate, either against random sequences that are generated by shuffling the input sequences or against another set of sequences provided by the user, the statistical significance of the motifs by computing the Chi^2, the mean $\bar{X}$, the standard deviation s and the $Z_{score}(m) = \frac{\bar{X} - \mu_m}{s}$ where μ_m is the number of occurrences of the motif in the input sequences. All the above is computed using the number of occurrences of the motif in the input sequences randomised r times or alternatively, the number of of occurrences of the motif present in the second set given as input. For the work presented here, we always proceed with an evaluation against the input sequences randomised.

ii MEME

In this section, we briefly explain the basis of the algorithm of MEME as described in [Bailey and Elkan, 1995]. It is important to note that, since its first publication, MEME has evolved and many tools from the MEME SUITE [Bailey et al., 2015] are a variation of MEME enabling more features.

MEME stands for Multiple EM for Motif Elicitation and is based on an unsupervised machine learning approach with the algorithm Expectation Maximization (EM) originally

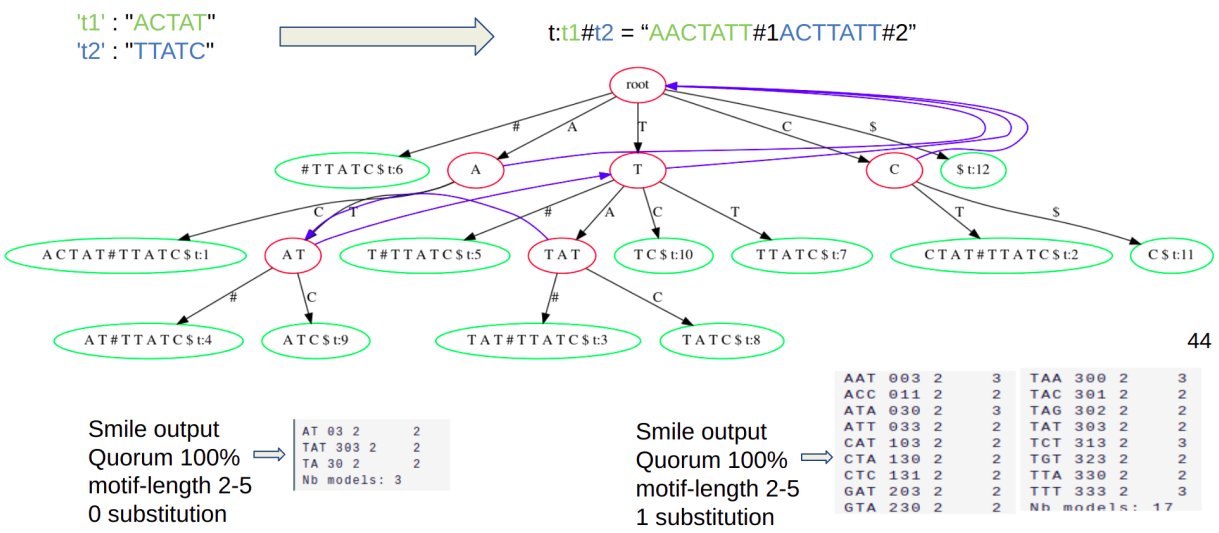


Figure B.5 – Example of a SMILE process on two example queries using the same input sequences $ACTAT$ and $TTATC$. The constructed suffix tree is displayed, with the internal process which concatenates all input sequences together. Both queries ask for motifs with a quorum of 100% which forces the motifs to be in the two sequences, and with a length from 2 to 5 nucleotides. In the first query no substitution is allowed ($e = 0$) whereas the second allows for one substitution ($e = 1$).

introduced by Lawrence and Reilly in [Lawrence and Reilly, 1990]. Expected maximization works iteratively to find the best local set of parameters of a problem. It begins from an initial parameter and then alternates between a step of expectation and a step of maximization until a convergence is achieved, *i.e.* the difference between both steps is smaller than a threshold, which corresponds to a locally best parameter.

In the case of MEME (depicted in Figure B.6), the EM set of parameters is the probability matrix of a motif p. The initial value matrix is set from a k-mer in the input sequences. However the k-mer cannot be directly converted to a probability matrix simply by setting the probability of each nucleotide to 1, otherwise the EM algorithm would converge to the same matrix. Therefore, the trick is to set each probability X of the nucleotide of the k-mer to $0.4 < X < 0.8$ and the other one to $\frac{1-X}{|\mathcal{A}|-1}$ where $|\mathcal{A}|$ is the cardinality of the alphabet $\mathcal{A}$. The expectation step takes the motif p of iteration $i-1$ or the initial motif if it is the first iteration and then computes occurrences of this motif and stores them in a matrix called z. The maximization step does the opposite and infers the motif p from the matrix z. This cycle of expectation and maximization iterates until the difference between p from the iteration $i-1$ and p from the current iteration i is smaller than a threshold ϵ.

The resulting MEME algorithm takes as parameter the number of motifs n to be found and their length W. To reduce the computation cost and since the EM algorithm converges quickly enough, MEME is run for 1 iteration (not until convergence) on each W-mer of the input sequences. The motif with the highest likelihood is selected and the EM algorithm is executed until convergence is reached. Formally, the likelihood is the probability of the data given the model and MEME defines it as the sum of the specificity of each nucleotide of the subsequence W-mer. The specificity $spec_{ij}$ of the letter i at

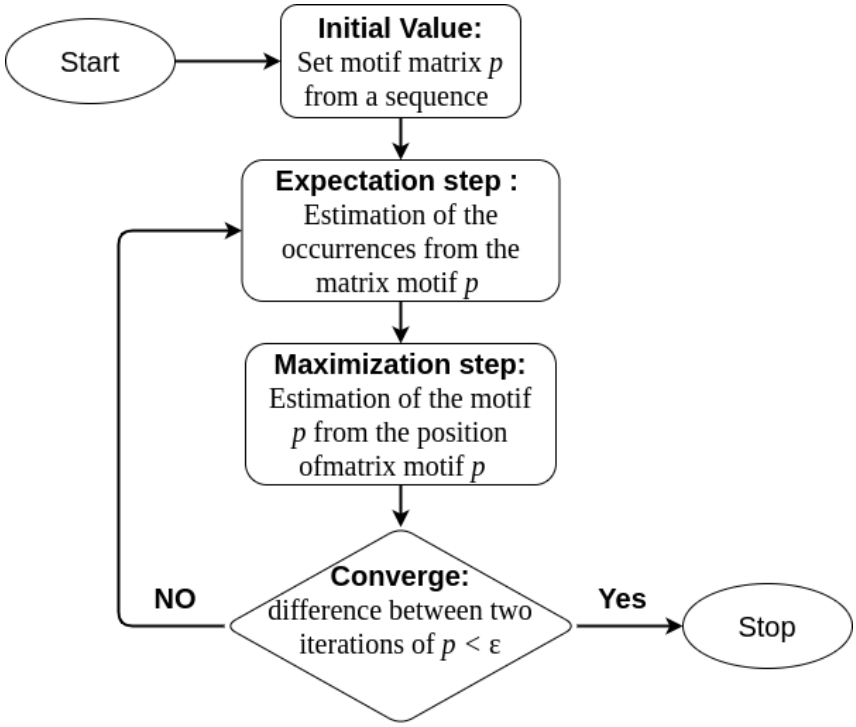


Figure B.6 – Schematic view of the Expectation Maximization Algorithm.

position $j \in 1..W$ is given by:

$$spec_{ij} = log(\frac{p_{ij}}{p_{iO}}) \tag{B.7}$$

where p_{ij} is the probability of the letter i at position j of the motif and p_{iO} is the probability of the letter i in all non-motif positions. This notion tries to infer the likelihood that a subsequence W-mer is part of the motif versus part of the background.

In order to find other motifs, MEME partially erases the occurrences of the previously inferred motifs of the dataset in function of the probability associated for each nucleotide of the motif. For instance, if a nucleotide inside a motif has a very high probability then it is strongly erased from all occurrences of this motif and has a low chance of being used in another motif; on the contrary, if the nucleotide of a motif has a very low probability then it is weakly erased from all occurrences of this motif and has a higher chance of being used in another motif. MEME continues on the curated dataset until n motifs are inferred or until no more starting W-mer is available from the input sequences.

Chapter C

Contribution: Materials

C.1 The winding path towards getting precise interaction data

To test different hypotheses and tools, we first have to find or to create reliable datasets of validated interactions. The path towards getting precise interaction data is often winding. In this chapter, we present the one we follow, that started as explained in this section by a navigation of three databases.

Navigating MirTarBase, miRBase and Ensembl

In all cases, we started by focusing our attention on the species *Mus musculus* and *Arabidopsis thaliana*. Indeed, both benefit of a substantial number of miRNA-mRNA interactions already studied (see Table C.1). Initially, we put aside looking into human data as there appeared to be in this case less interactions that had been strongly validated. Notice that *Arabidopsis thaliana* is also studied for its interaction with the fungus *Botrytis cinerea* [Weiberg et al., 2013]. This fungus is capable of transferring small RNAs to plants, such as *Arabidopsis thaliana*, which can silence genes involved in immunity.

Once the species of interest have been selected, databases can be searched to retrieve the relevant data. An important database that we considered is miRTarBase [Hsu et al., 2011] which provides information on the interactions between miRNAs and mRNAs, including the specific pairs involved in each interaction.

These concern various species. We chose *Mus musculus* due to the high number of available interactions (around 50000 interactions from which 1700 are labeled as strong). This database only provides the gene (entrez id) and miRNA names. A small amount of additional work was then required to get the transcript id and their sequences. There are several nomenclatures for transcript ids: Ensembl id, UCSC id, Refseq id and Gencode id.

We chose the Ensembl id and obtained our transcript from Biomart ensembl [Kinsella et al., 2011] and we developed a script which queries the NCBI database to get a transcript id from the gene id provided by miRTarBase. The sequences of the miRNAs were then retrieved from the database mirBase [Griffiths-Jones et al., 2007].

This led us to the creation of two datasets which contain the experimentally validated

miRNA–mRNA interactions from miRTarBase [Hsu et al., 2011]. The first one contains all interactions, whether weak or strong, and the second contains only the strong interactions (see Table C.1).

Species	**validation**	**#miRNAs**	**#mRNA**	**#interactions**	**Mean (Min,Max) # mRNA targeted by one miRNA**	**Mean (Min,Max) #miRNA that bind on one mRNA**
Mus musculus	All	959	7238	49830	51.96 (1, 1174)	6.88 (1, 84)
Mus musculus	Strong	292	815	1627	5.571 (1, 49)	1.99 (1, 29)
Arabidopsis thaliana	All	40	70	106	2.65 (1, 11)	1.51 (1, 7)
Arabidopsis thaliana	Strong	8	8	15	1.875 (1, 4)	1.87 (1, 4)

Table C.1 – Overview of the first datasets created.

We used these datasets to start the PhD journey and to test accessibility with some methods (see Section D). We were rapidly stuck on two points. First, each gene can produce multiple transcripts, and it is not always clear which transcript is being targeted by a given miRNA. In the literature, one common approach to address the issue of multiple transcripts is to consider the longest transcript as the target of the miRNA. This is because the longest transcript usually contains the complete coding sequence and is thus more likely to be translated into protein. However, this approach may not always be appropriate as the other transcripts may have important regulatory roles and could also be targeted by miRNAs. Second, we lack precision on where exactly a miRNA binds on the mRNA. Indeed, an mRNA can be several thousand nucleotides long and the interaction is usually not longer than 20-30 nucleotides. We tried to use different prediction tools to locate the exact interaction site. However, the tools did not have a consensus target site for each miRNA/mRNA couple.

Nonetheless, despite these limitations, valuable insights were extracted from these datasets and were subsequently validated in our newly generated dataset from CLASH (see next section). Notably, this dataset revealed the number of mRNAs that are targeted by a single miRNA. This number of target sites per miRNA can go up to a thousand with a mean of 52 or 5.5 depending on the strength of the experimental validation. This information is also rarely used for target site prediction/inference. We explore this idea in Chapter E

C.2 Dataset creation from CLASH

Since we hit a wall with our first approach of creating a dataset from the available databases, we looked for other experimental validations and we discovered the paper "Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding" [Helwak et al., 2013] which satisfied our needs with their provided dataset. Notice that this time, the data concerns human as indicated by the title of the paper.

The CLASH experimental procedure is explained in section ii and has several advantages. It produces a large amount of results that provide precise information about the nucleotides involved in miRNA-mRNA interactions, including non canonical interactions.

While CLASH has many advantages in identifying the precise nucleotides involved in miRNA-mRNA interactions, it has been criticized for focusing on the interaction sites

rather than the overall function of the miRNA. It should be noted that not all miRNA-mRNA interactions lead to regulation or any other function, and some mRNAs may act as a sponge, depleting the miRNA and preventing its normal function. However, studying the mechanism of interactions is also crucial for understanding the miRNA function.

i First CLASH dataset

The information provided by the CLASH data as presented in [Helwak et al., 2013] contains 27 columns (see Table C.2), together with the precise placement of 399 miRNAs onto 7390 mRNA targets, representing a total of 18514 miRNA-mRNA attachment sites.

Column number	Column name	Column description
Column 1	seq_ID	unique ID of interaction
Column 2	**microRNA_name**	**miRNA name from miRBase release 15**
Column 3	**miRNA_start**	**start coordinate in miRNA**
Column 4	**miRNA_end**	**end coordinate in miRNA**
Column 5	**miRNA_seq**	**sequence of miRNA (start – end)**
Column 6	**mRNA_name**	**gene name and mRNA transcript name from ENSEMBL release 60**
Column 7	**mRNA_start**	**start coordinate in mRNA transcript**
Column 8	**mRNA_end_extended**	**end coordinate in mRNA transcript, bioinformatically extended by 25 nt**
Column 9	**mRNA_seq_extended**	**sequence of mRNA fragment (start – end+25 nt)**
Column 10	chimeras_decompressed	number of chimeric reads from experiments E1-E6 supporting the interaction
Column 11	experiments	number of experiments in which reads supporting the interaction were found
Column 12	experiments_list	names of experiments in which reads supporting the interaction were found
Column 13	microRNA_first	1 if the order of fragments in chimera was "miRNA-mRNA"; 0 otherwise
Column 14	two_way_merged	1 if both "miRNA-mRNA" and "mRNA-miRNA" chimeras were found; 0 otherwise
Column 15	seed_type	type of seed (defined as in Figure 2A)
Column 16	num_basepairs	number of predicted basepairs between miRNA and mRNA
Column 17	seed_basepairs	number of predicted basepairs between nts 2-7 of miRNA and mRNA
Column 18	folding_energy	predicted minimum energy of interaction between miRNA and mRNA
Column 19	5'UTR	if mRNA fragment of chimera overlaps the 5' UTR; 0 otherwise
Column 20	CDS	if mRNA fragment of chimera overlaps the CDS; 0 otherwise
Column 21	3'UTR	if mRNA fragment of chimera overlaps the 3' UTR; 0 otherwise
Column 22	folding_class	miRNA-mRNA folding class
Column 23	conservation_score	Phylogenetic conservation of interacting region based on the PhyloP scores
Column 24	log2_target_enrichment	Enrichment of mRNA following depletion of 25 miRNAs in Hafner et al. (2010)
Column 25	CLASH_single_reads_ovlp	Number of non-chimeric reads overlapping mRNA fragment
Column 26	CLASH_cluster_ovlp	if cluster of single reads overlaps mRNA fragment, NA otherwise
Column 27	PAR_CLIP_cluster_ovlp	1 if cluster of reads from Hafner et al.

Table C.2 – CLASH dataset columns description with the most useful ones in bold; Source [Helwak et al., 2013]

Most miRNAs target more than one mRNA, or more than one position in a same mRNA, which explains why the number of attachment sites is much higher than the actual number of miRNAs or of mRNAs.

We used the first 9 columns in combination with the full sequences of the genome GRCh37 release 60. We generated the sequences of our test dataset by matching the mRNA name in the real sequences with the mRNA_name (Column 6), and matching the position mRNA_start (Column 7) and the mRNA_seq_extended (Column 9).

We generated two datasets using the first 9 columns. The first one is composed of the complete mRNA sequences. From the version of the human genome used in the CLASH paper, namely *GRCh37 release 60* downloaded from the *Ensembl* database, we selected only the sequences of interest from the cDNA by matching the name of the mRNA from CLASH (C.2–Column 6), and then extracting the mRNA sequences which contain an interaction region situated exactly where CLASH placed it (C.2–Column 7 and 9).

Figure C.1 – Extension of 200 nucleotides on each side of the target sites given by CLASH. Some interactions were not included because they are too close to the 3 or 5'end.

If several candidates from the cDNA corresponded to one same mRNA in CLASH, we selected the one with the longest sequence. In the process, two mRNAs from CLASH were not found and we ended up with a dataset of 18497 regions of interaction from 399 miRNAs targeting 7388 mRNAs.

The second dataset contains the precise regions of interaction as given by CLASH with an extension of 200 nucleotides before the start of the interaction region and 200 after its end. When such extension was not possible because the region of interaction was closer than 200 nucleotides from the 5' or the 3' end of the mRNA, we eliminated the sequence from the dataset. In the end, we have 15755 interactions (see Figure C.1).

The interaction sites were further divided into different sets, either based on the region of the mRNA where they are located, or on the class of interaction to which they belong. In relation to the regions, it is important to notice that although most often only the 3'UTR regions of the mRNAs are considered by the *in silico* methods for inferring the miRNA targets, approximately 5% of the interaction sites were reported within the 5'UTR regions and, surprisingly, 60% within the CDS regions, as compared to 35% within the 3'UTR.

In the same way, the interaction sites may be divided into five classes, three of which, namely Classes I, II and III, involve a seed at the 5' end of the miRNA. This involvement is either strict (Class I) or concerns the seed plus a region, of respectively 4 or 5 bases, at a shorter (Class II) or slightly longer distance from the seed (Class III). As concerns the other two classes, as indicated in the CLASH paper [Helwak et al., 2013], the binding is either limited to a region located in the middle or 3' end of the miRNA (Class IV), or is distributed and less stable (Class V). The proportion of targets in each class varies between approximately 17.8% for the smallest class (Class II) and 25% for the biggest one (Class III), see Figure C.2.

We wish to call attention to the fact that while the numbers obtained for the classes add up, as expected, to the number of miRNA-mRNA interaction sites considered, this is not necessarily the case when we consider the regions instead of the classes. The reason is that some interaction sites may be located between two regions, namely 5'UTR and CDS or CDS and 3'UTR, or may be found in non-coding RNAs that arise from pseudogenes as indicated in the paper on CLASH. When there exists a difference, it is a small one.

The number of sequences for each of the two datasets above, in total and divided by regions and classes, is given in Table C.3.

The same observation as previously made in our first datasets on *Mus musculus* and *Arabidopsis thaliana* (see C.1), namely that many miRNAs have several target sites can

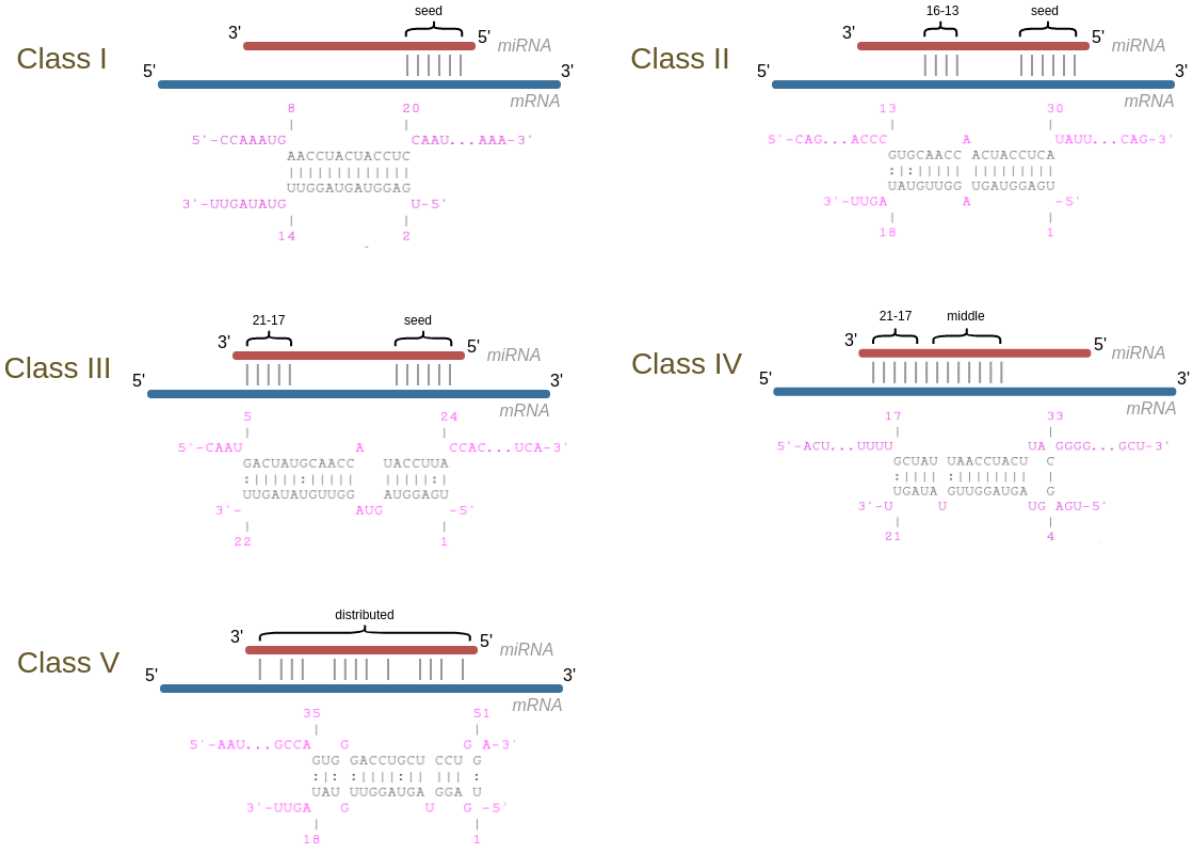


Figure C.2 – Schema of the five Human classes with examples of interactions on the miRNA *let-7a*. Interactions were computed using IntaRNA.

Number of sequences for dataset composed of all complete mRNAs								
Total	5'UTR	CDS	3'UTR	Class I	Class II	Class III	Class IV	Class V
18497	868	11112	6093	3590	3290	4629	3382	3606
Number of sequences for dataset composed of all CLASH interaction regions +-200 nucleotides								
Total	5'UTR	CDS	3'UTR	Class I	Class II	Class III	Class IV	Class V
15754	291	10304	4916	3095	2833	3970	2890	2966

Table C.3 – Number of sequences for each of the two datasets chosen, in total and divided by regions and classes.

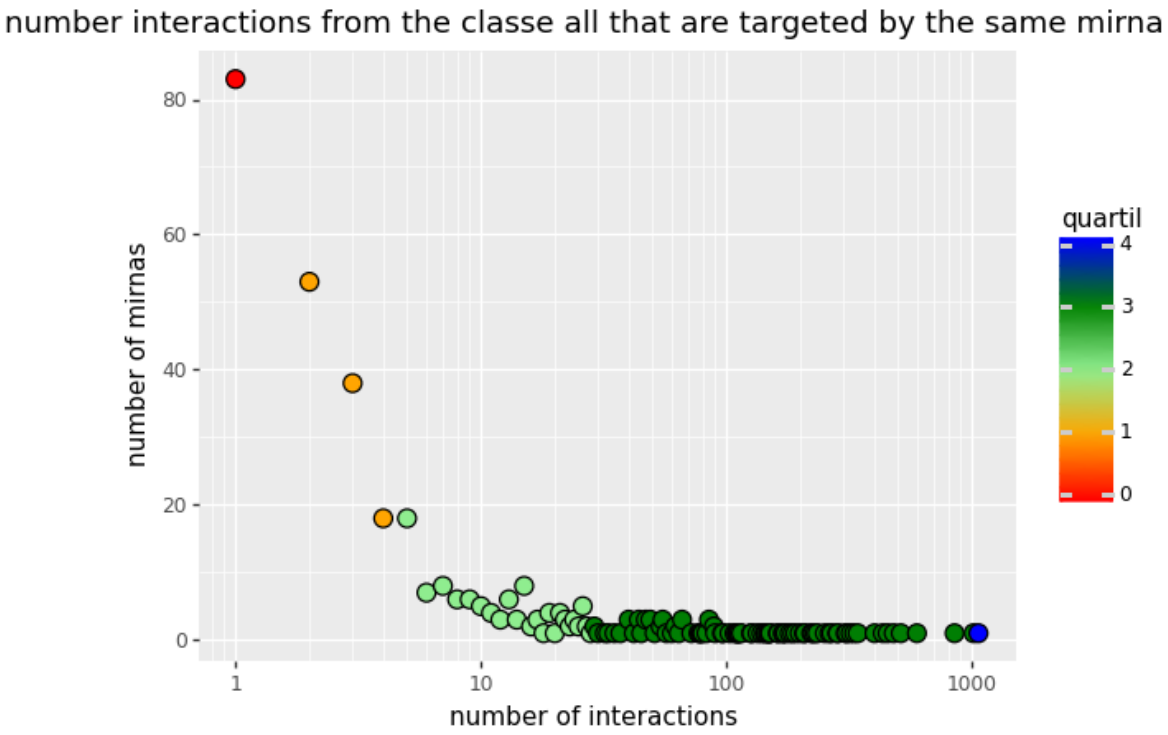


Figure C.3 – Histogram of the number of miRNAs by number of interactions.

be observed here, see Figure C.3. When considering all interactions from our CLASH dataset, we have an average of 46.40 interactions per miRNA with a 25%, median, 75% quantile respectively at 2, 5 and 28.5. This reinforces the interest of the idea of using intra-species conservation for target prediction.

An intriguing discovery with this dataset is the number of interactions that do not canonically bind on the seed region (Classes IV and V). Moreover, there is a majority of targets in the CDS region.

The authors of the paper on CLASH in human [Helwak et al., 2013] evaluate the function of each interaction using the database mirTarBase [Hsu et al., 2011]. The result is in harmony with the current knowledge that interactions with a seed match and interactions situated in the 3'UTR are the most upregulated. Nonetheless, interactions without a canonical seed match or interactions situated in the CDS are still half as efficient.

ii Second CLASH dataset

For further validation, we searched for other free datasets that precisely indicate the region of interaction on the mRNA and the miRNA. These constraints are, as far as we know, only available with the CLASH experimental method. We thus reviewed several published CLASH datasets [Moore et al., 2015, Bullard et al., 2019, Gay et al., 2018, Fields et al., 2021] and selected one of the datasets of *Mus musculus* from Moore [Moore et al., 2015]. Indeed, the other datasets were too specific and provided interactions for some miRNAs only, for instance *miRNA-320* in Fields [Fields et al., 2021], Kaposi's sarcoma-associated herpesvirus (KSHV)-involved miRNA in Gay [Gay et al., 2018], and Murine gammaherpesvirus-68 (MHV68) in Bullard [Bullard et al., 2019].

There are actually two datasets in [Moore et al., 2015]: the first one on *human* miRNA-mRNA interactions which provides 32712 interactions of 542 miRNAS on 7070, and the second one on *Mus musculus* with 130119 interactions of 614 miRNAs on 11748 genes. We selected the dataset on *Mus musculus* due to the higher number of available interactions compared to those available for humans.

However, the interactions are aligned only in the chromosome region and not on the mRNA as in the case of the Helwack CLASH datasets. Moreover, some interactions are located in intergenic regions or inside exons. Therefore, a process of finding the matching mRNA from the annotation data and then cleaning the interaction not found is mandatory (see Table C.4). We proceeded by retrieving the sequence of the interactions through a matching of the chromosome name and position on the full-chromosome sequence (version 54 [Kinsella et al., 2011]). We then retrieved the matching transcript names with annotation data that link the transcript names and their position on the chromosome (from version 54 [Kinsella et al., 2011]). We obtained for each transcript name, their complementary DNA (cDNA) sequences (from Biomart version 54 [Kinsella et al., 2011]) and made sure that the sequence interaction retrieved from the full chromosome could be identically aligned on the transcript sequence.

Column name	Column description
cluster.ID	unique chimera cluster ID in format chromosome_start_end_miRNA
chr	chromosome
start	chromosome start position of chimera cluster (hg18)
end	chromosome end position of chimera cluster (hg18)
N	"cluster size", number of individually cloned events comprising a chimera cluster
strand	chromosome strand
region	transcript region of interaction
gene.id	Entrez gene ID for cluster
gene.symbol	Offical gene sybol for cluster
miRNA	ligated microRNA for chimera cluster
miR.map	RNAhybrid duplex structure prediction for miRNA, where "|"= Watson-Crick base-pairing; "W": G-U wobble pairing; "-": mismatched or unpaired sites
target.map	RNAhybrid duplex structure prediction for target
MFE	RNAhybrid predicted minimum free energy for duplex structure
duplex.start	Predicted start position of duplex structure in 75 nucleotide target region (where position 1 is the first nucleotide of the chimera cluster)
seed match	Strongest seed match within 75 nucleotide target region (where position 1 is first nucleotide of chimera cluster). "mm": mismatch and "indel": insertion or deletion in target motif (resulting in bulged nucleotide)
seed position	Position of seed match within 75 nucleotide target region (where position 1 is first nucleotide of chimera cluster)
bulged miRNA site	Position in miRNA (where position 1 is miRNA 5' end) of bulged miRNA nucleotides. Only applicable to sites with seed matches with single nucleotide deletions.
bulged target site	Position with respect to miRNA of bulged target nucleotides (e.g. 2,3 indicates a bulged target nucleotide between miRNA positions 2 and 3). Only applicable to sites with seed matches with single nucleotide insertions.
mismatch seed site	Position in miRNA (where position 1 is miRNA 5' end) of seed mismatch, when present. For sites with bulged or mismatched motifs, involved nucleotides are indicated.
mismatched or bulged nucleotide	For bulged motifs, the bulged miRNA or target residue is indicated. For mismatched motifs, the atypical base pairing is indicated as "miRNA nucleotide,target nucleotide."
k.group	Interaction cluster, according to k-means clustering of duplex structures.

Table C.4 – Second CLASH dataset columns description, Source [Moore et al., 2015]

At this step, we had 13344 interactions linked with an average of 2.11 transcripts and a maximum of 24 transcripts for one interaction. The presence of multiple transcripts can be attributed to alternative splicing, which results in different isoforms of the same gene. To address this issue, we selected the longest transcript for each interaction.

A final cleaning step was to remove the interactions which were, according to Moore, located elsewhere that in the 3'UTR, CDS or 5'UTR. Interaction regions had to be one of:

✓ CDS

✓ CDS||5'UTR

✓ 5'UTR

✓ 3'UTR

✓ CDS||intron

✓ CDS||3'UTR

✓ 5'UTR||intron

✓ 3'UTR||intron

✓ 5'UTR||3'UTR

✓ CDS||3'UTR||intron

We excluded all of these regions:

✗ intron

✗ deep_intergenic

✗ downstream_10k

✗ exon_unclassified

✗ downstream_10k||upstream_10k

✗ upstream_10k

✗ exon_unclassified||downstream_10k

✗ CDS||5'UTR||3'UTR

✗ exon_unclassified||upstream_10k

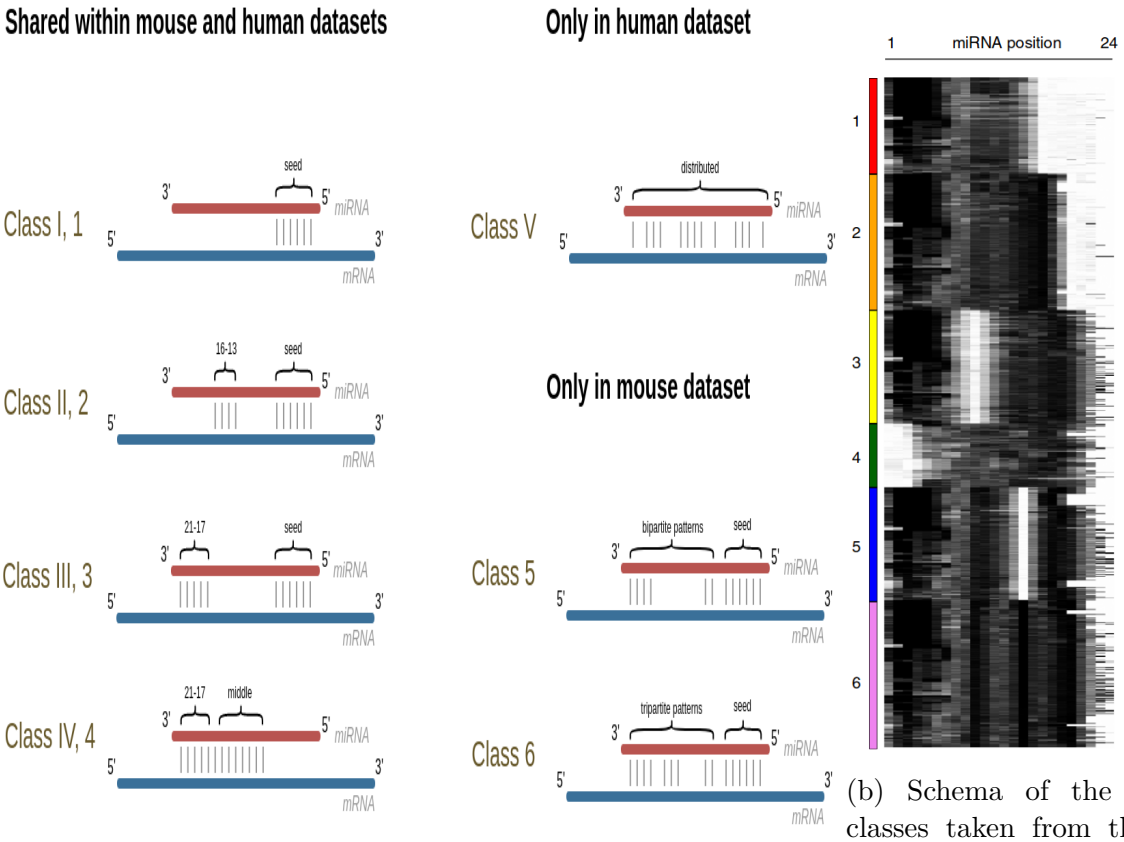
✗ CDS||5'UTR||intron

✗ exon_unclassified||downstream_10k||upstream_10k

One can argue about the exclusion of CDS||5'UTR||intron, intron and CDS||5'UTR||3'UTR, however this decision was made to have data similar to our first CLASH dataset. After the cleaning, the dataset on *Mus musculus* contains 13071 interactions of 378 miRNAs on 2462 mRNAs.

The authors of the dataset on *Mus musculus* clustered the interactions based on the interaction pattern as in the first CLASH dataset. Six different classes were found from the k-clustering where only Classes 1 to 4 are similar to Classes I to IV of our first CLASH dataset. Class V with a distributed and less stable interaction is not found in the Moore dataset. Classes 5 and 6 from Moore show interactions involving the seed and a bi or tripartite auxiliary pairing pattern (see Figure C.4. Class 5 exhibits a distinct non-paired gap positioned at the nucleotide 15 and around.

Similarly to the first CLASH data, two sub-datasets were created by gathering all the complete mRNAs and by extending by 200 nucleotides before the start of the interaction region and after the end of such region. Some interactions too close to the end cannot be extended and therefore we did not include them (see Figure C.1). Table C.5 summarises the final number of sequences.



(a) Similarities between human clash dataset classes and mouse classes.

(b) Schema of the 6 classes taken from the CLASH paper [Moore et al., 2015].

Figure C.4 – Clash Mouse dataset

Number of interactions										
Total	3'UTR	CDS	5'UTR	other regions	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
13071	4843	7136	187	905	1631	2672	2269	1117	2363	3019
Number of sequences for dataset composed of all CLASH interaction regions +-200 nucleotides										
Total	3'UTR	CDS	5'UTR	other regions	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
11581	3817	6917	102	745	1428	2366	1987	976	2129	2695

Table C.5 – Number of interactions for each of the two datasets chosen, in total and divided by regions and classes.

iii Dataset nomenclature

In order to facilitate the comprehension of the reader on which dataset we are probing, we defined the following nomenclature which is identical for both mouse and human datasets:

— Seq_E – refers to the site of interactions as given by the CLASH experiment.

— Seq_W – refers to the site of interactions extended by 200 nucleotides on both ends as described in Figure C.1

— Seq_C – refers to the complete mRNA.

Note that the amount of noise increases as we go from the first to the last of these datasets.

Chapter D

Contribution: Revisiting accessibility and seeds usage

The accessibility of mRNA sequences is an important factor in their interaction with miRNAs (see Section iv for details about accessibility).

The debate revolves around the exact mechanisms by which a miRNA and a mRNA interact, and that may influence the efficiency and specificity of such interaction.

One aspect of such debate centers on the role of the mRNA secondary structure in the miRNA–mRNA interactions. Many methods of miRNA–mRNA interaction prediction [Mann et al., 2017, Gumienny and Zavolan, 2015, Kertesz et al., 2007, Agarwal et al., 2015] use accessibility whereas others [Miranda et al., 2006, John et al., 2004] do not. A 2021 study [Kern et al., 2021] that compared 88 miRNA target predictors concluded that miRanda (that predicts miRNA–mRNA interactions without accessibility) and TargetScan (that predicts miRNA–mRNA interactions with accessibility) outperformed other tools. The question of the accessibility in the miRNA–mRNA prediction thus persists and we investigated its importance in this chapter by comparing the predictions of three different tools: miRanda [John et al., 2004], Pita [Kertesz et al., 2007], and IntaRNA [Busch et al., 2008, Mann et al., 2017]. The last two do use accessibility in their prediction while as stated above, miRanda does not.

Another area of debate concerns the importance of the miRNA seed sequence complementarity to target mRNA sites. While it is generally accepted that the seed is critical for target recognition and binding, most of the studies are performed in the 3'UTR region and on partially biased data that rely on *in silico* predictions. Moreover, according to both the mouse and human data sets, there is around one fifth of the number of interactions that are seedless. We therefore decided to explore further this aspect also.

This chapter appeared in its majority in the *Genes* paper accessible in open access at this ULR: `https://www.mdpi.com/2073-4425/14/3/664`. It is organised following a classical paper structure with first the methods part presented followed by the results. Notice that the chapter contains additional results that were not included in the paper.

D.1 Methods

In this section, we present the methods used to conduct the different studies we did. One such study revolves around benchmarking some of the tools for predicting the miRNA–mRNA interaction sites with an emphasis on the usage of both the accessibility and seed features. The other study analyses accessibility directly on mRNAs with different approaches.

i Revisiting accessibility through CLASH data

There are many methods for inferring miRNA–mRNA interactions. However, as we wanted to conduct our study using experimentally validated data, namely the CLASH human and mouse data presented in Section C.2, this implied that, for any given method, we could run it ourselves on such data, which meant that the method should be publicly available. Furthermore, since our objective was to focus exclusively on two main features, namely accessibility and the seed, we decided to choose among such methods those that do not use many more features or ones that would introduce noise. Three main methods, which are also among the better-known ones, were thus identified and chosen. These are miRanda [John et al., 2004], Pita [Kertesz et al., 2007], and IntaRNA [Busch et al., 2008, Mann et al., 2017]. Notice that the last one, IntaRNA, is not specific to miRNA–mRNA interactions. It was indeed originally developed to identify sncRNA–mRNA interactions in bacteria. A fourth method, also very well known, could have been considered, as it published in its original paper a study based on the same CLASH data as we are using here. This is TargetScan [Agarwal et al., 2015]. The method, however, uses many other features, namely 26, including inter-species evolutionary conservation. We therefore put it aside to allow us to exclusively focus on the importance of the seed match and accessibility within our results.

Before briefly presenting the main characteristics of each of the methods selected, it is important to remember that we have one main motivation for this study, which is the issue of accessibility: is this an interesting feature or not? Two of the methods chosen use this feature, namely Pita, whose development was indeed strongly motivated by its importance for the authors (as reflected in its name, which stands for Probability of Interaction by Target Accessibility), and IntaRNA. The third method, miRanda, is thus the only one that does not explicitly compute the accessibility of the mRNA.

Moreover, we would like to observe that all methods, including the three that we chose to use, often include more general tools as modules within their algorithm; these tools are publicly available. Many belong to the ViennaRNA Package [Lorenz et al., 2011] and enable addressing some basic problems. These are related to the main features that are considered when trying to infer miRNA–mRNA interactions.

These main features concern:

— The seed match. We recall that this is the problem of finding the best binding site between two RNAs. It is addressed by adopting a sequence alignment method, albeit with some adaptation, the main of which is that sequence complementarity and not sequence identity is considered. We recall that in the case of miRNAs, the seed starts at nucleotide 2 from the 5′ end of the miRNA and is of a length between 6 and 8.

— The minimum free energy of the interaction, or what was called the stability of the duplex formed between the miRNA and the mRNA in the introduction. This corresponds to the secondary structure between miRNA and target mRNA that has the lowest value of free energy required to unfold it. Some methods further work within a nearest-neighbour free energy model, where dinucleotides instead of single nucleotides are considered [Turner and Mathews, 2010]. The way free energy is computed may differ importantly between different methods [Kruger and Rehmsmeier, 2006, Tafer and Hofacker, 2008].

— The accessibility of the interaction site. This is related to the free energy that would be required to open the secondary structure of the mRNA [Kertesz et al., 2007, Marin and Vanicek, 2011].

We then present below the main characteristics of the three methods that were used in our study:

— PITA: Among the methods for predicting the targets of a miRNA, PITA was a pioneer in considering accessibility when attempting to carry out such predictions. With that aim in sight, after identifying miRNA interaction sites using complementary sequences to the seed, it computes an energy score $\Delta\Delta G$, which is equal to the difference between ΔGduplex and ΔGopen, the first being the energy for binding both RNAs together, which is computed with a modified version of RNAduplex, while the second is the energy required to open a region and is computed using RNAfold [Hofacker et al., 1994] and a sliding window of 70 additional nucleotides before and after the predicted target region.

— INTARNA: Similar to PITA, INTARNA predicts RNA–RNA interactions using the accessibility of the target sites. INTARNA actually combines two energy scores, the first one representing the target accessibility computed with RNAplfold [Hofacker et al., 1994], and the second one corresponding to the energy of the RNA–RNA hybridisation computed as in RNAhybrid [Kruger and Rehmsmeier, 2006]. Again, as with PITA, INTARNA does look for complementary sequences to the seed; however, contrarily to PITA, it does this in a third step only.

— MIRANDA: MIRANDA is the only method among the three chosen that does not consider the accessibility of the target region. The method proceeds by first searching for complementary sequences with an emphasis on binding the 5' region of the miRNA (region of the seed) and then by computing the energy of the pairing using RNAduplex.

In order to revisit accessibility and seed match, we first evaluated whether or not the methods considering these features were better predictors than those not including them. In a second step, we focused on INTARNA, which was used alone to allow us to consider or not each one of these features while computing predictions of miRNA–mRNA interactions.

ii Performance Evaluation

For each method, the performances were evaluated in terms of precision, recall and F-score metrics, considering the different numbers of chimeric reads overlapping the interaction sites. As a reminder, each chimeric read contains the mRNA and miRNA sequences

involved in the interaction. On the other hand, we wanted to be very precise and thus decided to initially consider –in both cases of datasets– a position as a **strong true positive**, TP_{strong} for short, when the interaction site predicted by the method is **fully included within the site given by CLASH**. It may also happen that the predicted site overlaps the site given by CLASH without being fully included in it. This may be seen as a weaker type of true positive, and we therefore chose to also consider it separately. We named it a **weak true positive**, TP_{weak} for short.

We then called false positives, FP for short, all predictions made that are not TP_{strong}. Note that FP is the same for both TP_{weak} and TP_{strong}.

In the same way, there are different types of false negatives that may be considered for each miRNA–mRNA couple given as input. One will be a strong false negative in the sense that no interaction site whatsoever is detected. We denoted it by No-Prediction for short. We also have weaker types of false negatives when a prediction is made that does not fully fall within the interaction site as given by CLASH; it either overlaps such a site or is outside it (however by force corresponds to the right couple). These weaker false negatives plus the strong ones add up to what we named **all false negatives**, FN_{all} for short.

Given these distinctions between strong and weak true positives, we computed two types of precision (Equation D.1), recall (Equation D.2) and F-score (Equation D.3).

$$\text{Precision-strong} = \frac{TP_{strong}}{TP_{strong} + FP} \quad \text{Precision-overlap} = \frac{TP_{strong} + TP_{weak}}{TP_{strong} + FP} \tag{D.1}$$

$$\text{Recall-strong} = \frac{TP_{strong}}{TP_{strong} + FN_{all}} \quad \text{Recall-overlap} = \frac{TP_{strong} + TP_{weak}}{TP_{strong} + FN_{all}} \tag{D.2}$$

$$\begin{aligned} \text{f-score-strong} &= 2\frac{\text{recall-strong} \times \text{precision-strong}}{\text{recall-strong} + \text{precision-strong}} \\ \text{f-score-overlap} &= 2\frac{\text{recall-overlap} \times \text{precision-overlap}}{\text{recall-overlap} + \text{precision-overlap}} \end{aligned} \tag{D.3}$$

iii mRNA accessibility

We used two approaches to compute the mRNA accessibility. In the first one, we defined mRNA accessibility to be related to the number of times a nucleotide was unpaired in all the optimal structures it belongs to, with each structure being computed inside a window of 150 nt sliding along Seq_W. In order to compute such structures, we used RNAfold from the ViennaRNA package. The output of RNAfold gives the minimum free energy of the optimal secondary structure as well as the structure in bracket notation. Using this notation, we can then count the number of times a nucleotide is paired or not when we slide the window, moving it nucleotide per nucleotide until the end of the mRNA sequences given as input. The second approach adopts another well-used definition of accessibility, which is implemented in RNAplfold [Hofacker et al., 1994]. The accessibility values of both methods were normalized in the Section D.2 below to be in the range of 0 to 100.

The first approach of accessibility is a naive one compared to the current state of art such as RNAplFold. However, it is meant to explore a simpler and step-by-step process of computing accessibility, that leaves room for potential improvement in computing power and precision. In that sense and also hoping for such improvement, we compared the behaviour of the minimal free energy (mfe), the partition function (pf) and the maximum expected accuracy (mea) which are gradually more computationally expensive. The minimal free energy corresponds to the most energetically favourable structure and hence the most probable one. The maximum expected accuracy, as defined by [Do et al., 2006] takes into consideration very frequent sub-optimal structures instead of systematically the best pairing one. The partition function explores all possible structures. We refer to Section i for more computational details and explanations about these three approaches for computing accessibility.

The comparison of the mfe, mea and pf parameters is represented in Figure D.1, where we display for each nucleotide the average free energy of all sliding windows with the three different structures originated from the different accessibility calculations. The free energy of the window from i to j is represented in the middle position $\frac{i+j}{2}$. The three approaches are very similar. If we study the difference between mfe and pf position by position, we have a nearly constant difference. The mean of the differences range from 1.441 to 1.453 according to the classes while the standard deviation of the differences ranges from 0.009 to 0.012. We thus judged these three approaches to be close enough to only consider one of them and chose the one that is the fastest, namely mfe.

Another question concerned the size of the sliding windows inside which the secondary structure is computed. All the methods that compute or use the concept of accessibility tend to adopt various window sizes depending on whether to favour speed when computing accessibility on the entire length of the mRNA or precision with a potentially longer-range base-pairing. This question is the focus of the papers [Lange et al., 2012, Tafer et al., 2008] which use RNAplfold. RNAplfold denotes by W the length of the window, L the length of the maximum span between two base pairings and u the width of the accessible site. The paper [Tafer et al., 2008] finds the most significant window length to be 80, the maximum span range to be 40 with the width of the accessible site being either 8 or 16. The paper [Lange et al., 2012] on the other hand concludes with these statements:

— The optimal base pairing span is $L = 150$ or $L = 100$ depending on the dataset. However, the difference in performance is small when L varies from 100 to 150 nucleotides.

— The window size equivalent to the length of the base pairing span introduces a bias and should be discouraged. However, this problem is partially resolved when using $W = L + 50$

We decided to follow the results presented in [Lange et al., 2012] which are closer to the default parameters of IntaRNA. We thus chose for our usage of RNAplfold a window length of $W = 150$ and a base pairing span of $L = 100$.

D.2 Results

In this section, we present the results of the study of the accessibility and seed anchoring.

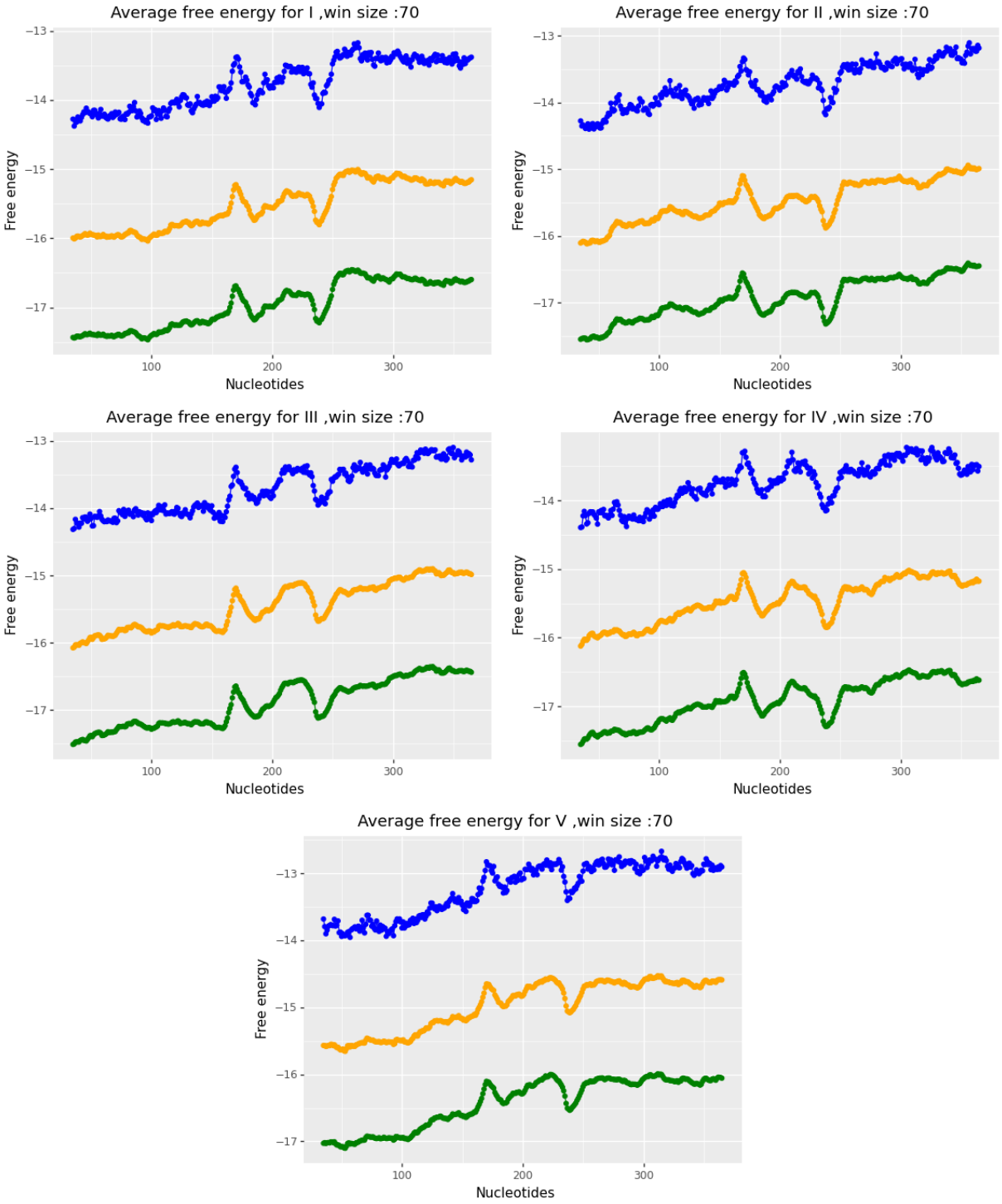


Figure D.1 – Comparison of the average free energy of a sliding window of 70 nucleotides of the minimum free energy in orange, the partition function in green and the maximum expected accuracy in blue computed on the human Seq_W classes.

i Interaction prediction from the CLASH datasets using three methods

The results obtained for the human and mouse CLASH datasets (introduced in Chapter C.2) in each of the two cases of input considered, namely Seq_C or Seq_W, are presented in Figure D.2 for all regions and classes together and in Figures D.3,D.4,D.5 and D.6 for each region and each class.

We observe that the results obtained by any method on the Seq_C dataset are always worse in terms of precision or recall than those obtained when considering Seq_W, and this occurs in a rather important way (see Figures D.2). This is expected, as the amount of noise in the first case is, as mentioned in C.2, much higher. There indeed, we consider the whole sequence of the mRNAs for which at least one miRNA–mRNA interaction site has been validated by CLASH, while in the second, we consider only a small region around such interaction sites.

For the mouse dataset, the results improve as the number of reads increases.

Considering classes and regions, the poorest results overall are, in general, obtained for Classes IV and V for the human data as well as Classes 4 and 5 for the mouse data (see Figures D.3,D.4,D.5 and D.6). This is expected as these classes are supposed to correspond to miRNAs that bind to the mRNA not through a seed, not even necessarily in a localised or strongly stable manner. For both datasets, the 3'UTR regions always exhibit the best results in terms of F-score.

Interestingly, concerning accessibility, miRanda, which does not consider this as one of its features, has, with a few exceptions, the worst behaviour in relation to both Pita and IntaRNA, more substantially in terms of F-score and recall, whatever the type, than in terms of precision.

This appears to support the idea that accessibility is an important feature, although some caution should be taken as there are other differences among the methods analysed here, even when considering the same feature as the exact parameters used for such may differ. The order in which the features are considered, which is not always the same for all three methods, may also have an influence on the results. In any case, as concerns accessibility, this explains why we decided to refine our study, as presented in the next section.

ii Study of the accessibility of the interaction site and seed match with IntaRNA

Before proceeding with a direct study of the accessibility of the interaction site, we decided to refine one of the studies made in the previous section, related to the method that in almost all cases produces the best results on the CLASH data, namely IntaRNA.

Such refinement concerned eliminating accessibility or seed matching. This is indeed possible with the new version of IntaRNA published in 2017 [Mann et al., 2017]. Results are presented in Figure D.7. Here, over-start and over-end indicate the number of predictions made that were not fully within the interaction site but overlapped it, either at the start or at the end. The term wo in the figure stands for without. To facilitate the comparison, we indicated again the results obtained by IntaRNA with its default parameters.

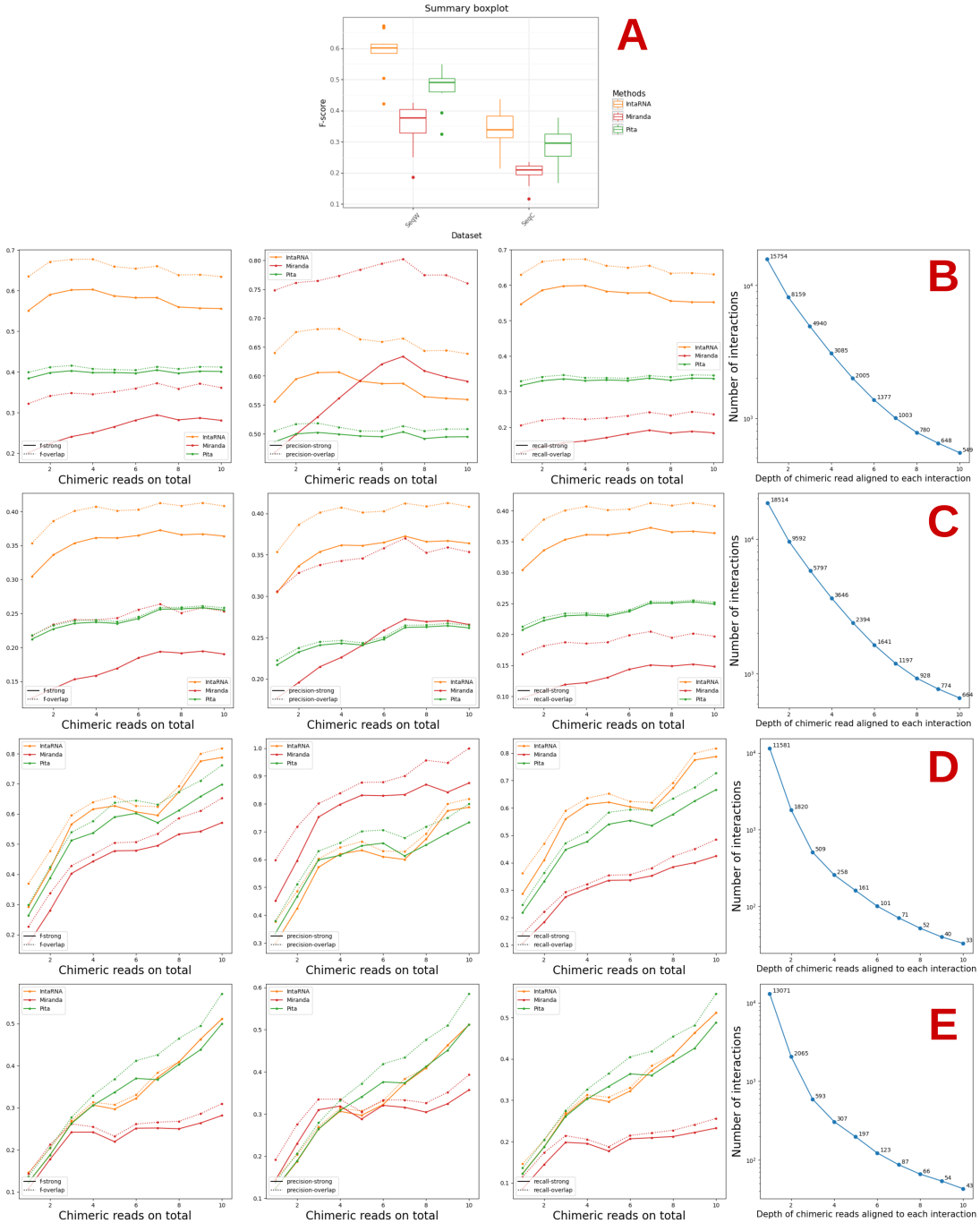


Figure D.2 – F-score, precision and recall for miRanda, Pita, and IntaRNA used with default parameters. (**A**) summarises the results of the combined human and mouse datasets. (**B**,**C**) and, resp., (**D**,**E**) show the results on the Seq_W and Seq_C human, resp. mouse, datasets.

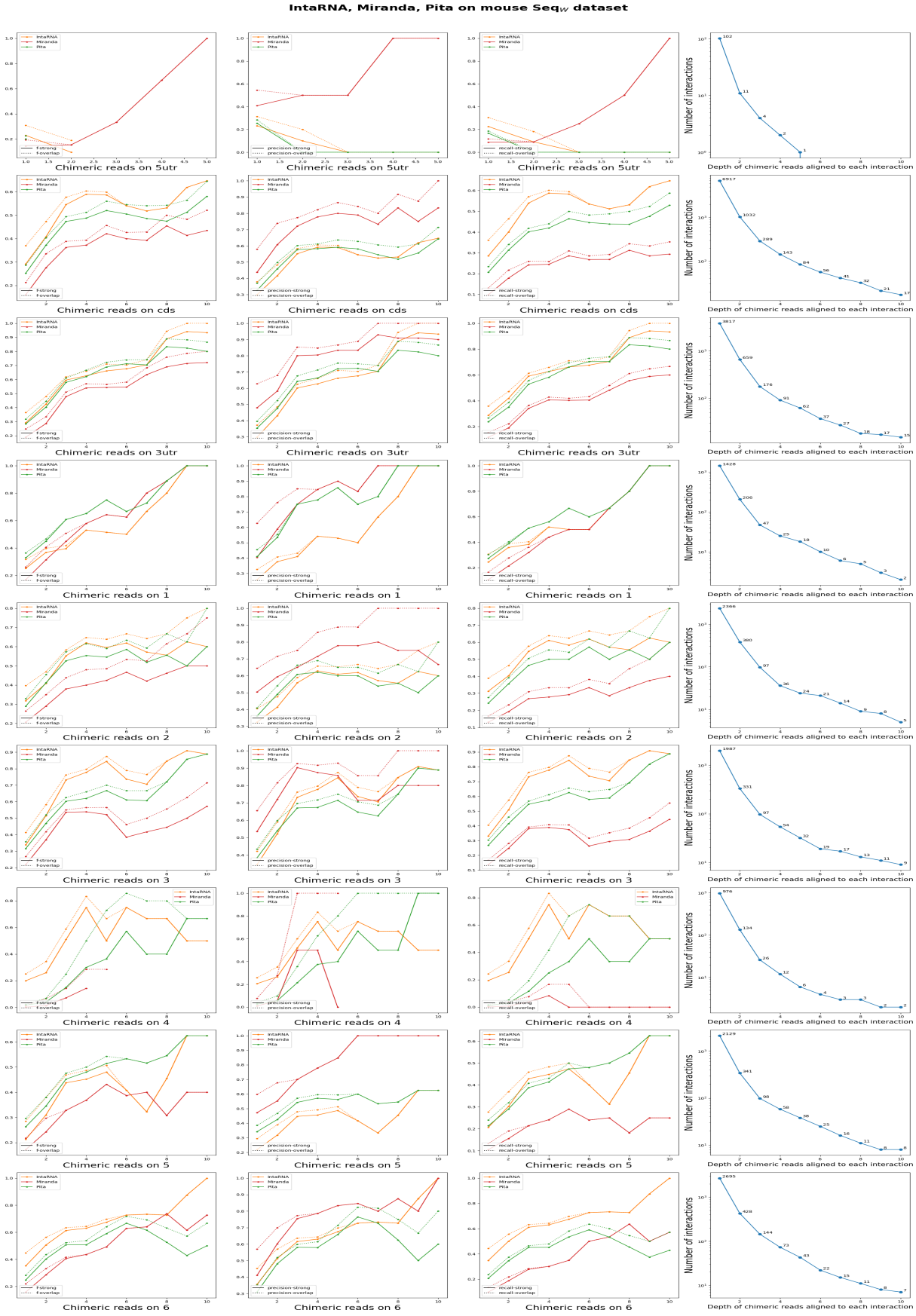


Figure D.3 – Performance of the three methods considering accessibility and seed match on the mouse Seq$_W$ datasets.

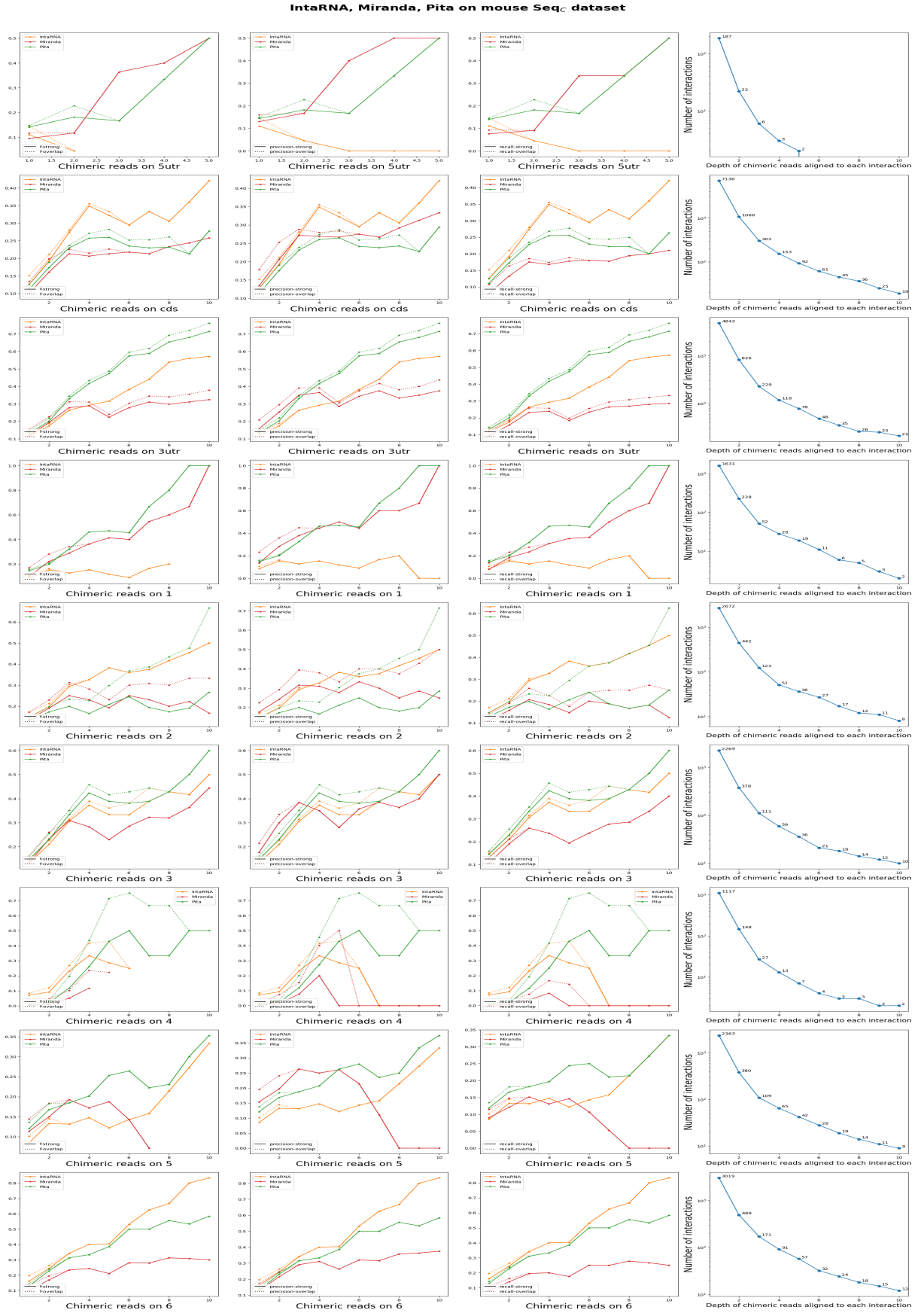


Figure D.4 – Performance of the three methods considering accessibility and seed match on the mouse Seq$_C$ datasets.

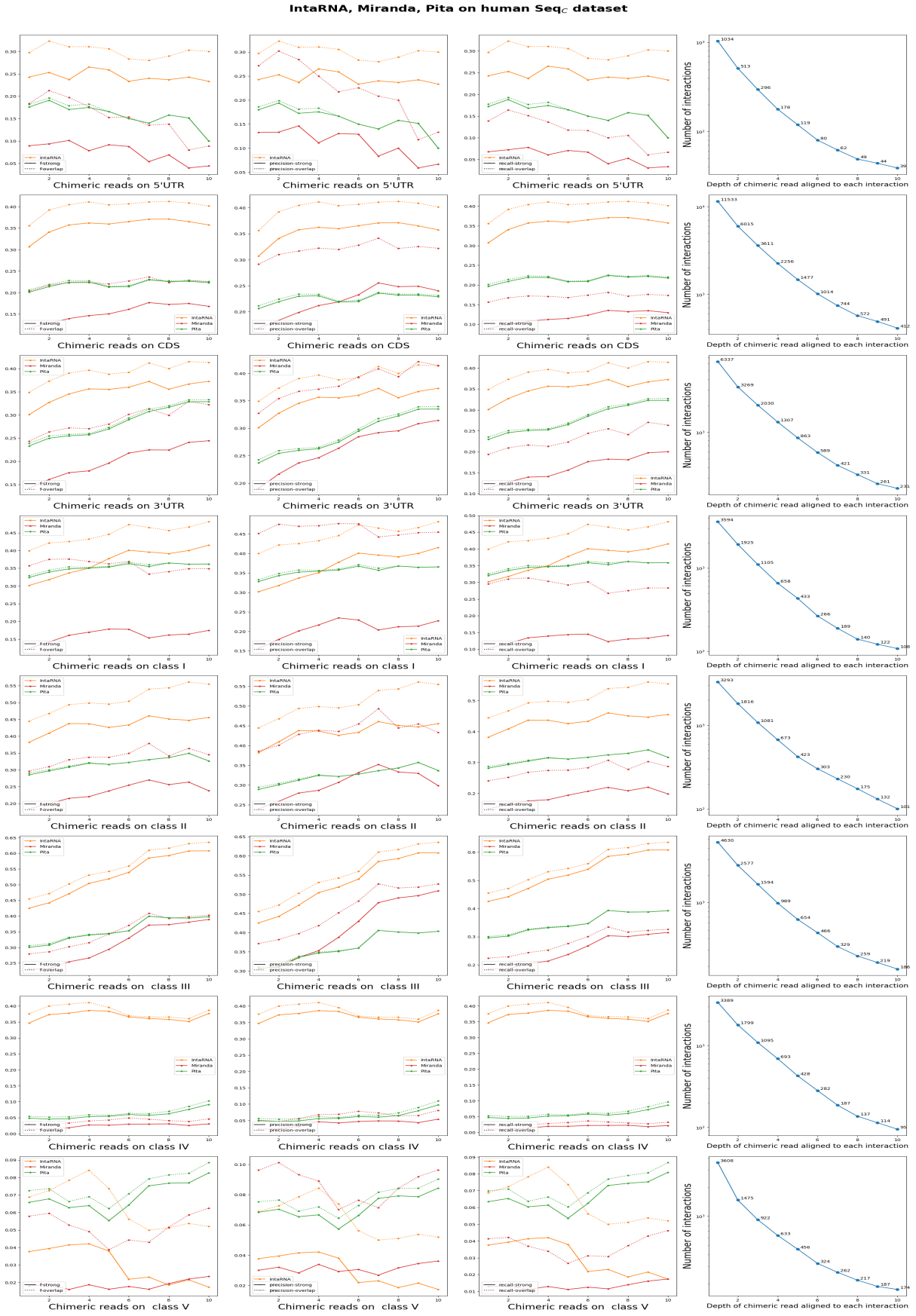


Figure D.5 – Performance of the three methods considering accessibility and seed match on the human Seq$_W$ datasets.

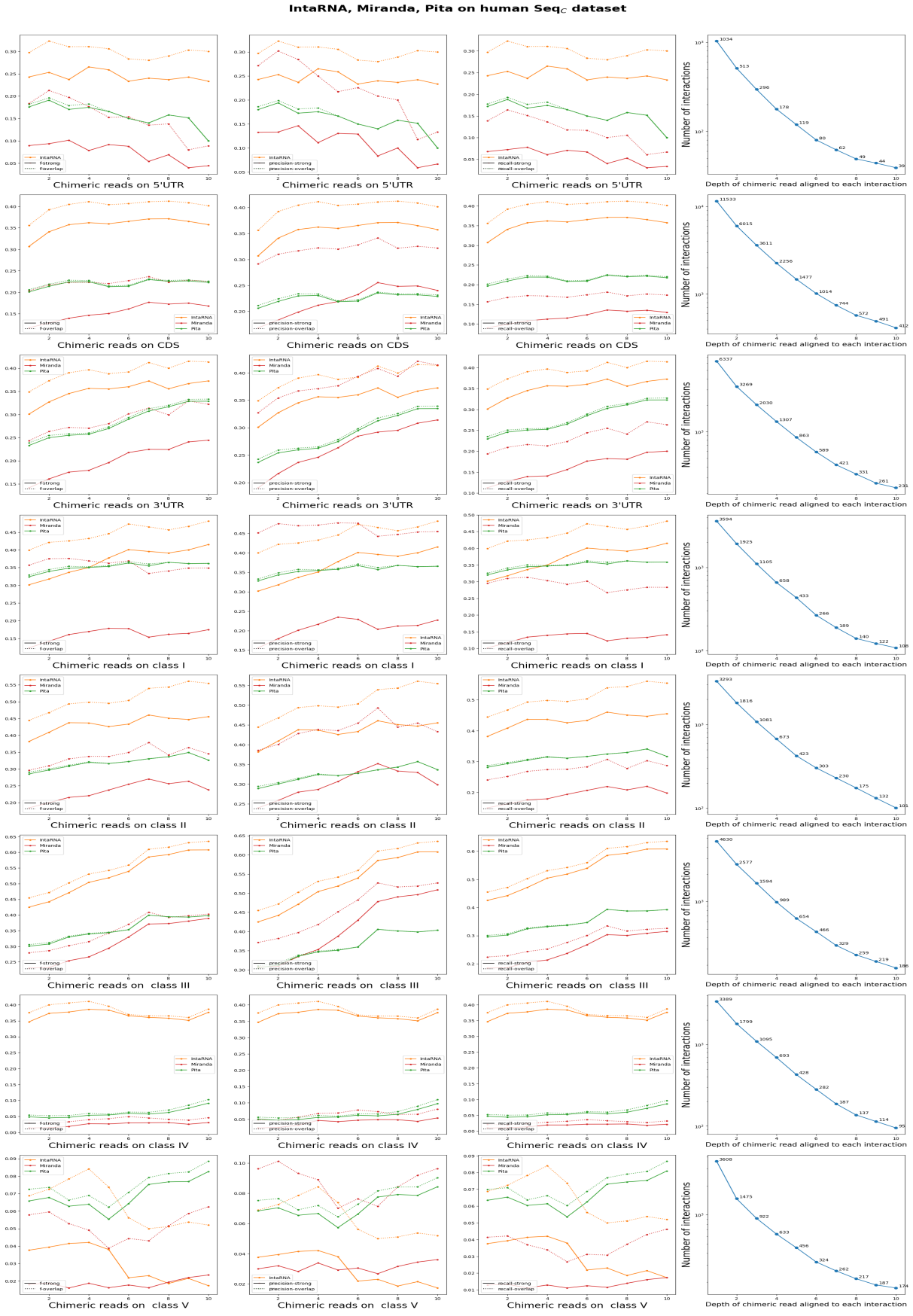


Figure D.6 – Performance of the three methods considering accessibility and seed match on the human Seq$_C$ datasets.

For both the human and the mouse datasets, the best results are obtained when considering the accessibility (Figure D.7). There are rare exceptions in the case of the human data for Seq_C (Precision-overlap for the 5'UTR region) and five for Seq_W (Recall-strong for Total, 5′UTR, CDS, 3′UTR, and Class V; see Figures D.8,D.9,D.10 and D.11). At least in relation to IntaRNA, this reinforces the idea that accessibility is an important feature and should be taken into account in the prediction of the interaction sites.

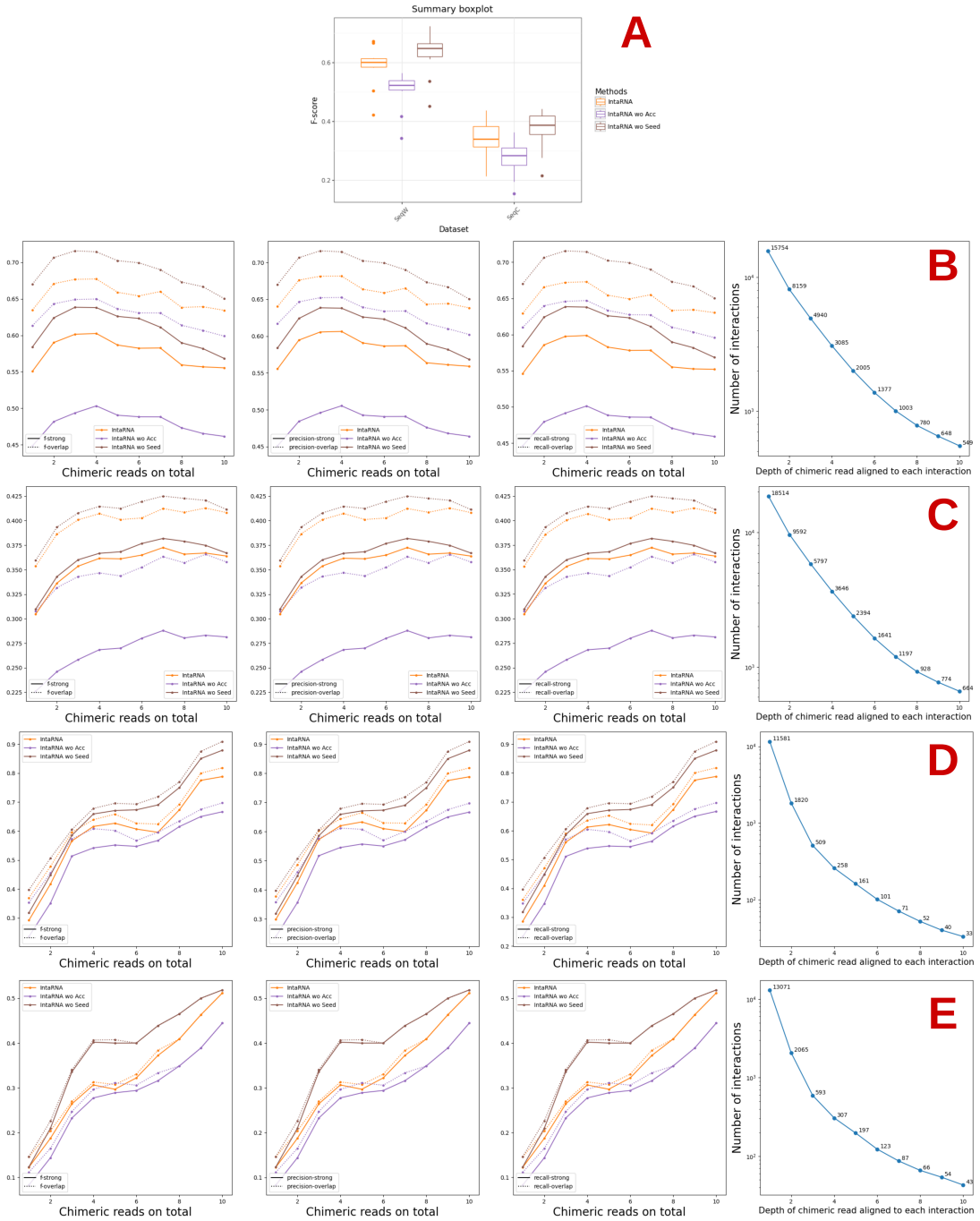


Figure D.7 – Performance of IntaRNA considering or not accessibility and seed match. The term wo in the Figure stands for WithOut. (**A**) summarises the results of the combined human and mouse datasets. **B**,**C**) and, resp., (**D**,**E**) show the results on the Seq_W and Seq_C human, resp. mouse, datasets.

While the worst results are obtained when not considering the accessibility, not considering the seed gives either the same results as IntaRNA with default parameters or, in some cases, better ones. There are a few exceptions for Classes 1 and 6 on Seq_W of the mouse data, where the results without the seed became the worst when considering interactions supported by at least 6 and 8 chimeric reads, respectively (see Figures D.8

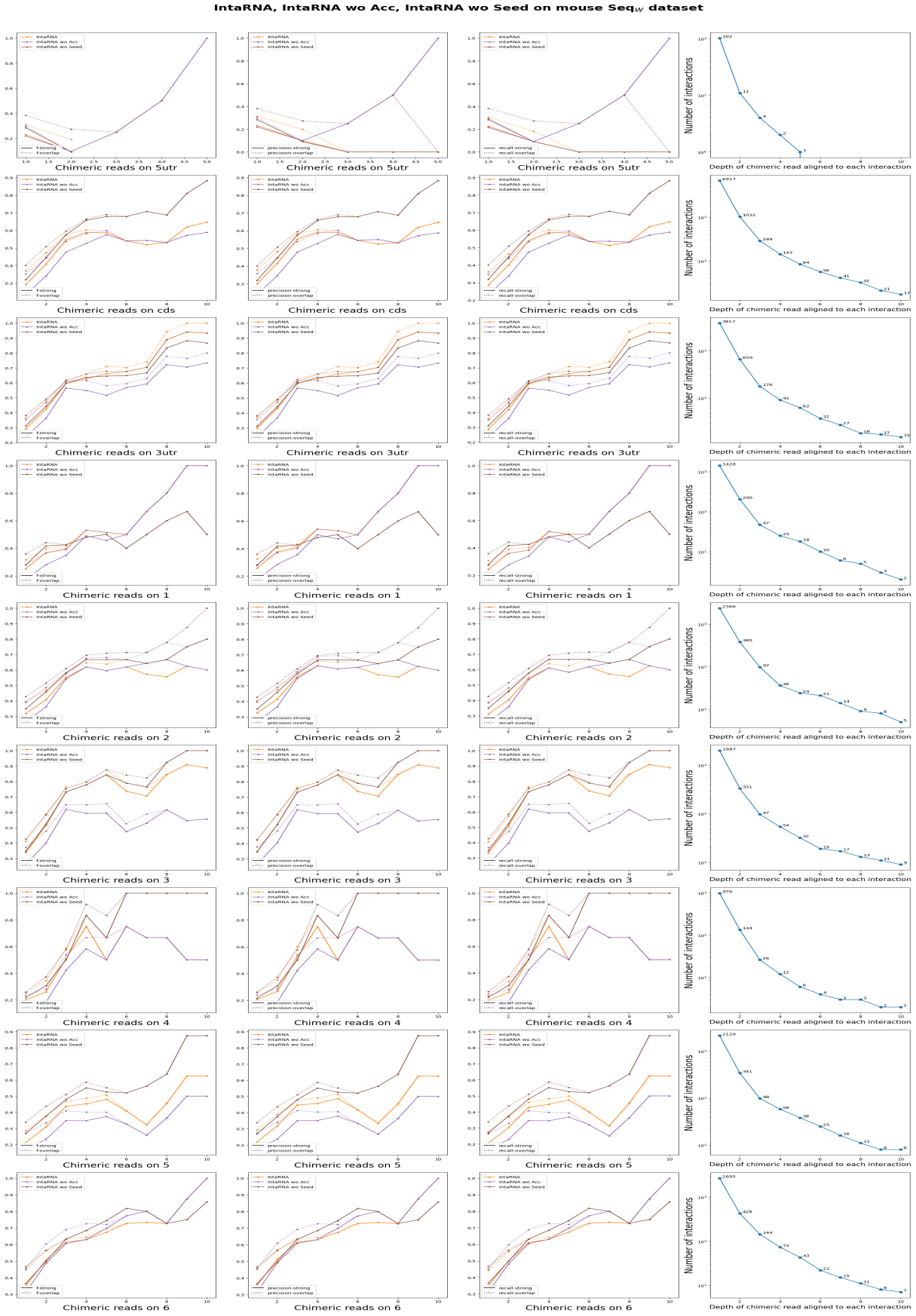


Figure D.8 – Performance of IntaRNA considering or not accessibility and seed match on the mouse Seq$_W$ datasets. The term wo in the Figure stands for WithOut.

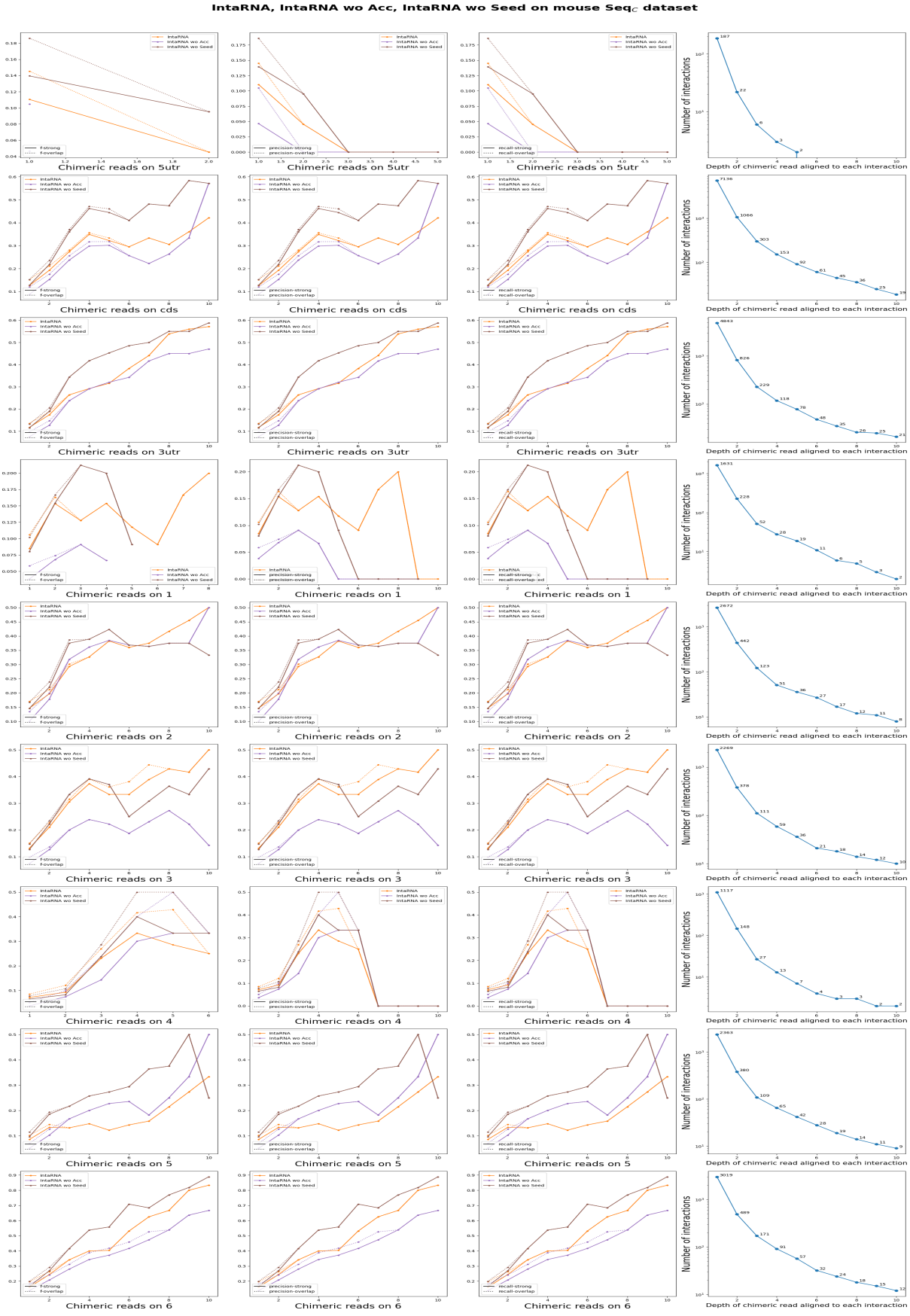


Figure D.9 – Performance of IntaRNA considering or not accessibility and seed match on the mouse Seq$_C$ datasets. The term wo in the Figure stands for WithOut.

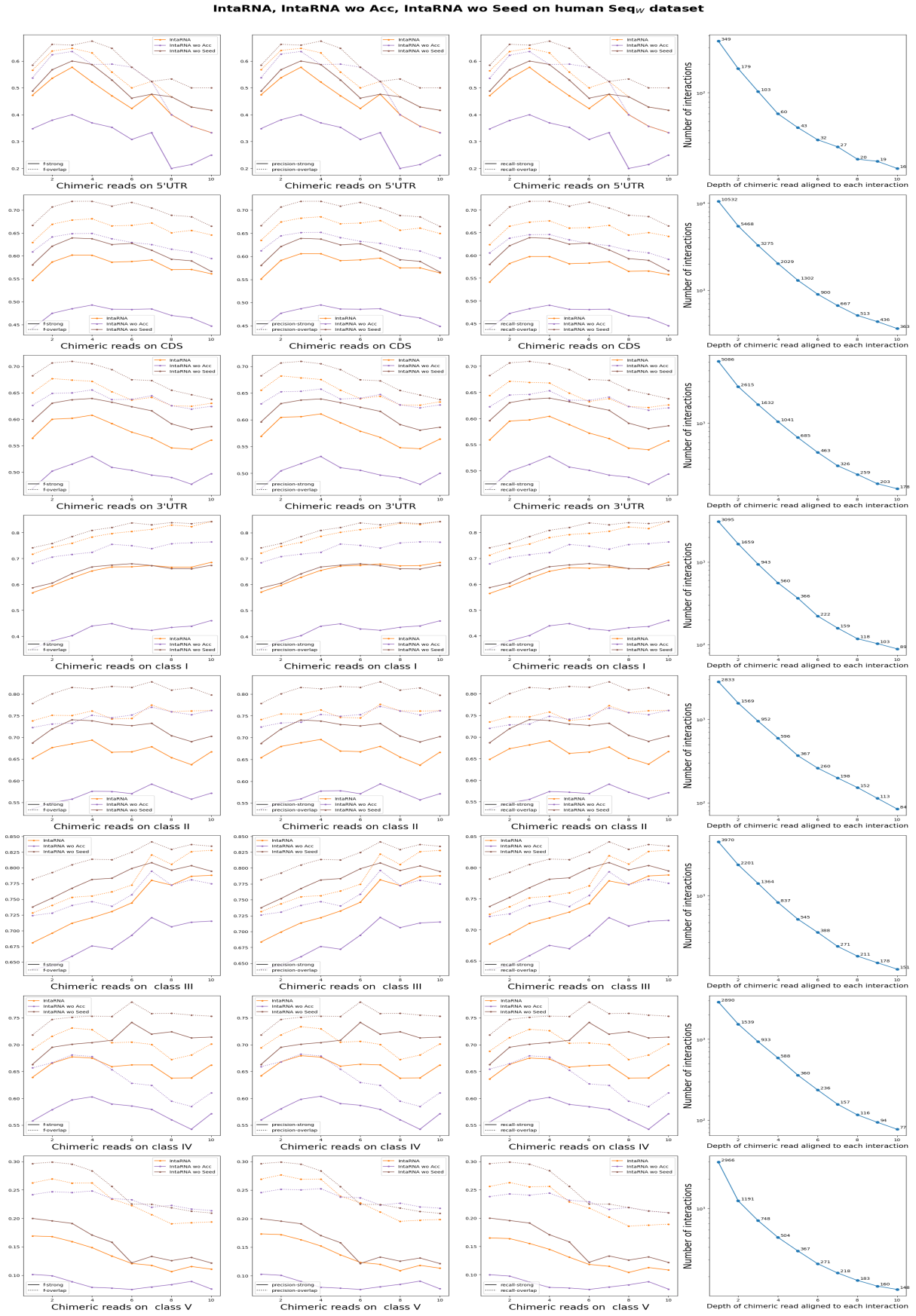


Figure D.10 – Performance of IntaRNA considering or not accessibility and seed match on the human Seq$_W$ datasets. The term wo in the Figure stands for WithOut.

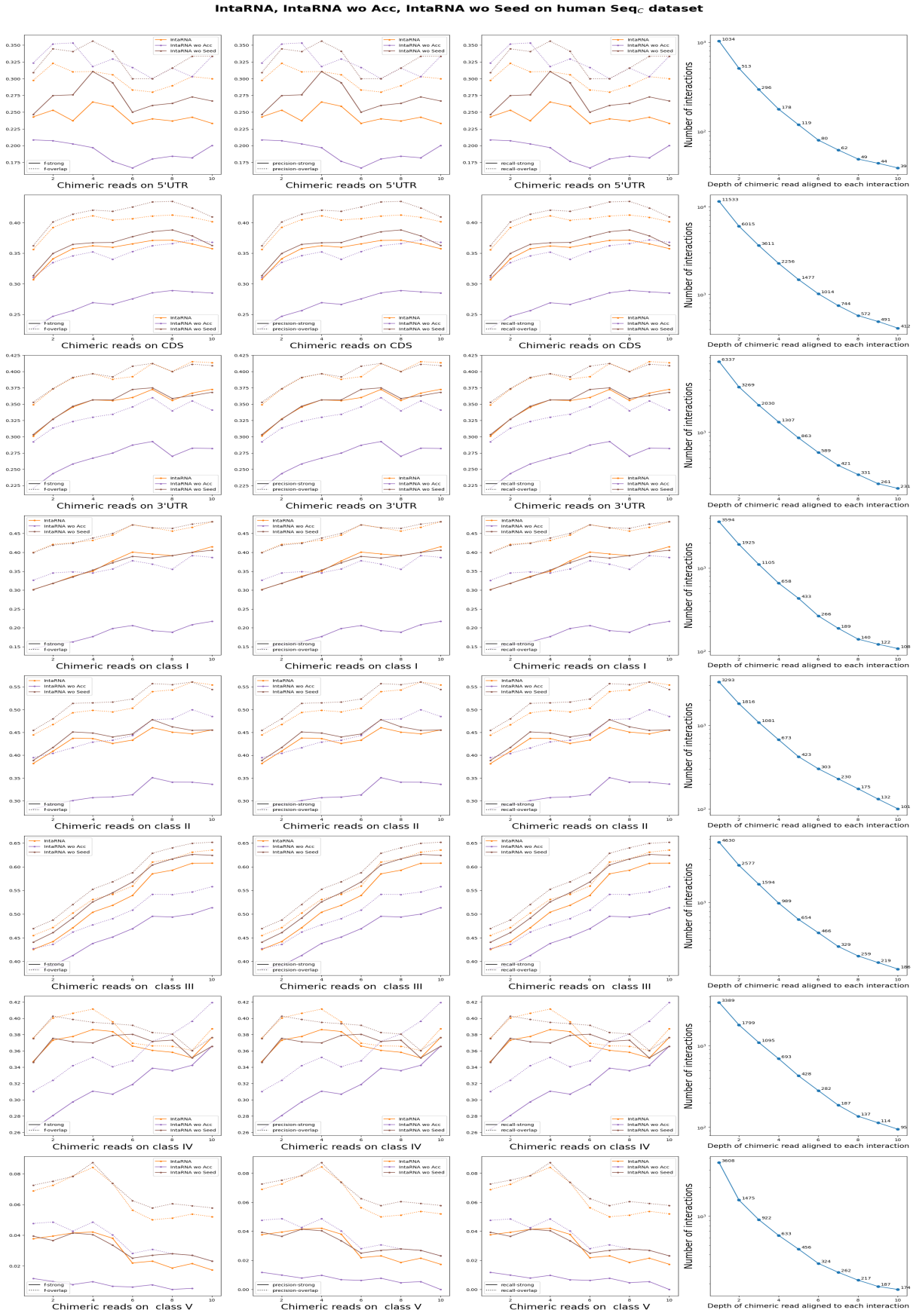


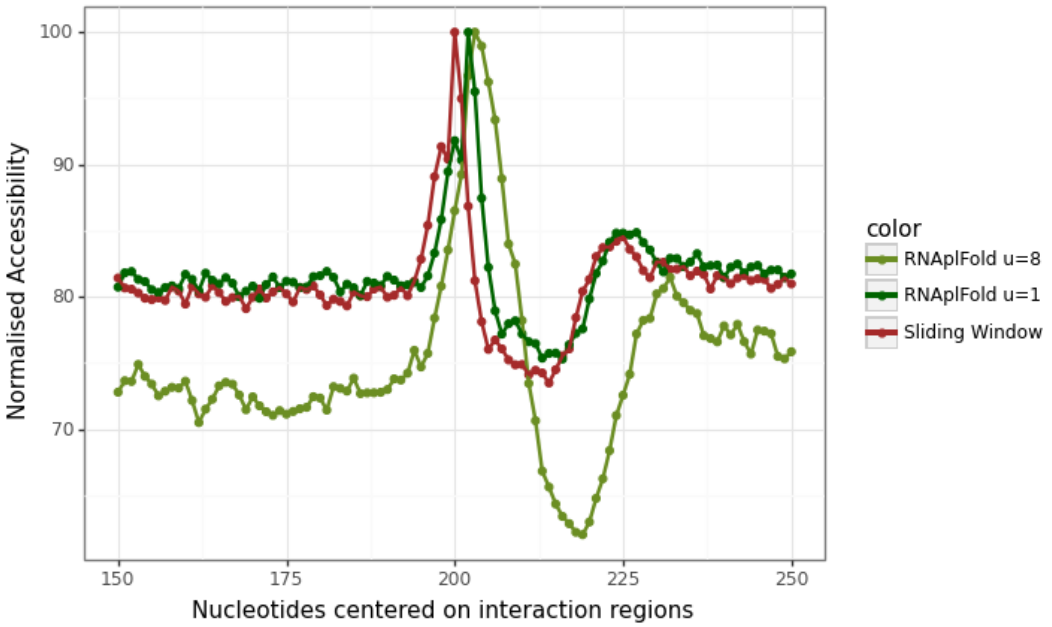
Figure D.11 – Performance of IntaRNA considering or not accessibility and seed match on the human Seq$_C$ datasets. The term wo in the Figure stands for WithOut.

and D.10).

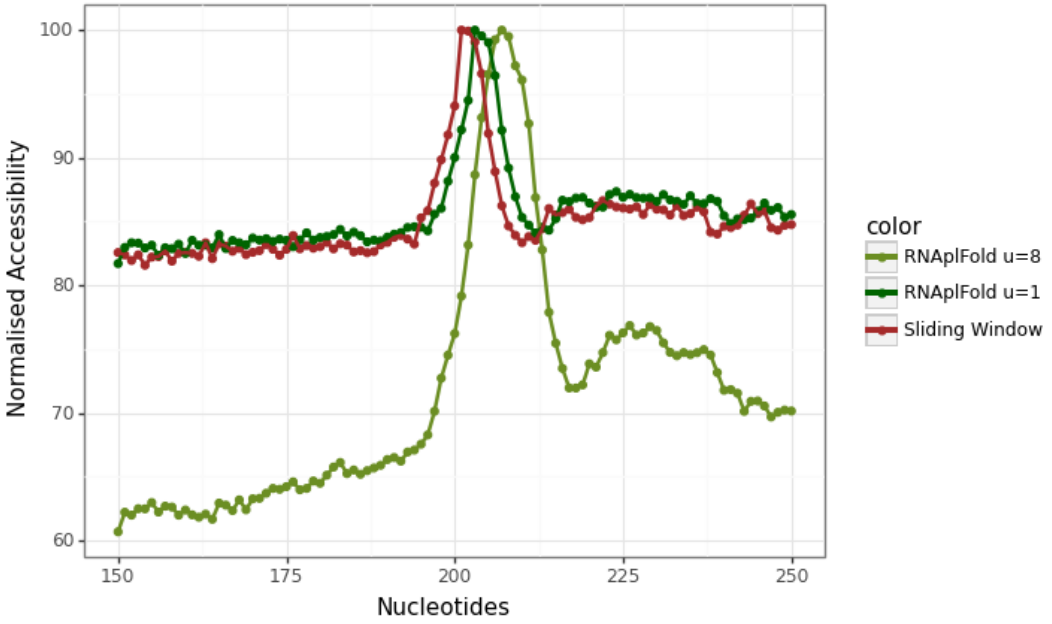
When the result is better without the seed, the difference is never very substantial. Such results are nevertheless somewhat unexpected and surprising. One possible explanation might be that IntaRNA considers the seed in the last step of its procedure. It would have been nice to be able to conduct the same test with Pita or miRanda, both of which start by checking for a match to the seed. There is, however, no way to fully eliminate such a feature, as is the case with IntaRNA. The only possibility, for instance, with Pita, would be to enter a very low value as a parameter for the seed width (in this case, 3), but this is a rather artificial way of proceeding and could thus generate other artefacts or bias the results in unexpected ways.

iii Refinement of the accessibility study

Although the results of the plots presented in Figure D.7 in Section ii reinforce the importance of considering the accessibility of an interaction site, we wished to further extend our exploration of this feature by using RNAfold and RNAplFold. Given a window of length l, a given nucleotide in a given sequence may thus belong to a different structure at most l times, except for those at positions less than l or greater than $L-l+1$ of the mRNA, where L is the latter's length. Considering all mRNAs of which there are N, a nucleotide may then be part of at most l times N structures. Following the literature, we fixed l at 150, which is the size of the window used in IntaRNA. As concerns N, it is equal to 15,754 for the human dataset and 11,581 for the mouse dataset, as indicated in Tables C.3 and C.5. The accessibility is then measured by a scoring procedure based on the computation of the average number of times each nucleotide is unpaired. We also computed accessibility as implemented in RNAplFold [Hofacker et al., 1994]. We parameterized RNAplFold with a window size (W) of 150 nts, inside which the accessibility is computed, a maximum base pairing span (L) of 100 nts and a continuous accessible sequence (u) of 1 and 8. Figure D.12 shows the results obtained considering all 15,754 mRNAs. Each class is displayed separately in Figures D.13 and D.14. For the sake of clarity, the graphic is centered on the interaction site, and values are shown from position 150 to 250.



(a) Accessibility computed for all Seq_W on human dataset



(b) Accessibility computed for all Seq_W on mouse dataset

Figure D.12 – Accessibility computed for all Seq_W on mouse and human datasets. In red is the accessibility when considering only the optimal structure with a sliding window. In dark green and in olive green are the accessibility computed with RNAplFold with an accessible window u of 1 and 8.

For both methods, we observe a rather distinct positive peak at the central position where the site is predicted to start by CLASH. This positive peak covers the start of the interaction and indicates a site potentially more accessible and thus more prone to beginning an interaction. However, if we look now at the same type of graphic but this time consider each mRNA sequence individually, it is difficult to observe any accessible region

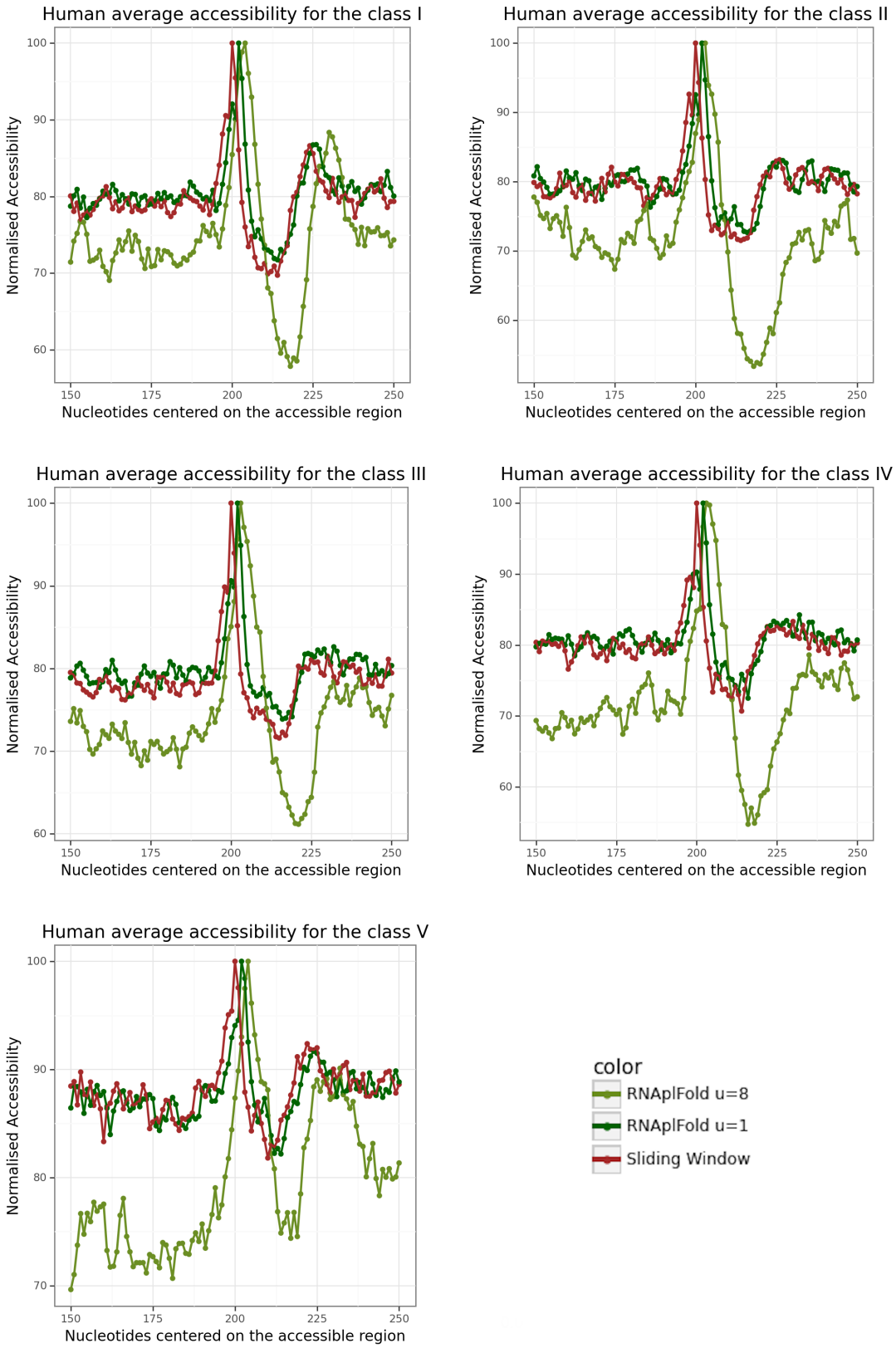


Figure D.13 – Normalised accessibility of all classes on Seq_W of the human dataset. In red, the normalised accessibility when considering only the optimal structure with a sliding window. In dark green and in olive green, the normalised accessibility computed with RNAplFold with an accessible window u of 1 and 8.

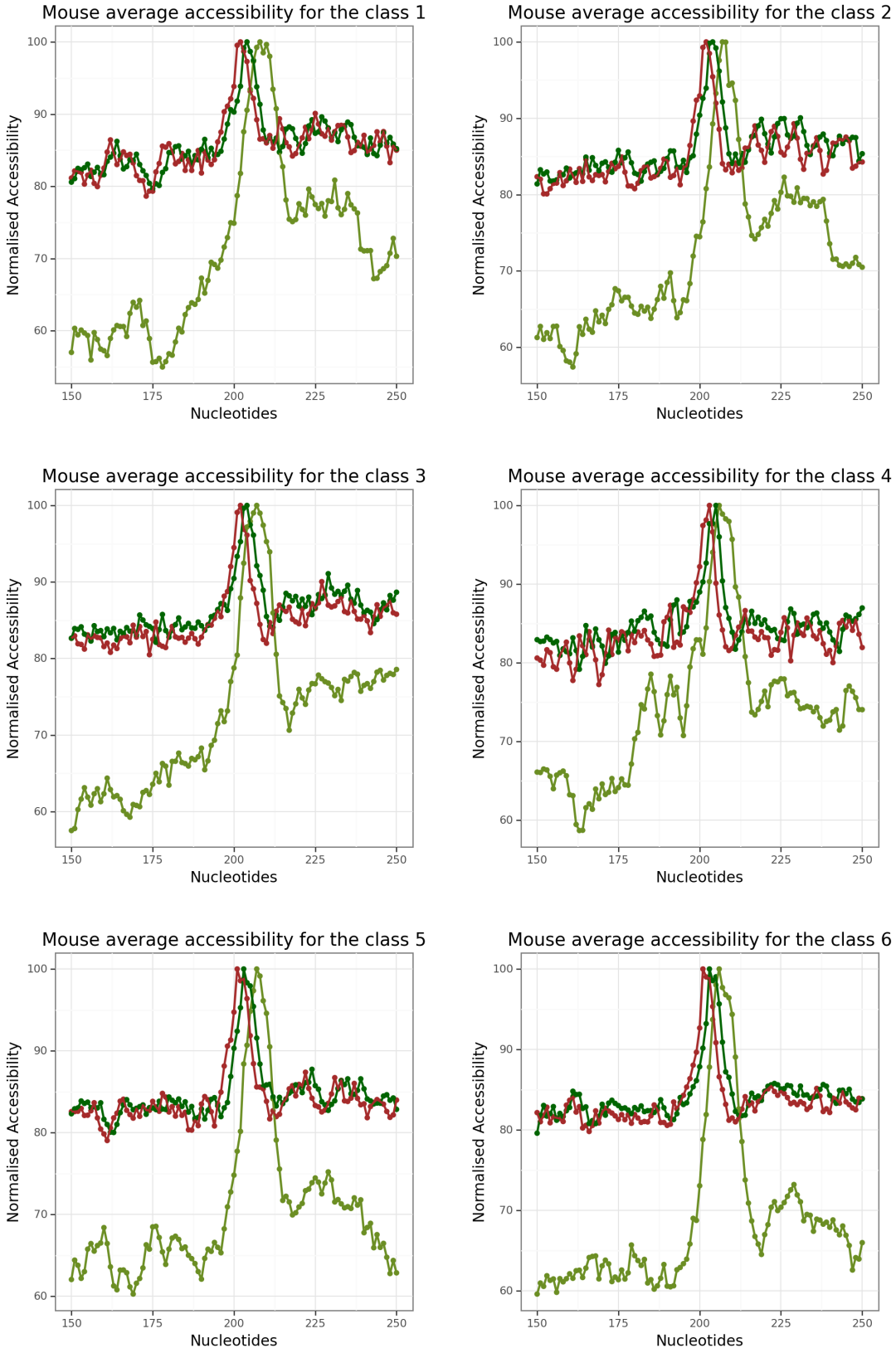


Figure D.14 – Normalised accessibility of all classes on Seq_W of the mouse dataset. In red, the normalised accessibility when considering only the optimal structure with a sliding window. In dark green and in olive green, the normalised accessibility computed with RNAplFold with an accessible window u of 1 and 8.

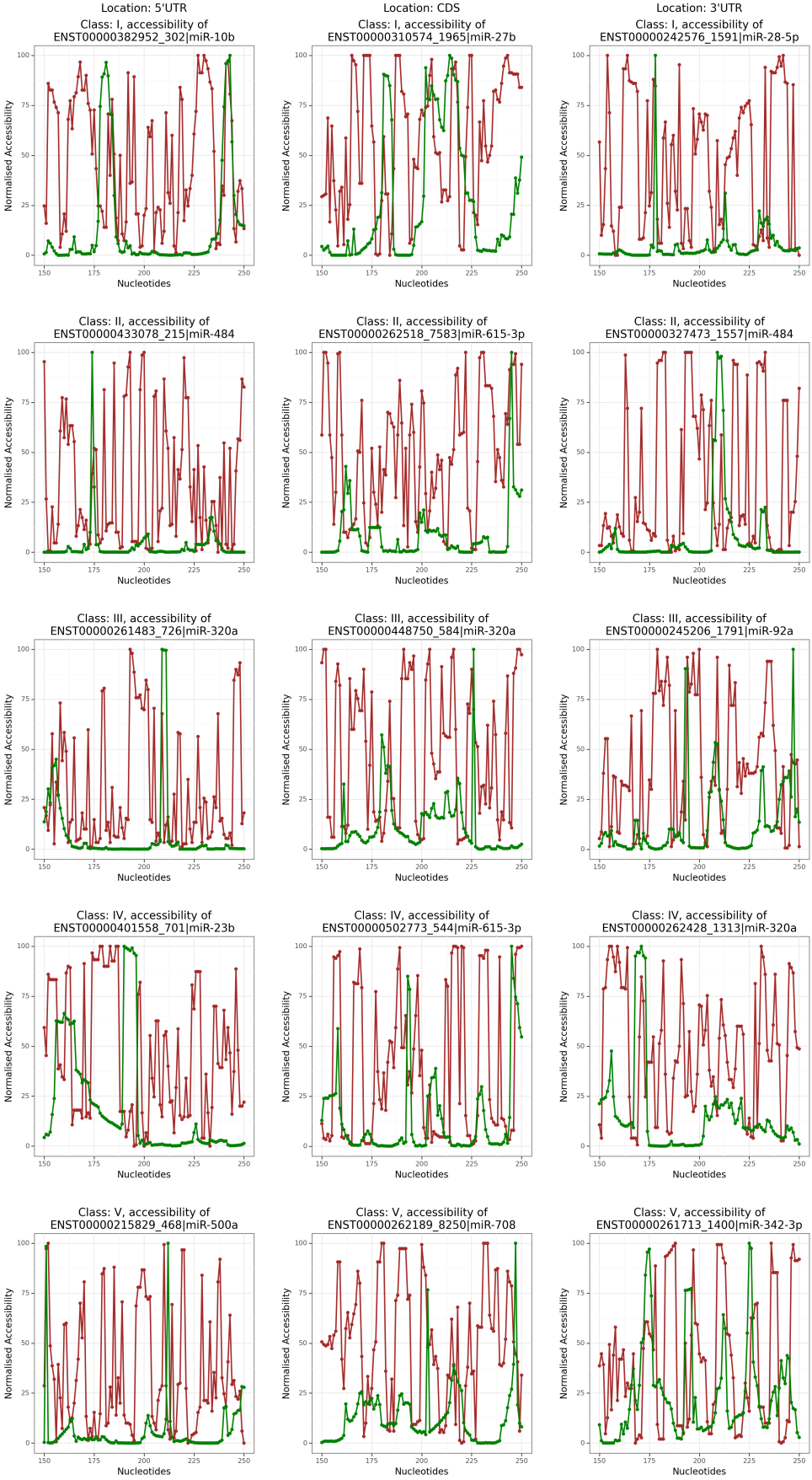


Figure D.15 – One randomly selected interaction for each class on each location on the human Seq$_W$ dataset. In red, the normalised accessibility when considering only the optimal structure with a sliding window. In green, the accessibility computed with RNAplFold with an accessible window u of 8.

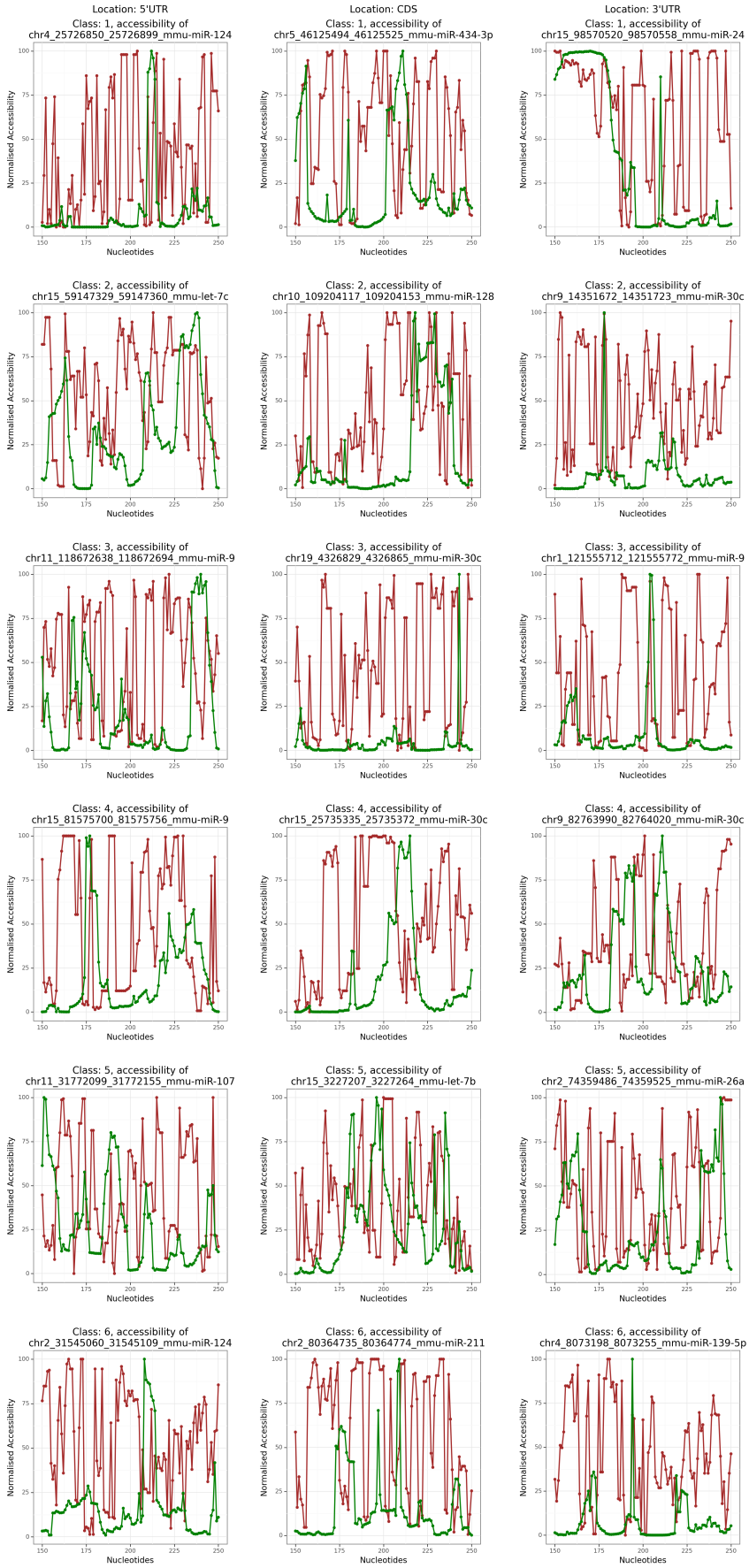


Figure D.16 – One randomly selected interaction for each class on each location on the mouse Seq_W dataset. In red, the normalised accessibility when considering only the optimal structure with a sliding window. In green, the accessibility computed with RNAplFold with an accessible window u of 8.

as presented in Figures D.15 and D.16. Although the results obtained by RNAplfold with $u = 8$ appear to contain much less noise, still in this case also the peak of accessibility is often situated outside of the experimental interaction sites.

Notice that the accessibility approach using a sliding window (Red) only considers one optimal structure for each window, whereas the approach using RNAplFold takes into account all suboptimal structures (see Figure D.12). Both approaches exhibit similar results. In the case of the human Seq_W dataset, all four classes have the same level of accessibility when considering the noise baseline and the positive peak of accessibility. In the case of the fifth class, the positive peak of accessibility is less noticeable from the noise floor. In the case of the mouse Seq_W dataset, there are no real differences between all six classes.

We also investigated if there is a noticeable difference between the interactions situated in the 3'UTR, CDS or 5'UTR regions. We show the results of this comparison in Figures D.17 and D.18 on the mouse and human Seq_W datasets. We have similar peaks as those observed with the classes. In the case of both species, the CDS region has the higher ratio of positive peaks (200) versus the noise floor. This is also the case for the negative peak at position 225. The low number of interactions located in the 5'UTR regions is the reason for the big variations observed, which make it difficult to estimate the floor noise. However, the peak remains very discernible. Only in the case of the human Seq_W dataset, there is a second positive peak at position 230. This second positive peak is only visible when using the method RNAplFold with u set to 8.

It is interesting to note that, on average, a nucleotide in the 5'UTR region is more often paired with another nucleotide than those in the CDS and the same statement can be made concerning the nucleotides in the CDS region which are more often paired than those in the 3'UTR region. Therefore, the 3'UTR region is on average slightly more accessible.

Altogether, these results support the idea that accessibility is in general an important feature.

D.3 Discussion and perspectives

We have come to the point where we can comfortably say that accessibility is truly an important factor to consider for miRNA–mRNA interactions. Nonetheless, there are still many things to discover and develop. Indeed, we only considered the accessibility of the mRNAs and not of the miRNAs, whereas although the miRNAs are shorter, they could also be folded. Indeed, it has been proposed that secondary structures adopted by mature miRNAs might play an important role in order to fulfill their biological activity [Gangemi et al., 2020].

Moreover, the accessibility of both could be influenced by a variety of factors such as RNA-binding proteins and epigenetic modifications. Indeed, in [Lin et al., 2019] for instance, it is shown that there is a significant association between miRNA regulation and hypermethylation, hence a possible modification of the accessibility of the miRNA target.

The fact that the peak of accessibility at the site of interaction in the CDS is more noticeable than in the 3'UTR region makes accessibility an even more important feature to consider when predicting interactions in the 3'UTR.

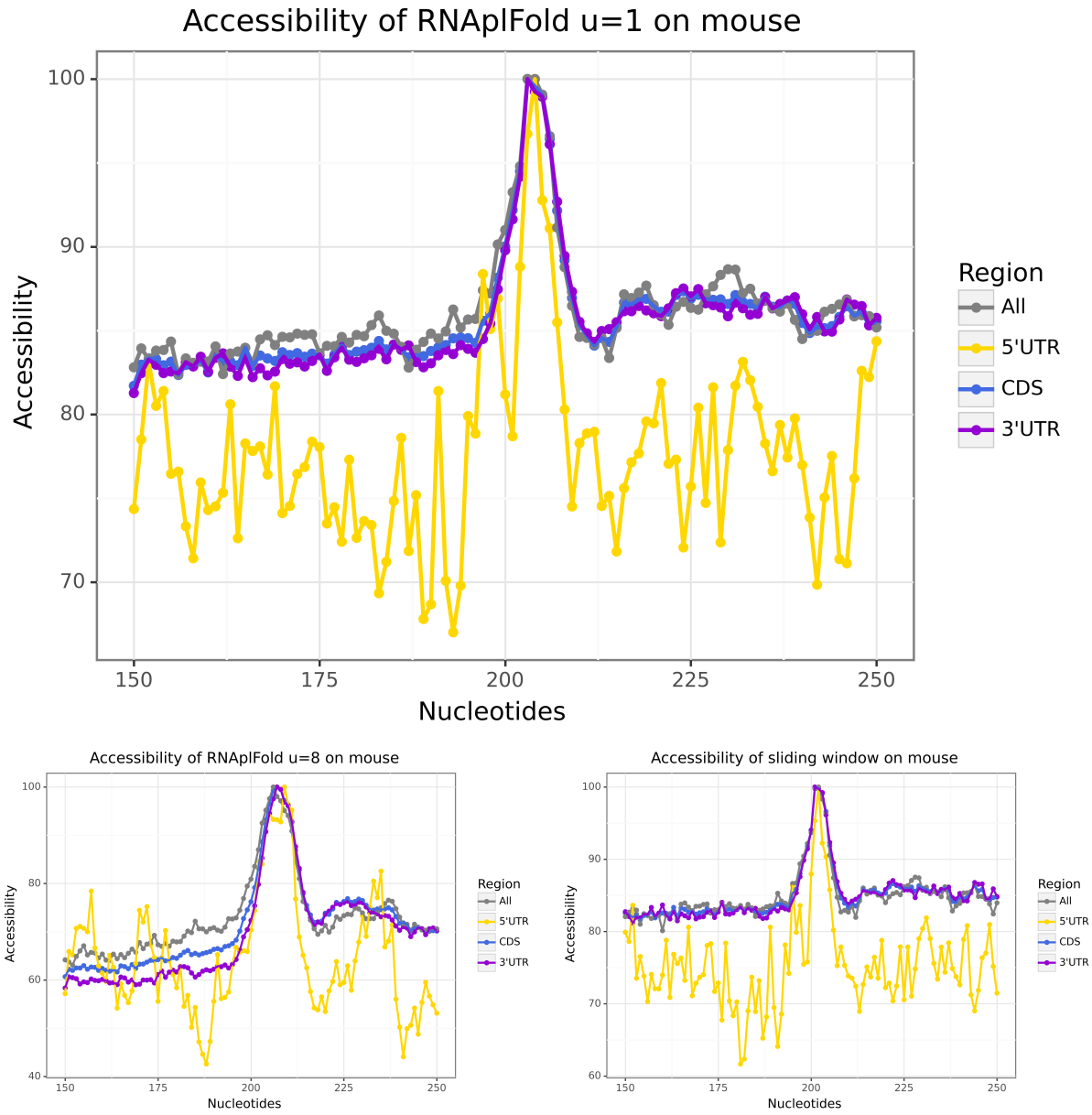


Figure D.17 – Accessibility on each region of the mouse Seq_W dataset. From top to bottom left and then bottom right, we have the computation of the accessibility with RNAplFold $u = 8$, RNAplFold $u = 1$ and then our naive method that count the average number of bases paired on a sliding window.

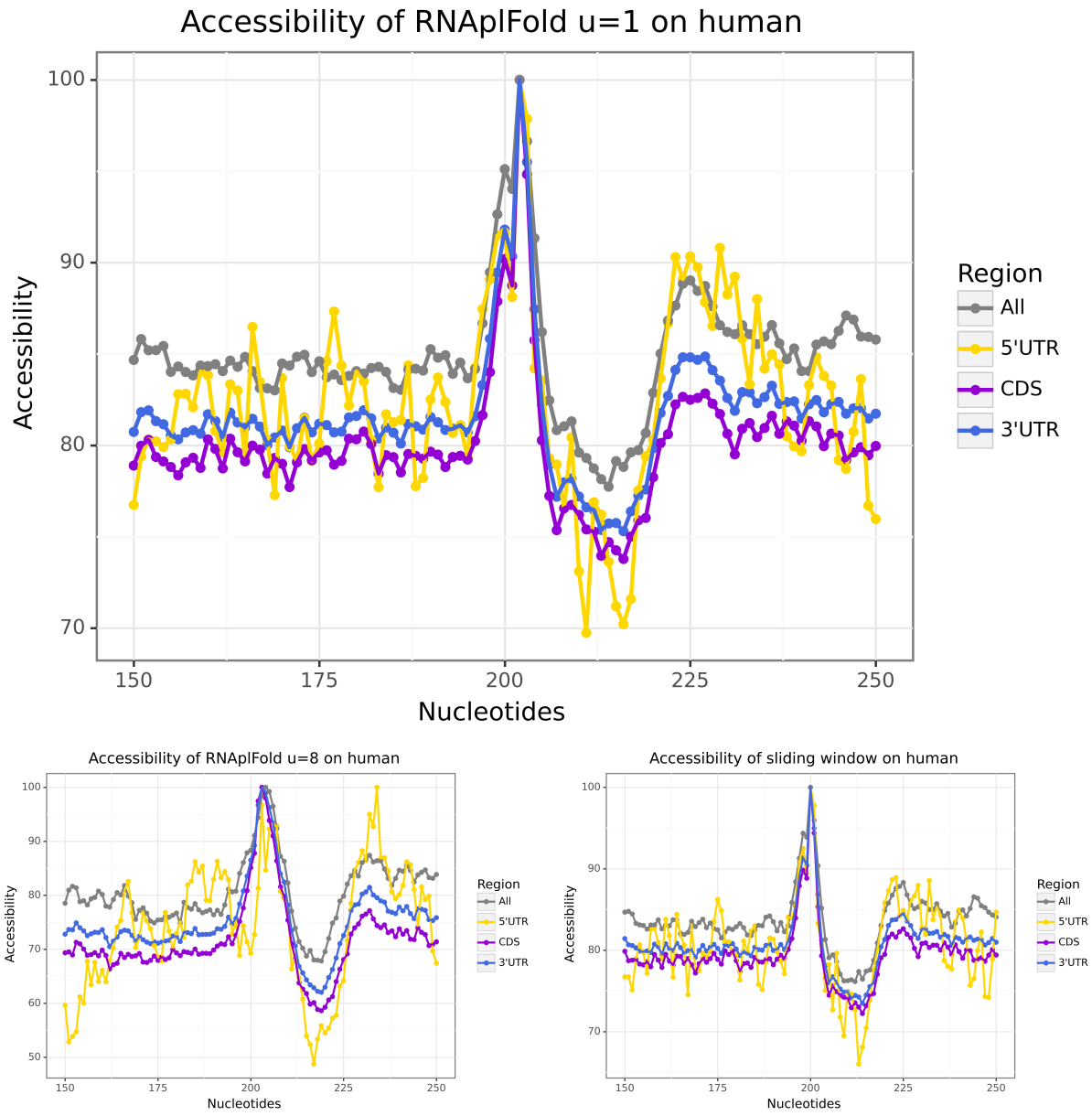


Figure D.18 – Accessibility on each region of the human Seq_W dataset. Accessibility on each region of the mouse Seq_W dataset. From top to bottom left and then bottom right, we have the computation of the accessibility with RNAplFold $u = 8$, RNAplFold $u = 1$ and then our naive method that count the average number of bases paired on a sliding window.

The fact that the negative peak at position 225 that follows the positive peak at position 200 is visible in both datasets, all classes and regions considered, is for now not fully explained and should be studied further.

Chapter E

Contribution: Conservation of motifs intra-species

The possible conservation or over-representation of motifs intra-species has been examined already in the literature [Miranda et al., 2006, Gumienny and Zavolan, 2015] and notably in one of the papers presenting one of the two CLASH data that we are using here [Helwak et al., 2013]. Besides the fact that such studies have been quite high-level and have focused on patterns that originate from the seed or from the sequences of miRNAs, there was never the idea that such conservation might be considered as an extra feature in the identification of interaction sites. Indeed, given that some miRNAs target many different mRNAs, one would then expect that, in such cases at least, taking into account such additional information could help improve the discovery of new interactions for an already known miRNA, or even point to the existence of new miRNAs not yet identified. Clearly, such an idea will not concern all miRNAs, only those that have multiple targets.

In this chapter, we present a study of this approach in chronological order. The study thus began with a first attempt to identify some signal when searching for conserved motifs. This signal was then analysed, which led to the discovery of a sequence composition bias before and inside the interaction sites. The following step was a more in-depth look at the base pairing in the interaction sites to see whether a pattern appeared that we could then use to improve the inference of the conserved or over-represented motifs. To test the validity of such pattern, we established a procedure to validate the motifs inferred based on the placement of their occurrences and their similarity to the corresponding miRNAs. This validation process was then carried on with another method of motif inference. In the final step, using only the idea of conservation, we managed to identify some of the targeted sites in complete mRNAs.

We used two tools to search for over-represented motifs: Smile [Sagot, 1998] and Meme from the Meme Suite [Bailey et al., 2009]. Both take as input a set of sequences from which motifs are inferred. They accept sequences on various alphabets, corresponding to DNA, RNA, or amino acids. In the case of this chapter, the sequences are either mRNAs or miRNAs which are on the alphabet (A,C,G,T).

Smile was developed internally in the team and enables to infer motifs composed of one or more boxes (parts) that are at a fixed or variable distance from one another. These motifs can be further constrained by fixing a minimum or maximum number of time each of the letters of the alphabet may appear in the motifs to be inferred. Smile is also

capable of giving an estimation of the over-representation or under-representation of the motifs by comparing the number of occurrences of those inferred with such number as observed in the input sequences shuffled. A positive Z-score indicates that a motif is found more often in the real sequences than in the shuffled ones. A negative Z-score indicates that the motif is found less often. The latter case is also of interest in general. However, it is in contradiction with the concept of conservation and will not be considered in this thesis. Based on a Gaussian distribution, we can convert a Z-score to a p-value with the normal cumulative distribution function ($normcdf$). If the Z-score is positive we have that $p\text{-}value = 1 - normcdf(Z\text{-}score)$, and directly that $p\text{-}value = normcdf(Z\text{-}score)$ if the Z-score is negative.

The Meme Suite is a collection of several well-known methods that are specialised in the inference and comparison of motifs. The master piece of the collection is the Meme algorithm which, at first sight, resembles Smile. Nonetheless, Meme differs from Smile as it is based on a heuristic approach. It provides as answer a number of motifs as specified by the user or until it exceeds one of its other thresholds such as runtime. There is also Streme from the Meme Suite that serves the same purpose of inferring motifs. It is better suited for larger datasets. We compared Streme and Meme on our datasets and the results favoured Meme. Despite the massive advantage in terms of time efficiency of Streme as compared to Meme, the motifs inferred by Streme on the interaction sites and on the Seq_W dataset are less similar to the miRNAs than those inferred by Meme. Indeed, according to the paper [Bailey, 2021], the success rate of Meme is better than Streme on datasets containing less than a few hundred sequences which moreover are relatively short. A final point in favour of using Meme is that the location of the occurrences of each motif is returned by Meme while it is not by Streme which requires an additional tool such as Fimo in order to do this, thus adding an extra layer of approximation.

Both Smile and Meme together with their associated parameters are detailed in Section B.4.

E.1 First signal and sequence composition bias

We began our investigation by exploring the human Seq_W dataset (see Chapter C) where we searched for motifs that might correspond to the mRNA sites interacting with a miRNA seed. In our case, we concentrated our attention on the seeds that would interact with at least 10 sites. Therefore, we set the parameters of Smile in order to infer motifs of length 7 with a quorum of at least 10 occurrences. We expected to find motifs situated at the position of the interaction sites (namely, position 200 for the Seq_W dataset) with a positive Z-score.

To better visualise the results, we created histograms that show, for each position of the input sequences, how many motifs were found at that position. For instance, a motif of length 7 found on a sequence that covers the positions 100-106 will be counted at position 100. A motif can be found several times at the same position but in different sequences. For instance, a motif m can cover the positions 100-106 in sequence 1 and also the positions 100-106 in sequence 2.

We present the results with two different visualisations:

— The first one counts a unique motif at each position so that the histogram would count 1 for motif m at position 100.

— The second visualisation counts the number of sequences that are occurrences of a motif at a given position; this would give us a count of 2 for motif m at position 100.

In both cases, without any filter (that is, without any selection made based on the Z-score), SMILE finds 16384 motifs, which corresponds to all possible motifs of size 7 (indeed, we have that $4^7 = 16384$).

We present in Figure E.1 the histograms for the first case, considering all motifs first, then only those with a positive Z-score, and finally only those with a negative Z-score. In all three histograms of this figure, we observe a negative peak in the region both down and upstream of position 200 with a small positive peak centered on the exact location of the interaction sites for the case of all motifs and more so for the case of the motifs with a positive Z-score. The negative peaks might be explained by the fact that if the regions around the sites of interaction are indeed conserved, we would expect to have fewer motifs covering it. The small positive peaks, that are more noticeable for the motifs with a positive Z-score only, may indicate that at the precise location of such interaction sites, there are more motifs that are statistically significant, something that is not observed in the histogram (Figure E.1c) corresponding to the motifs with a negative Z-score only.

The same observation can be made with the histograms of Figure E.2 with however this time a much higher positive peak for the motifs with a positive Z-score (Figure E.2a). In the case of motifs with a negative Z-score on the other hand (Figure E.2b), there is also an unexpected positive peak, which furthermore this time is located before the position of interaction.

In all cases, using a positive Z-score improves the signal. However, the ratio between signal and noise (motifs that are not situated at the positions 200-220 of the interaction) is however small and even with a strong filter (such as all motifs with a Z-score superior to 30, as shown in Figure E.2c), we obtain a ratio of 1.25 only.

Moreover, as the negative peak remained puzzling and even more so the positive peak upstream of the position of the interaction sites observed in Figure E.2b, we decided to investigate the sequence composition around the interaction sites.

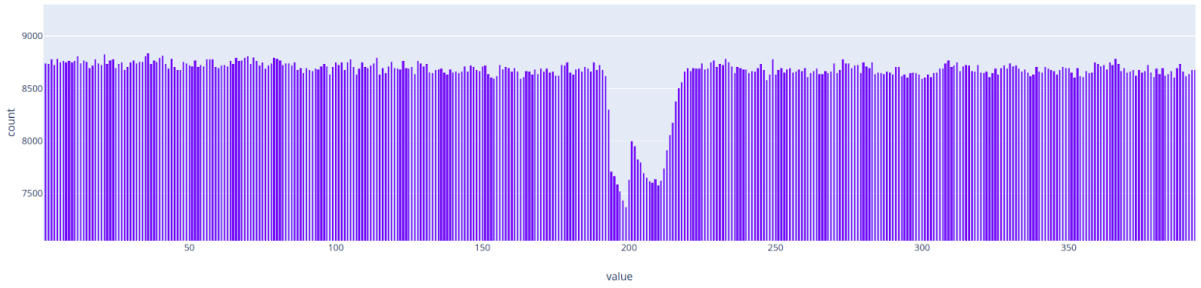
i Sequence composition around the interaction sites

In order to do such investigation, we decided to represent the sequences with a logo that stacks the percentage of times (probability) that a letter appears at each position in the sequences. The letters are ranked from the most frequently found at the bottom to the least frequently found at the top (see Figure E.3).

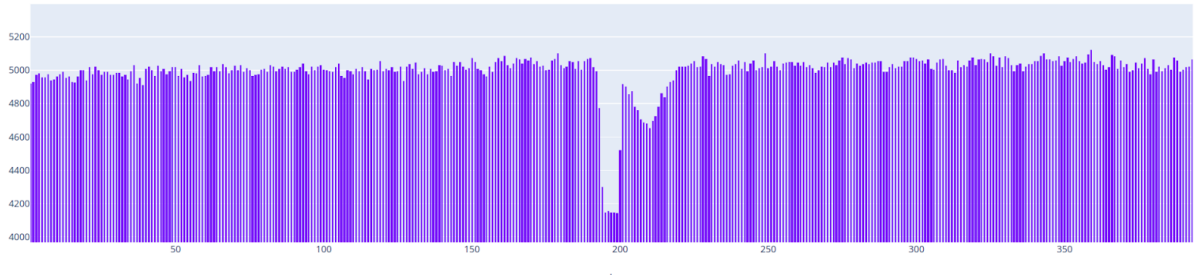
From the sequence logos, we were able to notice two regions. One is situated from 15 nucleotides before the interaction to the interaction point. We denote this region by R_1. We also found a second region from the interaction point to 20 nucleotides after which will be denoted by R_2, which is also the region of interaction.

On the human dataset, we have that:

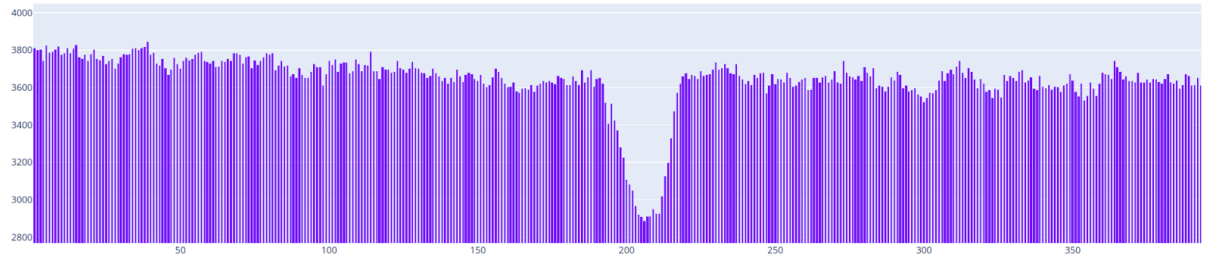
— Region R_1 exhibits a slightly higher concentration (approx. 52%) of the nucleotides C and A and of the di-nucleotides CC, AG, and CA. We observe the same for all classes except for Class V (see Figure E.4);



(a) Results without filter of Z-score (16383 motifs)

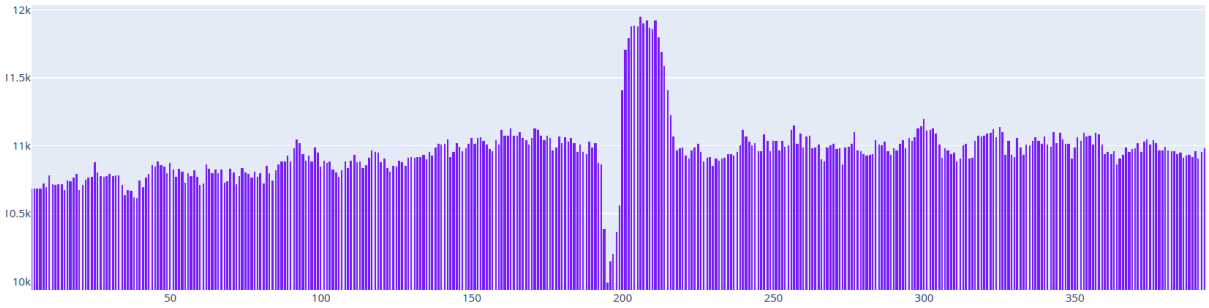


(b) Only the motifs with positive Z-scores (6625 motifs)

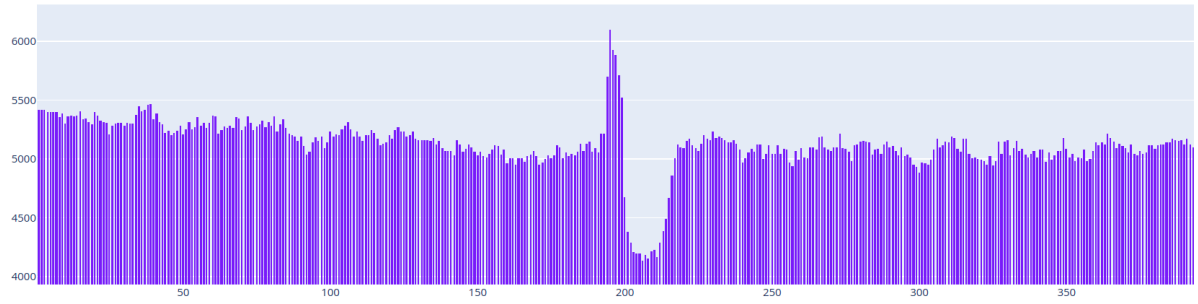


(c) Only the motifs with negative Z-scores (9753 motifs)

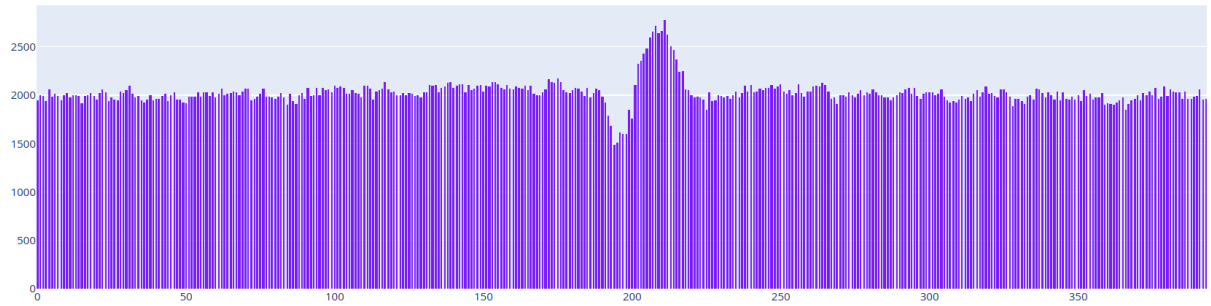
Figure E.1 – Results of SMILE on the human Seq_W dataset with a unique motif visualisation (1).



(a) Only motifs with positive Z-scores



(b) Only motifs with negative Z-scores



(c) Only motifs with $Z - score > 30$

Figure E.2 – Results of SMILE on the human Seq_W dataset. Visualisation of the number of sequences that have a motif at each position (2).

— Region R_2 begins with A in 50% of the cases and then has a higher frequency (55–60%) of the nucleotides G and C. The higher GC content is observed in all the classes except in Class V (see Figure E.4). This observation is surprisingly in opposition to what is presented in the current literature [Grimson et al., 2007,Nielsen et al., 2007] which pinpoints a higher AU content upstream and downstream of the interaction site. Indeed, when distinguishing the region of interaction, there is a higher AU (or AT) upstream of the interaction sites only for the 3'UTR regions. In the 3'UTR regions, the higher AT content (approx. 60%) can clearly be observed upstream and downstream of the interaction sites (see Figure E.5). However, this is not the case for the CDS regions which have a higher CA content downstream, then the first 7 nucleotides have a higher number of ATs followed by a higher number of ACs.

On the mouse dataset, we have that:

— Region R_1 has a slightly higher concentration (approx. 52%) of ATs. This is especially the case for Classes 1 and 3. However, Class 6 displays a higher AC content in this region (see Figure E.4);

— Region R_2 starts with either a T or a C and exhibits a higher content (60–70 %) of ATs (52–65 %) (or AUs in RNA language) for the first 7 nucleotides, then a slightly higher (approx. 52%) content of ACs can be seen. We have similar observations on the different classes and in the 5'UTR and CDS regions. There is a very distinct higher content (55-70%) of AUs downstream and upstream of the interaction sites in the 3'UTR regions (see Figure E.5).

The differences between our observations on the interactions in the CDS regions and what appears in the literature is a reminder that the majority of the miRNA–mRNA interactions in the literature is focused on the 3'UTR regions, whereas the two CLASH datasets we used evidence twice more interactions in the CDS regions than in the 3'UTR ones.

We then decided to go back to the inference of motifs using however more appropriate parameters.

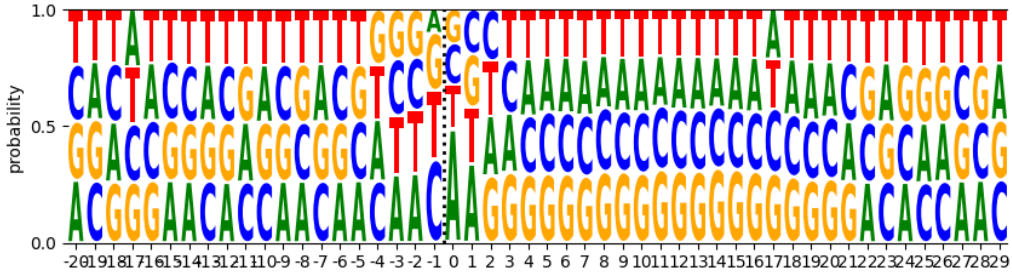
ii Characterisation of the CLASH interaction patterns

Indeed, given that there are different classes of interaction sites, we decided to refine the parameters of Smile to fit such parameters according to the characteristics of each class.

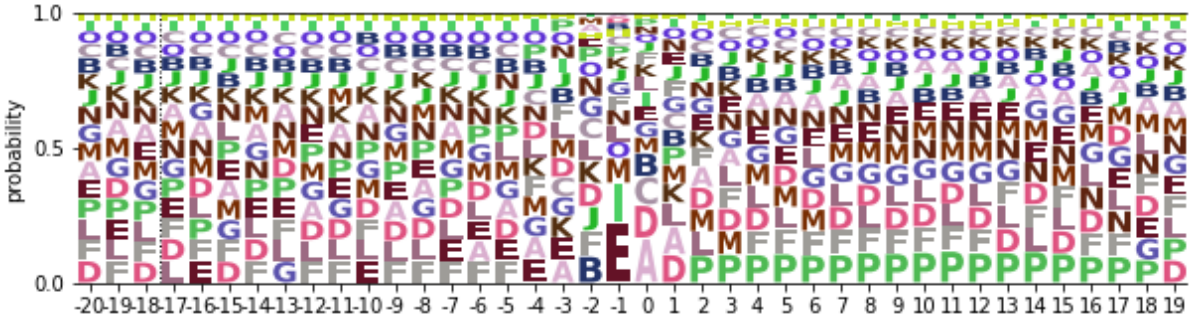
We also came up with the idea of an experiment that would more precisely define what a valid motif is according to two criteria, one related to the position and the other to how the motif matches the mRNA sequence.

As a reminder, with our human Seq_W dataset, we know the couple miRNA-mRNA as well as the region of interaction on both. However, this region is still 20 to 100 nucleotides long while we are looking for a more precise location corresponding to the interaction site.

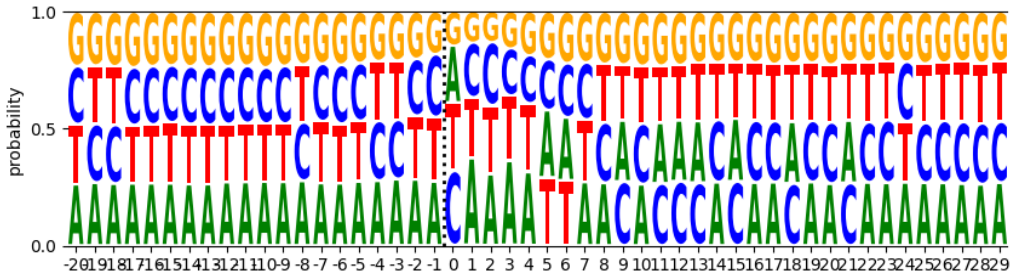
To be able to identify such location using the human CLASH data presented in [Helwak et al., 2013], we decided to illustrate the base pairing predictions from CLASH with a sequence logo (Figure E.6) where each letter represents the percentage of times that the letter at each position of the sequence is paired. The sequence logo of all interactions



(a) On the human dataset: Single nucleotide composition around the interaction sites situated at position 0.

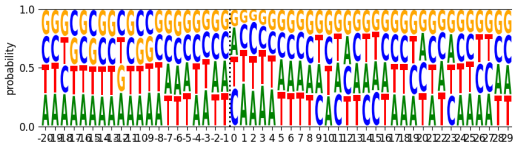


(b) On the human dataset: sequence logo in di-nucleotides: 'AA': 'A', **'AC': 'B'**, 'AT': 'C', 'AG': 'D', **'CA': 'E'**, **'CC': 'F'**, 'CT': 'G', 'CG': 'H', 'TA': 'I', 'TC': 'J', 'TT': 'K', **'TG': 'L'**, 'GA': 'M', 'GC': 'N', 'GT': 'O', **'GG': 'P'**

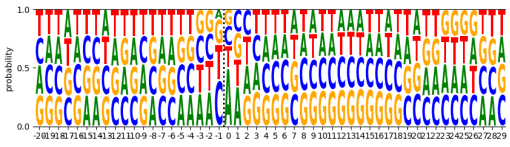


(c) On the mouse dataset: Single nucleotide composition around the interaction sites situated at position 0.

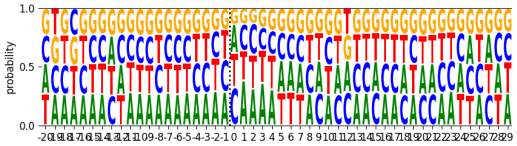
Figure E.3 – Analysis of the single and di-nucleotides around the point of interaction situated at position 0.



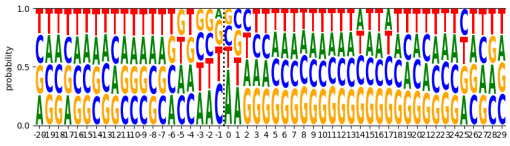
(a) Class 1



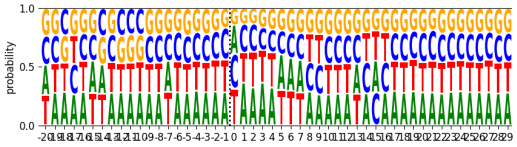
(b) Class I



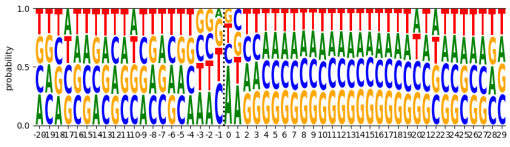
(c) Class 2



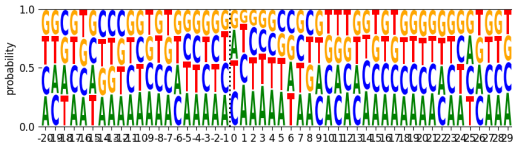
(d) Class II



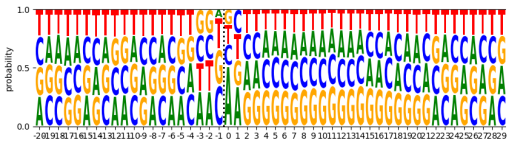
(e) Class 3



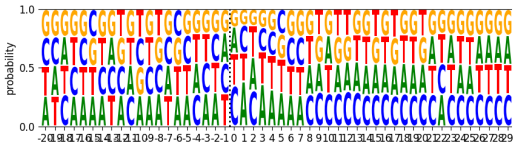
(f) Class III



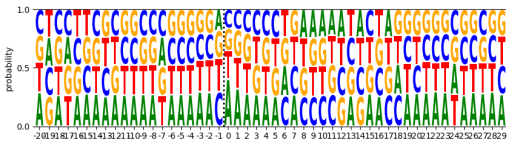
(g) Class 4



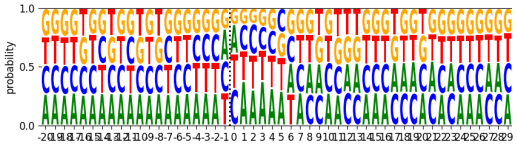
(h) Class IV



(i) Class 5

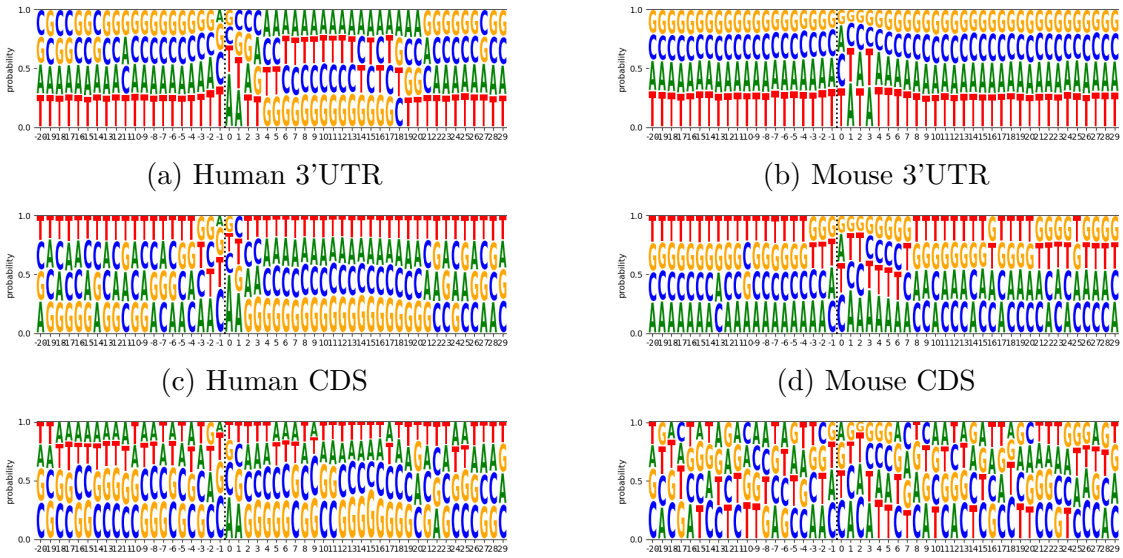


(j) Class V



(k) Class 6

Figure E.4 – Human dataset on the right and mouse dataset on the left: Analysis of the composition of single nucleotides around the point of interaction situated at position 0 for each class.



(a) Human 3'UTR

(b) Mouse 3'UTR

(c) Human CDS

(d) Mouse CDS

(e) Human 5'UTR

(f) Mouse 5'UTR

Figure E.5 – Human dataset on the right and mouse dataset on the left: Analysis of the composition of single nucleotides around the point of interaction situated at position 0 for each region

shows that the first nucleotide of the targets is paired in less than 10% of the interactions, then this percentage starts to grow until it reaches a peak of 60% of nucleotides paired at positions 12 to 14, and finally, it reduces again after position 14 to reach 10% at position 40.

The correct motifs should therefore be located within positions 1 to 40. Moreover, the motifs themselves should present different characteristics depending on the class of interaction considered. Thus, in the case of human, we should have that the motif is composed of:

— Class I: A single box with a motif size of 9–10 nucleotides.

— Class II: Two boxes separated by a space of 4–5 nucleotides, with the first box of width 4–5 nucleotides and the second box of width 4 nucleotides.

— Class III: Two boxes separated by a space of 8–9 nucleotides, with the first box of width 4 nucleotides and the second box of width 3–4 nucleotides.

— Class IV: One box of width 12 nucleotides.

The motifs are then further validated by matching them on the reverse complement of the miRNAs. In the case where substitutions are allowed or of motifs composed of more than one box (see Section i), searching for the reverse complement of the miRNAs is less immediate. In these two cases, we thus decided to proceed by looking for all motifs on the miRNA using the same parameters except for the quorum that is fixed to 1, and then by checking whether the motifs found on the set of mRNAs match the reverse complement of one of those found on the miRNA.

Finally, to make the computation easier and to eliminate some noise, the sequences from Seq_W are divided into four groups according to the interaction class of the miRNA targeting each sequence. We therefore have 4 different datasets for each class: Seq_{W_I},

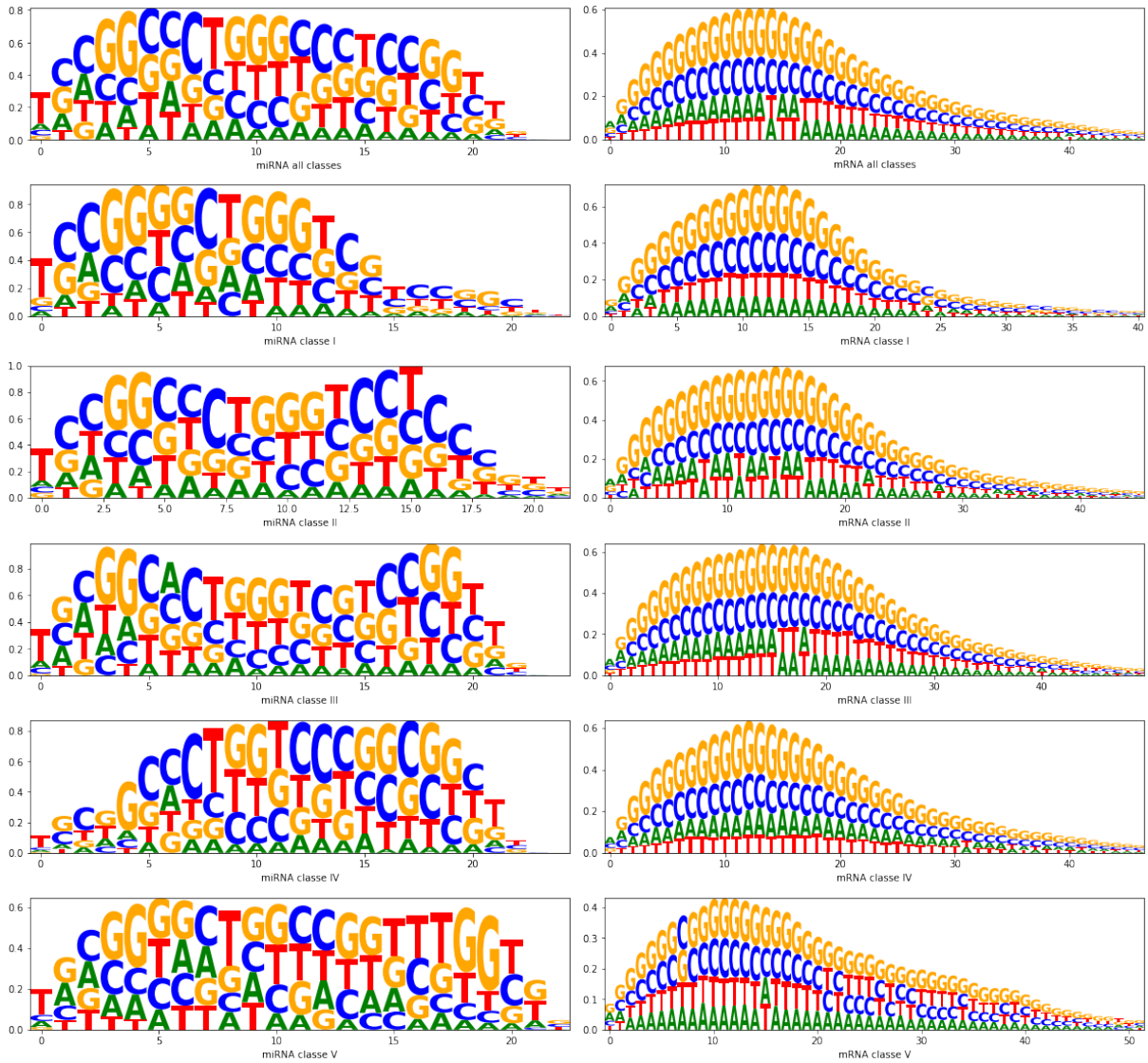


Figure E.6 – Sequence logo of the miRNAs and mRNAs of the interaction between a couple miRNA/mRNA. Letters represent the percentage of times that each letter at that position of the sequence is paired.

Class	#miRNAs	#mRNAs	#interactions
I	221	3095	3594
II	232	2833	3293
III	248	3970	4630
IV	213	2890	3389

Table E.1 – Reminder of the number of miRNAs, mRNAs and interactions per class on the human dataset

Class	#motifs	#motifs valid	%miRNAs	%mRNAs	%interactions	Precision
I	793,544	1714	98%	34%	35%	0.002
II	315,780	6279	100%	73%	99%	0.02
III	386,162	3759	100%	68	83%	0.01
IV	1,039,636	44	16%	1%	1 %	4.23×10^{-5}

Table E.2 – Results of the validation procedures of the motifs inferred in each class of the human dataset. The inference is performed without any substitutions allowed.

$\text{Seq}_{W_{II}}$, $\text{Seq}_{W_{III}}$ and $\text{Seq}_{W_{IV}}$.

The results, without estimation of the over-representation by Z-score, are displayed in Table E.2 and as a reminder, we displayed in Table E.1 the number of miRNAs, mRNAs and interactions per class.

In the end, we still find too many motifs that are not valid. By computing the ratio of valid motifs divided by the total number of motifs, we observe that we have the worst precision for the Classes IV and I with a respective precision of 4.23×10^{-5} and 0.002. The motifs composed of two boxes have a slightly better precision with 1% for Class III and 2% for Class II.

To better understand the range of the motifs inferred, notice that we count the number of miRNAs, mRNAs and interaction sites which have at least an occurrence of a valid motif. We count a miRNA as retrieved if there is at least an occurrence of a valid motif that matches the reverse complement of the miRNA sequence. We count an interaction as retrieved if there is at least an occurrence of a valid motif that is placed between positions 200 and 240 of the mRNA. Finally, we count an mRNA as retrieved if this mRNA has at least one of its interaction sites covered by a motif. As a reminder, there may be several interaction sites on the mRNAs that are targeted by one or more miRNAs.

Without applying any filter based on the Z-score, in the first three classes, nearly all miRNAs have a matching motif. In the second and third classes, nearly all interactions and more than half the mRNAs have an occurrence of a motif placed in the valid position.

However, when applying a filter that selects motifs with a p-value smaller than 0.05 (which is roughly equivalent to a Z-score > 6), the number of motifs drops of about one third (see Table E.3). Even with this reduction, the precision is quite low with a maximum of 3% for Class III. Concerning the miRNAs, mRNAs and interactions covered, they remain still very high with 100%, 99% and 85% of the miRNAs retrieved for Classes II, III and I.

The problem with Class IV might be that we are searching for motifs without sub-

Class	#motifs selected	# motifs valid	% miRNAs	% mRNAs	%interactions	Precision
I	123,504	889	85%	25%	25 %	0.007
II	110,660	3,590	100 %	72%	84%	0.03
III	128,564	2,388	99%	67%	82%	0.02
IV	950,610	34	13 %	1%	<1%	3.6×10^{-5}

Table E.3 – Results of the validation procedure of the motifs that have a p-value <= 0.05 inferred in each class of the human dataset. The inference is performed without any substitutions allowed.

stitutions (see Section i) which might not be appropriate for that class. However, when substitutions are allowed, we get too many motifs even with only one substitution and the computation of the p-values may then become very expensive. Indeed, with one substitution allowed – in the case of two boxes the substitution is allowed in either the first or the second box – we have 1216314, 327680, 409600, and 13016003 motifs respectively for Classes I–IV. With this number of motifs, it is obvious that we find more matches of mRNAs and miRNAs, and indeed Class IV has now 738 valid motifs that cover 91% of the miRNAs, 19% of the mRNAs and 18% of the interaction sites. However, the precision remains low, namely 5.7×10^{-5}.

There is still an interesting observation to be made, and this is that it is preferable to have a substitution in the first seed box rather than in the second one when dealing with motifs composed of two boxes. More of the motifs found are valid in that case.

We also tried the idea of selecting only the best n motifs. This leads to identifying many low-complexity motifs such as *AAAAAAAAA* and *GTGTGTGTG* which are indeed over-represented compared to the shuffled sequences, yet they have poor similarity with the majority of the miRNAs.

Blindly searching for motifs does not give us good enough results, and we therefore had to re-consider our motif inference process.

In addition, the computational task of determining if a motif is statistically significant using shuffling, as Smile and Meme do, is only an approximation with potential biases.

In the case of Smile which is an exact algorithm, this may also be expensive in terms of computation time. Because it is a heuristic, Meme on the other hand infers only a few probabilistic motifs making it less time consuming to then select only the most significant one(s). Therefore, we chose for now to continue to explore the idea of motifs and intra-conservation with Meme and with datasets that are miRNA and class-specific.

E.2 Pattern discovery and similarity with the miRNA

We considered separately, for each miRNA in both CLASH datasets, the set of mRNAs it targets. In the first experiment, we gave as input to Meme, for each miRNA, the exact interaction sites on its mRNA targets. This dataset is denoted by Seq_{E_i}. In the second experiment, we gave as input to Meme, for each miRNA, the exact site plus 200 nucleotides down and upstream of that site for all its mRNA targets. This dataset is denoted by Seq_{W_i}. In all cases, we selected only the motif with the lowest combined match p-value [Bailey and Gribskov, 1998] which we considered as the best one found by Meme.

Note that for each class, region or combination of both, we had to create a different file that contains either the Seq_{E_i} or the Seq_{W_i} dataset, hence the result obtained on a dataset of each class is not representative of the result obtained on the full dataset which contains all interactions of all classes.

As concerns the number of miRNAs, of the 399 in the human CLASH dataset, only 316 were taken into account as the remaining 83 ones (approximately 21% of the miRNAs) have only one predicted interaction site. When considering miRNAs that have at least 10 predicted interaction sites, the number goes down to 162 (approximately 41% of the total). Similarly in the mouse dataset, only 299 of the 378 miRNAs have 2 or more targets (approximately 79% of the miRNAs) and 147 have at least 10 or more identified interaction sites (approximately 39% of the miRNAs).

i MEME parameterisation

We used the parameters by default from the tool as available on the website (meme-suite.org), except for the site distribution and the width of the motifs.

As concerns the site distribution indeed, we tested with both the default one which searches for Zero or One Occurrence Per Sequence (zoops) and with the one that searches for One Occurrence per sequence (oops). To validate which motif distribution suits the best our idea of conservation, we searched with both models on the mouse seq_{E_i} dataset and then compared the motifs found with their miRNA sequence using for this TOMTOM (see Figure E.7 and Section ii for details about the method used for the comparison). It shows that both models are very similar one to another and we therefore decided to continue with the oops model. However, it is interesting to note that a search for each individual class with the zoops model was able to find one distinct motif for each of the groups with an average coverage (sequences where the motif is found divided by the number of sequences in the group) of 90%. If we exclude Class V with a lower coverage of 83%, the coverage of MEME is of 95%. However, when we take the motifs found by MEME for the same dataset as previously with only 10 or more interactions, the global coverage drops to 82% which is due to the fifth class having a coverage of 62% while the 4 other classes have a coverage of around 94%. The latter numbers are obtained when searching for patterns without separating the regions. The high coverage in this case means that there is indeed a shared pattern between the interactions and even between the regions (3'UTR, CDS, 5'UTR).

As concerns now the width of the motif, by default, MEME sets it between 6 and 50 nucleotides. We changed this to between 6 (minimum length of a seed) and 23 (maximum length of a miRNA). The positions on each sequence for every motif occurrence were recovered directly from the output of MEME.

Using the One Occurrence Per Sequence (oops) motif distribution, we found motifs with an average width of around 11 and 10 nucleotides for Seq_{E_i} and around 13 and 14 for Seq_{W_i} of, respectively, human and mouse (see Table E.4).

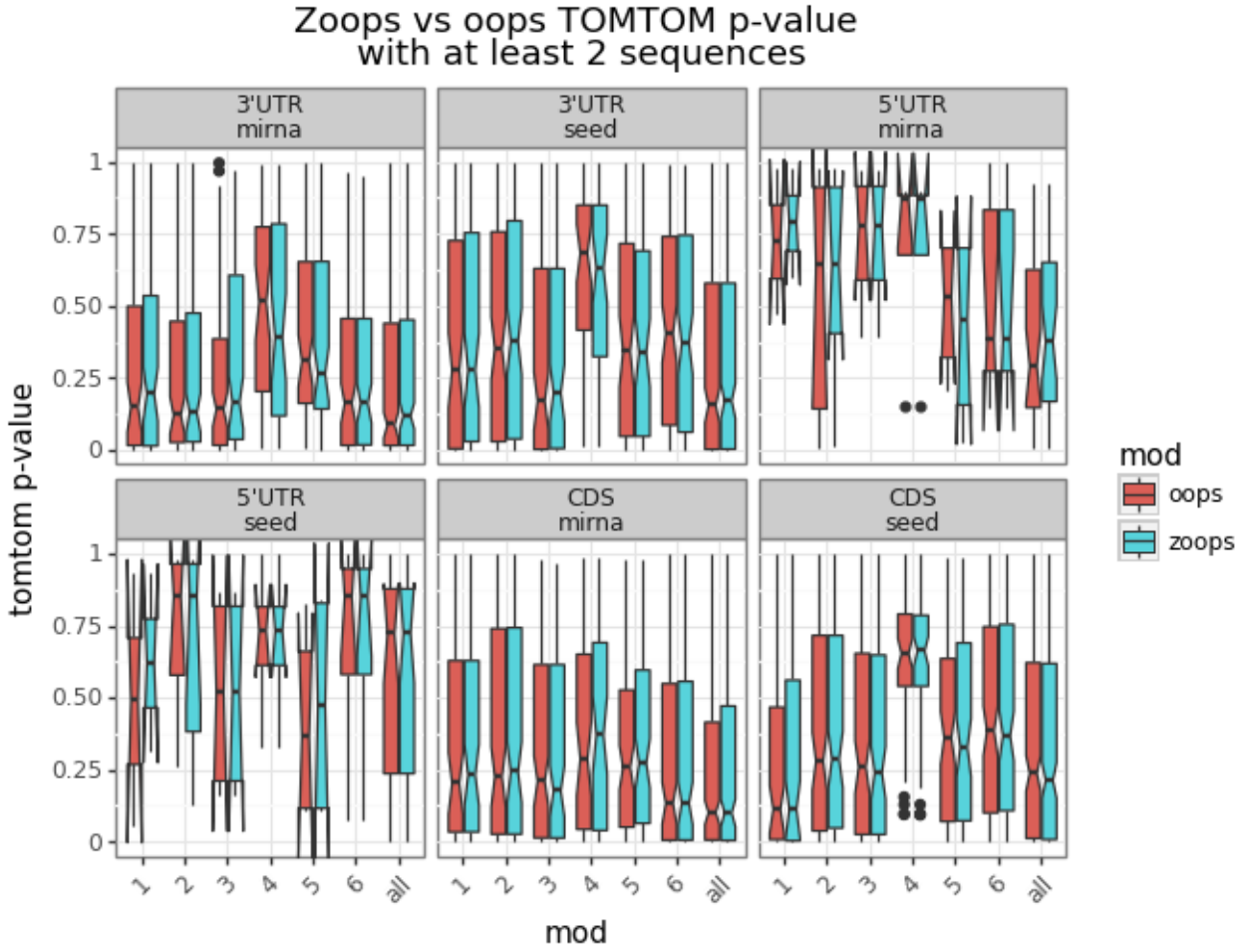


Figure E.7 – Comparison of the ZOOPS and OOPS motif distribution mode from MEME. We searched motifs with MEME in the mouse seq_{E_i} dataset and then compared the motif with the sequence of the miRNA i and the seed of this miRNA i using TOMTOM. Note that the boxplots' notches that concern the region 5'UTR are bigger than the boxplot itself. This is due to the low number of interactions in this region.

Table E.4 – Average width (column "Av. width") of the best motif of width between 6 and 23 found by MEME using the oops site distribution on the sequences given by CLASH. The column "#Seqs" corresponds to the total number of sequences for each subset.

Human Dataset				
	Seq_{E_i}		Seq_{W_i}	
Classes	**Av. width**	**#Seqs**	**Av. width**	**#Seqs**
all	11.31	18431	14.57	15670
I	10.36	3531	13.77	3030
II	11.50	3224	14.83	2770
III	11.84	4560	15.11	3907
IV	11.0	3318	14.29	2828
V	10.15	3534	14.25	2898
Mouse Dataset				
	Seq_{E_i}		Seq_{W_i}	
Classes	**Av. width**	**#Seqs**	**Av. width**	**#Seqs**
all	10.0	12992	14.72	11500
1	9.45	1556	13.18	1357
2	9.77	2588	13.40	2287
3	9.54	2182	14.01	1908
4	8.92	1061	14.35	922
5	9.90	2298	13.62	2057
6	9.42	2947	14.09	2624

Note that we did a test search with SMILE on Seq_{E_i} of the miRNA LET-7A looking for motifs of length 7 with a quorum of 65% and allowing for up to 1 substitution and found four motifs, namely $CTACCT$, $TACCTC$, $CCTAC$, and $CAACCT$ that are small variations of a same motif $CTACCTC$. When searching on Seq_{W_i}, we found 3,772 motifs which represent mostly noise.

ii Similarity of the MEME motifs with the miRNA and the seed.

We then verified if the best motifs detected by MEME matched the seed region or the miRNA sequence by comparing the matrices of sequences with TOMTOM [Gupta et al., 2007] from the MEME suite. For this purpose, the seed region or miRNA sequence is transformed into a matrix in which for each position we include a 1 for the nucleotide present at that position and 0 for the others.

Figures E.8c and E.9c show boxplots of the p-values given by TOMTOM when comparing either the seed or the complete miRNA sequence to the motifs predicted by MEME with the site distribution oops (one or plus) when the input has gradually more noise by using Seq_{E_i} then Seq_{W_i} and with miRNAs having either more than 2 targets or more than 10 targets.

When using only the regions detected by CLASH relative to the miRNAs having more than 10 targets, the classes for which the seed is very well defined (I, II and III in human and 1, 2, 3 and 6 in mouse) showed a better correspondence between the best motif

detected and both the seed and the miRNA. Although there is still a correlation between seed and motifs in Class IV of human, this trend is lost when using the complete miRNA sequence and not seen at all in both cases for Class V. The same is not observed in mouse.

Nevertheless, it is interesting to note that by adding more noise to the dataset by either using miRNAs with 2 or more targets or by extending the regions with 200 nucleotides down and upstream (Figures E.8c and E.9c), the correlation between motifs and seed, or between motifs and miRNA decreases considerably.

These results could be an indication as to why the seed match might be seen ambiguously as a good or bad feature when predicting targets, especially since we usually do not have the *a priori* information of neither the number of targets for a given miRNA nor the correct interaction sites with most sequencing experiments performed for miRNA detection and target prediction.

iii Location on the miRNA of motifs found in the mRNAs

Consensus motifs are at first glance dissimilar to the miRNAs and seeds

We saw that there is a correlation detected by TOMTOM between the seed and the motif. However, this is not clear for all motifs and for all classes. We continued to investigate this idea by considering the motifs individually and displaying them with a sequence logo.

This new sequence logo representation differs from the two presented in Section E.1, and corresponds instead to the classical sequence logo that represents the sequence conservation of nucleotides as defined in [Schneider and Stephens, 1990] and which are found for instance on the web server of MEME. This sequence logo is created from a set of aligned sequences (the occurrences of our motifs) and has for aim to show the frequency of the nucleotides at each position. This frequency of appearance of a nucleotide is close to the idea of conservation and is called residue in the original paper of [Schneider and Stephens, 1990]. If several letters are possible for a position, they are stacked at that position. However, in that case the residues are scaled in function of their frequency in order to represent by the height of the stacked letters the amount of conservation information in bits. For instance, no information is zero bits therefore, in the case of DNA, all four letters are equiprobable to appear in the sequence. On the contrary, the level 2 bits is the highest information and occurs when only one letter is found in that position on all the sequences.

The residue at position i is computed with:

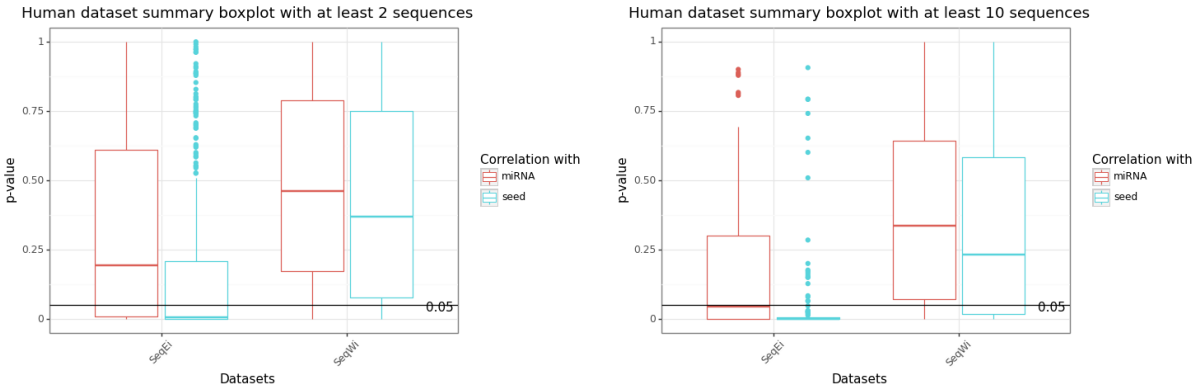
$$R_i = log_2(4) - (H_i + e_n) \text{ where } H_i = -\Sigma_{b=1}^{t} f_{b,i} log_2(f_{b,i}) \tag{E.1}$$

where $f_{b,i}$ is the frequency of the letter b at position i and e_n corresponds to the small-sample correction. An approximation of the latter is given by:

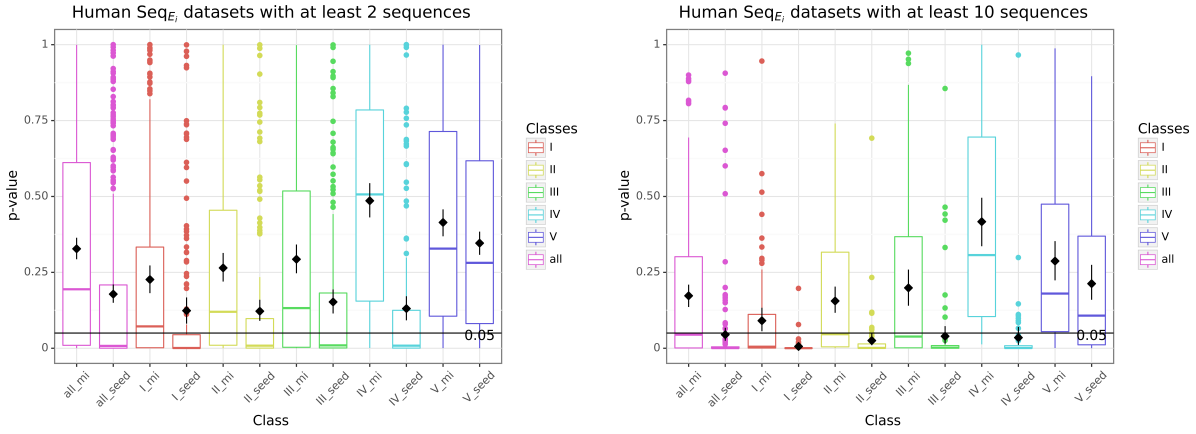
$$e_n = \frac{(4-1)/2n}{log_2(2)} \tag{E.2}$$

where n is the number of occurrences of this motif.

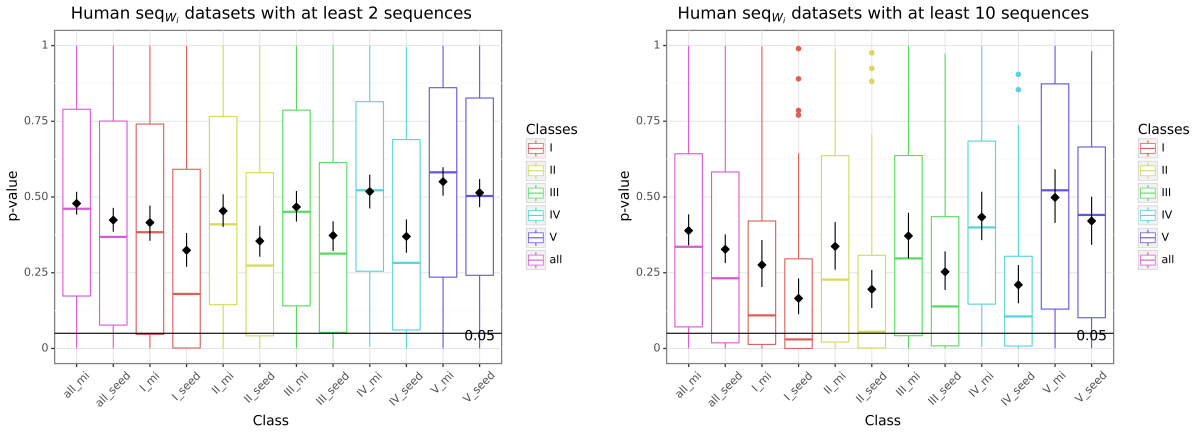
The sequence logos (Figure E.10) indicate that the likeness between the seed or RCseed (reverse complement) and the consensus motif is for most motifs not visible. Indeed, in



(a) Correlation with the miRNA and the seed summarized.

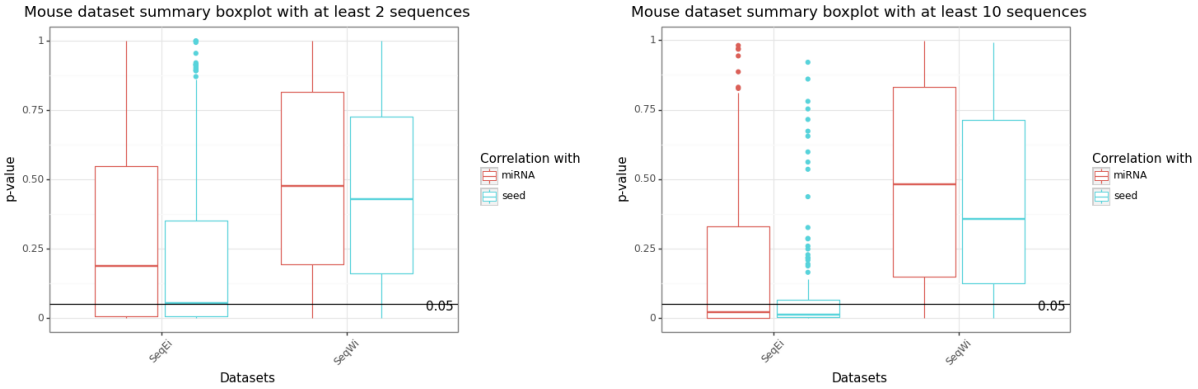


(b) P-value from the comparison of the motifs with the sequences given by CLASH. The boxplot indicates the distribution of the p-values for each class when comparing the best motif predicted by MEME with either the seed (seed) or the miRNA (mi) on datasets with more than 2 or 10 sequences.

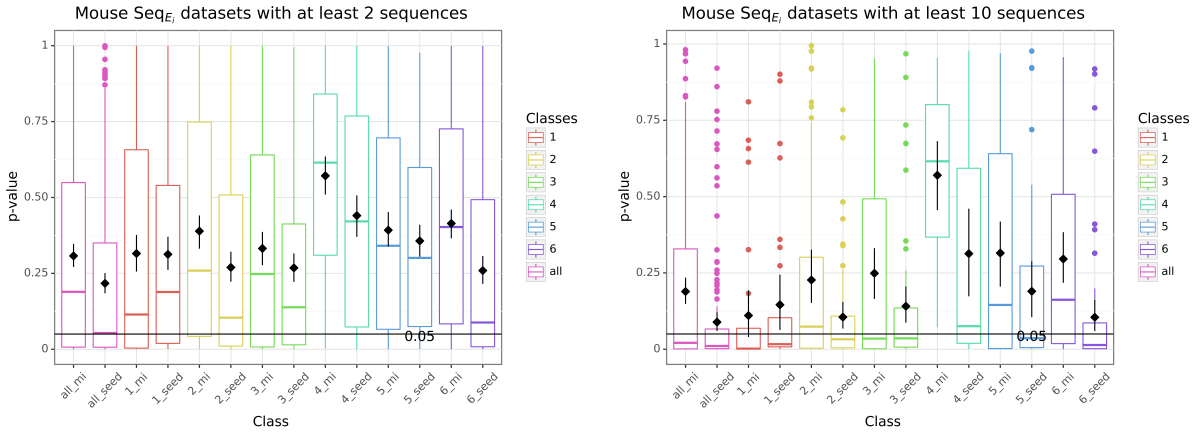


(c) P-value from the comparison of the motifs with Seq$_{W_i}$ datasets. The boxplot indicates the distribution of the p-values for each class when comparing the best motif predicted by MEME with either the seed (seed) or the miRNA (mi) on datasets with more than 2 or 10 sequences.

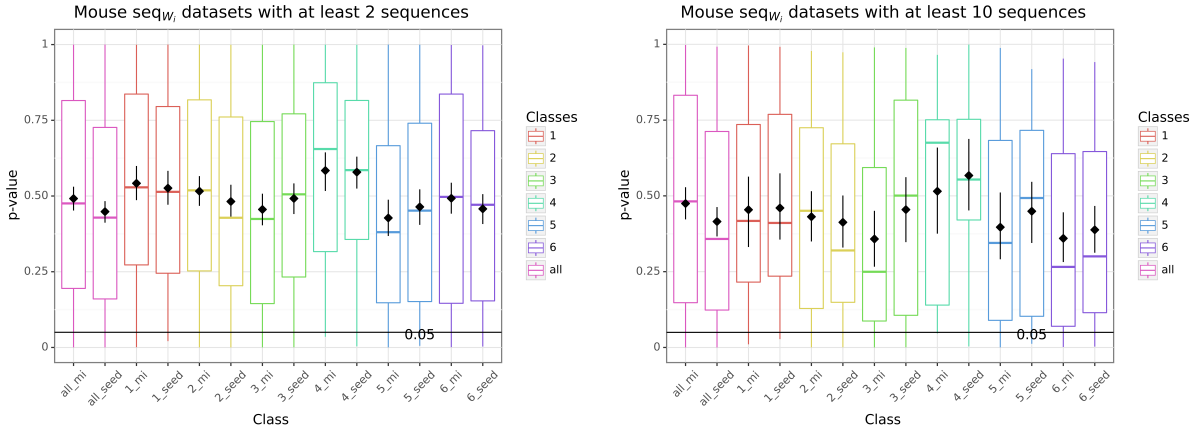
Figure E.8 – Boxplots on the human dataset.



(a) Correlation with the miRNA and the seed summarized.

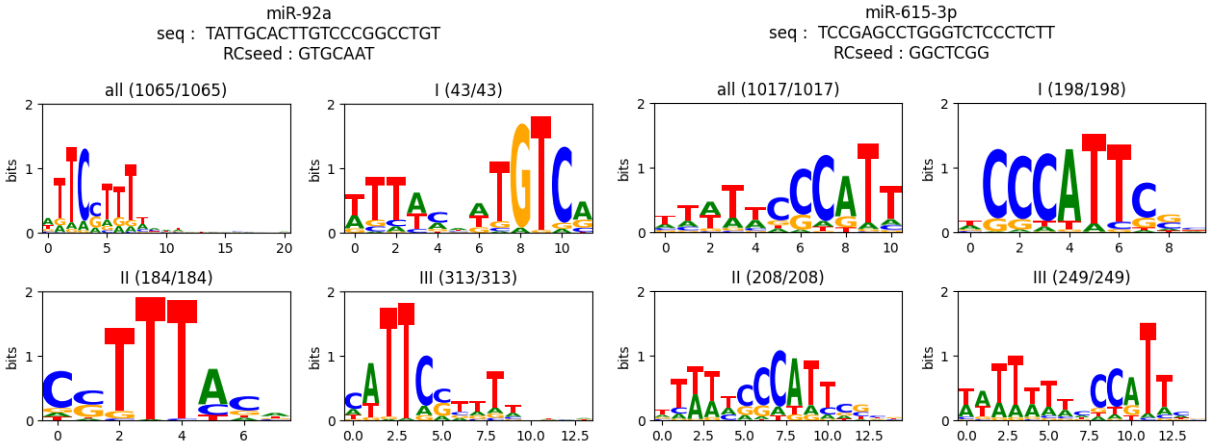


(b) P-value from the comparison of the motifs with the sequences given by CLASH. The boxplot indicates the distribution of the p-values for each class when comparing the best motif predicted by MEME with either the seed (seed) or the miRNA (mi) on datasets with more than 2 or 10 sequences.

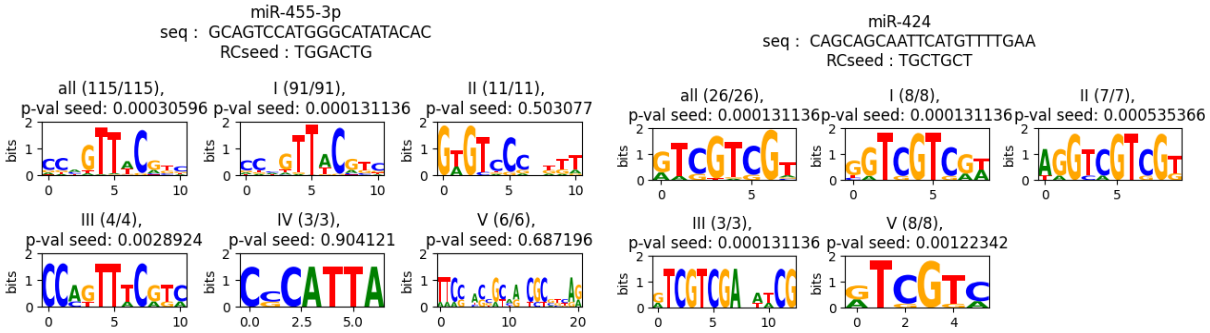


(c) P-value from the comparison of the motifs with Seq$_{W_i}$ datasets. The boxplot indicates the distribution of the p-values for each class when comparing the best motif predicted by MEME with either the seed (seed) or the miRNA (mi) on datasets with more than 2 or 10 sequences.

Figure E.9 – Boxplots on the mouse dataset.



(a) Motif generated from $\text{Seq}_{E_{mir-92}}$ and $\text{Seq}_{E_{mir-615-3p}}$. Mir92 is the miRNA with the most interactions in the human dataset. Mir615-3p is a randomly chosen miRNA.



(b) Motif generated from $\text{Seq}_{E_{mir-455}}$ and $\text{Seq}_{E_{mir-424}}$ which are the best and second best when comparing the motif found in Class I with the seed.

Figure E.10 – Sequence logo of the motifs found on $\text{Seq}_{E_{mir-92}}$, $\text{Seq}_{E_{mir-615-3p}}$, $\text{Seq}_{E_{mir-455}}$ and $\text{Seq}_{E_{mir-424}}$. These miRNAs are respectively the the one having the largest number of interactions, randomly chosen, the best, and finally the second best when comparing the motif found in Class I with the seed. The number of sequences is written in parentheses where the numerator is how many occurrences of this motif is found and the denominator is the total number of sequences of Seq_{W_i}. The model oops was used so both numbers are always the same. RCSeed stands for the reverse complement of the seed. Note that we used TOMTOM with default parameters which automatically compares two motifs together in both normal and in reverse complement form.

any of the motifs presented in Figure E.10 with the exception of *mir-424p*, there are only around two letters from the consensus motif related to the seed. It is even the case with *mir-455-3p* which has the best correlation between the seed and the motif according to TOMTOM.

However, in a few cases we have a fully or partially recognisable seed. For instance, the motif inferred from $\text{seq}_{E_{mir-424p}}$ (Figure E.10b) has $GTCGTCGT$ for consensus and the complement of the seed of the miRNA *mir-424p* is $TCGTCGT$ which is exactly the consensus motif minus the first letter.

This difference of best correlation between the seed and the motif detected by TOMTOM and what we can directly observe with the sequence may be explained by the fact that TOMTOM uses the number of sequences from which the motif is inferred, therefore a smaller consensus motif similarity with motifs inferred from 100 occurrences will usually have a better score than another motif with higher consensus motif similarity inferred from only 10 occurrences.

Motif location on the miRNA

We then wanted to better understand the link between motifs and miRNAs and the previous results incited us to perform the analysis with another comparison method. We chose for this Fimo [Grant et al., 2011] which scans sequences to find individual matches of a motif. Fimo was initially put aside as it could not precisely compare motifs with the seed because of the size of the latter.

We now therefore searched with Fimo where is placed on the miRNA i the motif inferred on Seq_{E_i}. When several matches were found, we selected the best. It should be noted that not all motifs have a match with the miRNA i. We show the number of motifs with a match in Table E.5.

We displayed this placement with a heatmap where a coloured position indicates the position of the motif on the miRNA. The colour varies from black to red and indicates the amount of conservation in bits: black means nearly no conservation and red means high conservation (see the definition and computation of conserved information based on the residue E.1).

The heatmap displaying all interactions (Figure E.11) emphasises that more than one third of the motifs are located on the seed of a miRNA, while the other two thirds are distributed evenly along the remaining of the miRNA.

When looking at the heatmap separated by class (see the left column of Figures (E.13 and E.12), we have a similar proportion of motifs placed on the seed except for the expected seedless Class IV and Class 4 on, respectively, human and mouse where we still see around $\frac{1}{4}$th of the motifs placed on the seed position. Fimo gives us the placement on the miRNAs and also a p-value of motif occurrence. This p-value is defined as the probability of getting an equivalent or better score by a match to a random sequence of the same length as the motif with a position on the sequences [Grant et al., 2011].

Motifs matching miRNAs with a p-value equal or below $5e^{-5}$ (see the right column in Figures (E.12 and E.13) show a remarkable resemblance to the theoretical interaction placement on the miRNAs for each class.

Motifs selected by their p-value as shown in the right column of the Figures (E.12 and E.13 represented a first small success in trying to infer interaction sites based on conservation. Indeed, with the motifs having a p-value equal or below $5e^{-5}$, we were able

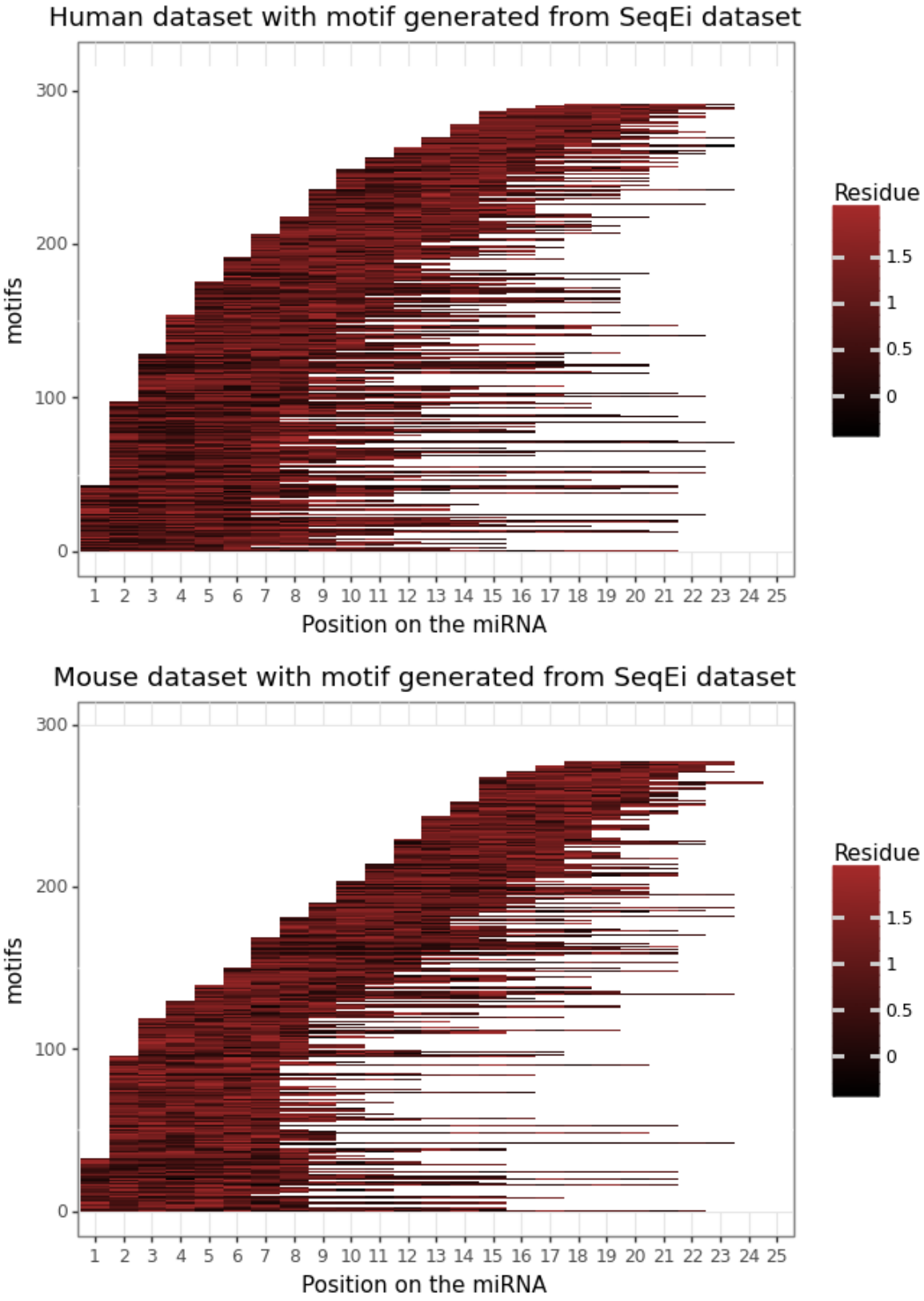


Figure E.11 – Location on the miRNAs of the motifs inferred on all interactions. Red means that this position is very conserved.

Human dataset			
Class	#miRNAs	#motifs with a match	#motif with p-value <= $5e^{-5}$)
All	316	292	109
I	158	150	81
II	163	152	57
III	178	162	49
IV	142	131	50
V	201	183	6
Mouse dataset			
Class	#miRNAs	#motifs with a match	#motif with p-value <= $5e^{-5}$)
All	299	277	31
1	129	118	2
2	166	154	12
3	150	140	11
4	96	90	4
5	131	118	3
6	178	167	22

Table E.5 – Reminder of the number of miRNAs per class, then the number of motifs inferred on SeqE_i with a match on miRNA i, and finally the number of motifs inferred on SeqE_i with a match on miRNA i.

to retrieve miRNA/mRNA interaction patterns as defined by the clustering of each class (see Section C.2) that are not the seed by searching only on the mRNAs. This is rather encouraging since the methods based on pattern recognition (see Section iii) were focused on the seed.

Specificity of the motifs inferred at the interaction sites of each given miRNA

Until now, we were always comparing the motifs found on Seq_{E_i} with their respective miRNA i. We were able to find a specific interaction pattern, however another question appeared: are we able to link the motifs inferred from Seq_{E_i} with the miRNA i or the motif has a better score with another miRNA?

To answer this question, we used Fimo again to compare the motifs with the sequences of all miRNAs, including the one (miRNA i) experimentally identified. We only selected the best match per miRNA when Fimo found more than one occurrence. We then sorted all miRNAs by their p-values and considered the miRNA i as retrieved when it was ranked first by Fimo.

In the majority of the tests, the correct miRNA i is not the closest to the motif found when looking for all occurrences. However, when selecting occurrences with a p-value below $5e^{-5}$, we partially managed to retrieve miRNA i from all the miRNAs in the human dataset especially for Class I. In the mouse dataset, the number of times we retrieved the correct miRNA was very low, even selecting occurrences with a p-value below $5e^{-5}$ (see Tables E.6).

In the human dataset, the results for Class V are the worst. This is expected due

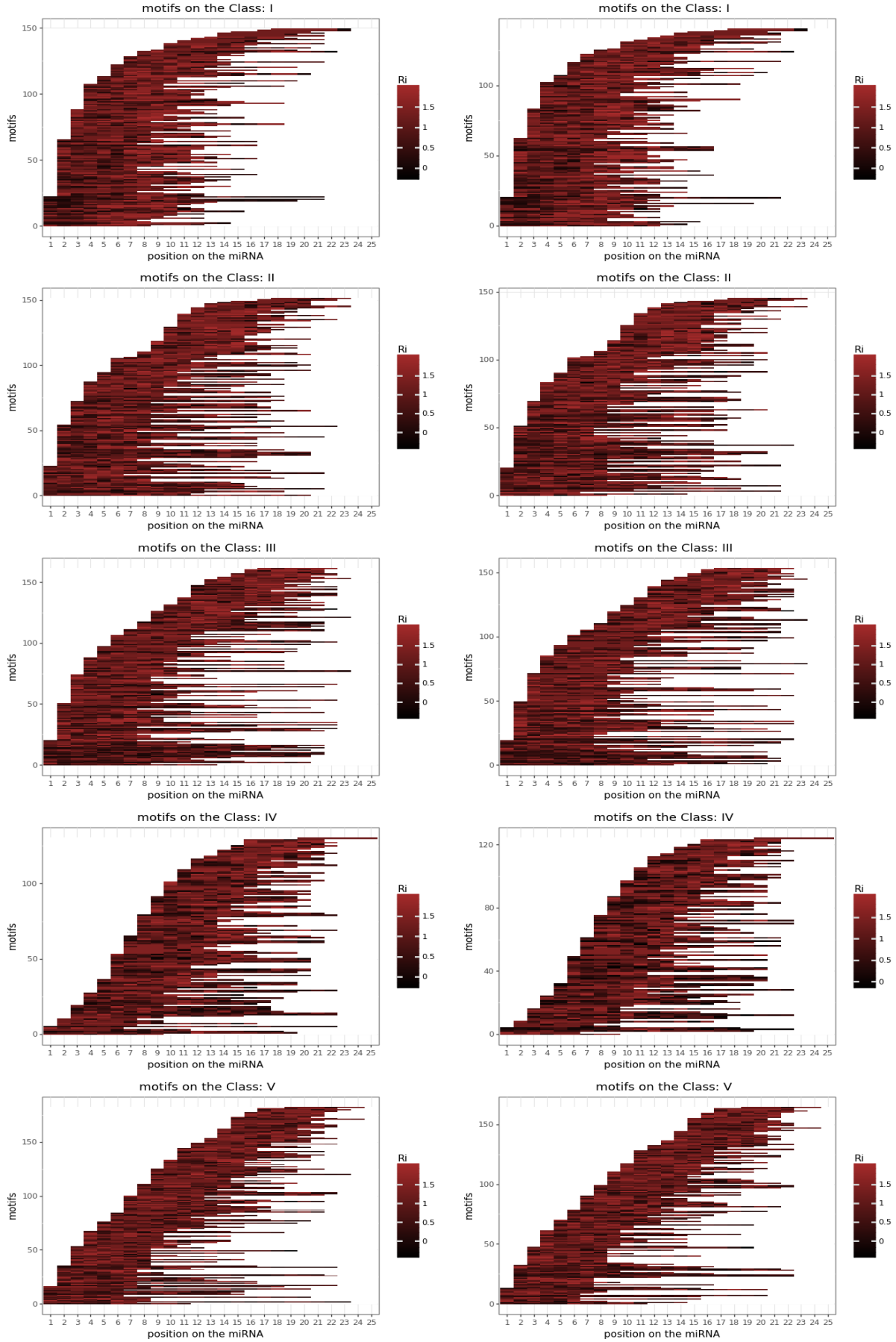


Figure E.12 – Heatmap for each class of motifs found on the human Seq_{E_i} dataset and their placement on their respective miRNA i. On the left column: all motif placements on their respective miRNA. On the right column: only the motifs that have a corresponding p-value as given by FIMO that is less or equal to $5e^{-5}$.

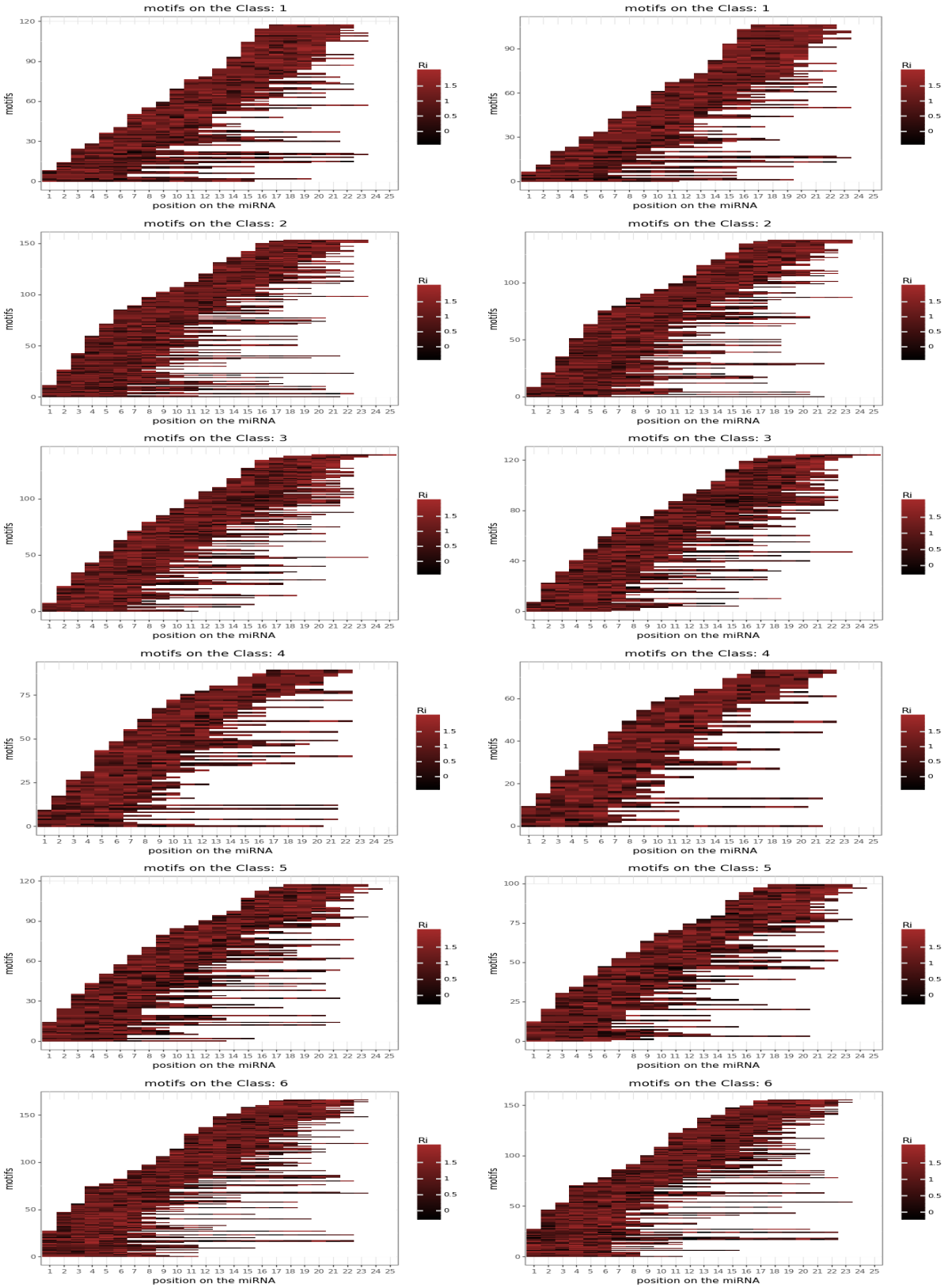


Figure E.13 – Heatmap for each class of motifs found on the mouse Seq_{E_i} and their placement on their respective miRNA i. On the left column: all motif placements on their respective miRNA. On the right column: only the motifs that have a corresponding p-value as given by FIMO that is less or equal to $5e^{-5}$.

Table E.6 – Number of times the motifs generated from Seq_{E_i} on the mouse and the human datasets have the best similarity with miRNA i.

	Human dataset					
	All occurrences			Occurrences with p-value <= $5e^{-5}$		
Class	**first rank of miRNA i**	**Top 10 of miRNA i**	**#motifs**	**first rank of miRNA i**	**Top 10 of miRNA i**	**#motifs**
Total	83	152	316	76	122	183
I	68	106	158	63	90	100
II	45	94	163	40	68	90
III	46	90	178	38	61	92
IV	40	87	142	31	61	73
V	3	27	201	2	8	60
	Mouse dataset					
	All occurrences			Occurrences with p-value <= $5e^{-5}$		
Class	**first rank of miRNA i**	**Top 10 of miRNA i**	**#motifs**	**first rank of miRNA i**	**Top 10 of miRNA i**	**#motifs**
Total	33	103	299	17	50	119
1	7	38	129	2	5	21
2	19	57	166	9	19	46
3	14	44	150	8	18	44
4	9	18	96	4	4	9
5	7	25	131	0	8	30
6	31	57	178	18	28	52

to the distributed nature of the interaction pattern of this class. The worst class in the mouse dataset is also Class 5, which could be explained by the non-paired pattern of the nucleotide 15 and around.

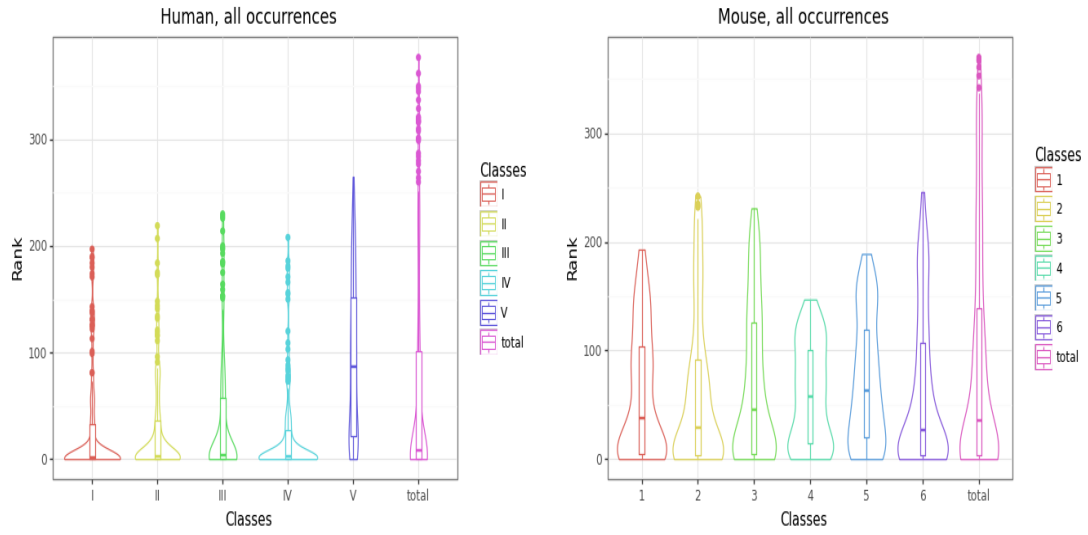
It is worth mentioning that the miRNAs in human are far better retrieved than those in mouse. In the mouse dataset, the correct miRNA is rarely ranked first. However, it is often among the best top 10. The results are shown in Table E.6 and the distribution of the rank of the miRNA is presented in Figure E.14.

E.3 Prediction of the interaction sites using intra-species conservation

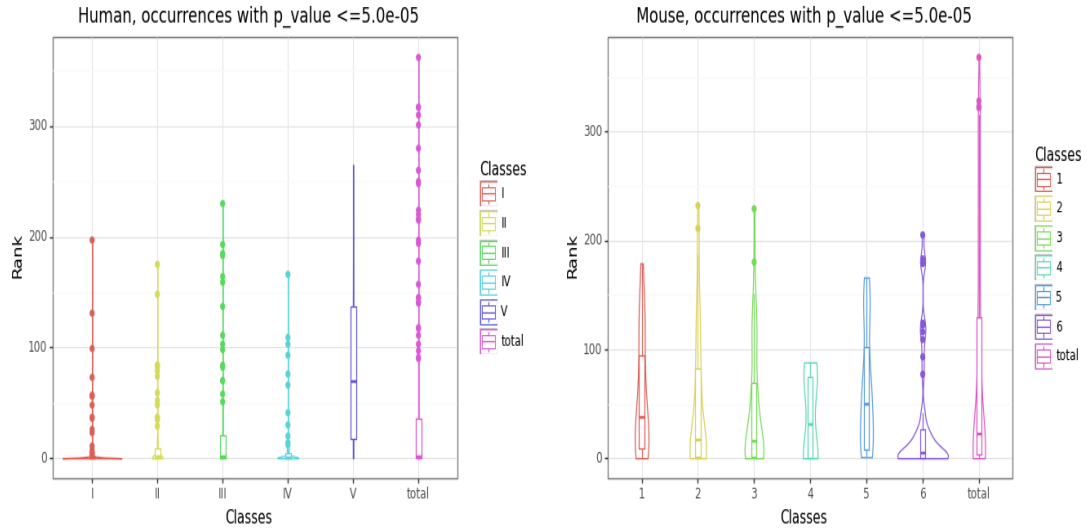
In this section, we are gradually injecting the previous knowledge gathered in order to help improve the discovery of new interactions for an already known miRNA, or even point to the existence of new miRNAs not yet identified.

The performance of the motif predictions was also evaluated using the precision, recall and F-score metrics for the miRNAs with, respectively, at least two and at least ten targets, and we considered the occurrence of a given motif predicted by Meme as a **strong true positive**, TP_{strong} for short, when it is **fully included within the site given by CLASH**. When the occurrence of a predicted motif overlapped the site given by CLASH without being fully included in it, this was considered as a weaker type of true positive and we therefore chose to also consider it separately. We named it **weak true positive**, TP_{weak} for short.

Considering the parameters used by Meme for motif prediction and the constraint of having only one solution, the false positives, FP for short, and false negatives, FN for short, have the same value. Notice that in consequence, the recall and the precision metrics also have the same value. We decided to only show the F-score.



(a) On human dataset, with all occurrences (b) On mouse dataset, with all occurrences



(c) On human dataset, only occurrences with a p-value $<= 5e-5$ (d) On mouse dataset, only occurrences with a p-value $<= 5e-5$

Figure E.14 – Distribution of the rank of the miRNA i against all other miRNAs when comparing the motif inferred in Seq_{E_i}

i Searching in noisier sequences for the interaction sites

In a similar way to the comparison of methods carried out in Chapter D, we created different Seq_{W_i} datasets according to the numbers of chimeric reads supporting each interacting sequence.

We then computed on the Seq_{W_i} datasets the F-score for all occurrences of all motifs for each class and region of the mouse and for each class for the human. We gathered the results in Figure E.15. Notice that in this experiment, we are scoring the placement of the occurrences independently of the motif.

The results obtained are independent of the class, region or depth of the reads. It is around 10%. An F-score of 0.10 indeed corresponds to 10% of occurrences that are inside the interaction sites. Moreover, the average width of the interaction sites is roughly 45 nucleotides so the width of a sequence of Seq_{W_i} is around 445 nucleotides, which means that a random occurrence has a 10% probability of being well placed.

All F-scores in relation to the mouse dataset are around 0.1, therefore not better than a randomly picked position. In relation to the human dataset, the F-scores are around 0.2 with a peak at around 0.3 for Class I which is only slightly better than random picking. Class V in the human is nonetheless around 0.1. Overall, these results are equivalent to a random picking and we decided to put aside the depth of the reads in the following experiments since it massively increases the number of datasets and with it the computational cost without adding any information at the current stage of the study.

Additionally, we wanted to understand from the previous results if the true positives are coming from the occurrences of a few motifs or rather globally from the few occurrences coming from all motifs that are true positives. In other words, we wanted to analyse the performance of the motifs instead of their occurrences.

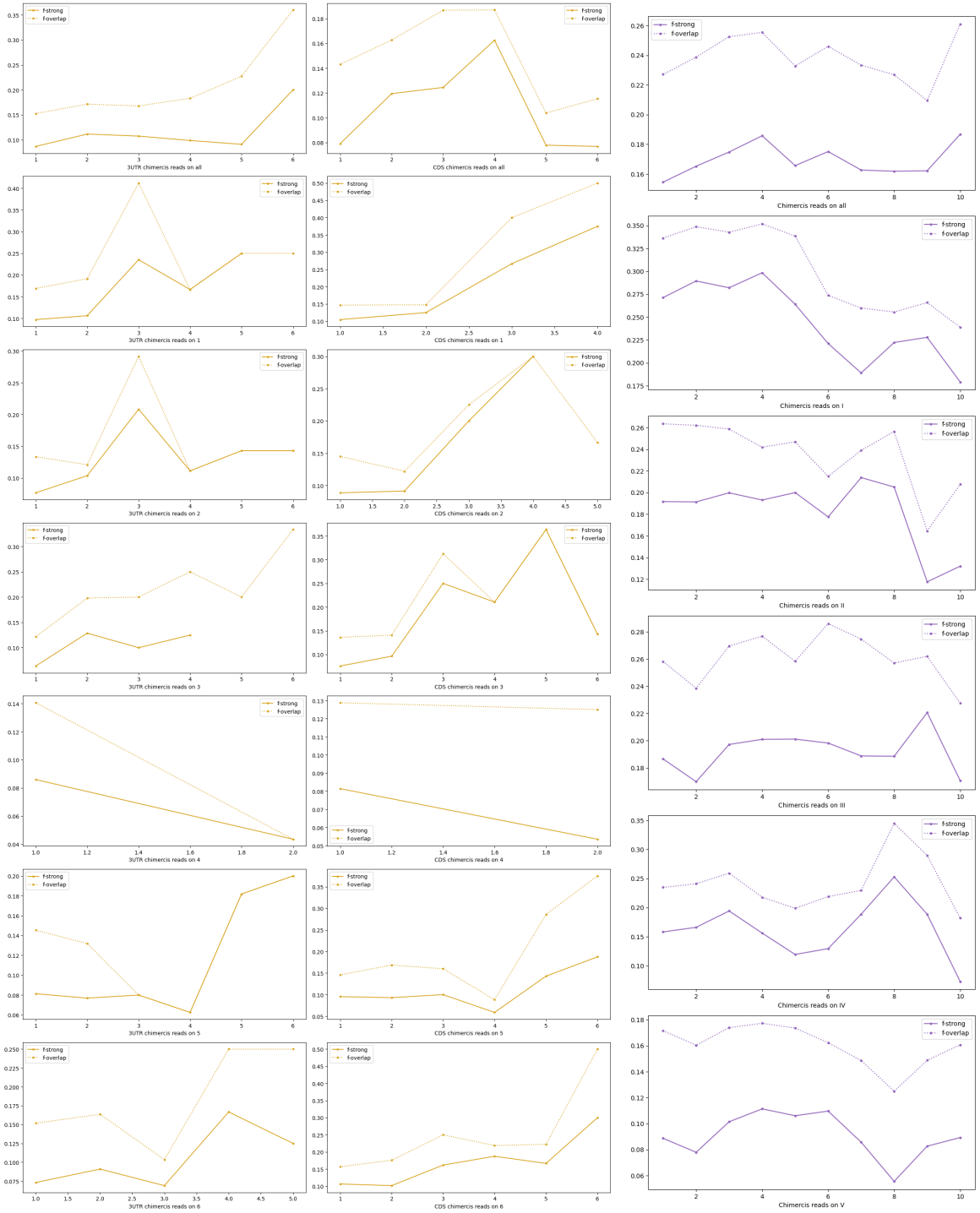
Therefore, we computed the F-score for each motif individually to verify the distribution of these metrics (see Figure E.16). We checked the F-score computed on the occurrences of a motif, which emphasises the fact that the majority of the motifs, as expected, do not correspond to the interaction sites but to other conserved parts of the sequences. There are, however, a few motifs that have near perfect F-scores, which means that they correspond at least partially to interaction sites. These conserved motifs are as expected once again to be found more in the first three classes of the human dataset (because of the seed) and less so in relation to Classes 1 and 2 in the mouse dataset.

ii Performance of the best motif

To further explore the approach of intra-species, we needed to refine the experiment and test if focusing only on the best motifs inferred on Seq_{E_i} and that retrieve the miRNA i when compared to all other miRNAs as described in Section iii would increase the F-score in relation to the Seq_{W_i} dataset. Furthermore, we used only motifs that have at least an occurrence with p-value $<= 5e^{-5}$. We call such motifs θ_i

The results of this evaluation are displayed in Figure E.17.

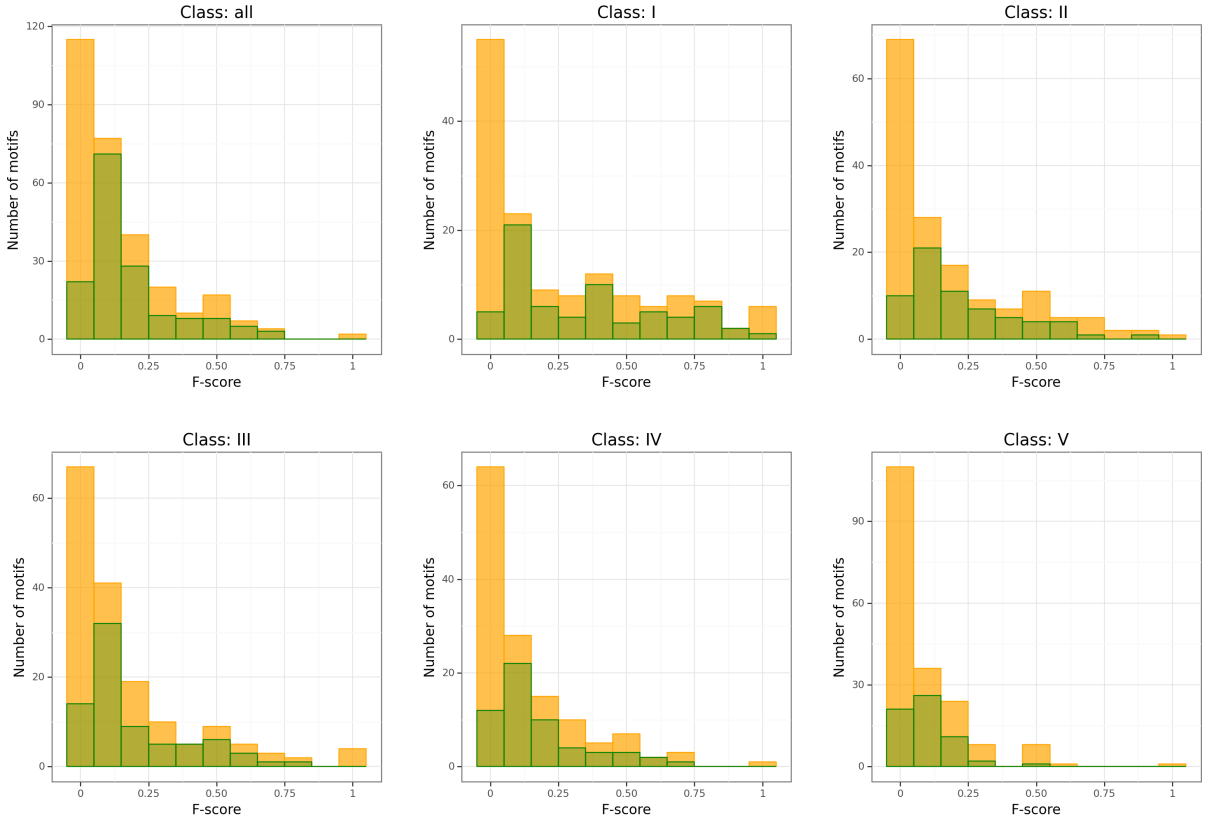
In the mouse dataset, there are only a handful of motifs θ_i (see Table E.5). However, five of them had a perfect F-score with two in the Class total, one in Class 3 and two in Class 6. In Class 6, while the F-score of the motifs θ_i is not perfect it is rather good (around 0.8). In the other classes 2, 3 and 5, they are not very good, with an F-score around 0.5 or worse.



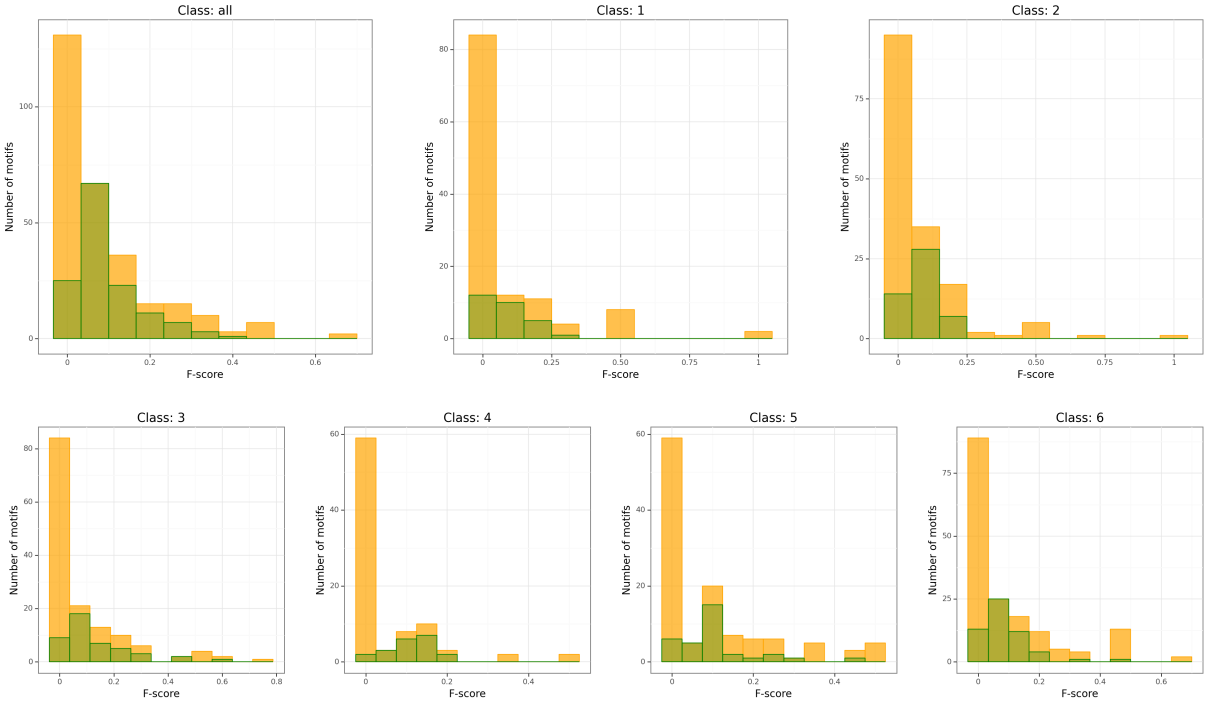
(a) MEME motifs inferred in the mouse Seq_{W_i} for each class and region.

(b) MEME motifs inferred in the human Seq_{W_i} for each class.

Figure E.15 – MEME motifs inferred in Seq_{W_i}.

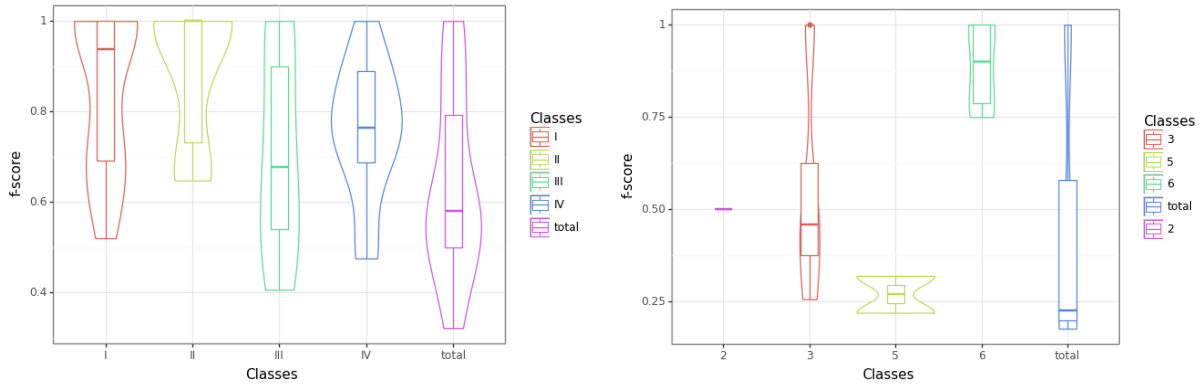


(a) Human dataset



(b) Mouse dataset

Figure E.16 – Histogram representing the number of motifs by F-score on Seq_{W_i} for each dataset and for each class. In orange are the datasets with at least 2 targets, and in green those with at least 10 targets. Note that the green dataset is a subset of the orange one which implies that the green number of miRNAs is equal or smaller than the orange one.



(a) Human dataset

(b) Mouse dataset

Figure E.17 – Distribution of the F-score when searching on Seq_{W_i} with motifs θ_i.

In the human dataset, there are more motifs θ_i compared to the mouse one and this analysis identifies 21 motifs θ_i with a perfect F-score (respectively three, eight, six, three, and one in the Class total and then Classes I to IV. When they are not perfect, their scores are still very good, around 0.9 for Classes I and II and around 0.75 for Classes III and IV.

iii Performance on complete mRNAs

Our final test was to assess the performance of the motifs θ_i on an even noisier dataset. Hence, in relation to the dataset Seq_{W_i}, we created the dataset Seq_{C_i} that gathers all the complete mRNAs that are targeted by the miRNA i.

The result of this test is displayed in Figure E.18.

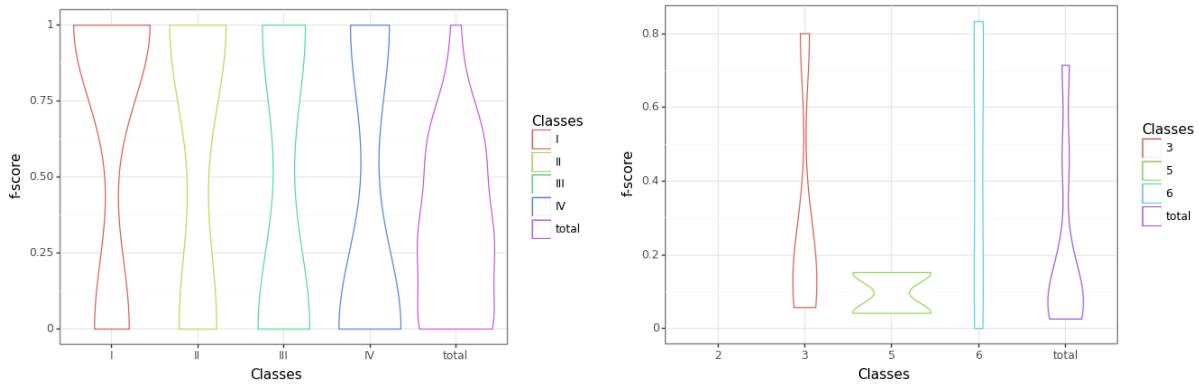
We have intriguing results on the human dataset. There are 30 motifs (respectively one, twelve, seven, five, and five in Class total and in Classes I to IV) that have a perfect F-score of 1.0, which is more than the 21 on Seq_{W_i}. It is puzzling that we find interaction sites in a noisier environment. However, one possible reason to explain this behaviour could be related to the way how FIMO computes the p-value according to the input sequences. As we only take the best occurrence for each sequence, changing the p-value of some occurrences might change what are the best occurrences.

In the mouse dataset, the results are less surprising with no perfect score and with a best F-score of 0.83.

In the end, these results demonstrate a possible, yet difficult usage of this approach even on complete mRNAs. It is interesting to note that the classification in different classes to search the initial motif θ_i is mandatory because motifs θ_i inferred on all interaction classes (Class total) have lower results than those inferred from each of the individual classes.

E.4 Perspectives

The results obtained in this first analysis of the possible usefulness of exploiting intra-species conservation and the associated over-representation of motifs to infer new inter-



(a) Human dataset

(b) Mouse dataset

Figure E.18 – Distribution of the F-score when searching on Seq_{C_i} with motifs θ_i.

action sites and thus also new miRNAs show some promise. However, there is still a lot of work to be done in order to be able to benefit from such feature. It is also obvious that this feature will have to be used together with others.

The global and deterministic approach of Smile could also be useful to infer motifs composed of more than one box that could target specific classes of interaction sites, but this would require also that we improve the results provided by Smile through a clustering of the motifs found, especially when substitutions are allowed.

There are several other paths that we could explore. One would be to limit the input of any motif inference method to only the highly accessible mRNA regions, thereby potentially reducing the amount of noise.

Although we put it aside for now, it might be interesting also to study under-represented motifs as inferred by Smile.

Finally, since we developed a process to validate the fitness of a motif with an interaction, it might be interesting to use this process with classical machine learning algorithms such as genetic ones to obtain the optimal parameters that would ideally retrieve only motifs that are specific to this interaction, if these exist.

Chapter F

Contribution: Web-service for finding motifs

The main purpose of this chapter is to give a second life to the motif inference method Smile [Sagot, 1998, Marsan and Sagot, 2000] previously developed in the team. Although it is quite old, it has features that current methods of motif inference, such as those from the Meme Suite, handle in a different way or lack altogether. These include, for instance, allowing to impose some constraints on the alphabet of the motifs, or to look for motifs composed of several boxes. Many such methods are furthermore heuristics – this is in particular the case of the methods from the Meme Suit – while Smile is an exact algorithm.

Given the vast diversity of the persons working in computational biology, a tool has to be accessible to someone not used to command lines. The best solution for this is to make it available on a website. We started with a desktop application based on Electron. Electron is a framework for building cross-platform (Mac, Linux, Windows) using web codes such as HTML, CSS and JavaScript. From this desktop application, we migrated to a website which is now running and available at `http://134.214.213.44/`. However, it is in a prototype stage that still requires work together with the implementation of new features in order to be made fully public. This website is presently contained in a virtual machine hosted on the *Cloud girofle* cluster inside the LBBE laboratory.

The whole development of this web-service was pursued with the idea of easy scalability, low maintenance and simplicity of usage for users from both the biological and computer science communities.

We divided this chapter in two sections that are related to the two traditional Methods and Results sections of a paper. The first one briefly explains the framework and organisation of this application. The second section presents its current functionalities and usages.

F.1 Web-service framework

In this section, we will rapidly present the global framework that supports this application.

There are four parts that compose such framework, they are schematised in Figure F.1. These parts are:

— The client wants to analyse sequences and then access the results in a reasonable

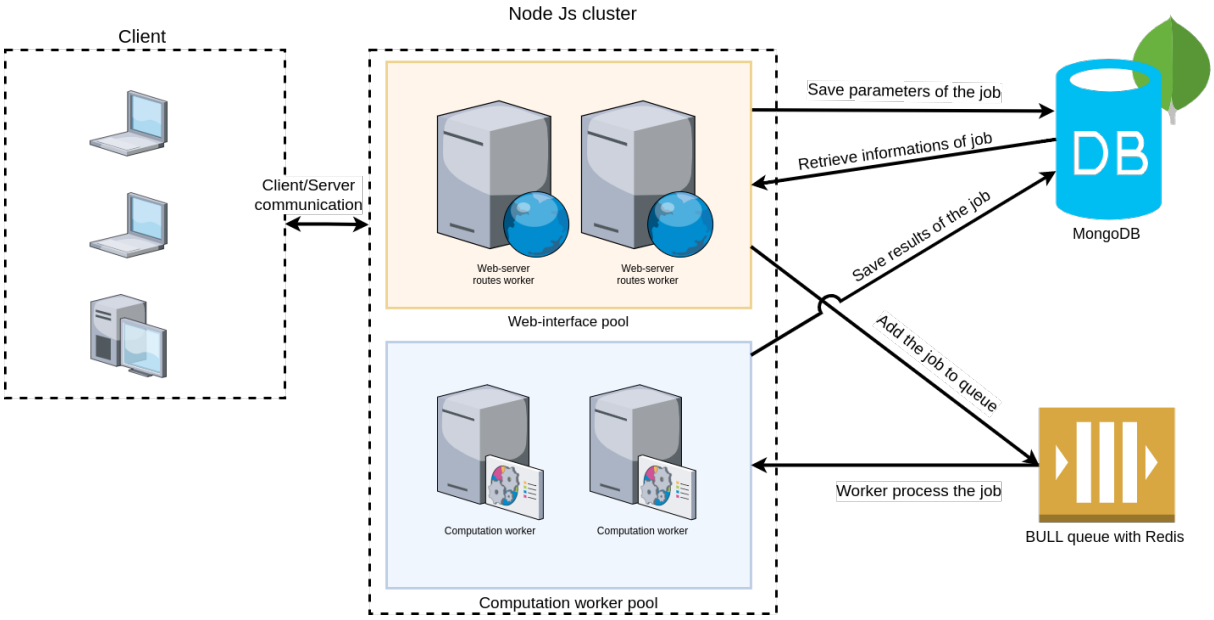


Figure F.1 – Diagram of the framework of the Web-service.

time. He/she might also want to share the results obtained.

— The Database MongoDb stores the job parameters and job results.

— The queue from the API Bull based on Redis works as a buffer that stores uncompleted jobs until a computation worker is available to process the job.

— The NodeJs Cluster contains several workers split in two types: web-server route workers and computation workers. The web-server route workers handle the communication with the client by taking requests such as parameters and input sequences and delivering the results of the job. The computation workers are the task force of the whole application. They handle the CPU intensive calculus and are divided in functions called processors which compute a single task, for instance the motif inference by Smile. The separation between Web and CPU workers prevents slowness and stuttering of the application responsiveness. Both worker types interact with the database and the Bull queue. The web-server route workers feed jobs to the queue whereas the computation workers eat the jobs from the queue.

The choice for the server environment was NodeJs. There are several reasons that support this decision. One of them is that the initial iteration of this project was the desktop application with Electron which is built on top of NodeJs to provide access, among other things, to the file system. NodeJs is also very popular and used by *e.g.* Netflix, LinkedIn, IBM...

i Virtual machine and dockers

A virtual machine, or VM for short, is the emulation of a computer inside another computer. It enables a strong division of several standalone instances for security purposes and convenience. Each instance of a VM can have its own operating system and therefore prevent minor and even major bugs or malicious behaviour to affect other VMs. To reduce

the hardware cost and for the other advantages mentioned previously, VMs are frequently used on a big computer called cluster. The current virtual machine hosted on the cluster has the technical specificity of being a decent private computer with 16GB of RAM, 8 CPU cores and 100 GB of memory. However, all the blocks of this application can easily be upgraded with more processing power or with cores to allow parallel computation by adding more workers.

In order to facilitate a possible future migration, we ensured a minimal required package list for this application to work on a virtual machine. The required packages are:

— Fail2ban: a package that prevents malicious persons to brute force the ssh connection.

— Nginx: a package that serves as a reverse proxy and links the local ip of the application with the public ip.

— Python3: Python is required to simplify usage of current and future analytic scripts.

— IntaRNA: this is a miRNA-mRNA interaction predictor which can also serve to align two RNA sequences.

— Docker and Docker-compose: These two packages ensure that dockers can run and that different docker modules can communicate.

Notice that Docker is an open source platform that helps developers to package, deploy and run applications in containers. These containers embody everything that is needed for the execution of an application, and can consequently be executed on any machine with Docker installed. They are basically minimalist VMs with comparable advantages of security, scalability, consistency, isolation and efficiency.

Our whole application is inserted in 3 different Docker containers which consist of Nodejs, Redis and MongoDb. They communicate together via Docker-Compose which is a tool that allows the cooperation of several Dockers by launching or closing all Dockers at the same time with a command.

ii Storage implementation with MongoDb and Redis

There are two types of memory:

— Volatile Memory which is a temporary memory that will be emptied when the machine is powered off. Random Access Memory (RAM) is the common volatile memory. Despite being expensive, RAM is very time efficient in writing and reading.

— Non volatile Memory which is a long term memory and retains the data even when the machine is powered off. Hard drives and SSDs are the common non volatile memories. They are cheaper, yet slower in terms of reading and writing time than RAM.

Databases can use one type of memory or both, depending on their use-case. For instance, MongoDb uses only non volatile memory and is aimed for long term persistence with an implementation of various mechanisms that improve the reading and writing delay. Redis, on the contrary, uses both types of memory and aims to store hot-data (data that needs frequent or fast access). Redis mainly stores data in RAM, however there are mechanisms that ensure the backup of the data in case of power shortage. These backup data have to be written on non volatile memory in order to survive a power shortage.

Queue storage organisation

The finite amount of computing power available in our application and the possible high number of requests from clients coerce a prioritisation in the order in which the requests are processed. A simple and fair method of the first-in, first-out approach implemented by a queue is a way to address this.

The chosen implementation of the queue is managed by an API called Bull https://github.com/OptimalBits/bull which stores data in a Redis database as the queue. Bull ensures the atomicity of the queue movement. In other words, Bull guarantees that once a transaction of the queue is started, it will end and update accordingly the database. This atomicity prevents the partial update of the database which can produce jobs stuck inside the queue, or even worse with jobs stuck in a computing phase which take slots that would be available for other jobs.

While the job is being computed, the results are parsed and stored in mongoDB.

Persistence storage organisation

Despite the persistent memory, MongoDb is very efficient for searching inside a large volume of data. This is very useful in our case, for instance for server-side tables that display only 10, 50, 100... motifs instead of all the 200,000 in a single time.

We opted for the implementation of MongoDB due to its free open-source nature and the fact that it is one of the most popular database tools available.

MongoDB is a document-oriented database which means that MongoDB uses a hierarchical structure that can be compared to the organisation of an XML or JSON. There are three kinds of structures inside a MongoDb instance that can be created to organise and store different data types: the structure *database* contains one or several *collections* which contain one or several *items*.

We organised our instance of MongoDb Docker with two databases.

Database-1 stores small organisational data that are mainly useful for the correct execution of the application; we call this type "job metadata". Database-1 contains one collection with several items which are the job metadata. An example of metadata is shown in Listing F.1.

Listing F.1 – Item of the database-1 that contains jobs-metadata as saved in MongoDb

```
{
  _id: ObjectId("6419cd90eb0f39a41a42698f"),
  param: {
    origin_input_fasta_name: 'extended_both_class-III.fa',
    cl_input_file: 'server_tmp_file/1b2b9cf889ebc94631a258a24cf98531',
    status: 'S-Parced',
    alphabet_symb: [ 'A', 'C', 'G', 'T' ],
    alphabet: 'server_tmp_file/5acfe60eb727945f53beb1a99f54aa93f97946c9',
    email: 'nicolas.homberg@inria.fr',
    custom_name: 'III',
    quorum: '0.06347594261774787',
    boxes: '2',
    b1_min_l: '4',
```

```
    b1_max_l: '5',
    b1_sub: '1',
    b1_min_spa_len: '10',
    b1_max_spa_len: '10',
    b2_min_l: '3',
    b2_max_l: '4',
    b2_sub: '1',
    len_min: '7',
    len_max: '9',
    substitution: '1',
    task_id: '41'
  },
  lastupdate: ISODate("2023-03-21T15:54:37.846Z")
}
```

The job ID (as *_id*) is provided by MongoDb when inserting the new job and is then used during all the process to identify this job. This ID is also carried over to the URL of the results page. Parameters (denoted by *param*) contain the parameters of Smile that are adapted for the usage of the application.

It contains the information about:

— *origin_input_fasta_name* which is the original name of the input fasta file as given by the client. This information is helpful in order to give it back to the client as a reminder of his input file. The key *custom_name* serves a similar purpose and allows the user to name its job. The fasta input file name has to be changed internally because two clients could have the same filename with different sequences (for instance, '*test.fa*'). At the reception of the client input fasta file and to prevent redundancy of several identical fasta files, we compute the *md5checksum* of the file and compare it with all currently saved files. If this file already exists, we use it directly and if it is not present, we save this file with its *md5checksum* result as the name. This file path is stored in the *cl_input_file* key where *cl* stands for "command line." Using the *md5checksum* results as the name allows for a fast comparison with all existing files without having to recompute the *md5checksum* for them.

— *Status* which is an internal code for the job state. *Status* can contain the following code:

  — *created* indicates that only the input fasta file was given by the user. Indeed, the fasta file is sent once it is filled-in by the user whereas all other parameters are only sent when the user clicks the "Run smile" or "Run smile + validation" buttons. This behaviour allows to secretly compute the *md5checksum* and ensures the ability to read potentially huge fasta files.

  — *Pending* indicates that this job is waiting in the queue.

  — *Computing* indicates that this job is being computed.

  — *S-Parced* indicates that the results of Smile are being computed and stored in the database.

  — *V-Parced* indicates that the results of the validation of Smile are being computed and stored in the database.

— *alphabet_symb* which indicates the letters of the alphabet to be used. It is needed for custom alphabets provided by the client. SMILE reads the alphabet as a file, therefore, these letters are written in a file if it does not exist already and the path is saved in the key *alphabet*.

— *email* which is a key that serves to store the email address and then to send an email when the job is completed.

— *task_id* which is provided by the BULL queue when a job enters the state 'computing'. It allows the retrieval of the job progress while it is being computed by the computation worker processors.

— *lastupdate* which contains the date of the last update of this item in the database, for instance when the job-status changes. It is essential to keep track of the age of the data so that we can delete old ones frequently. The deletion of items based on a date is a feature of MONGODB. We set the lifetime of an item to one week. We added more layers of data cleaning by setting a process similar to CRON-TAB that, every day at 2a.m., automatically iterates over all existing files (fasta files and alphabet files) and deletes them if they are not used in any item of Database-1. The other keys store the parameters of SMILE.

Database-2 on the other hand stores the result data useful for the client and these results can rapidly reach an important size. It is organised in a collection per job named after the job ID with each collection containing motifs stored as items. An example of a job collection data is shown in Listing F.2.

Listing F.2 – SMILE results data as saved in MONGODB

```
[
  {
    _id: 0,
    motif: 'AAAA',
    seq: 2,
    occ: 3,
    seq2pos: { '0': [ 1, 0 ], '1': [ 0 ] },
    Z_1occ: 0.99,
    Z_All: 0.99,
    p_1occ: 0.16108706105163917,
    p_All: 0.16108706105163917
  },
  {
    _id: 1,
    motif: 'AAAT',
    seq: 1,
    occ: 1,
    seq2pos: { '1': [ 1 ] },
    Z_1occ: 0.99,
    Z_All: 0.99,
    p_1occ: 0.16108706105163917,
    p_All: 0.16108706105163917
  }
```

]

For each motif, we store an ID that corresponds to the order in which the motif is parsed. We have then the number of sequences and occurrences stored in the keys *seq* and *occ*. *seq2pos* is a dictionary that stores the position of the occurrences in the sequences. This dictionary alone rapidly contains thousands of occurrence positions spread over several hundreds of sequences. This is for only one motif whereas we may have thousands of motifs. Both the number of occurrences per motif and the number of motifs accumulated can produce a collection having a size close to 1 GB for just one one motif inference job. *Z_1occ*, *Z_All*, *p_1occ* and *p_All* store the Z-score and p-value computed for, respectively, only one occurrence per sequence and for all occurrences.

iii Modifications done to Smile

Smile is the main pillar of this application and is composed of several modules:

— *p_bloc* with the executable name *x-smile* which finds motifs using a suffix tree.

— *Sigstat* with the executable name *e-smile_shuffling* or *e-smile_Ushuffling*. Both take as input the motifs found by *x-smile* and evaluate if a motif is significant by comparing its number of occurrences with the number of occurrences of the motif in sets of sequences obtained by shuffling the input ones.

We brought some modifications to those modules. We transformed *e-smile_shuffling* into *e-smile_Ushuffling* by changing the shuffler used in the 1992 version of Smile to the faster and more recent U-shuffle [Jiang et al., 2008], notably used in the Meme Suite. We also added another implementation that helps dealing with the massive amount of data which is the modification of the output of *x-smile* from a file output to a stream into the standard output (stdout). We simultaneously changed the input of *e-smile_Ushuffling* to read from the standard input (stdin). With these modifications, we can pipe the output of *x-smile* directly to *e-smile_Ushuffling* and remove the slow process of writing and reading several times from a file. In addition, between *x-smile* and *e-smile_Ushuffling*, we can add an intermediate process that will simultaneously parse the data of *x-smile* while transmitting these to *e-smile_Ushuffling*. The parser sends the data processed to the MongoDb Database-2 in a bulk of 1000 motifs for better performance.

We also modified the output of *e-smile_Ushuffling* to be a stream. The small gain in performance is due to the same reasons as for the output modification of *x-smile*.

iv Graphical optimisation

The front-end of an application is everything that is sent to and interacts with the client. It is opposed to the back-end that composes everything that is server-related. The front-end code has to endorse the responsibility of dealing with a vast range of possible hardware and software. Indeed, despite screen sizes ranging from a phone to an ultra-wide screen, the front-end implementation should be comfortable to wield and arrange the layout of the screen component accordingly. Additionally, as the different web browsers such as Safari, Google Chrome, Opera, are not always compatible with all features, we implemented a popular framework called Bootstrap which provides a template for customisable components such as navigation menus, buttons and typography that adapt to the screen size and to various devices.

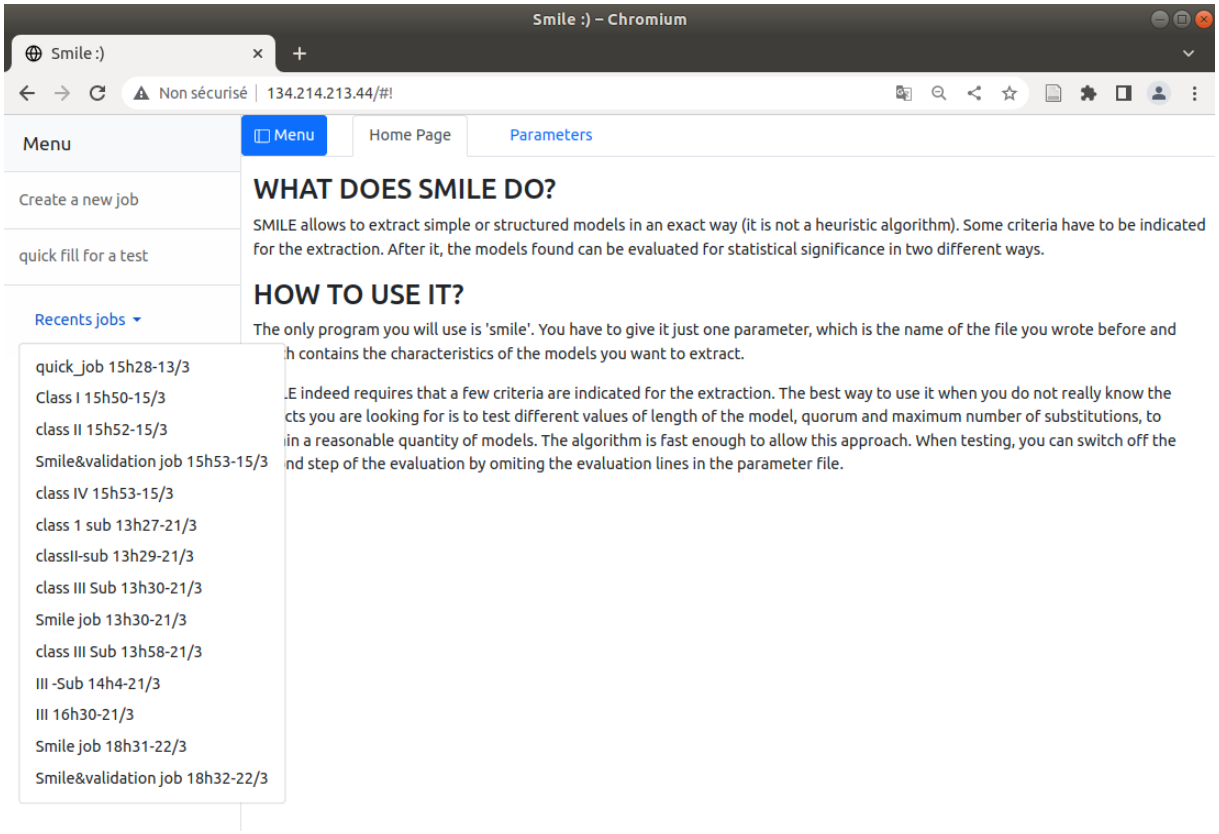


Figure F.2 – Home page

We decided that the best way to display the SMILE data and results is with tables and plots. The table is created with the help of the package *data-table* which requires J-QUERY. *Data-table* allows the user to sort, select and even to delete data in the table and enable a server-side table. A server-side table makes use of a database hosted on the server side. This prevents the whole table from being sent to the user. It is the server that retrieves part of the wanted data and sends it to the front-end code that will enable the rendering of the table.

F.2 Usage of the web-service

In this section, we demonstrate the current state and features of the web-service. It is in a functional state, yet can still be vastly improved in many ways.

There are currently only two pages or states for this web-service, the main page or home page permits the input of the parameters and launches a job which executes SMILE alone or SMILE followed by a step of validation of the motifs inferred. Once a job is launched, the web-service redirects the user to the results page.

We present the home page and the range of the possible parameters in the first section and then present the features of the results page in the second section below.

i Input of the parameters of SMILE

The home page of the web-service is displayed in figure F.2. It currently contains a small description of the tool SMILE that will have to be updated. There are two menus, one on the side and one at the top. Both menu items vary depending on the home page or the result page. The side-menu which can be hidden by pressing the blue Menu button or if the size of the screen is too small, contains three elements: 1. *Create a new Job*, 2. *Quick fill for a test* and 3. *Recent jobs*.

Recent jobs (see Figure F.2) store all the previous results page launched by the user. They are named after the custom name given by the user with the hour and date of creation. If there is no custom name provided, the jobs are named with either *Smile job* or *Smile&validation job*.

Both *Create a new Job* and *Quick fill for a test* redirect to the parameters view, also accessible with the top menus. *Quick fill for a test* will additionally fill out in advance some parameters.

The parameters page is displayed in Figure F.3. Each input field is followed by information tags, which provide a succinct help and description of the parameter. Only the first two fields are optional. They provide the possibility to name the job and to be alerted by email once the job data are available.

The next box corresponds to the input fasta file. Once this field is filled-in, it returns some basic information such as the number of sequences and their average lengths (see Figure F.4). This feedback allows the user to make sure that the web site can read the input file. The number of sequences is further used in the field quorum for the conversion between percentage and number of sequences.

The alphabet field currently proposes, in a drop-down, a list of three choices of alphabets, among DNA, RNA and custom. The "Custom" option allows the user to enter another alphabet by separating each letter by a space or a new line. Additionally, more than one letter may be considered at a given position. For instance, a user can enter AC that is interpreted as either A or C at that position, resulting in TAG and TCG being considered as the same motif. Furthermore a joker letter $*$ is available. The latter only makes sense when used conjointly with a constraint on the maximum number of such jokers.

The frame entitled *General motif information* and the following *Multiple boxes search* (see Figure F.5) set the constraints for the motifs. *Multiple boxes search* can be checked, which creates another field asking for the number of boxes, which will then create for each box, fields for the minimum and maximum length of each box, the space between each box and the number of substitutions allowed in each box. The three fields: *Total motif minimum length*, *Total motif maximum length*, and *Total substitutions* set the general minimum and maximum length and number of substitutions for the motifs including all boxes. If we have three boxes each having one substitution allowed and we set a general substitution of one in *Total substitutions*, the resulting motif will only contain one substitution in either of the three boxes. In a similar manner, we can set a composition constrain for the general motif or for each box individually. When clicking the blue button "add composition", it creates three inline components below. The first field is a drop-down menu which lets the user select one of the letters of the alphabet, the second asks for the maximum number of times of the selected letter in the previous field, and the third component is a button that removes this line.

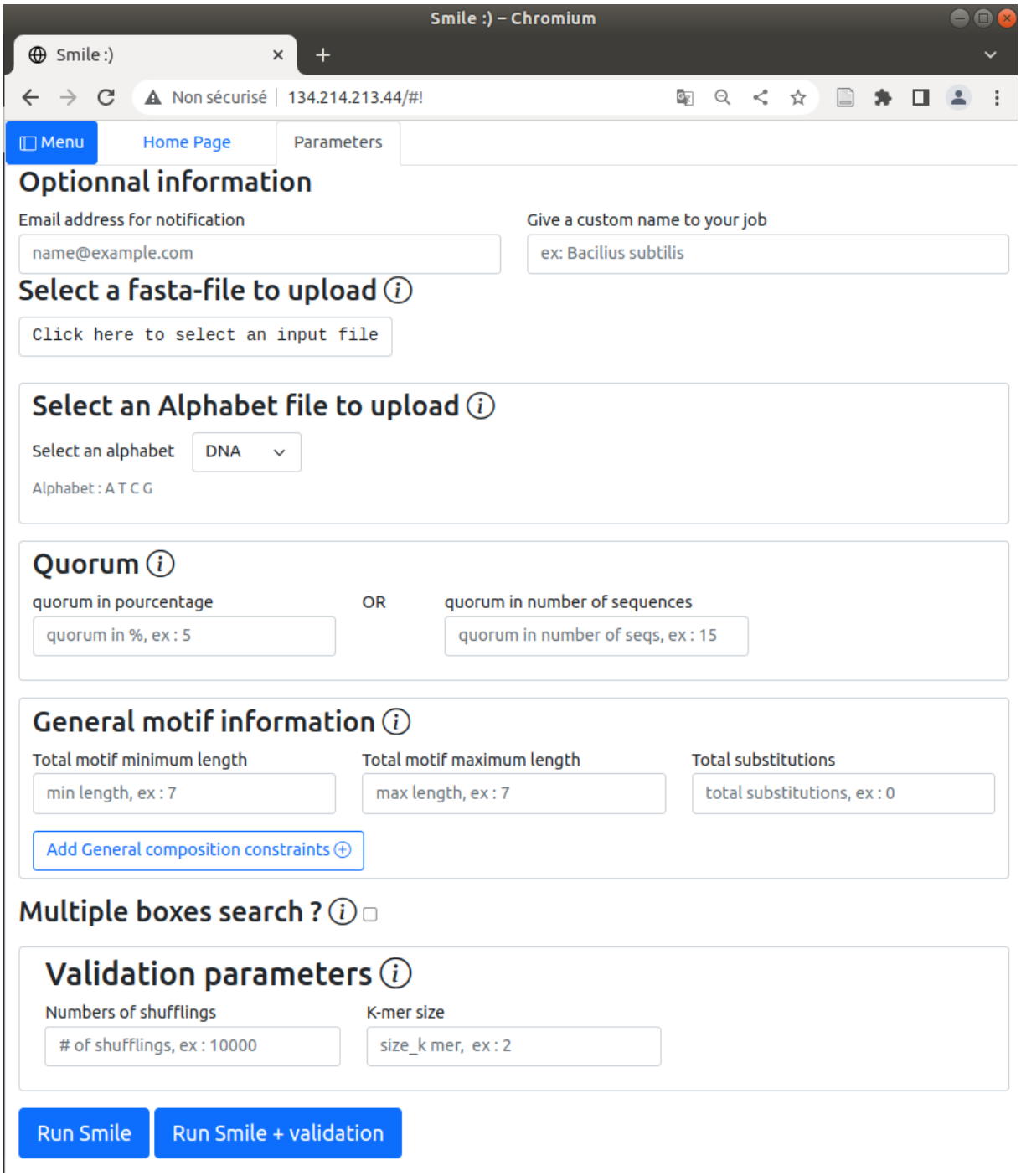


Figure F.3 – Input parameters page

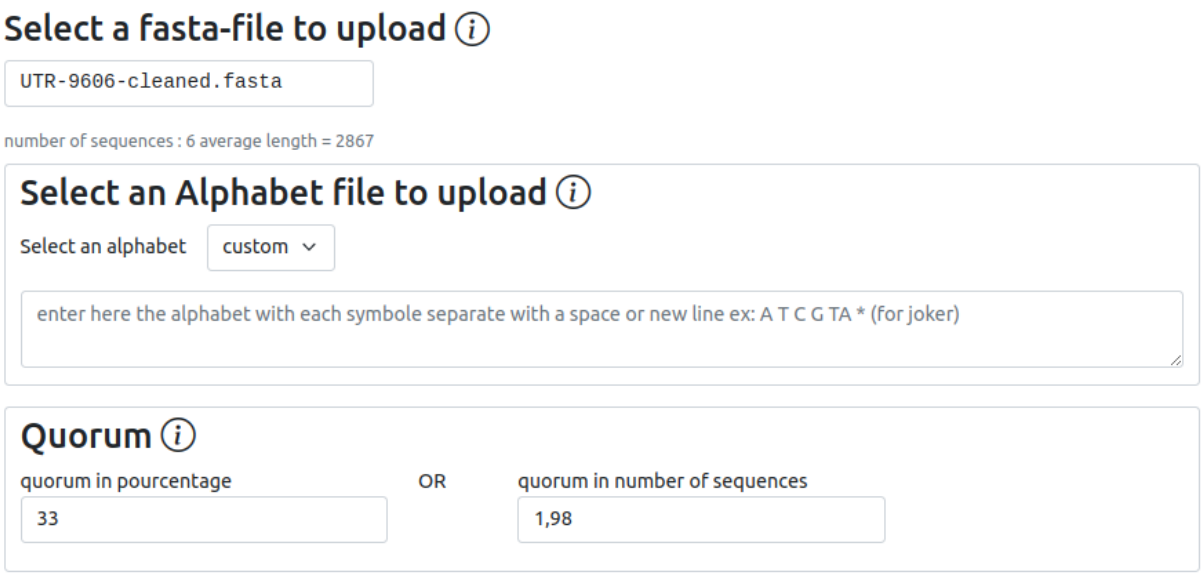


Figure F.4 – Input fasta files, custom alphabet and quorum conversion between number of sequences and percentage.

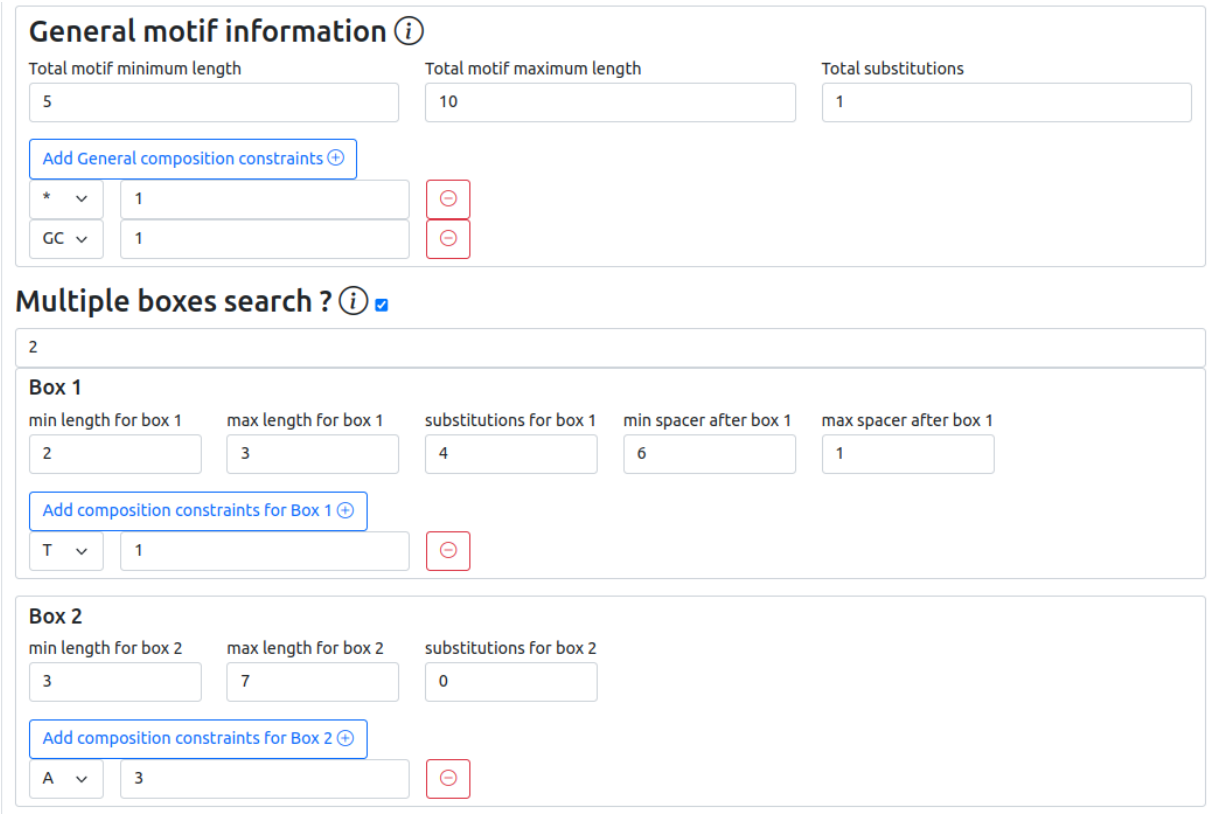


Figure F.5 – Constraint parameters. In this example, we have defined a custom alphabet of A, T,GC and $*$

Finally, the last two fields take the number of times the input data are shuffled and the size k of the k-mers whose number should be conserved during the shuffling. Once the job is launched, the user is redirected to the results page.

ii Presentation and plots of the results

The result page show first the job information (displayed in Figure F.6) which covers two aspects. The first is the status of the job and the time remaining for each step of the computation, while the second recalls the parameters which were entered for this job. The status can display 4 different statuses: 'Pending','Computing','S-parsed' and 'V-parsed' where S and V stand for respectively SMILE and validation.

Once the result of SMILE is available, a new top menu 'Result' is created and the user is further redirected to the results (Figure F.7, even if the validation is still in the process of being computed. The result table is then updated with the validation results once these are available.

The side menu allows to come back to the parameterisation page with *Create a new Job* and *Create a new job with the same parameters*. The latter will pre-fill each with the same parameters of this job. There is also the choice to delete this job with *Delete this job* and to export the results of the job. We separated the export of the placement of all occurrences with their sequences ('Seq2pos' in the database–2 F.2) and the main results of the motifs with their number of occurrences, sequences, Z-score and p-value, presented in the table. *export JSON of each motif-ID to sequences-ID/position* corresponds to the latter and *export to csv* corresponds to the former. Both exports are compressed in Gzip.

There is, for now, only one plot which is created with the library PLOTLY (see Figure F.8) and displayed in an interactive histogram that shows, for each nucleotide of the input sequences, how many motifs were found on each position of the sequence. For instance, a motif of length 7 found in a sequence that covers the nucleotides 10-16 will be counted at position 10. A motif can be found several times in the same position but in different sequences. For instance, a motif m can be found at the positions 100-106 in sequence 1 and also at the positions 100-106 in sequence 2. There are two possible visualisations. The first one is to count a unique motif at each position so that the histogram would count 1 for motif m at position 100. The second visualisation is to count the number of sequences that are occurrences of a valid motif at a given position; this would give us a count of 2 for motif m at position 100. This visualisation is most useful with sequences of similar width. Additionally, we can select rows of the table manually or all motifs with a p-value smaller than 0.05 with a click and only display in the histogram those selected motifs.

Another feature that is partially implemented is the possibility to compute an alignment with a set of sequences provided in a fasta file. For now, it is only able to align the motifs found with the file of the sequences of miRNAs used in the human CLASH data C.2 and to search with INTARNA for perfect complementary matches. However, this needed feature requires additional work such as a better parameterisation and also a placement in a separate queue, instead of the web server worker as is the case for now.

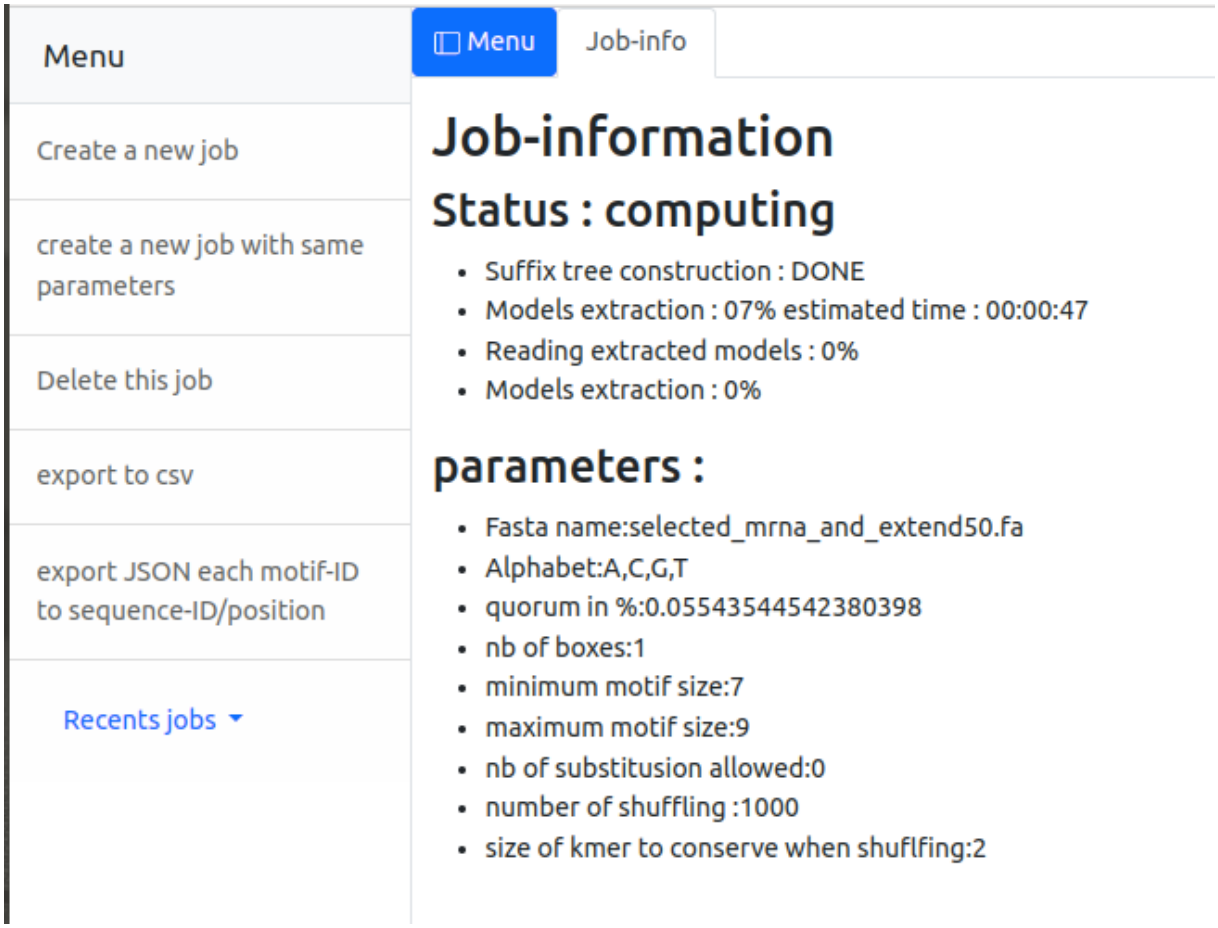


Figure F.6 – The result page with the job information and time estimation.

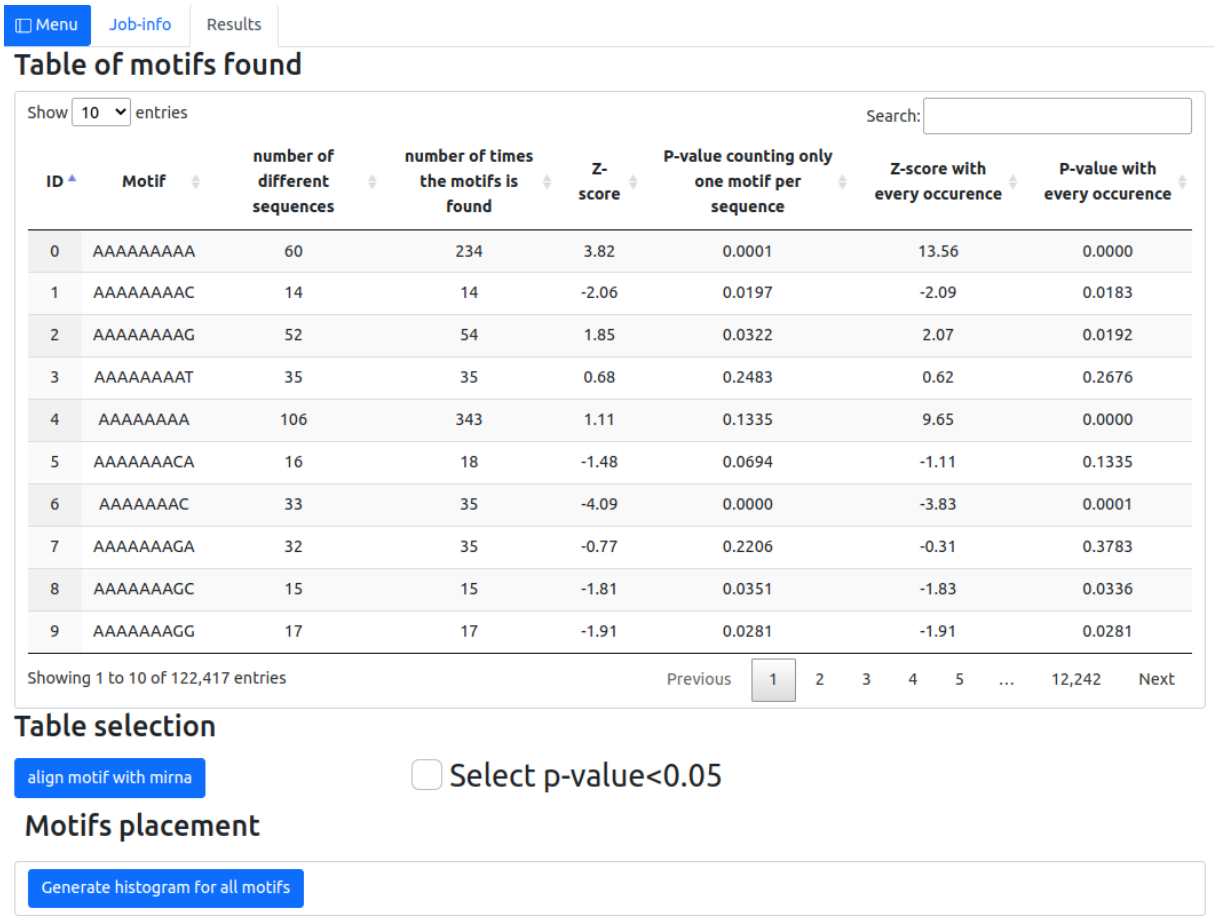


Figure F.7 – Result table.

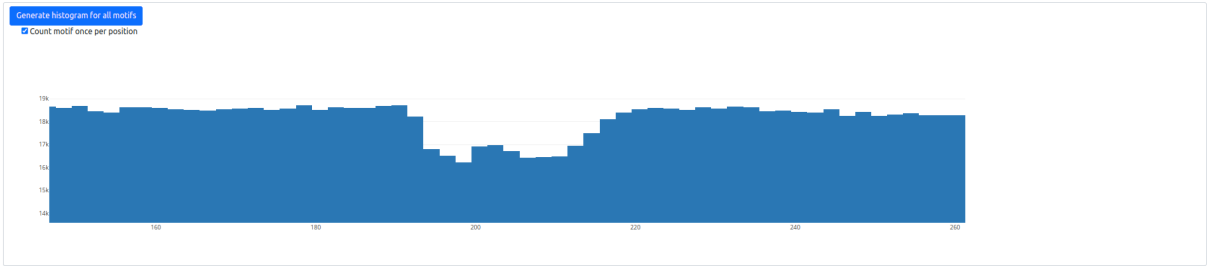


Figure F.8 – Example of the interactive histogram implemented.

F.3 Perspectives

Since the beginning, we have undeniably improved our knowledge on web development and with hindsight, we would have changed the initial choice of some environment packages such asNojeJs and bootstrap.

Indeed, NojeJs is a leader in server code for fast and scalable applications. However, its features of real-time updates and notification are underused in our project. Moreover, it is not the easiest to learn and to maintain. This is especially the case for people who are more comfortable and used to the Python language instead of JavaScript. Django or Flask, both server-oriented Python environments, would have made a better starting point and would have helped the future development and maintenance of Smile.

A renovation of the whole design of the Web site would help make it more user-friendly. We have plans to implement new functionalities such as:

— Cluster the motifs found by Smile. One easy model would be to merge the shorter motifs that are included in a longer one in the case of a search for motifs of different lengths. Another idea would be to search for a motif of maximum length still verifying the other parameters which could be done in an efficient approach when reading the suffix tree used by Smile. A maximum motif, or maxmotif as define in the implementation [Arimura and Uno, 2007] is the longest recurring motif in a sequence, the implementation of Arimura support a certain amount of joker as letter which can be any letter from the alphabet.

  We could, furthermore, provide a probabilistic visualisation of the motifs in the same way as Meme.

— Provide a visualisation of a selected motif placement on each mRNA. This is already partially feasible with the current histogram implementation. However, we only have information from all input sequences and we lack the possible knowledge of individual sequence. In other terms, there is never the option to select some sequences and see which motifs are inferred on them.

— Implement all features of Smile such as validation on a user-provided sequence instead of the shuffled sequences and the Delta parameter which computes for motifs with several boxes the best space gap between the boxes.

Besides the above, we could also implement the recent and promising post-inference selection method *SEISM* recently developed by a PHD student of the team, Antoine Villié.

Additionally, there is a list of minor quality-of-life improvements that should be fast to implement:

— The options *Quick fill* and *Create a new job with the same parameters* should respectively use a test file and re-use the same file entered by the user.

— Add more options in the drop-down list for alphabets such as amino acids to infer motifs from proteins.

— Add more filters to select the motif. For instance, we could have a two-point slider which would set the minimum and maximum Z-score for either a band-pass or notch/band-reject filter.

— The security of this application could easily be improved by using the HTTPS protocol instead of HTTP.

Conclusion and Perspectives

In the introduction of this thesis, we outlined the current biological mechanisms involved in the interaction of a miRNA with an mRNA, and introduced the fundamental features used by the computational approaches for identifying the sites of such interactions, knowing that for some features, there is no current consensus on which approach is the most appropriate. In the case of some features, there is not even a consensus whether it should be considered.

In Chapter D, we aimed to investigate two key features in miRNA-mRNA interactions: accessibility of mRNA sites and the seed region. The former is a feature that is among those debated among the scientific community, while the latter is widely acknowledged as the most important feature in such interactions. All our research focuses on the mechanisms underlying miRNA-mRNA interactions, without specifically considering the functional consequences of these interactions. Our research showed interesting results where accessibility appears clearly observed on average in the interaction sites, regardless of the interaction pattern or region. However, the accessibility of individual mRNA sites is on the other not clearly observable.

Regarding the seed, it is surprising to see that some methods predict the interaction better without taking it into account. However, this may be because we are only considering the interaction mechanism and not the functional outcome of these sites.

In chapter E, we investigated the concept of intra-conservation, with a focus once again on the underlying mechanisms of miRNA-mRNA interactions rather than considering the functional consequences of these interactions. Our analysis revealed many identical sequences located throughout the mRNAs. However, we were able to extract specific motifs that resemble the interaction base-pairing patterns, and thus to identify some interaction sites in complete mRNAs. Further analyses are however required. For instance, we could test the compatibility of the motifs inferred in one species with those inferred in another species, thereby considering both intra and inter-species conservation, thereby selecting only the regions that are conserved in both cases which might enable to reduce the noise.

In chapter F, we presented an ongoing work of a web service that allows for the inference of motifs using Smile. Although it is currently functional and available online, it still lacks additional features that would make it a more user-friendly and useful tool. In the future, we plan to integrate the two main research directions of this thesis by enabling the inference of motifs only in the mRNA regions identified as accessible.

Bibliography

[Agarwal et al., 2015] Agarwal, V., Bell, G. W., Nam, J.-W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4:e05005.

[Agarwal et al., 2018] Agarwal, V., Subtelny, A. O., Thiru, P., Ulitsky, I., and Bartel, D. P. (2018). Predicting microRNA targeting efficacy in Drosophila. *Genome Biology*, 19(1):152.

[Akhtar et al., 2016] Akhtar, M. M., Micolucci, L., Islam, M. S., Olivieri, F., and Procopio, A. D. (2016). Bioinformatic tools for microRNA dissection. *Nucleic Acids Research*, 44(1):24–44.

[AkÃğakaya et al., 2011] AkÃğakaya, P., Ekelund, S., Kolosenko, I., Caramuta, S., ÃŰzata, D. M., Xie, H., Lindforss, U., Olivecrona, H., and Lui, W.-O. (2011). miR-185 and miR-133b deregulation is associated with overall survival and metastasis in colorectal cancer. *International Journal of Oncology*, 39(2):311–318. Publisher: Spandidos Publications.

[Alexiou et al., 2009] Alexiou, P., Maragkakis, M., Papadopoulos, G. L., Reczko, M., and Hatzigeorgiou, A. G. (2009). Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, 25(23):3049–3055.

[Ambros and Horvitz, 1987] Ambros, V. and Horvitz, H. R. (1987). The lin-14 locus of Caenorhabditis elegans controls the time of expression of specific postembryonic developmental events. *Genes & Development*, 1(4):398–414. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[Andronescu et al., 2008] Andronescu, M., Bereg, V., Hoos, H. H., and Condon, A. (2008). RNA STRAND: The RNA Secondary Structure and Statistical Analysis Database. *BMC Bioinformatics*, 9(1):340.

[Arimura and Uno, 2007] Arimura, H. and Uno, T. (2007). An efficient polynomial space and polynomial delay algorithm for enumeration of maximal motifs in a sequence. *Journal of Combinatorial Optimization*, 13(3):243–262.

[Arvey et al., 2010] Arvey, A., Larsson, E., Sander, C., Leslie, C. S., and Marks, D. S. (2010). Target mRNA abundance dilutes microRNA and siRNA activity. *Molecular Systems Biology*, 6(1):363.

[Atkins and Paula, 2010] Atkins, P. and Paula, J. (2010). *ATKINS' PHYSICAL CHEMISTRY*. Oxford University Press.

[Bailey, 2021] Bailey, T. L. (2021). STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, page 7.

[Bailey et al., 2009] Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server):W202–W208.

[Bailey and Elkan, 1995] Bailey, T. L. and Elkan, C. (1995). Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. *machine learning journal*, page 33.

[Bailey and Gribskov, 1998] Bailey, T. L. and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14(1):48–54.

[Bailey et al., 2015] Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Research*, 43(W1):W39–W49.

[Bartel, 2009] Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, 136(2):215–233.

[Bartel, 2018] Bartel, D. P. (2018). Metazoan MicroRNAs. *Cell*, 173(1):20–51.

[Benner, 1988] Benner, S. A. (1988). Extracellular âĂŸcommunicator RNAâĂŹ. *FEBS Letters*, 233(2):225–228.

[Bloch et al., 2017] Bloch, S., WÄŹgrzyn, A., WÄŹgrzyn, G., and Nejman-FaleÅĎczyk, B. (2017). Small and SmallerâĂŤsRNAs and MicroRNAs in the Regulation of Toxin Gene Expression in Prokaryotic Cells: A Mini-Review. *Toxins*, 9(6):181. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[Borer et al., 1974] Borer, P. N., Dengler, B., Tinoco, I., and Uhlenbeck, O. C. (1974). Stability of ribonucleic acid double-stranded helices. *Journal of Molecular Biology*, 86(4):843–853.

[Bullard et al., 2019] Bullard, W. L., Kara, M., Gay, L. A., Sethuraman, S., Wang, Y., Nirmalan, S., Esemenli, A., Feswick, A., Hoffman, B. A., Renne, R., and Tibbetts, S. A. (2019). Identification of murine gammaherpesvirus 68 miRNA-mRNA hybrids reveals miRNA target conservation among gammaherpesviruses including host translation and protein modification machinery. *PLOS Pathogens*, 15(8):e1007843.

[Busch et al., 2008] Busch, A., Richter, A. S., and Backofen, R. (2008). IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856.

[Cai et al., 2018] Cai, Q., He, B., Kogel, K.-H., and Jin, H. (2018). Cross-kingdom RNA trafficking and environmental RNAi -nature's blueprint for modern crop protection strategies. *Current Opinion in Microbiology*, 46:58–64.

[Chalfie, 1981] Chalfie, M. (1981). Mutations that lead to reiterations in the cell lineages of C. elegans. *Cell*, 24(1):59–69.

[Chen, 2012] Chen, X. (2012). Secreted microRNAs: a new form of intercellular communication. *Cell Press*, 22(3):8.

[Chi et al., 2009] Chi, S. W., Zang, J. B., Mele, A., and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNAâĂŞmRNA interaction maps. *Nature*, 460(7254):479–486. Number: 7254 Publisher: Nature Publishing Group.

[Condon et al., 2004] Condon, A., Davy, B., Rastegari, B., Zhao, S., and Tarrant, F. (2004). Classifying RNA pseudoknotted structures. *Theoretical Computer Science*, 320(1):35–50.

[de Pontual et al., 2011] de Pontual, L., Yao, E., Callier, P., Faivre, L., Drouin, V., Cariou, S., Van Haeringen, A., GeneviÃÍve, D., Goldenberg, A., Oufadem, M., Manouvrier, S., Munnich, A., Vidigal, J. A., Vekemans, M., Lyonnet, S., Henrion-Caude, A., Ventura, A., and Amiel, J. (2011). Germline deletion of the miR-17âĹij92 cluster causes skeletal and growth defects in humans. *Nature Genetics*, 43(10):1026–1030. Number: 10 Publisher: Nature Publishing Group.

[de Rie et al., 2017] de Rie, D., Abugessaisa, I., Alam, T., Arner, E., Arner, P., Ashoor, H., ÃĚstrÃűm, G., Babina, M., Bertin, N., Burroughs, A. M., Carlisle, A. J., Daub, C. O., Detmar, M., Deviatiiarov, R., Fort, A., Gebhard, C., Goldowitz, D., Guhl, S., Ha, T. J., Harshbarger, J., Hasegawa, A., Hashimoto, K., Herlyn, M., Heutink, P., Hitchens, K. J., Hon, C. C., Huang, E., Ishizu, Y., Kai, C., Kasukawa, T., Klinken, P., Lassmann, T., Lecellier, C.-H., Lee, W., Lizio, M., Makeev, V., Mathelier, A., Medvedeva, Y. A., Mejhert, N., Mungall, C. J., Noma, S., Ohshima, M., Okada-Hatakeyama, M., Persson, H., Rizzu, P., Roudnicky, F., SÃętrom, P., Sato, H., Severin, J., Shin, J. W., Swoboda, R. K., Tarui, H., Toyoda, H., Vitting-Seerup, K., Winteringham, L., Yamaguchi, Y., Yasuzawa, K., Yoneda, M., Yumoto, N., Zabierowski, S., Zhang, P. G., Wells, C. A., Summers, K. M., Kawaji, H., Sandelin, A., Rehli, M., Hayashizaki, Y., Carninci, P., Forrest, A. R. R., and de Hoon, M. J. L. (2017). An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nature Biotechnology*, 35(9):872–878. Number: 9 Publisher: Nature Publishing Group.

[DeVoe and Tinoco, 1962] DeVoe, H. and Tinoco, I. (1962). The stability of helical polynucleotides: Base contributions. *Journal of Molecular Biology*, 4(6):500–517.

[Do et al., 2006] Do, C. B., Woods, D. A., and Batzoglou, S. (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98.

[Doench, 2004] Doench, J. G. (2004). Specificity of microRNA target selection in translational repression. *Genes & Development*, 18(5):504–511.

[Eyking et al., 2016] Eyking, A., Reis, H., Frank, M., Gerken, G., Schmid, K. W., and Cario, E. (2016). MiR-205 and MiR-373 Are Associated with Aggressive Human Mucinous Colorectal Cancer. *PLOS ONE*, 11(6):e0156871. Publisher: Public Library of Science.

[Fang and Rajewsky, 2011] Fang, Z. and Rajewsky, N. (2011). The Impact of miRNA Target Sites in Coding Sequences and in 3âĂšUTRs. *PLOS ONE*, 6(3):e18067. Publisher: Public Library of Science.

[Ferguson et al., 1987] Ferguson, E. L., Sternberg, P. W., and Horvitz, H. R. (1987). A genetic pathway for the specification of the vulval cell lineages of Caenorhabditis elegans. *Nature*, 326(6110):259–267. Number: 6110 Publisher: Nature Publishing Group.

[Fields et al., 2021] Fields, C. J., Li, L., Hiers, N. M., Li, T., Sheng, P., Huda, T., Shan, J., Gay, L., Gu, T., Bian, J., Kilberg, M. S., Renne, R., and Xie, M. (2021). Sequencing of Argonaute-bound microRNA/mRNA hybrids reveals regulation of the unfolded protein

response by microRNA-320a. *PLOS Genetics*, 17(12):e1009934. Publisher: Public Library of Science.

[Fridrich et al., 2019] Fridrich, A., Hazan, Y., and Moran, Y. (2019). Too Many False Targets for MicroRNAs: Challenges and Pitfalls in Prediction of miRNA Targets and Their Gene Ontology in Model and Non-model Organisms. *BioEssays*, 41(4):1800169.

[Friedman et al., 2008] Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. (2008). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105.

[Fu et al., 2019] Fu, Y., Chen, J., and Huang, Z. (2019). Recent progress in microRNA-based delivery systems for the treatment of human disease. *ExRNA*, 1(1):24.

[Gangemi et al., 2020] Gangemi, C. M. A., Alaimo, S., Pulvirenti, A., GarcÃŋa-ViÃśuales, S., Milardi, D., Falanga, A. P., FragalÃă, M. E., Oliviero, G., Piccialli, G., Borbone, N., Ferro, A., DâĂŹUrso, A., Croce, C. M., and Purrello, R. (2020). Endogenous and artificial miRNAs explore a rich variety of conformations: a potential relationship between secondary structure and biological functionality. *Scientific Reports*, 10(1):453. Number: 1 Publisher: Nature Publishing Group.

[Garcia et al., 2011] Garcia, D. M., Baek, D., Shin, C., Bell, G. W., Grimson, A., and Bartel, D. P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nature Structural & Molecular Biology*, 18(10):1139–1146. Number: 10 Publisher: Nature Publishing Group.

[Gay et al., 2018] Gay, L. A., Sethuraman, S., Thomas, M., Turner, P. C., and Renne, R. (2018). Modified Cross-Linking, Ligation, and Sequencing of Hybrids (qCLASH) Identifies Kaposi's Sarcoma-Associated Herpesvirus MicroRNA Targets in Endothelial Cells. *Journal of Virology*, 92(8):e02138–17. Publisher: American Society for Microbiology.

[Gelhausen and Raden, 2018] Gelhausen, R. and Raden, D. M. (2018). Constrained RNA-RNA interaction prediction. *Master thesis*, page 76.

[Grant et al., 2011] Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.

[Greenberg, 1979] Greenberg, J. R. (1979). Ultraviolet light-induced crosslinking of mRNA to proteins. *Nucleic Acids Research*, 6(2):715–732.

[Griffiths-Jones et al., 2007] Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2007). miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36(Database):D154–D158.

[Grimson et al., 2007] Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. (2007). MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Molecular Cell*, 27(1):91–105.

[Grosswendt, 2014] Grosswendt, S. (2014). Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions. *Molecular Cell*, page 13.

[Gu et al., 2017] Gu, H., Zhao, C., Zhang, T., Liang, H., Wang, X.-M., Pan, Y., Chen, X., Zhao, Q., Li, D., Liu, F., Zhang, C.-Y., and Zen, K. (2017). Salmonella produce microRNA-like RNA fragment Sal-1 in the infected cells to facilitate intracellular survival. *Scientific Reports*, 7(1):2392.

[Gumienny and Zavolan, 2015] Gumienny, R. and Zavolan, M. (2015). Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. *Nucleic Acids Research*, 43(3):1380–1391.

[Guo, 2012] Guo, F. (2012). Chapter Five - Drosha and DGCR8 in MicroRNA Biogenesis. In Guo, F. and Tamanoi, F., editors, *The Enzymes*, volume 32 of *Eukaryotic RNases and their Partners in RNA Degradation and Biogenesis, Part B*, pages 101–121. Academic Press.

[Gupta et al., 2007] Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biology*, 8(2):R24.

[Ha and Kim, 2014] Ha, M. and Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology*, 15(8):509–524.

[Hafner et al., 2010] Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1):129–141.

[Helwak et al., 2013] Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. *Cell*, 153(3):654–665.

[Hofacker et al., 1994] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte for Chemie Chemical Monthly*, 125(2):167–188.

[Homberg et al., 2023] Homberg, N., GalvÃčo Ferrarini, M., Gaspin, C., and Sagot, M.-F. (2023). MicroRNA Target Identification: Revisiting Accessibility and Seed Anchoring. *Genes*, 14(3):664. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[Horvitz and Sulston, 1980] Horvitz, H. R. and Sulston, J. E. (1980). Isolation and genetic Characterization of cell-lineage mutants of the nematode Caenorhabditis elegans. *Genetics*, 96(2):435–454.

[Hoy and Buck, 2012] Hoy, A. and Buck, A. (2012). Extracellular small RNAs: what, where, why? *Biochemical Society Transactions*, 40(4):886–890.

[Hsu et al., 2011] Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T., Chen, G.-Z., Lee, C.-J., Chiu, C.-M., Chien, C.-H., Wu, M.-C., Huang, C.-Y., Tsou, A.-P., and Huang, H.-D. (2011). miRTarBase: a database curates experimentally validated microRNAâĂŞtarget interactions. *Nucleic Acids Research*, 39(suppl_1):D163–D169.

[Hughes et al., 2011] Hughes, A. E., Bradley, D. T., Campbell, M., Lechner, J., Dash, D. P., Simpson, D. A., and Willoughby, C. E. (2011). Mutation Altering the miR-184 Seed Region Causes Familial Keratoconus with Cataract. *The American Journal of Human Genetics*, 89(5):628–633. Publisher: Elsevier.

[Hui, 1992] Hui, L. C. K. (1992). *Lecture Notes in Computer Science*. Lecutre Notes in Computer Science. Springer.

[Jiang et al., 2008] Jiang, M., Anderson, J., Gillespie, J., and Mayne, M. (2008). uShuffle: A useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, 9(1):192.

[Jo et al., 2015] Jo, M. H., Shin, S., Jung, S.-R., Kim, E., Song, J.-J., and Hohng, S. (2015). Human Argonaute 2 Has Diverse Reaction Pathways on Target RNAs. *Molecular Cell*, 59(1):117–124. Publisher: Elsevier.

[John et al., 2004] John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human MicroRNA Targets. *PLoS Biology*, 2(11):e363.

[Kedde and Agami, 2008] Kedde, M. and Agami, R. (2008). Interplay between microRNAs and RNA-binding proteins determines developmental processes. *Cell Cycle*, 7(7):899–903. Publisher: Taylor & Francis _eprint: https://doi.org/10.4161/cc.7.7.5644.

[Kern et al., 2021] Kern, F., Krammes, L., Danz, K., Diener, C., Kehl, T., KÃijchler, O., Fehlmann, T., Kahraman, M., Rheinheimer, S., Aparicio-Puerta, E., Wagner, S., Ludwig, N., Backes, C., Lenhof, H.-P., vonÂăBriesen, H., Hart, M., Keller, A., and Meese, E. (2021). Validation of human microRNA target pathways enables evaluation of target prediction tools. *Nucleic Acids Research*, 49(1):127–144.

[Kertesz et al., 2007] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39(10):1278–1284.

[Kim and Kim, 2007] Kim, Y.-K. and Kim, V. N. (2007). Processing of intronic microRNAs. *The EMBO Journal*, 26(3):775–783. Publisher: John Wiley & Sons, Ltd.

[Kinsella et al., 2011] Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., and Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, 2011(0):bar030–bar030.

[Kong et al., 2012] Kong, Y. W., Ferland-McCollough, D., Jackson, T. J., and Bushell, M. (2012). microRNAs in cancer management. *The Lancet Oncology*, 13(6):e249–e258.

[Kruger and Rehmsmeier, 2006] Kruger, J. and Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Research*, 34(suppl_2):W451–W454.

[Lange et al., 2012] Lange, S. J., Maticzka, D., MÃűhl, M., Gagnon, J. N., Brown, C. M., and Backofen, R. (2012). Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Research*, 40(12):5215–5226.

[Lawrence and Reilly, 1990] Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins: Structure, Function, and Bioinformatics*, 7(1):41–51. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.340070105.

[Lee et al., 1993] Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854.

[Lewis et al., 2005] Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*, 120(1):15–20.

[Lewis et al., 2003] Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of Mammalian MicroRNA Targets. *Cell*, 115(7):787–798.

[Lewohl et al., 2011] Lewohl, J. M., Nunez, Y. O., Dodd, P. R., Tiwari, G. R., Harris, R. A., and Mayfield, R. D. (2011). Up-Regulation of MicroRNAs in Brain of Human Alcoholics. *Alcoholism: Clinical and Experimental Research*, 35(11):1928–1937. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1530-0277.2011.01544.x.

[Lin et al., 2019] Lin, C., Yuan, G., Hu, Z., Zeng, Y., Qiu, X., Yu, H., and He, S. (2019). Bioinformatics analysis of the interactions among lncRNA, miRNA and mRNA expression, genetic mutations and epigenetic modifications in hepatocellular carcinoma. *Molecular Medicine Reports*, 19(2):1356–1364. Publisher: Spandidos Publications.

[Liu et al., 2016] Liu, S., da Cunha, A., Rezende, R., Cialic, R., Wei, Z., Bry, L., Comstock, L., Gandhi, R., and Weiner, H. (2016). The Host Shapes the Gut Microbiota via Fecal MicroRNA. *Cell Host & Microbe*, 19(1):32–43.

[Long et al., 2007] Long, D., Lee, R., Williams, P., Chan, C. Y., Ambros, V., and Ding, Y. (2007). Potent effect of target structure on microRNA function. *Nature Structural & Molecular Biology*, 14(4):287–294. Number: 4 Publisher: Nature Publishing Group.

[Lorenz et al., 2011] Lorenz, R., Bernhart, S. H., Honer zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26.

[Maes et al., 2009] Maes, O. C., Chertkow, H. M., Wang, E., and Schipper, H. M. (2009). MicroRNA: Implications for Alzheimer Disease and other Human CNS Disorders. *Current Genomics*, 10(3):154–168.

[Mann et al., 2017] Mann, M., Wright, P. R., and Backofen, R. (2017). IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Research*, 45(Web Server issue):W435–W439.

[Marin et al., 2013] Marin, R. M., Sulc, M., and Vanicek, J. (2013). Searching the coding region for microRNA targets. *RNA*, 19(4):467–474.

[Marin and Vanicek, 2011] Marin, R. M. and Vanicek, J. (2011). Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Research*, 39(1):19–29.

[Marsan and Sagot, 2000] Marsan, L. and Sagot, M.-F. (2000). Algorithms for Extracting Structured Motifs Using a Suffix Tree with an Application to Promoter and Regulatory Site Consensus Identification. *Journal of Computational Biology*, 7(3-4):345–362.

[MarÃŋn and VanÃŋÄĐek, 2012] MarÃŋn, R. M. and VanÃŋÄĐek, J. (2012). Optimal Use of Conservation and Accessibility Filters in MicroRNA Target Prediction. *PLOS ONE*, 7(2):e32208. Publisher: Public Library of Science.

[Mathews et al., 1999] Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure11Edited by I. Tinoco. *Journal of Molecular Biology*, 288(5):911–940.

[McCaskill, 1990] McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119.

[MencÃŋa et al., 2009] MencÃŋa, Ã., Modamio-HÃÿybjÃÿr, S., Redshaw, N., MorÃŋn, M., Mayo-Merino, F., Olavarrieta, L., Aguirre, L. A., del Castillo, I., Steel, K. P., Dalmay, T., Moreno, F., and Moreno-Pelayo, M. Ã. (2009). Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nature Genetics*, 41(5):609–613. Number: 5 Publisher: Nature Publishing Group.

[Millar and Waterhouse, 2005] Millar, A. A. and Waterhouse, P. M. (2005). Plant and animal microRNAs: similarities and differences. *Functional & Integrative Genomics*, 5(3):129–135.

[Miranda et al., 2006] Miranda, K. C., Huynh, T., Tay, Y., Ang, Y.-S., Tam, W.-L., Thomson, A. M., Lim, B., and Rigoutsos, I. (2006). A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes. *Cell*, 126(6):1203–1217.

[Mockly and Seitz, 2019] Mockly, S. and Seitz, H. (2019). Inconsistencies and Limitations of Current MicroRNA Target Identification Methods. In Lagana, A., editor, *MicroRNA Target Identification*, volume 1970, pages 291–314. Springer New York, New York, NY.

[Moore et al., 2015] Moore, M. J., Scheel, T. K. H., Luna, J. M., Park, C. Y., Fak, J. J., Nishiuchi, E., Rice, C. M., and Darnell, R. B. (2015). miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. *Nature Communications*, 6(1):8864. Number: 1 Publisher: Nature Publishing Group.

[Needleman and Wunsch, 1970] Needleman, S. B. and Wunsch, C. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal molecular Biology*.

[Nielsen et al., 2007] Nielsen, C. B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., and Burge, C. B. (2007). Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, 13(11):1894–1910. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[Pasquinelli et al., 2000] Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., MÃijller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., and Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89. Number: 6808 Publisher: Nature Publishing Group.

[Phua et al., 2015] Phua, Y. L., Chu, J. Y. S., Marrone, A. K., Bodnar, A. J., Sims-Lucas, S., and Ho, J. (2015). Renal stromal miRNAs are required for normal nephrogenesis and glomerular mesangial survival. *Physiological Reports*, 3(10):e12537. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.14814/phy2.12537.

[Pruitt et al., 2007] Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(Database):D61–D65.

[Qureshi et al., 2014] Qureshi, A., Thakur, N., Monga, I., Thakur, A., and Kumar, M. (2014). VIRmiRNA: a comprehensive resource for experimentally validated viral miRNAs and their targets. *Database*, 2014:bau103.

[Raden et al., 2018] Raden, M., Mohamed, M. M., Ali, S. M., and Backofen, R. (2018). Interactive implementations of thermodynamics-based RNA structure and RNAâĂŞRNA interaction prediction approaches for example-driven teaching. *PLOS Computational Biology*, 14(8):e1006341.

[Reda El Sayed et al., 2021] Reda El Sayed, S., Cristante, J., Guyon, L., Denis, J., Chabre, O., and Cherradi, N. (2021). MicroRNA Therapeutics in Cancer: Current Advances and Challenges. *Cancers*, 13(11):2680.

[Rehmsmeier et al., 2004] Rehmsmeier, M., Steffen, P., HÃűchsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[Romao et al., 2011] Romao, J. M., Jin, W., Dodson, M. V., Hausman, G. J., Moore, S. S., and Guan, L. L. (2011). MicroRNA regulation in mammalian adipogenesis. *Experimental Biology and Medicine*, 236(9):997–1004. Publisher: SAGE Publications.

[Sagot, 1998] Sagot, M. F. (1998). Spelling approximate repeated or common motifs using a suffix tree. In Goos, G., Hartmanis, J., van Leeuwen, J., Lucchesi, C. L., and Moura, A. V., editors, *LATIN'98: Theoretical Informatics*, volume 1380, pages 374–390. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

[Salmena et al., 2011] Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. (2011). A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? *Cell*, 146(3):353–358.

[Schirle et al., 2014] Schirle, N. T., Sheu-Gruttadauria, J., and MacRae, I. J. (2014). Structural basis for microRNA targeting. *Science*, 346(6209):608–613. Publisher: American Association for the Advancement of Science.

[Schnall-Levin et al., 2010] Schnall-Levin, M., Zhao, Y., Perrimon, N., and Berger, B. (2010). Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3âĂšUTRs. *Proceedings of the National Academy of Sciences*, 107(36):15751–15756. Publisher: Proceedings of the National Academy of Sciences.

[Schneider and Stephens, 1990] Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100.

[Sheu-Gruttadauria et al., 2019] Sheu-Gruttadauria, J., Pawlica, P., Klum, S. M., Wang, S., Yario, T. A., Schirle Oakdale, N. T., Steitz, J. A., and MacRae, I. J. (2019). Structural Basis for Target-Directed MicroRNA Degradation. *Molecular Cell*, 75(6):1243–1255.e7.

[SheuâĂŘGruttadauria et al., 2019] SheuâĂŘGruttadauria, J., Xiao, Y., Gebert, L. F., and MacRae, I. J. (2019). Beyond the seed: structural basis for supplementary micro RNA targeting by human Argonaute2. *The EMBO Journal*, 38(13).

[Stark et al., 2007] Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., Ruby, J. G., Brennecke, J., Hodges, E., Hinrichs, A. S., Caspi, A., Paten, B., Park, S.-W., Han,

M. V., Maeder, M. L., Polansky, B. J., Robson, B. E., Aerts, S., van Helden, J., Hassan, B., Gilbert, D. G., Eastman, D. A., Rice, M., Weir, M., Hahn, M. W., Park, Y., Dewey, C. N., Pachter, L., Kent, W. J., Haussler, D., Lai, E. C., Bartel, D. P., Hannon, G. J., Kaufman, T. C., Eisen, M. B., Clark, A. G., Smith, D., Celniker, S. E., Gelbart, W. M., and Kellis, M. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–232. Number: 7167 Publisher: Nature Publishing Group.

[Sugimoto et al., 2012] Sugimoto, Y., KÃűnig, J., Hussain, S., Zupan, B., Curk, T., Frye, M., and Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biology*, 13(8):R67.

[Tafer et al., 2008] Tafer, H., Ameres, S. L., Obernosterer, G., Gebeshuber, C. A., Schroeder, R., Martinez, J., and Hofacker, I. L. (2008). The impact of target site accessibility on the design of effective siRNAs. *Nature Biotechnology*, 26(5):578–583.

[Tafer and Hofacker, 2008] Tafer, H. and Hofacker, I. L. (2008). RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, 24(22):2657–2663.

[Taylor and Gercel-Taylor, 2013] Taylor, D. D. and Gercel-Taylor, C. (2013). The origin, function, and diagnostic potential of RNA within extracellular vesicles present in human biological fluids. *Frontiers in Genetics*, 4.

[Turner and Mathews, 2010] Turner, D. H. and Mathews, D. H. (2010). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 38(suppl_1):D280–D282.

[Ule et al., 2005] Ule, J., Jensen, K., Mele, A., and Darnell, R. B. (2005). CLIP: A method for identifying proteinâĂŞRNA interaction sites in living cells. *Methods*, 37(4):376–386.

[Ule et al., 2003] Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. B. (2003). CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science*, 302(5648):1212–1215. Publisher: American Association for the Advancement of Science.

[Uzuner et al., 2022] Uzuner, E., Ulu, G. T., GÃijrler, S. B., and Baran, Y. (2022). The Role of MiRNA in Cancer: Pathogenesis, Diagnosis, and Treatment. In Allmer, J. and Yousef, M., editors, *miRNomics: MicroRNA Biology and Computational Analysis*, Methods in Molecular Biology, pages 375–422. Springer US, New York, NY.

[Vasudevan, 2012] Vasudevan, S. (2012). Posttranscriptional Upregulation by MicroRNAs: Posttranscriptional Upregulation by MicroRNAs. *Wiley Interdisciplinary Reviews: RNA*, 3(3):311–330.

[Wang et al., 2011] Wang, Z., Rao, D. D., Senzer, N., and Nemunaitis, J. (2011). RNA Interference and Cancer Therapy. *Pharmaceutical Research*, 28(12):2983–2995.

[Waterman and Smith, 1978] Waterman, M. S. and Smith, T. F. (1978). RNA secondary structure: a complete mathematical analysis. *Mathematical Biosciences*, 42(3):257–266.

[Weiberg et al., 2013] Weiberg, A., Wang, M., Lin, F.-M., Zhao, H., Zhang, Z., Kaloshian, I., Huang, H.-D., and Jin, H. (2013). Fungal Small RNAs Suppress Plant Immunity by Hijacking Host RNA Interference Pathways. *Science*, 342(6154):118–123.

[Wightman et al., 1993] Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–862.

[Wilhelm and Smibert, 2005] Wilhelm, J. E. and Smibert, C. A. (2005). Mechanisms of translational regulation in Drosophila. *Biology of the Cell*, 97(4):235–252. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1042/BC20040097.

[Zhang et al., 2004] Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E., and Filipowicz, W. (2004). Single Processing Center Models for Human Dicer and Bacterial RNase III. *Cell*, 118(1):57–68. Publisher: Elsevier.

[Zuker and Stiegler, 1981] Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148.

Abstract:
MicroRNAs (miRNAs) are non-coding RNAs present in eukaryotes that regulate the expression of messenger RNAs (mRNAs) up or down. These miRNAs have significant potential in future treatment of cancer and other diseases. The miRNA-mRNA interactions are intricate and involve various mechanisms, such as sequence complementarity, accessibility, and conservation. This thesis focuses on two such mechanisms, namely accessibility and intra-species conservation of the site of interaction, using experimental data from Cross-linking, Ligation And Sequencing of Hybrids (CLASH). Although the accessibility of interaction sites on mRNAs is generally observed, it is not consistent for all interactions. Intra-species conservation is a rare feature, which we explore by inferring conserved motifs from mRNA interaction sites. Although the results are noisy, in some specific cases, we manage to retrieve some mRNA interaction sites from the inferred motifs.

Résumé :
Les microARNs (miARNs) sont de petit ARNs non codant présents dans tous les eucaryotes qui régulent, positivement ou négativement, l'expression des ARN messagers (ARNms). Les miARNs ont un potentiel important pour de futurs traitements du cancer et d'autres maladies. Les interactions miARN-ARNm dépendent d'une variété de mécanismes complexes, tels que la complémentarité des séquences, l'accessibilité et la conservation. Cette thèse se concentre sur deux de ces mécanismes, à savoir l'accessibilité et la conservation intra-espèce du site d'interaction, en utilisant des données expérimentales de Cross-linking, Ligation And Sequencing of Hybrids (CLASH). Bien que l'accessibilité des sites d'interaction sur les ARNms soit généralement observée, cela n'est pas le cas pour toutes les interactions. La conservation intra-espèce est un mécanisme peu considéré que nous avons étudiée au travers la recherche de motifs conservés dans les ARNms. Bien que les résultats obtenus soient bruités, il est possible de retrouver via ces motifs certains sites d'interaction sur les ARNms.