UNIVERSITÉ DE LORRAINE | BIBLIOTHÈQUES UNIVERSITAIRES

AVERTISSEMENT

Ce document est le fruit d'un long travail approuvé par le jury de soutenance et mis à disposition de l'ensemble de la communauté universitaire élargie.

Il est soumis à la propriété intellectuelle de l'auteur. Ceci implique une obligation de citation et de référencement lors de l'utilisation de ce document.

D'autre part, toute contrefaçon, plagiat, reproduction illicite encourt une poursuite pénale.

Contact bibliothèque : ddoc-theses-contact@univ-lorraine.fr
(Cette adresse ne permet pas de contacter les auteurs)

LIENS

Code de la Propriété Intellectuelle. articles L 122. 4
Code de la Propriété Intellectuelle. articles L 335.2- L 335.10
http://www.cfcopies.com/V2/leg/leg_droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Structural characterization of RNA binding to RNA recognition motif (RRM) domain using data integration, 3D modeling and dynamic simulation

THÈSE

présentée et soutenue publiquement le 11 July 2023

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Hrishikesh Dhondge

Composition du jury

Président :	Julie Thompson	DR CNRS, Strasbourg
Rapporteurs :	Olga Kalinina	Professeur, Université de Sarrebruck
	Julie Thompson	DR CNRS, Strasbourg
	Stanisław Dunin-Horkawicz	Researcher, Université de Varsovie
Examinateur :	Alain Denise	Professeur, Université Paris Saclay
Encadrants :	Isaure Chauvot de Beauchêne	CR CNRS, Nancy
	Marie-Dominique Devignes	CR CNRS (HDR), Nancy
Membre invité :	Wim Vranken	Professeur, Vrije Universiteit Brussel

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

Mis en page avec la classe thesul.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813239.

Acknowledgements

As the journey of my PhD comes to an end, I find myself in a flashback that transports me back to the series of interviews with which it all started back in April 2019. From April 2019 to the present day, countless individuals have generously contributed to my journey, both directly and indirectly, helping me achieve this significant milestone. I want to seize this moment to express my heartfelt gratitude to each and every one of them.

First and foremost, I express my sincere and deepest sense of gratitude to my thesis supervisors, **Isaure Chauvot de Beauchêne** and **Marie-Dominique Devignes**, for giving me an opportunity to work under their guidance. The knowledge and insights I have gained from my supervisors are incredibly valuable. Their guidance, advice, unwavering support, and constant encouragement have been the pillars that made this journey remarkably smooth. They extended their assistance not only in the research but also in handling administrative tasks and navigating the intricacies of the complex French system. Throughout these years, I was always looking forward to our meetings. Meetings full of interesting ideas to find the best approach to tackle the scientific problem we were facing. I extend my heartfelt gratitude to both of you for helping in my personal and professional growth.

I am grateful to the entire CAPSID team for their consistent support. We had some wonderful hikes and had the opportunity to learn more about one another. The tea-time sessions were also great and a perfect place to know more about other's work. A special thanks to **Anna Kravchenko** for your exceptional support and being the saviour when dealing with the French administration. My sincere thanks go to **Antoine Moniot** and **Dominique Mias-Lucquin** for their roles as both friends and mentors, always there when I needed guidance. Thanks to **Kamrul Islam**, **Diego Amaya ramirez**, and **Athénaïs Vaginay**, your unwavering camaraderie, intellectual collaboration, and shared experiences have been invaluable. My sincere thanks to Athénaïs Vaginay for being the lab companion during the last few months of this journey. Thanks to **Malika Smail-Tabbone**, **Sabeur Aridhi**, **Hamed Khakzad**, **Yasaman Karami** and **Sjoerd de Vries**, your diverse perspectives, insightful discussions, and collective dedication have made this journey not only educational but also truly enjoyable. I am genuinely grateful to have had the privilege of working with such an exceptional team.

I would like to thank **Prof. Wim Vranken** and Bio2Byte group for hosting me at VUB, Belgium for my secondment. A special thanks to **Joel Roca-Martinez** for his support and for being a great lab buddy during my secondment.

I am extremely grateful to the RNAct consortium. Thank you all for your friendship, support, and shared experiences throughout this journey. The laughter and chats during the workshops gave me moral and scientific boost, reassuring that I'm on the right track. A special thanks to other two members of the "Trio LaLaLa", **Anna Pérez i Ràfols** and **Roswitha Dolcemascolo** for standing by my side throughout.

I would like to thank all the members of the jury for accepting our invitation and participating in the evaluation of this work: **Prof. Olga Kalinina**, **Dr. Julie Thompson**, **Prof. Alain Denise**, and **Dr. Stanislaw Dunin-Horkawicz**.

Thanks to the engineering team at LORIA, especially **Philippe Noel**, **Jonathan Alcuta**, and **Patrice Ringot** for their help during development and deployment of the InteR3M database.
Thanks to **Antoinette Courrier** and **Isabelle Herlich** who were helpful for all the administrative work throughout this period. The list will be incomplete without the laboratory staff from reception to catering, in particular **Isabelle**, **Florianne**, and **Caro**.

To end, I would like to thank my family and friends back in India for always supporting and believing in me.

Finally, I render my sincere thanks to all those who lent their hands directly and indirectly to accomplish this endeavour.

Contents

1 Introduction **1**
- 1.1 Central Dogma of Molecular Biology . . . 2
- 1.2 The importance of RNA Recognition Motif . . . 3
- 1.3 Thesis Aims and Contributions . . . 4

2 Structural Bioinformatics of RRM: State of the art **7**
- 2.1 Proteins Bioinformatics . . . 8
  - 2.1.1 Protein sequence . . . 8
  - 2.1.2 The Four Levels of Protein Structure . . . 9
  - 2.1.3 Protein Alignments . . . 11
- 2.2 Protein Classification and Domain Databases . . . 14
  - 2.2.1 The concept of protein domain . . . 14
  - 2.2.2 Sequence-based classification . . . 14
  - 2.2.3 Structure-based classification . . . 15
  - 2.2.4 Integrated classification . . . 16
- 2.3 RNA-binding domains and available resources . . . 16
  - 2.3.1 Different RNA Binding Domains . . . 16
  - 2.3.2 Available databases for RNA-binding domains . . . 18
- 2.4 Protein-RNA complexes . . . 19
  - 2.4.1 Molecular Interactions . . . 19
  - 2.4.2 Thermodynamic parameters . . . 20
  - 2.4.3 Experimental methods to Study Binding Affinity . . . 21
- 2.5 Modelling 3D structures of protein and protein-RNA complex . . . 22
  - 2.5.1 Protein 3-D Structure prediction . . . 23
  - 2.5.2 Critical Assessment of protein Structure Prediction (CASP) . . . 24
  - 2.5.3 Modeling Protein-RNA complexes . . . 26
- 2.6 Assessment of modelled 3D structures . . . 28
  - 2.6.1 Overview of Molecular Dynamics . . . 28
  - 2.6.2 Free energy computation . . . 31
  - 2.6.3 Example studies . . . 33

3 InteR3M: The Database for Interactions of RNA and RRM **35**
- 3.1 Introduction . . . 37
- 3.2 Scope & Requirements . . . 37
  - 3.2.1 Mission Statement . . . 37
  - 3.2.2 Mission Objectives . . . 37
  - 3.2.3 Delineation of a correct set of RRM families . . . 38

3.2.4 Use cases . . . 41
3.3 Database Design . . . 44
3.3.1 Conceptual Database Design . . . 44
3.3.2 Logical Database Design . . . 45
3.3.3 Physical Database Design . . . 45
3.4 Implementation . . . 45
3.4.1 Data Collection . . . 47
3.4.2 Database Implementation . . . 49
3.4.3 Implementation of User Interface . . . 50
3.4.4 Testing . . . 50
3.5 Using InteR3M . . . 51
3.5.1 Search functionality . . . 51
3.5.2 RRM instance display . . . 51
3.5.3 Protein display . . . 51
3.5.4 Ligand instance display . . . 51
3.5.5 Experiment display . . . 53
3.5.6 List-of-contacts display . . . 53
3.5.7 Multicriteria Search . . . 53
3.6 Strategies to update InteR3M database . . . 54
3.7 Results . . . 55
3.8 Discussion & Conclusion . . . 62

4 CroMaSt: A workflow for assessing domain classification by cross-mapping of structural instances between protein domain databases 64
4.1 Introduction . . . 65
4.2 Approach . . . 66
4.3 Methods . . . 70
4.3.1 Selection of data sources . . . 70
4.3.2 Retrieve the domain structural instances . . . 70
4.3.3 Compare sets/lists of domain structural instances . . . 71
4.3.4 Cross-mapping of the unique domain structural instances . . . 71
4.3.5 Computation of average structures . . . 72
4.3.6 Structural alignments . . . 73
4.3.7 Implementation . . . 74
4.4 Results . . . 75
4.5 Discussion . . . 79
4.6 Conclusion . . . 80
4.7 Future Perspectives . . . 80

5 Data driven modeling of RRM-RNA complexes 82
5.1 Introduction . . . 83
5.2 Deciphering RRM-RNA Recognition Code . . . 83
5.2.1 General Approach . . . 83
5.2.2 RRM Master alignment . . . 83
5.2.3 Mapping Contacts onto the Alignment . . . 85
5.2.4 Similarity among RRM-RNA complexes . . . 87
5.2.5 Computation of scoring matrices . . . 87
5.2.6 Discussion . . . 88
5.3 Modeling 3D structures of RRM domains . . . 89

5.3.1 Methodology . . . 89
5.3.2 Testing RRMpip . . . 91
5.3.3 Results & Discussion . . . 91
5.4 Modeling RRM-RNA complexes . . . 93
5.4.1 Anchored Docking: Definition and requirements . . . 93
5.4.2 Extraction and Clustering of Anchoring Patterns . . . 93
5.4.3 Resulting Anchoring patterns . . . 95
5.4.4 Docking with Anchoring patterns . . . 96
5.5 Evaluating RRM-RNA complexes . . . 99
5.5.1 Stability of an RRM-RNA complex . . . 99
5.5.2 Free Energy Computation . . . 110

6 Conclusions & Perspectives 117
6.1 Summary of the main contributions . . . 118
6.2 InteR3M Database . . . 118
6.3 CroMaSt Workflow . . . 119
6.4 RRMScorer . . . 120
6.5 RRM-RNA Dock . . . 120
6.6 Evaluating 3D structures of RRM-RNA complexes . . . 121
6.7 Future Directions . . . 121

A Structural Inspection 124
A.1 Classification codes and Alignment scores . . . 124
A.1.1 SCOP classification . . . 124
A.1.2 CATH classification . . . 124
A.1.3 Kpax alignment scores . . . 125
A.2 Structural Inspection of Pfam families . . . 126
A.3 List of RRM instances having bound and unbound structures . . . 146

B CroMaSt Data 147
B.1 Average structures at family level . . . 147
B.2 Comparison of CroMaSt results with structure-based domain databases 151
B.3 RRM clan in Pfam . . . 151
B.4 Starting CroMaSt with different Pfam family . . . 151
B.5 List of obsolete and inconsistent structural instances . . . 155
B.5.1 Inconsistent structural instances . . . 155
B.5.2 Inconsistencies in Pfam and CATH . . . 156
B.5.3 Demonstration of Residue-mapping . . . 156

C Binding Free Energy Computations 158
C.1 D42A point mutation . . . 158
C.2 D48A point mutation . . . 162
C.3 S54A point mutation . . . 166
C.4 F59A point mutation . . . 170
C.5 Q88A point mutation . . . 174

D Summary of the thesis (in French) 192

Chapter 1

Introduction

Summary

1.1	**Central Dogma of Molecular Biology**	**2**
1.2	**The importance of RNA Recognition Motif**	**3**
1.3	**Thesis Aims and Contributions**	**4**

Computational biology has accelerated the pace of advancements in biological sciences. It has opened up new areas of research including data organization, integration, modelling and simulating biological systems, and prediction of interactions between two biomolecules to help us understand the mechanisms of biological processes, especially diseases, and discover therapeutics. The classical goals of computational biology are to distinguish between noise and signals, obtain and quantify trends, and put these together, so that we are able to figure out how the information flows, how the processes are regulated, and what goes wrong in disease [Nussinov, 2013]. More advanced goals consist in designing new molecules and therapeutic approaches, by modelling and simulation methods taking advantage of accumulated knowledge [Hosseinzadeh et al., 2017, Coluzza, 2017]. In any case, computational biology remains a pluridisciplinary field in which computer science results need to be confronted to experimental validation.

1.1 Central Dogma of Molecular Biology

The flow of genetic information is explained by the fundamental theory of central dogma developed by Francis Crick in 1958. The central dogma suggests that DNA contains the information needed to make all of our proteins, and that RNA is a messenger that carries this information to the ribosomes, where the information is 'translated' into the functional product, protein (Figure 1.1).

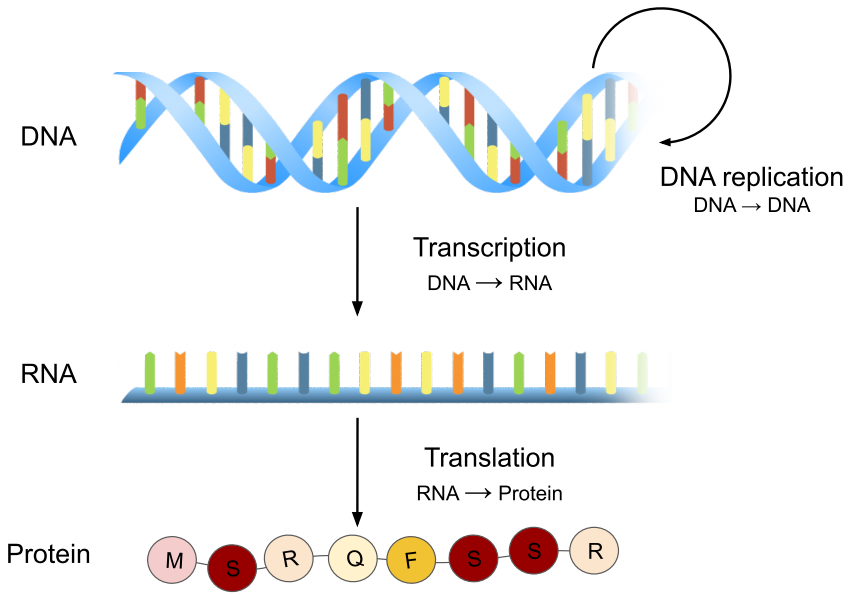


Figure 1.1: Central dogma of molecular biology, flow of information between DNA, RNA and protein.

Protein's functions are best understood when particular domains are found in their structure. A protein domain is generally defined as a conserved structure identified by conserved residues in a multiple sequence alignment across different types of proteins often sharing similar functions. Moreover, it is generally assumed that the protein domains can fold independently from the rest of the protein [Batey and Clarke, 2008].

While RNA was originally believed to be only a carrier of genetic information, this belief was challenged by subsequent work from recent years with discoveries of both new classes of RNAs (e.g., noncoding RNAs) and new RNA-based mechanisms of gene regulation (e.g., microRNA and RNAi silencing) [Holt and Schuman, 2013]. This increased the understanding of many fascinating mechanisms of RNA and of its role as a central player in cellular regulation. A large number of proteins capable of binding to and regulating RNA help facilitate RNA function. RNA-binding proteins (RBPs) regulate numerous aspects of co- and post-transcriptional gene expression (including RNA splicing, polyadenylation, capping, modification, export, localization, translation and turnover). Sequence-specific associations between RBPs and their RNA targets are typically mediated by one or more RNA-binding domains (RBDs), such as the RNA recognition motif (RRM) and hnRNP K-homology (KH) domain [Ray et al., 2013].

1.2 The importance of RNA Recognition Motif

In eukaryotes, the RNA recognition motif is one of the most abundant protein domains. In humans, 497 proteins containing at least one RRM have been identified. Assuming about 20000 - 25000 human genes, the RRM would therefore be present in about 2% of gene products. RRMs are well-studied and are extremely versatile in their RNA recognition capability, which can even be modulated allosterically [Ryder et al., 2012]. RRM domains often occur in multi-domain RBPs, with their modular association allowing the recognition of separate RNA motifs that are sequentially remote [Maris et al., 2005]. A protein domain in such abundance is necessarily biologically important and associated with many functions in the cell. The RRM domain plays an important role in several key biological processes including post-transcriptional gene regulation, formation of amyloid-like aggregates [Berchowitz et al., 2015], abnormal cell proliferation [Chen et al., 2019], maintenance of stem cells and telomerase activity [Xie et al., 2021]. The engineering of these RRMs have many applications for the creation of new synthetic biological pathways and for the discovery of new treatments for RNA-asssociated diseases [Shotwell et al., 2020].

RNAct is an MSCA-ITN[1] project with a research focus of designing novel RRM proteins for exploitation in synthetic biology and bio-analytics. Ten Early Stage Researchers (ESRs) have combined computational and experimental methods to achieve this goal within the RNAct project (Figure 1.2). I am ESR3 within the RNAct project, focusing mainly on the diversity and characterization of RRM domains using computational techniques.

Understanding RNA binding is essential for designing RRM that can inhibit or bind to specific RNAs. To understand the RNA binding patterns of RRM domain, it is necessary to collect and analyze all the available information about RRMs. While several studies have been carried out trying to decipher the RNA binding code for single and multi-RRM domains [Birney et al., 1993, Allain et al., 1996, Bauer

[1]Marie Skłodowska-Curie Innovative Training Network

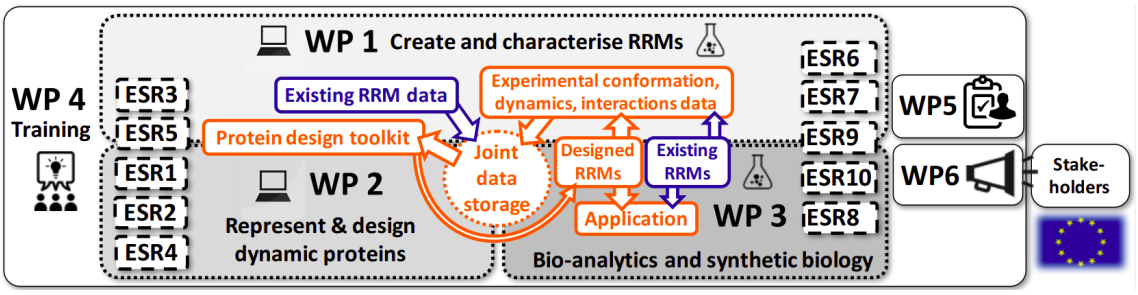


Figure 1.2: Overview of the work packages (WP) and ESR involvement within the RNAct project. Image by W. Vranken taken from the RNAct application.

et al., 2012, Lukavsky et al., 2013, Wang et al., 2014], there are very few successful examples of RRM design [Blakeley and McNaughton, 2014, Chen et al., 2016]. The reason behind this is likely because RRM domains bind RNA with diverse and less predictable binding modes and specificities compared to other RBPs.
Thus, it is important to study and understand the diversity displayed by RRM domains and their binding modes to successfully design novel RRM domain with desired activity.

1.3 Thesis Aims and Contributions

This thesis aims to describe the diversity of RRM domains and of their binding modes to RNA based on experimental reports, and to exploit such data for deciphering the RRM-RNA recognition code and for improving modeling of RRM-RNA complexes. Figure 1.3 illustrates the various steps of this work.

Chapter 2 covers a general introduction to protein sequences and structures, protein alignments, protein domains, protein functions, and protein-RNA interactions. Moreover, this chapter helps to understand the state of the art methods and resources that will be useful for this thesis.

A key step in this work consisted in developing a complete and comprehensive database with RRM information. This database stores domain information from domain databases, sequence information from UniProt, structural information from Protein Data Bank (PDB), and binding information retrieved from literature. The collected data was analyzed to study the diversity of RRM which helped us to differentiate the RRM fold from the non-RRM fold. The database provides an easy way to explore and exploit the existing knowledge about the RNA binding capacity of RRM domains. *Chapter 3* describes a new database called 'InteR3M' (Interactions of RNA and RNA Recognition Motif) including it's design, implementation and data collection procedures.

Different generalist domain databases employ different classification systems leading to inconsistent data for a given domain type. For example, domain families corresponding to the same type of domain in two different databases may not contain the same domain instances. We encountered such inconsistencies during the data collection procedure for the InteR3M database. To solve this problem, we developed the 'CroMaSt' workflow for assessing domain classification by cross-mapping of structural instances between protein domain databases. Our

cross-mapping approach provides an easy way to identify the domains classified by only one classification system and not the other. Further analysis of these domains (classified by only one method) can lead us to identify wrongly classified domains or the domains not recognized by a method. *Chapter 4* presents the CroMaSt workflow including the details for conceptualization and implementation of the method, and usage for RRM domain type.

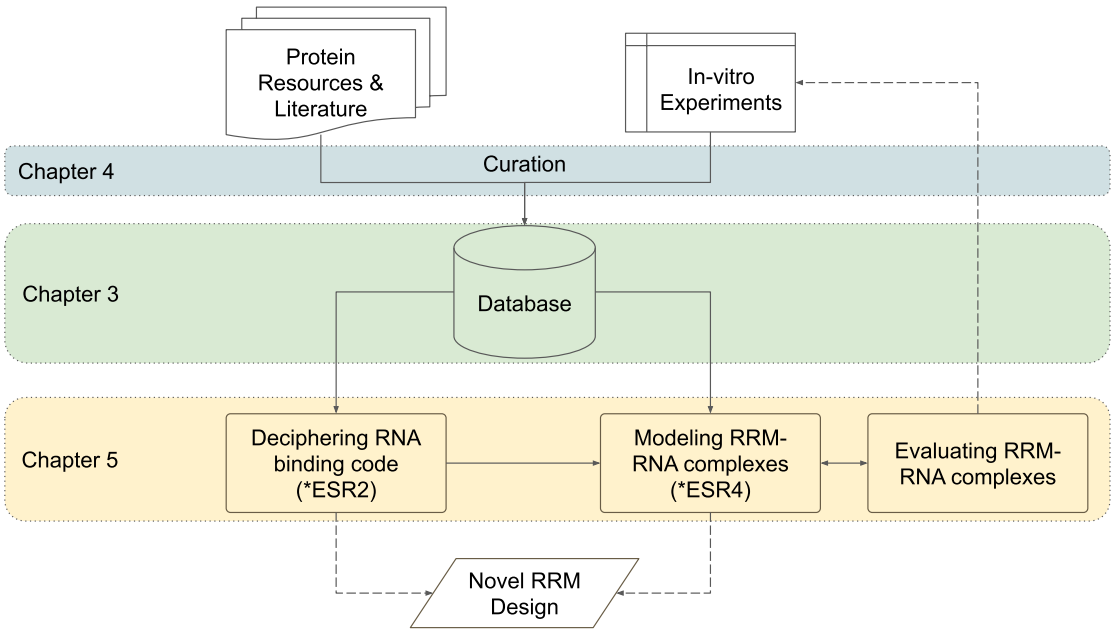


Figure 1.3: My role from computational perspective within the RNAct project. The dotted boxes represent chapters of this thesis. The dashed lines indicate experimental tasks within the RNAct project.
* indicates the collaboration with other ESRs within the RNAct project.

Chapter 5 describes the computational approaches developed using the data from InteR3M database. The sequence and structural data stored in the database can be used to generate an alignment covering all the RRM domains. With the help of an alignment, one can easily explore and retrieve the residue-specific characteristics from the RRM domain. The RRM alignment and binding information can be used jointly to identify the residues from interface and different binding modes of RRM leading towards the better understanding of RRM binding patterns (Section 5.2).
Moreover, other computational approaches, like modeling 3D structures of RRM from sequences and modeling RRM-RNA 3D complexes, can be developed using information from this database. Such modeling approaches can be used to bridge the gap between RRM sequences and available structures. We can start with testing of a simple comparative modeling approach to model the 3D structures of RRM from the sequence. In the frame of RRM-RNA modeling, the goal of this objective is to model diverse structures of an RRM from a given sequence so that at least few structures will be close to the RNA-bound form of the RRM (Section 5.3).
The resulting 3D structures of an RRM can be used to model RRM-RNA complexes by docking protocols. In addition, we can incorporate the binding information stored in the database into the docking protocol (data-driven approach) (Section 5.4).
Then the modelled 3D structures of RRM-RNA complexes can be evaluated further

to distinguish between strongly bound and weakly bound RRM-RNA complexes (Section 5.5).

At the end, *Chapter 6* summarises the contributions of this thesis, and it presents some future possible developments along with scientific prospects.

Chapter 2

Structural Bioinformatics of RRM: State of the art

Summary

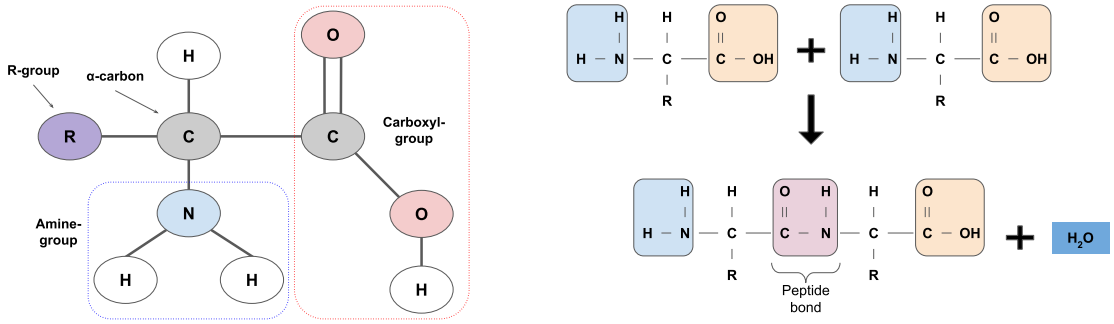
2.1 Proteins Bioinformatics **8**
2.1.1 Protein sequence 8
2.1.2 The Four Levels of Protein Structure 9
2.1.3 Protein Alignments 11
2.2 Protein Classification and Domain Databases **14**
2.2.1 The concept of protein domain 14
2.2.2 Sequence-based classification 14
2.2.3 Structure-based classification 15
2.2.4 Integrated classification 16
2.3 RNA-binding domains and available resources **16**
2.3.1 Different RNA Binding Domains 16
2.3.2 Available databases for RNA-binding domains 18
2.4 Protein-RNA complexes **19**
2.4.1 Molecular Interactions 19
2.4.2 Thermodynamic parameters 20
2.4.3 Experimental methods to Study Binding Affinity 21
2.5 Modelling 3D structures of protein and protein-RNA complex **22**
2.5.1 Protein 3-D Structure prediction 23
2.5.2 Critical Assessment of protein Structure Prediction (CASP) 24
2.5.3 Modeling Protein-RNA complexes 26
2.6 Assessment of modelled 3D structures **28**
2.6.1 Overview of Molecular Dynamics 28
2.6.2 Free energy computation 31
2.6.3 Example studies 33

2.1 Proteins Bioinformatics

Proteins are one of the main macromolecular components of life and are central players in the myriad of cellular functions in all living organisms. Proteins are responsible for nearly every task of cellular life, including cell shape and inner organization, product manufacture and waste cleanup, and routine maintenance. Proteins also receive signals from outside the cell and mobilize intracellular response. They are the workhorse macromolecules of the cell and are as diverse as the functions they serve.

2.1.1 Protein sequence

Proteins are made up of a sequence of amino acids translated from the sequence of nucleotides in a gene. In total, there are 20 naturally occurring amino acids in living organisms that create all the proteins. Each amino acid corresponds to a triplet of nucleotides according to the genetic code [Crick, 1968]. Every amino acid has an acidic carboxyl group, a basic amino group and a variable R group (side-chain) (Figure 2.1a).



(a) Amino Acid structure

(b) Formation of Peptide Bond

Figure 2.1: The amino acid structure and formation of peptide bond between two amino acids.

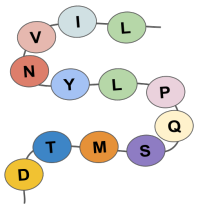
The amino acids are bound together by a peptide bond, in which the carboxyl group from one amino acid joins the amino group of another amino acid with a loss of water molecule (Figure 2.1b). Multiple amino acids are added in similar way next to each other to form a complete protein. Thus, protein has two ends, one with free amino group called N-terminus and the other with a free carboxyl group called C-terminus.

Universal Protein Knowledgebase (UniProtKB): The Universal Protein KnowledgeBase (UniProtKB) is the main publicly available resource for protein sequences and associated metadata [The-UniProt-Consortium, 2023]. UniProtKB is the central resource that combines a reviewed protein set from UniProtKB/Swiss-Prot and an unreviewed protein set from UniProtKB/TrEMBL, which gathers all possible protein sequences translated from all sequenced genomes so far.
All the protein entries from Swiss-Prot are linked to a summary of experimentally

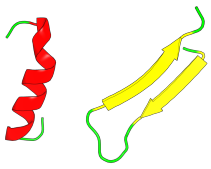
verified, or computationally predicted, functional information added by the experts, whereas all the protein entries from TrEMBL are annotated computationally by automated systems.
UniProtKB integrates data from other resources to add biological knowledge, making the UniProtKB a central hub for proteins. In addition, UniProtKB provides links to all those external resources from which the information was integrated.
The UniProtKB release 2023_01 contains a total of 246,440,937 entries. UniProtKB/Swiss-Prot contains 569,213 sequence entries, curated from 291,046 unique literature references, whereas UniProtKB/TrEMBL contains a total of 245,871,724 sequence entries.

2.1.2 The Four Levels of Protein Structure

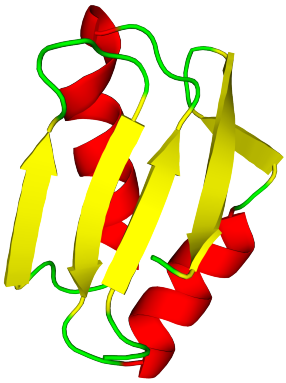
The specific sequence of amino acids in a protein determines its unique structure at different levels, i.e. primary, secondary, tertiary and quaternary structure. The primary structure of a protein is its linear sequence of amino acids. The secondary structure of a protein refers to regular, local structures of the protein backbone, stabilised by intra-molecular and sometimes inter-molecular hydrogen bonds. Common types of secondary structures are alpha helix and beta sheets. The tertiary structure is the overall three-dimensional shape of the protein determined by interactions like hydrogen bonding, ionic interactions, van der Waals forces, and hydrophobic packing. The quaternary structure of the protein is the arrangement of multiple polypeptide chains in a structural assembly (protein complex) formed by the interactions between these subunits (Figure 2.2).



(a) Primary Structure



(b) Secondary Structure



(c) Tertiary Structure



(d) Quaternary Structure

Figure 2.2: Structural levels of proteins.

Experimental Methods to Study Protein Structure

Several techniques can be used to study 3D structure of proteins, including X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM). These techniques allow to determine the precise position of all the atoms present in a protein, which can provide valuable insights into protein's function.

X-ray Crystallography: X-ray crystallography can provide very detailed atomic information, showing every heavy atom in a protein or nucleic acid along with atomic details of ligands, inhibitors, ions, and other molecules that are incorporated into the crystal. The protein is purified and crystallized, then the formed crystal is subjected to an intense beam of X-rays. The molecules in the crystal diffract the X-ray into a characteristic pattern of spots. These spots are then analyzed to determine the distribution of electrons in the protein, resulting in electron density map. This electron density map is interpreted to determine the location of each atom.

NMR Spectroscopy: NMR spectroscopy is used to determine the structure of proteins in solution rather than locked in a crystal or bound. To determine the coordinates of each atom from protein, the protein is labelled with an isotope (^{15}N, ^{13}C, or deuterium), then purified, placed in a strong magnetic field, and then probed with radio waves. A distinctive set of observed resonances may be analyzed to give a list of atomic nuclei that are close to one another, and to characterize the local conformation of atoms that are bonded together. This list of restraints is then used to build a model of the protein that shows the location of each atom.
As large proteins present problems with overlapping peaks in the NMR spectra, this method is limited to small or medium proteins.

Electron Microscopy (EM): EM is used to determine 3D structures of large macromolecular assemblies. A beam of electrons and a system of electron lenses is used to image the biomolecular particles directly. The most commonly used technique today involves imaging of many thousands of different single particles preserved in a thin layer of non-crystalline ice (cryo-EM). Provided these views show the molecule in myriad different orientations, a computational approach akin to that used for computerized axial tomography (CAT) scans in medicine will yield a 3D mass density map. With a sufficient number of single particles, the 3D electron microscopy (3DEM) maps can then be interpreted by fitting an atomic model of the macromolecule into the map, just as macromolecular crystallographers interpret their electron density maps [Zardecki et al., 2022].

Results from experimental studies of protein 3D structure determination are stored in a dedicated database, the Protein Data Bank (PDB). This data resource serves as the single most comprehensive repository for archiving protein 3D structures and provides valuable 3D structural data to support a wide range of scientific investigations.

Protein Data Bank (PDB) is a core data resource serving as the single global repository for atomic-level, 3D structural data of experimentally determined structures of biomolecules including their complexes with metal ions, drugs, and

other small molecules [Berman et al., 2007]. Since 2003, the PDB has been managed jointly by the Worldwide Protein Data Bank (wwPDB) consortium to ensure universal open access and compliance with FAIR[1] principles. The wwPDB consortium includes US Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) [Burley et al., 2021], Protein Data Bank in Europe (PDBe) [Armstrong et al., 2020], Protein Data Bank Japan (PDBj) [Bekker et al., 2022], and Biological Magnnetic Resonance Bank (BMRB) [Hoch et al., 2023].
As of 14-March-2023, the PDB contains 202,292 3D structures, 61,535 of which are from human sequences and 15,736 are structures containing nucleic acid.

2.1.3 Protein Alignments

Protein alignments attempt to establish residue-to-residue correspondence between two (pairwise) or more (multiple) proteins.

Why to align Proteins? Protein alignments have a wide range of applications including phylogenetic analysis, similarity search, function prediction etc. There are many algorithms and tools developed for each type of alignment over the past few years but the scope of this section is not to review the complete landscape of protein alignment tools. In particular we focused only on multiple protein alignment tools.
A set of proteins can be aligned either using sequence-based alignments (MUSCLE, T-coffee, Clustal Omega) or structure-based alignments (Kpax, FATCAT), or by incorporating sequence and structural information together (PROMALS3D, Expresso).
All of the above-mentioned alignments result into aligned sequences from input proteins with or without additional information like alignment score, sequence similarity, or superimposed structures.

Sequence-based Alignments

An alignment between two protein sequences is calculated by optimizing the alignment scoring function. The alignment score is usually a sum of the gap penalties that depend linearly on the gap lengths, and the pairwise substitution scores that depend on the matched residue types [Marti-Renom et al., 2004].
The sequence-based alignments can be performed through sequence-to-sequence alignments, sequence-to-profile alignments, or profile-to-profile alignments. A protein sequence profile lists the preference for all the 20 standard amino acid residue types at each position in a given multiple sequence alignment.

Clustal Family: Clustal family is comprised of a series of tools, starting from Clustal V [Griffin et al., 1994], Clustal W [Thompson et al., 1994], Clustal X [Thompson et al., 1997], to Clustal Omega [Sievers et al., 2011], to align protein and nucleic acid sequences. Clustal Omega is the latest addition to the Clustal family. Clustal Omega produces multiple sequence alignments using seeded guide trees and hidden markov model (HMM) profile-profile alignments. It supports

[1]Findable, Accessible, Interoperable, Reusable

alignments of protein, DNA, and RNA. Clustal Omega can also add sequences to an existing alignment. Clustal Omega uses a modified version of mBed [Blackshields et al., 2010] to produces guide trees that are just as accurate as those from conventional methods. In addition, it uses HHalign [Söding, 2005] for aligning HMM profiles that led Clustal Omega to achieve higher accuracy than earlier programs from the Clustal family [Sievers et al., 2011].

T-coffee: T-coffee (Tree-based Consistency Objective Function for alignment Evaluation) is a multiple sequence alignment tool that can align protein, DNA and RNA sequences. T-coffee pre-processes a data set of global and local pairwise alignments between the input sequences (using ClustalW [Thompson et al., 1994] and Lalign [Huang and Miller, 1991] programs, respectively) resulting in a library of alignment information to guide the progressive alignment. Thus, T-coffee uses this approach of combining local and global pairwise alignments to generate the final multiple alignments [Notredame et al., 2000].

MUSCLE: MUSCLE (MUltiple Sequence Comparison by Log-Expectation) is a program for creating multiple alignments of protein sequences. MUSCLE uses a progressive approach to align sequences by constructing a guide tree to estimate evolutionary relationships among the sequences. MUSCLE aligns a set of sequences in two stages using k mer and Kimura distances, respectively. In the first stage, the guide tree is built by computing the pairwise distances between all sequences using the k-mer distance metric. A k mer is a contiguous subsequence of length k. The k mer distance does not require an alignment, improving the speed significantly. In second stage, MUSCLE reestimates the guide tree using Kimura distance. For each aligned pair of sequences, MUSCLE computes the Kimura distance that will be used to produce binary tree using Unweighted Pair Group Method with Arithmetic Mean (UPGMA). MUSCLE uses this guide tree to guide the alignment, first aligning the two closest sequences in the guide tree and then proceeds to align the remaining sequences one by one [Edgar, 2004].

All these three programs are included in the EMBL-EBI sequence analysis tools [Madeira et al., 2022] and available at https://www.ebi.ac.uk/Tools/msa/.

Structure-based Alignments

Structure-based alignments does not need any prior knowledge of equivalent pair of residues, hence does not rely on the sequence alignment and the type of residues is ignored when the correspondence is established. Protein structural alignment can be either rigid or flexible. Rigid structural alignment does not allow any atoms in the protein structures to move relative to each other, so the alignment is performed by rotating and translating the entire protein structure as a single rigid body. The flexible alignment allows the rotations and translations between atoms of one structure making it possible to effectively align and compare protein structures even if they underwent structural rearrangements.

FATCAT webserver: The FATCAT structure alignment is formulated as an AFP (aligned fragment pair) chaining process, allowing flexibility in connecting

them. Aligned fragment pair is defined as a match of two fragments, one from each structure. A rotation and/or translation is introduced between two consecutive AFPs if it results in a substantially better superposition of the structures. FATCAT integrates simple extensions, gaps and twists into a unified scoring function and performs the alignment and hinge detection simultaneously using dynamic programming. Several post-processing steps are applied to refine the alignments. The significance of the similarity detected by FATCAT is evaluated by a P-value that measures the chance of getting the same similarity in two random structures [Ye and Godzik, 2004].

TM-align: TM-align is an algorithm to align protein structures pairwise. For two input protein structures, TM-align first generates optimized residue-to-residue alignment based on structural similarity using heuristic dynamic programming iterations. An optimal superposition of the two structures built on the detected alignment, as well as the TM-score value which scales the structural similarity, will be returned. TM-score is a metric for assessing the topological similarity of protein structures [Zhang and Skolnick, 2005].

PDBeFold: PDBefold is a structure alignment tool which can perform both pairwise and multiple structural alignments. PDBeFold assumes that multiple alignment preserves the connectivity of the structural elements and aims to identify optimal alignment for these structural elements. Structural element is defined as one or more secondary structural elements (SSE) found in a certain geometrical orientation with regard to each other and ordered in the same way along the protein sequence. The Q-score is used to evaluate the quality of pairwise alignments and guide the selection of SSEs that should be excluded from consideration [Krissinel and Henrick, 2005].

Kpax: Kpax uses gaussian overlap functions to score the local and spatial environment of each amino acid residue in a protein using dynamic programming approach to find the optimal global alignment for proteins based on their gaussian similarity scores. Kpax uses multiple C_α coordinate systems and a Gaussian peptide fragment scoring scheme to provide a sensitive structural similarity score. Kpax provides the functionalities to create a custom database for a set of given structures, and allows large structure databases to be searched or queried rapidly. Users can choose between rigid or flexible and pairwise or multiple structural alignments [Ritchie, 2016]. By default, this tool provides several metrics to assess the resulting alignments.

Alignments incorporating sequence and structural information

PROMALS3D: PROMALS3D (PROfile Multiple Alignment with predicted Local Structures and 3D constraints) is a progressive method that clusters similar sequences for easy alignments and applies more elaborate techniques to align the relatively divergent clusters. In the first alignment stage, PROMALS3D aligns similar sequences using a scoring function of weighted sum-of-pairs of BLOSUM62 scores. The first stage is fast and results in a number of pre-aligned groups that are relatively distant from each other. In the second alignment stage, one

representative sequence is selected for each group, and submitted to PSI-BLAST searches to retrieve additional homologs from UNIREF90 database and to PSIPRED secondary structure prediction. Then, an HMM of profile-profile alignments with predicted secondary structures is applied to pairs of representatives to obtain posterior probabilities of residue matches. These probabilities serve as sequence-based constraints that are combined with constraints derived from homologs with 3D structures or user-defined alignment constraints to derive a probabilistic consistency scoring function. The representative sequences are progressively aligned using such a consistency scoring function, and the pre-aligned groups obtained in the first stage are merged into the alignment of representatives to form the final multiple alignment of all sequences [Pei et al., 2008].

2.2 Protein Classification and Domain Databases

2.2.1 The concept of protein domain

A protein domain can be considered as an abstract class derived from the properties shared by multiple instances, where each instance is a well defined structural region of a protein. The protein domain class, from a sequence perspective, can be captured by a given HMM or PSSM (Position Specific Scoring Matrix) profile, generally called 'signature' (even with varying sequence length). From a structural perspective, it shares a common topology or fold but the size may vary in different instances. From a functional perspective, most often it associates with a specific, definitive function in the protein.
Domain instances can be characterized by the sequences of the corresponding proteins annotated with start and end positional boundaries. Each domain instance may have one or more 3D structures, i.e. domain **st**ructural **i**nstances abbreviated as StIs and further described in chapter 4.

There are several resources available that provide information about different protein domain types and their classification. In the rest of this thesis, we will use the term "domain" for the abstract class representing the set of domain instances satisfying the domain signature. All the domain databases can be grouped into three categories depending on the rationale used for classification.

2.2.2 Sequence-based classification

Among sequence-based domain databases, Pfam is certainly one of the most comprehensive resources.

Pfam database: Pfam classification starts with a seed alignment for a representative set of sequences. Structural data, when available, are used during the curation step to align domain boundaries with those provided by structure-based classification, such as SCOP [Bateman et al., 2002] or CATH [Finn et al., 2014]. Then, an HMM (Hidden Markov Model) profile is built based on the seed alignment which is then used to retrieve a full set of sequences matching this

domain from the UniProt reference proteomes, thus producing the Pfam entry full alignment. Pfam classifies all of the entries in one of the six ways:

1. **Family:** Collection of related protein regions
2. **Domain:** A structural unit
3. **Repeat:** A short unit which is unstable in isolation but forms a stable structure when multiple copies are present
4. **Motifs:** A short unit found outside globular domains
5. **Coiled-Coil:** Regions that predominantly contain coiled-coil motifs, regions that typically contain alpha-helices that are coiled together in bundles of 2-7
6. **Disordered:** Regions that are conserved, yet are either shown or predicted to contain bias sequence composition and/or are intrinsically disordered (non-globular).

Sets of Pfam entries that are thought to be evolutionarily related are grouped together into clans. This grouping is based on sequence similarity, structural similarity, functional similarity and/or profile-profile comparisons [Mistry et al., 2021].

2.2.3 Structure-based classification

SCOP database: The SCOP database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between proteins whose three-dimensional structure is known and deposited in the PDB. The classification of proteins in SCOP has been constructed mainly manually by visual inspection and analysis. In SCOP, entries are protein domains identified in PDB structures and organized into families and superfamilies and finally into structural folds and classes reflecting their secondary structure content. Domain boundaries are provided at both family and superfamily levels as the evolutionary relationships can sometimes span regions of different size between closely related proteins (family level) and more distantly related protein domains (superfamily level). In brief, the family domain boundaries can define conserved multi-domain regions whereas the superfamily domains span over the individual domains. The curated cross-references are indicated at the family level from SCOP to Pfam database in many cases [Andreeva et al., 2020]. SCOP2, a successor to the SCOP database was introduced in 2014 [Andreeva et al., 2014].

CATH database: The CATH project has developed a semi-automatic procedure to split 3D structures into the constituent domains defined as semi-independently folding globular units. These domains are then clustered into homologous superfamilies based on evolutionary relationships. The sequences of the CATH structural domains are then used to build HMM profiles in order to identify the domains in UniProt protein sequences for which no 3D structure is available. This effort is shared with the sister resource Gene3D. The lowest level of CATH is H for Homologous superfamilies and CATH provides structural superpositions of all representative domains of a superfamily. However, superfamilies can be sub-divided in functionally coherent groups named Functional Families (FunFams). Recently, CATH has created an additional class for non-globular domains [Sillitoe et al., 2021].

2.2.4 Integrated classification

InterPro database: InterPro is an integrated resource of predictive models or 'signatures' representing protein domains, families, regions, repeats and sites from major protein signature databases including CATH-Gene3D, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, PROSITE, SMART, SUPERFAMILY and TIGRFAMs. Thus, InterPro aims to combine the individual strength of each protein domain database without building domain models itself. Quality control is performed at InterPro when integrating new signatures by checking whether such signatures generate false positive matches. Hierarchical relationships are identified between InterPro entries to represent subfamilies displaying specific functions within larger families, or specific subclasses within certain classes of domains. The InterProScan software regularly calculates InterPro signature matches to UniProtKB [Blum et al., 2021].

CDD: The CDD (Conserved Domain Database) is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for domains and full-length proteins obtained from both NCBI projects (NCBI Protein Clusters collection, NCBIfam, CDD itself) and external sources (Pfam, SMART, COG, TIGRFAMs). These models are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST (Reversed Position Specific BLAST). Within CDD, the NCBIfam curated domains use 3D structure information to define domain boundaries explicitly and provide insights into sequence/structure/function relationships. CDD shares domain models with InterPro and contributes to enlarge InterPro with specific subfamilies [Lu et al., 2020].

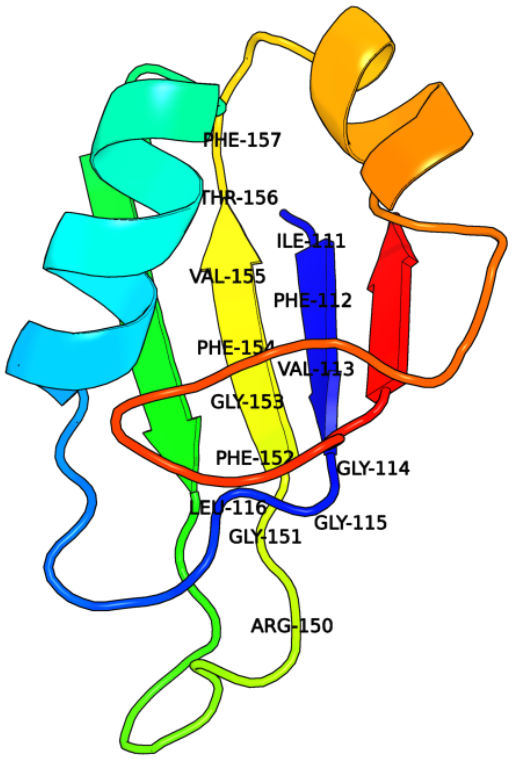
2.3 RNA-binding domains and available resources

RNA-binding domains are often in contact with only a few nucleotides, but there can be multiple RNA-binding domains within a single protein. The different categories of RNA-binding proteins (RBPs) and their structural specificities are presented below.

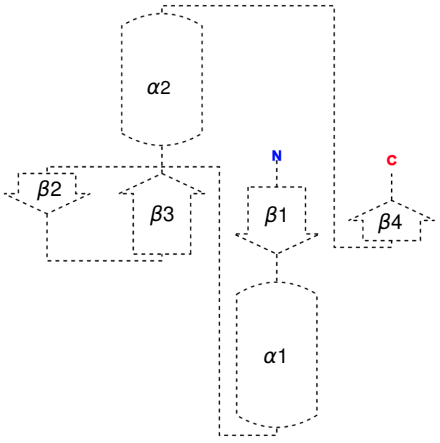
2.3.1 Different RNA Binding Domains

RRM Domain: RRM (RNA Recognition Motif) is the most abundant RNA binding domain and present in about 2% of gene products in human [Maris et al., 2005]. This domain can also interact with ssDNA and other protein partners, allowing it to perform various biological function [Cléry and Allain, 2012]. RRM domains are generally 90 amino acids long in sequence with two conserved motifs RNP1 and RNP2 consisting of 8 and 6 amino acids, respectively. The RNP1 and RNP2 consensus sequences are (R/K)-G-(F/Y)-(G/A)-(F/Y)-V-X-(F/Y) and (L/I)-(F/Y)-(V/I)-X-(N/G)-L, located on $\beta 3$ and $\beta 1$ sheets, respectively. An example of RRM domain is illustrated in Figure 2.3 with the topology from PDBe

Topology Viewer module[2]. RRM domains often occur in multi-domain RBPs, with their modular association allowing the recognition of separate RNA motifs that are sequentially remote [Maris et al., 2005].
A protein domain in such abundance is necessarily biologically important and associated with many functions in the cell. The RRM domain plays an important role in several key biological processes including post-transcriptional gene regulation, translation repression for meiosis activation in yeast [Berchowitz et al., 2015], abnormal cell proliferation [Chen et al., 2019], maintenance of stem cells and telomerase activity [Xie et al., 2021].



(a) Structure of RRM domain with conserved RNP sequences



(b) Topology of RRM domain

Figure 2.3: Typical characteristics of RRM domain illustrated with PDB entry 2MSS (Musashi1 RBD2, NMR). The amino acids of conserved motives RNP1 (RGFGFVTF) and RNP2 (IFVGGL) present in $\beta 3$ and $\beta 1$ segments, respectively have been indicated

KH Domain: The KH (K Homology) domain is of approximately 70 amino acids and typically found in multiple copies. There are two different versions of the KH motif, named type I and type II KH folds. Both type I and type II share a minimal KH motif ($\beta 1 \alpha 1 \alpha 2 \beta 2$), in addition to an extra α and β strand [Valverde et al., 2008].

Pumilio Homology Domain: This domain generally consists of eight 36 amino acids long repeats. The complete domain forms a curved structure allowing

[2]https://github.com/PDBeurope/pdb-topology-viewer

interaction with 8 to 10 nucleotides. Due to the very good understanding of the interaction between these domains and RNAs, it is possible to artificially create these domains to bind specific sequences of RNA [Zhou et al., 2021, Wang et al., 2002].

Zinc Finger Domain: A typical Zinc finger domain is about 30 amino acids long with a $\beta\beta\alpha$ protein fold in which a β-hairpin is pinned with an α helix by a Zn^{2+} ion. In a protein, there may be several copies of zinc finger domains, which are frequently arranged in groups or clusters of tandem repeats [Brayer and Segal, 2008].

2.3.2 Available databases for RNA-binding domains

RBPDB: RBPDB (database of RNA-binding protein specificities) is a collection of RNA-binding proteins linked to a curated database of published observations of RNA binding. The data in RBPDB is separated into two sets: proteins and experiments. Each experiment is an observation of RNA-protein binding to a single sequence (e.g. in a gel shift or UC cross-linking experiment) or many sequences (e.g. a SELEX or RIP-chip experiment). Each experiment is linked to a single RNA-binding protein. RBPDB contains binding information of 1141 RBPs including 414 from Humans.

RNAct: RNAct computes Protein-RNA interaction propensity scores using catRAPID algorithm [Bellucci et al., 2011] with the fragmentation procedure. For each protein-RNA pair, the fragment with the maximum interaction propensity score is used to assess overall binding ability. The catRAPID method was trained on X-ray and NMR data, not on recent high-throughput data. The agreement between the original catRAPID approach and the eCLIP experiments [Van Nostrand et al., 2016] is strong (AUC=0.72). Currently, the RNAct covers human protein-RNA interactome with 20778 proteins and 199330 RNA transcripts [Lang et al., 2019].

POSTAR: POSTAR3 is a comprehensive database for exploring POST-trAnscriptional Regulation based on high-throughput sequencing data from 7 species, including human, mouse, zebrafish, fly, worm, Arabidopsis, and yeast. POSTAR3, previously known as CLIPdb, describes the RBP-RNA interactions based on publicly available CLIP-seq data sets for 351 RBPs with their domains, Gene Ontologies, binding sequence motifs, structural preferences, and binding sites. The CLIPdb module provides various annotations for the RNA-binding proteins, including RNA recognition domains, Gene Ontology, sequence motifs, and structural preferences. All the binding sites for the query RNA-binding protein and expression level for the target genes are also included. [Zhao et al., 2022]

RRMdb: The RRMdb (RNA Recognition Motif database) is an evolution-oriented database of RNA recognition motif sequences, published in 2019 but that is no more available on the web. The RRMdb was compiled from 57,000 collected representative RRM domain sequences, classified into 415 families.
A representative set of RRM core domain sequences (consisting of the

$\beta1\alpha1\beta2\beta3\alpha2\beta4$ regions) was constructed from the SCOP and Pfam databases. The representative set was searched against the NCBI non-redundant protein database using PSI-BLAST, and the resulting sequences were filtered to 90% sequence identity using CD-HIT. This yields in a database of 57471 sequences. All these sequences were clustered into 415 families based on all-vs.-all BLAST comparisons using the Markov Cluster (MCL algorithm). A multiple sequence alignment was generated for each family using MUSCLE [Edgar, 2004], and calculated a Hidden Markov Model (HMM) using hhmake after manual corrections in alignment. A network was generated from these HMM models where the nodes represent RRM families and edges denote significant similarities between them. [Nowacka et al., 2019]

2.4 Protein-RNA complexes

2.4.1 Molecular Interactions

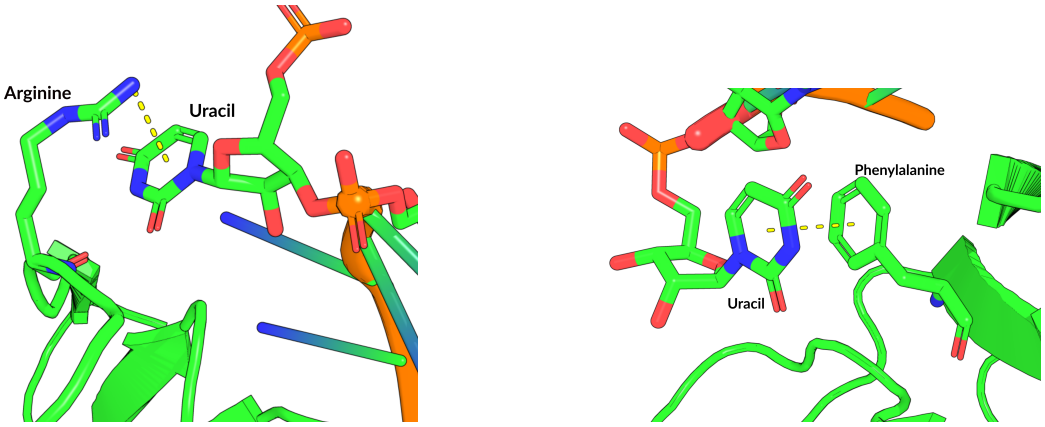
Molecular recognition is a term used to describe the selective binding between two or more molecules that is mediated by noncovalent interactions. Non-covalent interactions are relatively weak interactions formed between molecules that do not involve the complete sharing of electrons [Daze and Hof, 2016]. The function of proteins is determined at some level by their ability to form intra- or intermolecular non-covalent interactions. There are many different noncovalent interactions.

van-der-Waals interactions: These interactions arise when any two atoms approach each other closely, and create a weak, nonspecific attractive force. These nonspecific interactions result from the momentary random fluctuations in the distribution of the electrons of any atom, which give rise to a transient unequal distribution of electrons, that is, a transient electric dipole. If two non-covalently bonded atoms are close enough together, the transient dipole in one atom will perturb the electron cloud of the other. This perturbation generates a transient dipole in the second atom, and the two dipoles will attract each other weakly.

Ionic Bond: Ionic Bond is a strong non-covalent attraction between 2 charged molecules, a negatively charged (anion) and a positively charged (cation). As the ionic bonds involve fully-charged molecules, they're stronger than other non-covalent interactions like van der Waals interactions which only involve attractions between temporary charges.

Hydrogen Bond: The hydrogen bond is the attractive interaction between a hydrogen atom that is attached to a more electronegative (donor) atom and an acceptor atom bearing electrons available for sharing. Hydrogen bonds between hydrogen atom and nitrogen/oxygen atoms are most common in nature.

Cation-π Interaction: These interactions are attractive force between a positively charged cation and a negatively charged cloud of π systems [Dougherty, 2013]. The π electron cloud is formed by an aromatic ring that occurs in sidechains of aromatic amino acids or the nitrogenous base from nucleotides.



(a) Cation-π Interaction (b) π-π Interaction

Figure 2.4: π stacking interactions

π-π Interaction: These are attractive interactions between two aromatic π - π systems. These interactions can occur in one of the three common orientations: stacked (face-to-face or parallel), t-shaped (edge-to-face), and slip-stacked (slipped parallel) [Sinnokrot et al., 2002].

2.4.2 Thermodynamic parameters

Understanding of thermodynamic parameters like binding free energy (ΔG), enthalpy (ΔH), entropy (ΔS) is important to gain a deeper understanding of inter-molecular binding. A simple binding reaction between a protein (P) and ligand (L) to form a biomolecular complex (P:L) can be formulated as in Eq. 2.4.1.

$$P + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} P : L \qquad \text{(Eq. 2.4.1)}$$

Where, the k_{on} and k_{off} are the kinetic rate constants for the forward (binding) and backward (unbinding) reactions, respectively.
At equilibrium, the forward binding reaction is balanced by the backward unbinding reaction, and can be written as follows (Eq. 2.4.2):

$$k_{on}[P][L] = k_{off}[P : L] \qquad \text{(Eq. 2.4.2)}$$

where, the [P] and [L] represent the concentrations of protein and ligand, respectively. [P:L] represents the concentration of protein-ligand complex.
Then the association and dissociation constants for this reaction are defined by following equation:

$$K_a = \frac{k_{on}}{k_{off}} = \frac{[P : L]}{[P][L]} = \frac{1}{K_d} \qquad \text{(Eq. 2.4.3)}$$

where, K_a is association constant with the unit of M^{-1} and K_d is dissociation constant with the unit of M.
A fast association rate along with a slow dissociation rate will give a higher binding and lower dissociation constant, resulting in high binding affinity [Du et al., 2016]. Binding affinity is the strength of the binding interaction between

protein and ligand. Binding affinity is measured and reported using the dissociation constant (K_d). The smaller the K_d value, the greater the binding affinity of ligand for the protein and vice versa.

The capacity of a thermodynamic system is measured by the Gibbs free energy, to do maximum or reversible work at a constant pressure (isobaric) and temperature (isothermal) [Gilson and Zhou, 2007]. The free energy can be computed from the association constant (K_a) as follows:

$$\Delta G = -RTlnK_a \qquad \text{(Eq. 2.4.4)}$$

where, R is the gas constant, T is temperature, and K_a is association constant.

The Gibbs free energy can also be computed from enthalpic and entropic contributions of a thermodynamic system (Eq. 2.4.5).

$$\Delta G = \Delta H - T\Delta S \qquad \text{(Eq. 2.4.5)}$$

where, ΔH is change in enthalpy, ΔS is change in entropy, and T is temperature in Kelvin.
Free energy is a state function, i.e. defined by initial and final thermodynamic states, regardless of the pathway connecting these two states.

All the key driving forces in a binding reaction are compensatied by enthalpy and entropy. Enthalpy (H) is a measure of total energy of a thermodynamic system. A binding reaction has a negative value for the change in enthalphy (ΔH) for an exothermic process (forming energetically favorable interactions) and a positive value for an endothermic process (breaking of energetically favourable interactions).
The second law of thermodynamics determines that the heat always flows from higher temperature to lower temperature. This creates the disorder in the thermodynamic system and this disorder can be accounted with entropy. Entropy (S) is the measure of disorder or randomness in atoms and molecules in a system. The change in entropy (ΔS) can be positive for increase in degree of freedom and negative for the decrease in degree of freedom of the system.

2.4.3 Experimental methods to Study Binding Affinity

There are well established protocols/methods to study and investigate the protein-RNA binding [Ramanathan et al., 2019, Cozzolino et al., 2021]. These methods can help to investigate the contacts between protein and nucleic acid but can not specify the magnitude of free energy contribution made by the interaction.

Isothermal Titration Calorimetry (ITC): It is one of the physical technique that directly measures the heat released or absorbed all along a biomolecular process. This analytical method works on the basic principle of thermodynamics where contact between two molecules results in either heat generation or absorption, depending on the type of binding, that is, exothermic or endothermic. The measurements of the change in heat are used to determine the binding constants, enthalpy, entropy and the reaction stoichiometry [Srivastava and Yadav,

2019, Ladbury and Chowdhry, 1996]. ITC is the only approach to measure directly the heat exchange during complex formation at a constant temperature. ITC generally needs a large amount of sample, which limits its application to certain bio-macromolecules that are difficult to prepare in large quantities.

Surface Plasmon Resonance (SPR): This is an optical technique based on a microfluidic surface for detecting the molecular interaction of two molecules. SPR provides a robust platform for screening and compound optimization. A mobile molecule, called analyte, binds to an immobilized ligand on a thin gold film. Such interaction changes the refractive index of the film at the interface of the liquid sample and a surface with an immobilized sensor molecule. The corresponding kinetic parameters of the binding (i.e. the association rate constant (k_{on}) and the dissociation rate constant (k_{off})) and the affinity (K_d), and the thermodynamic parameters can be characterized. The most extensively used instruments to measure the binding kinetic are SPR biosensors with a microfluidic flow system and dextran surfaces. A simple fitting of the equation corresponding to a suitable model to the sensogram allows to determine the k_{on} and k_{off}, and K_d as the ratio of k_{off} to k_{on} [Patching, 2014].

Fluorescence Polarization (FP): It is a fluorescence-based technique used to detect the binding of 'fluorescent ligand' or 'fluorescence-labelled ligand' to a protein. This technique uses polarized light to excite the fluorophore, which then emits light that is detected by polarimeter. The polarization of the emitted light is dependent on the mobility of the labelled ligand, that affects the binding to protein. FP technique makes use of single fluorescent label strategy and does not involve the filtration or separation steps and, as thus, requires relatively fewer reagents, smaller amounts of protein, and relatively less expensive equipment than do SPR and ITC. The major advantage of FP over other methods is that it does not require the separation of bound and free ligand. FP is suitable for measurements of low-affinity interactions with higher (fast) dissociation rates [Moerke, 2009, Rossi and Taylor, 2011].

Electrophoretic mobility shift assay (EMSA): This method is widely used to detect protein complexes with DNA or RNA. In a classical assay, solutions of protein and nucleic acid are combined and the resulting mixtures are subjected to electrophoresis under native conditions through polyacrylamide or agarose gel. After electrophoresis, the distribution of species containing nucleic acid is determined, usually by autoradiography of 32P-labeled nucleic acid. In general, the electrophoretic mobility of protein-nucleic acid complexes is less than the free nucleic acid. This method is simple to perform and yet robust enough to accommodate a wide range of binding conditions [Hellman and Fried, 2007].

2.5 Modelling 3D structures of protein and protein-RNA complex

UniProtKB release 2023_01 has 246,440,937 protein sequences, while there are only 202,292 protein structures available in the Protein Data Bank (PDB) as of 14-March-2023, corresponding to 61,463 protein sequences. Three-dimensional

structure determination using experimental methods like X-ray crystallography and NMR is a quite time-consuming and complex procedure. Thus, many scientists focused on developing computational methods for modelling 3D structures from sequences to bridge this gap between protein sequences and protein structures.

2.5.1 Protein 3-D Structure prediction

All the computational methods for protein 3-D structure prediction can be grouped in two categories: template-based modeling and template-free modeling [Kuhlman and Bradley, 2019].

Template-based modeling

Template-based modeling uses previously determined 3-D structures of a homologous protein to model the unknown structure of the target protein. Template-based modeling is also called homology modeling or comparative modeling as it uses the homologous structure as a template to model the target protein. The basic principle of homology modeling is that proteins with similar amino acid sequences are likely to have similar structures. This is because the sequence of a protein determines its three-dimensional structure through the interactions between the side chains of the amino acids. Therefore, if a protein of interest has a high degree of sequence similarity to a protein with a known structure, it is likely that the two proteins will have similar structures. The homology modeling is usually divided into four steps:

1. Identify the template protein;
2. Alignment of the target-template sequences;
3. Model building and refinement;
4. Model evaluation.

There are several programs and web-servers available for each of the above-mentioned step of homology modeling.
The accuracy of the predicted structure can vary depending on the degree of similarity between the target and template proteins, the quality of the alignment, and model building steps. Furthermore, the accuracy of the predicted structure can be affected by the presence of point mutations in target protein.

Template-free modeling

Template-free modeling does not rely on the global similarity to a protein structure (previously determined) and can be applied to target proteins with novel folds. This approach of protein 3-D structure modeling uses large-scale conformational sampling and the application of physics-based energy functions. Fragment assembly is one of the most common approach used to tackle the issue of conformational sampling in which models are built from short and continuous backbone fragments retrieved from known protein structures. Template-free modeling is required when the target protein do not have any homologous protein with known 3-D structure.

When an homologous structure is available, template-based modeling provides more accurate models when compared to template-free prediction [Vallat et al., 2015].
The distinction between these two approaches has begun to blur, as template-based methods have incorporated energy-guided model refinement, and template-free methods have employed machine learning and fragment-based sampling approaches to exploit the information in the structural database [Hardin et al., 2002, Kuhlman and Bradley, 2019].

AlphaFold2

AlphaFold2 is a deep learning model that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, to predict the 3-D structure of protein [Jumper et al., 2021]. AlphaFold2 uses a multiple sequence alignment (MSA) as input and uses information about conservation and co-evolution of protein sequences from MSA. The key part of the workflow of AlphaFold2 is the *Evoformer*. The *Evoformer* consists of two transformer blocks capable of exchanging information to efficiently extract structural information from the MSA. The *Evoformer* passes the information to structure module that builds a 3D representation of the protein structure. AlphaFold2 also has a *Recycling* stage where it goes back and refines the prediction using the resulting 3D structure.
The 3D structure of multi-chain protein complexes can be modelled using AlphaFold-Multimer [Evans et al., 2021]. The success of AlphaFold2 led to development of AlphaFold DB providing open access to over 200 million protein structure predictions to help accelerate scientific research [Varadi et al., 2022].
The AlphaFold DB can be accessed at https://alphafold.ebi.ac.uk/.

RoseTTAFold

RoseTTAFold is another deep learning model that uses three-track neural network to obtain accuracy comparable to that of AlphaFold2. These three layers include 1D sequence, 2D distance map between residues and 3D atom coordinates of the protein structure. In this architecture, information from all three layers flows back and forth, allowing the network to collectively reason about the relationship between a protein's chemical parts and its folded structure.
In addition to the 3D structure of a single protein, RoseTTAFold has demonstrated the capacity to generate accurate 3D structure models for protein-protein complexes from sequence information [Baek et al., 2021].

2.5.2 Critical Assessment of protein Structure Prediction (CASP)

The Critical Assessment of protein Structure Prediction (CASP) initiative is aiming to help advance the methods of protein structure prediction from sequence. This experiment takes place every two year since 1994. CASP has shown continuous improvement in accuracy of protein 3-D structure prediction. CASP assessors have been using the measures from Global Distance Test (GDT) to quantify prediction performance since CASP3 in 1998 [Li et al., 2016]. GDT is a

structure similarity assessment test introduced by local global alignment (LGA) program [Zemla, 2003]. GDT provides two main scores/measures: 'Global Distance Test Total Score' (GDT_TS) and 'Global Distance Test High Accuracy' (GDT_HA). Both of these measures can be computed using Eq. 2.5.1 and Eq. 2.5.2, respectively[3].

$$GDT_TS = (GDT_P1 + GDT_P2 + GDT_P4 + GDT_P8)/4 \qquad \text{(Eq. 2.5.1)}$$

where, GDT_Pn denotes percentage of residues under distance cutoff <= nÅ.

$$GDT_HA = (GDT_P0.5 + GDT_P1 + GDT_P2 + GDT_P4)/4 \qquad \text{(Eq. 2.5.2)}$$

where, GDT_Pn denotes percentage of residues under distance cutoff <= nÅ.

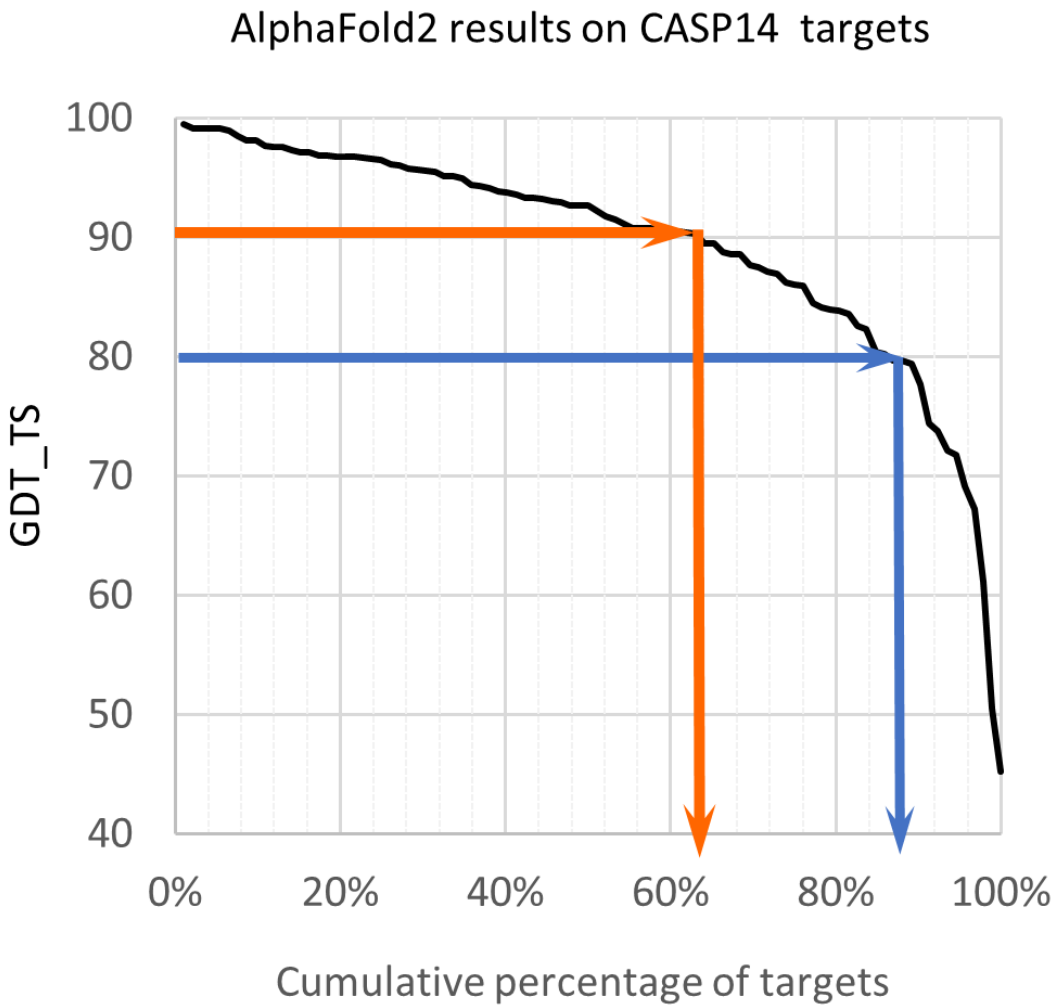


Figure 2.5: AlphaFold2 results on CASP14 targets (Black); this method modelled ∼2/3 of the targets with GDT_TS >90 (Orange), and ∼90% of targets with GDT_TS >80 (Blue). Figure taken from CASP14 (predictioncenter.org).

CASP14 marked an extraordinary increase in the accuracy of the computed three-dimensional protein structures with the emergence of the advanced deep

[3]https://predictioncenter.org/casp14/doc/help.html

learning method AlphaFold2. The accuracy of CASP14 models for template-based modeling targets (Figure 2.5) significantly superseeded accuracy of models that can be built by simple transcription of information from templates, and reached the level of GDT_TS=92 on average, which is significantly higher than the corresponding averages in previous two CASPs.

2.5.3 Modeling Protein-RNA complexes

The protein-RNA complexes play critical roles in various biological processes, such as gene expression and regulation. Understanding the 3D structures of these complexes is important to learn more about their interactions and functional mechanisms. 3D structure determination of protein-RNA complexes can be challenging and time consuming. Thus, computational methods, such as molecular docking, have become important tools for modeling the 3D structures of protein-RNA complexes.
The docking methods aim at predicting three-dimensional structures of macromolecular complexes, using the structure (atomic coordinates) of the individual molecules. Docking is performed when it is assumed that there is an interaction between two molecules. The larger molecule is usually referred to as the receptor, whereas the smaller one is called ligand.
The docking method consists of two basic steps. The first is sampling, i.e. search of the conformational space for possible relative orientations and conformations of the components resulting in sampled models (poses). The second step is scoring, i.e. assessment of the models from first step by a scoring function to distinguish correct (near native) poses from incorrect (non-native) ones [Vajda et al., 2013].

Sampling methods: The main goal of the sampling method is to have a sufficiently exahaustive sampling to explore conformations with local energy minima including the global minimum. There are several sampling methods used by different docking programs. Some of the widely used sampling methods are fast fourier transform used by Hex, monte carlo method used by Rosetta, and gradient descent used by HADDOCK and ATTRACT.

Scoring methods: The goal of the scoring function is to distinguish near native structures from non-native ones. Scoring functions involve approximating, rather than calculating, the binding affinity between the docked molecules [Meng et al., 2011]. Scoring functions can be divided in three groups: force-field based, empirical, and knowledge-based scoring functions [Kitchen et al., 2004].

Flexibility of macromolecules: Earlier, both the receptor and the ligand were considered as rigid while docking to reduce the computational cost. This approach is called rigid docking. Rigid docking might provide good results when experimentally determined 3D structures of the individual molecules (or their close homologs) are close to the bound form. If the available structures are very different from the bound form, the rigid docking will not work very well.
In such cases, the flexibility of these molecules can be considered in order to overcome such issues and achieve better results. Flexibility can be taken into

consideration for each partner in different ways. One popular approach is fragment-based docking. Fragment-based docking consists in chopping the ligand into fragments and performing multi-conformation docking of each fragment. Then the poses of fragments with compatible positions are assembled into a complete structure. Fragment-based docking permits the docking of a ligand onto a receptor without a known structure for the ligand. This method allows to model the local flexibility at the fragment level (multi-conformation docking) and global flexibility by assembling the fragments [de Beauchene et al., 2016].

The coarse-grained representation is a way to simplify the representation of macromolecules by grouping several atoms together using a 'pseudo atom'. This representation results in a relatively smooth surface representation containing fewer docking energy minima on the partners and allows for rapid and fully converged energy minimization compared to an atomic resolution representation. Another advantage of the coarse-grained representation is a shorter computation time as there are less atoms and less interactions to compute. But, one has to create and calibrate a new force field with the pseudo atoms. Calibrating a force field with pseudo atoms can be a tedious task. However, there are some coarse-grained representations with force fields used by the community like Martini [Periole and Marrink, 2013] and ATTRACT [Setny and Zacharias, 2011].

Data-driven docking: The docking can be improved by using experimental data to create constraints. There are many types of information that can be used in the docking including the overall shape of the complex, interface residues, and residue-residue contacts between partners. The study from de Beauchene et al. [2016] showed the use of data-driven docking in the form of anchoring contacts for ssRNA-RRM complexes to achieve better results compared to the state-of-the-art methods.

There are several docking programs available including ATTRACT, AutoDock [Forli et al., 2016], HADDOCK [Van Zundert et al., 2016], ZDOCK [Pierce et al., 2014], and Hex [Ghoorah et al., 2013]. Each of these programs support docking of specific biomolecules. Hex and ZDOCK are primarily dedicated to protein-protein docking, AutoDock is widely used for virtual screening of small molecules in addition to protein-peptide docking [Zhang and Sanner, 2019]. ATTRACT and HADDOCK can perform protein-protein and protein-nucleic acid docking.
In this thesis, we are interested in modelling 3D structures for protein-ssRNA complexes. The ssRNA-protein docking part of the ATTRACT program (we call it ssRNA'TTRACT) is being developed in our team. Thus, I will be focusing only on ATTRACT docking program.
For the scope of this thesis, we considered only the binding of ssRNA to RRM-containing proteins and therefore only 'RNA' term will be used instead of 'ssRNA', unless specified otherwise.

ATTRACT docking program: The ATTRACT docking program can be used for docking of proteins to protein [Zacharias, 2003], DNA [Piotr et al., 2012], RNA [Setny and Zacharias, 2011], and small ligands [May and Zacharias, 2005]. For RNA-protein docking, ATTRACT uses its coarse-grained representation and

the fragment-based approach proposed by de Beauchene et al. [2016]. ATTRACT uses the fragment library of tri-nucleotides created using ProtNAff [Moniot et al., 2022b]. The fragments of tri-nuclotides corresponding to the target RNA sequence are docked onto the protein (minimized from random positions according to the ATTRACT score using gradient descent). The redundant poses are eliminated and the poses with best score are kept for the assembly. Two poses are linked together if the RMSD between their two overlapping nucleotides is below the defined threshold. Finally, the scoring function is used to score all these assembled chains for the purpose of distinguishing between near native and non-native chains.

The modelled 3D structures of protein or protein-RNA complexes can be refined using a short energy minimization step or/and with molecular dynamics simulation.

2.6 Assessment of modelled 3D structures

2.6.1 Overview of Molecular Dynamics

Molecular dynamics (MD) simulations can be used to calculate the thermodynamic and energetic properties that will help the understanding of the conformational changes of molecule. There are several programs developed specifically to simulate the behavior of biomolecules. Some of the commonly used simulation programs are AMBER, CHARMM, GROMACS, and NAMD.

In MD, the forces on each atom are calculated as derivatives of potentials and substituted into Newton's equations of motion:

$$F_i = m_i a_i = m_i \frac{d^2 x_i}{dt^2} \qquad \text{(Eq. 2.6.1)}$$

where, m_i is the mass and a_i is the acceleration of atom i with the position of x_i along a single dimension during the time interval of t.

Force Fields are sets of potential functions and parameterized interactions that can be used to study physical systems. There are two main components of a force field: the set of potential energy functions and the set of parameters used in these functions.
In the simplest form, the potential energy of a molecule can be written as:

$$E = \sum_{bonds} E_{stretch} + \sum_{angles} E_{bend} + \sum_{dihedrals} E_{torsion} + \sum_{pairs} E_{nonbonded} \qquad \text{(Eq. 2.6.2)}$$

where $E_{stretch}$, E_{bend}, $E_{torsion}$, $E_{nonbonded}$ are the energy contributions from bond stretching, angle bending, torsional motion (rotation) around single bonds, and interactions between nonbonded atoms, respectively. The sums are over all the bonds, all the angles defined by any set of three bonded atoms A–B–C, all the dihedral angles defined by any set of four sequentially bonded atoms A–B–C–D, and all pairs of significant nonbonded interactions [Lewars, 2011].

All these energy functions can be grouped into two categories: bonded and non-bonded potential terms. The non-bonded terms are composed of electrostatic and non-electrostatic interactions between all pairs of atoms.

Coulomb's law is used to describe the electrostatic potential in MD. The point charges are assigned to the positions of atomic nuclei and atomic charges are derived using quantum mechanical (QM) methods with the goal to approximate the electrostatic potential around a molecule.

$$U_{elec} = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} \qquad \text{(Eq. 2.6.3)}$$

where, q_i and q_j are the charges on the atoms i and j, ϵ_0 is the permittivity of vacuum, ϵ_r is the relative permittivity, and r_{ij} is the distance between the pair of atoms.

Lennard-Jones potential is used to approximate the potential energy of non-electrostatic interaction between a pair of non-bonded atoms

$$U_{LJ} = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \qquad \text{(Eq. 2.6.4)}$$

where, ϵ is the well depth (a measure of how strongly two atoms attract each other), σ is the van der Waals radius (the distance at which the intermolecular potential between two atoms is zero), and r is the distance of separation between two atoms.

Bonded potential terms describe several types of interactions including bond stretching, angle bending, and torsion terms between the atoms within molecules. The bond potential is used to model the covalent interactions in a molecule. Bond stretch is approximated by a simple harmonic function describing oscillation about an equilibrium bond length r_0 with bond constant k_b.

$$U_{bond} = k_b\left(r_{ij} - r_0\right)^2 \qquad \text{(Eq. 2.6.5)}$$

The angle potential is defined for every triplet of bonded atoms (A–B–C) and describes the bond bending energy. It is approximated by a harmonic function describing oscillation about an equilibrium angle θ_0 with force constant k_θ:

$$U_{angle} = k_\theta\left(\theta_{ijk} - \theta_0\right)^2 \qquad \text{(Eq. 2.6.6)}$$

The torsion angle potentials are defined for every 4 sequentially bonded atoms. It describes the angular spring between the planes formed by the first three (A–B–C) and last three (B–C–D) atoms of a consecutively bonded atoms.

$$U_{Dihed} = k_\phi(1 + cos(n\phi - \delta)) \qquad \text{(Eq. 2.6.7)}$$

where, k_ϕ is the force constant, n defines periodicity, ϕ is the torsion angle, and δ is the phase shift angle.
The force constants for angle potential are about 5 times smaller than for bond stretching [4].

[4]Energy scale for the potential terms can be found at: https://computecanada.github.io/molmodsim-md-theory-lesson-novice/01-Force_Fields_and_Interactions/index.html

Periodic boundary conditions are a set of boundary conditions that makes possible to approximate a large (infinite) system using a unit cell. This unit cell in MD is called periodic box or simulation box. When one particle leaves the periodic box from one side, it reappears on the opposite side of the periodic box. Each simulation package provides several shapes of periodic boxes to choose from, like cubic box, truncated octahedron, rhombic dodecahedron. Some are more efficient than others because of their smaller volume including less extra (irrelevant) solvent [Wassenaar and Mark, 2006].

Molecular dynamic simulations are performed as close to experimental conditions as possible by controlling the factors like pressure, temperature. An ensemble describing a system consists of a large number of copies of a system representing a set of possible states that a real system might be in.
A thermodynamic ensemble is a statistical ensemble that is in statistical equilibrium. It provides a way to derive the properties of a real thermodynamic system from the laws of classical and quantum mechanics. Usually one of the following thermodynamic ensembles are used in MD simulations:

NVE or microcanonical ensemble: number of particles (N), volume (V), and the energy (E) of the system are kept constant (conserved).

NVT or canonical ensemble: number of particles (N), volume (V), and the temperature (T) of the system are kept constant (conserved).

NPT or isothermal-isobaric ensemble: number of particles (N), pressure (P), and the temperature (T) of the system are kept constant (conserved).

Different algorithms are implemented by each simulation package that provide options to control the parameters for these thermodynamic ensembles. For example, to control pressure AMBER uses Monte-Carlo and Berendsen barostats while NAMD uses Langevin barostat along with Berendsen [5]. Barostat helps to regulate the pressure by adjusting the volume of the system.

A typical MD run consists of the following steps:

System setup : The system is prepared for the simulation. This includes getting initial structure, creating a topology and coordinate file, setting the periodic box, adding water and ions, if necessary.

Energy minimization : The system is relaxed to remove any steric clashes or unusual geometry.

Equilibration : In the initial segment of the simulation, the energy of the system changes rapidly, but it eventually settles into a reasonable approximation of an oscillation around a mean. At this point the simulation has achieved equilibration.

Production run : Properties are calculated from the averages over the ensemble of structures generated during this step.

Analysis : This step is the analysis of the output from simulation.

[5]More details here https://computecanada.github.io/molmodsim-md-theory-lesson-novice/08-barostats/index.html

2.6.2 Free energy computation

Free energy is one of the most important quantity in the thermodynamic system. The computation of free energies is one of the key objective of MD. There are several methods to compute the free energy difference like MM-PBSA/MM-GBSA, umbrella sampling, thermodynamic integration, or free energy perturbation.

MM-PBSA (molecular mechanics Poisson-Boltzmann surface area) is a method to estimate the free energy of a system. The average free energy ($\bar{G}$), of a system is computed using Eq. 2.6.8, after removing any solvent and ion molecules from the system.

$$\bar{G} = \bar{E}_{bond} + \bar{E}_{angle} + \bar{E}_{torsion} + \bar{E}_{vdw} + \bar{E}_{elec} + \bar{G}_{pol} + \bar{G}_{np} - TS \qquad \text{(Eq. 2.6.8)}$$

Where, the first five energy terms correspond to the bond, angle, torsion, van der waals and electrostatic terms in the molecular mechanical force field, evaluated with no nonbonded cutoff. These terms together result in the average molecular mechanical energy. $\bar{G}_{pol}$ and $\bar{G}_{np}$ are polar and non-polar contributions of the solvation free energies, and the last term (TS) is the solute entropy, i.e, absolute temperature, T, multiplied by the entropy, S, estimated by a normal mode analysis. $\bar{G}_{pol}$ is typically obtained by solving the Poisson-Boltzmann (PB) equation or by using the generalized Born (GB) model (**MM-GBSA** approach).

Both of these methods, MM-PBSA and MM-GBSA, can be used to estimate binding free energies for a ligand (RNA) in complex with a protein using Eq. 2.6.9.

$$\Delta G = \bar{G}_{complex} - \bar{G}_{protein} - \bar{G}_{ligand} \qquad \text{(Eq. 2.6.9)}$$

There are two ways to compute binding free energies using Eq. 2.6.9:

1. run separate simulations of complex, protein, and ligand, resulting in three different trajectories.
2. run a single simulation of complex, and use just the snapshots from the resulting trajectory on complex.

The option 2 assumes that the snapshots of protein and ligand taken from the trajectory of complex are of comparable free energy to those that would emerge from separate trajectories of protein and ligand. It is also much faster as only a single simulation of the complex is required for computing binding free energies [Kollman et al., 2000].
The MM-PBSA method was originally developed for the AMBER software [Miller III et al., 2012], later on the scripts for this method were also presented for other simulation softwares like GROMACS and NAMD [Kumari et al., 2014].

Umbrella Sampling is a method that uses bias potentials along a reaction coordinate (ξ) to drive a system from one thermodynamic state to another. This change in thermodynamic state is carried out via a series of windows by using bias potentials like harmonic potentials. A bias function ($U_i(\xi)$) is applied to keep the system close to the reference point $\xi_{0,i}$ in each window i.

$$U_i(\xi) = k_i(\xi - \xi_{0,i})^2 \qquad \text{(Eq. 2.6.10)}$$

where, $\xi_{0,i}$ is the reference position of harmonic restraint defining the center of window i, and k_i is the force constant that defines the width of the corresponding window.
These sampled distribution along the reaction coordinate can be used to calculate the free energy change in each window. This series of windows are then combined by post-processing methods like weighted histogram analysis method (WHAM) [Souaille and Roux, 2001] or umbrella integration [Kästner and Thiel, 2005] to get the global free energy profile.
You et al. [2019] used umbrella sampling to investigate the free energy changes during ligand unbinding from protein and examined the effect of data point selections along the reaction coordinate in umbrella sampling.

Thermodynamic Integration (TI) is a method used to compute free energy difference between two given states (λ_0 and λ_1). In TI, the free energy difference is calculated by defining a thermodynamic path between the states and integrating over the ensemble-averaged enthalpy changes along the path.

$$\Delta G_{TI} = \int_0^1 \left\langle \frac{\partial U(x)}{\partial x} \right\rangle_\lambda d\lambda \qquad \text{(Eq. 2.6.11)}$$

where x is the full set of configurational coordinates, U is potential energy, $\langle ... \rangle_\lambda$ denotes the ensemble average at a particular value of λ. The limiting factor for TI is that only a finite number of λ values can be simulated, and thus the integral must be approximated by sum [Ytreberg et al., 2006].

Free Energy Perturbation (FEP) is often referred as computational alchemy as it is used in protein design to compute the free energy changes due to "alchemical" perturbations introduced during MD. In the FEP approach, λ intermediate states wherein atoms that need to appear, disappear, or mutate between the two molecules are represented by a linear combination of end-state Hamiltonians.
Hamiltonians are complex mathematical expressions representing the total energy of a system (kinetic and potential energy) as a function of is momentum (p = mv) and the space coordinates (q). Usual notation is H(p,q)[6].

In a FEP, independent equilibrium simulations at each λ value (intermediate states) are performed and exponential averaging is used to determine the free energy difference between two neighboring λ states, then these differences are summed to obtain the total free energy between two end states.
The free energy difference between two systems (A and B), represented by Hamiltonia $\mathscr{H}_A$ and $\mathscr{H}_B$ can be expressed as:

$$\Delta G_{FEP} = G_B - G_A = -RT \ln \left\langle e^{-\Delta \mathscr{H}/RT} \right\rangle_A \qquad \text{(Eq. 2.6.12)}$$

where R is gas constant, T is temperature, $\Delta\mathscr{H} = \mathscr{H}_B - \mathscr{H}_A$, and $\langle \rangle_A$ refers to an ensemble average over a system represented by Hamiltonian $\mathscr{H}_A$. This equation is the fundamental equation for free energy perturbation calculations. This equation (Eq. 2.6.12) will not lead to sensible free energy, if the two systems A and B differ in more than a trivial way. To overcome this problem the simulation is performed

[6]https://en.wikipedia.org/wiki/Hamiltonian_mechanics

in small steps where the pathway for the hybrid molecule to go from one state to other is via the coupling parameter λ that allows smooth mixing of the two states. The λ is changed from 0 (state A) to 1 (state B) in a number of small steps called as FEP windows [Kollman, 1993]. Thus, the Eq. 2.6.12 can be generalized to:

$$\Delta G_{FEP} = -RT \sum_{\lambda=0}^{1} \ln \left\langle e^{-\Delta \mathscr{H}'/RT} \right\rangle_{\lambda} \qquad \text{(Eq. 2.6.13)}$$

where $\Delta \mathscr{H}' = \mathscr{H}_{\lambda+d\lambda} - \mathscr{H}_{\lambda}$ and $\mathscr{H}_{\lambda} = \lambda \mathscr{H}_B + (1-\lambda)\mathscr{H}_A$.

Free energy is a state function and its value is determined by the end state of the system and not by the path of transformation. Thus, if one wants to compute the difference of free energy between a receptor (protein) in complex with one or the other ligand (RNA), the following thermodynamic cycle can be used (Figure 2.6).

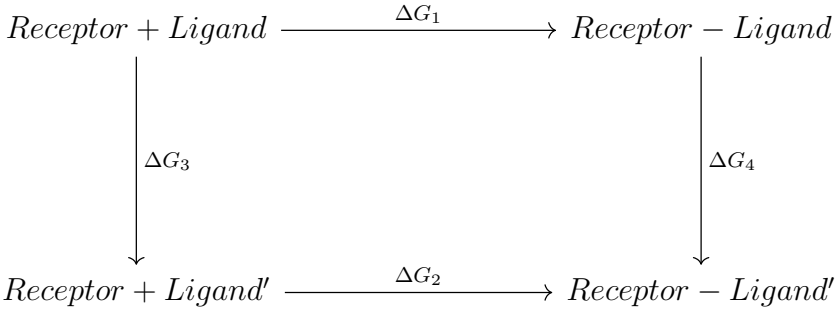


Figure 2.6: A thermodynamic cycle describing the binding of two ligands, Ligand and Ligand′, to a Receptor.

Using the above cycle (from Figure 2.6), the various free energy differences are related to each other by the Eq. 2.6.14, suggesting that only two transformations are required to get the $\Delta\Delta G$.

$$\Delta\Delta G = \Delta G_1 - \Delta G_2 = \Delta G_3 - \Delta G_4 \qquad \text{(Eq. 2.6.14)}$$

2.6.3 Example studies

All the computational approaches require agreements with experimental data up to a certain extent to be considered as reliable and accurate. In most cases (if not all), the accuracy of computational method is assessed by the experimental information. The experimental NMR parameters are generally used to assess the quality of new force fields [Huang and MacKerell Jr, 2013]. NMR structural data can also be used to create biasing potentials to improve the simulation stability [Li and Brüschweiler, 2014].

There are several experimental and computational studies performed with RRMs and RNAs to check their binding.
Basu et al. [2021] used molecular docking and molecular dynamic simulation to study the FUS RRM domain and a 12-nucleotides long RNA with 'GGUG' motif. This RRM domain lacks some of the conserved residues from RNP1 motif and has

an unusual lysine-rich loop (KK-loop) between $\alpha 1$ and $\beta 2$. They mutated the three lysine residues from this loop to alanine (K312A, K315A, and K316A from PDB entry 2LCW) and found that several native contacts from the KK-loop are lost in the mutant. In addition, they have also analyzed binding affinity using umbrella sampling method to conclude that the KK-loop is important for the stability of RRM-RNA complex.

Nolan et al. [1999] investigated the role of a conserved aromatic amino acid (F56 from RNP1 motif) forming stacking interactions using CD spectroscopy and gel mobility shift assays [Carey, 1991] in U1A-RNA complex (PDB ID: 1URN). Based on mutant studies, they found that the F56 residue contributes significantly to the RNA binding and the mutants were observed to bind with lower binding affinity. Based on these experimental calculations, Kormos et al. [2007] performed free energy computations using MM-GBSA protocol for wild type U1A-RNA complex and F56 mutant U1A-RNA complexes. The relative binding free energies obtained in this study follow the trend of the experimental binding free energies.

MD simulations and experimental studies can be used to complement each other. Krepl et al. [2016] demonstrated the MD simulation as a viable tool to complement the NMR studies using RRM-RNA complexes. They used Fox-1 RRM and SRSF1 RRM2 with their RNA targets. The simulations predicted an unanticipated role of Arg142 at protein-RNA interface of SRSF1 RRM2-RNA complex. This was then confirmed by NMR and ITC experiments by mutating Arg142 to Alanine. They also showed that the use of experimental NMR NOEs-based restraints in the early stages of the simulations (first 120 ns in this case) leads to more stable simulations.

Gapsys and de Groot [2017] developed a protocol for alchemical free energy calculations with nucleotide mutations in protein-DNA complexes. The results from their protocol are in agreement with the experimentally determined binding profiles. The program to generate the input files required for MD simulation with alchemical transformation can be found at http://pmx.mpibpc.mpg.de/dna_webserver.html.
Currently, there is no supporting force fields for the mutations in RNA. Developers of the 'pmx' package are working on it[7].

This chapter covered the important concepts to understand the contributions of this thesis, described in next chapters, but considering the diversity of methods, it was not possible to cover all the methods in detail. Whenever necessary I will give additional remarks on the methods I have used in the next chapters.

[7]https://bioexcel.eu/research/projects/alchemical-free-energy-calculations-in-biomolecules/

Chapter 3

InteR3M: The Database for Interactions of RNA and RRM

Summary

3.1 Introduction **37**

3.2 Scope & Requirements **37**

3.2.1 Mission Statement 37

3.2.2 Mission Objectives 37

3.2.3 Delineation of a correct set of RRM families 38

3.2.4 Use cases 41

3.3 Database Design **44**

3.3.1 Conceptual Database Design 44

3.3.2 Logical Database Design 45

3.3.3 Physical Database Design 45

3.4 Implementation **45**

3.4.1 Data Collection 47

3.4.2 Database Implementation 49

3.4.3 Implementation of User Interface 50

3.4.4 Testing 50

3.5 Using InteR3M **51**

3.5.1 Search functionality 51

3.5.2 RRM instance display 51

3.5.3 Protein display 51

3.5.4 Ligand instance display 51

3.5.5 Experiment display 53

3.5.6 List-of-contacts display 53

3.5.7 Multicriteria Search 53

3.6 Strategies to update InteR3M database **54**

3.7 Results **55**

3.8 Discussion & Conclusion **62**

3.1 Introduction

This chapter describes the design, implementation and data collection for the relational database of Interactions of RNA and RNA Recognition Motif (InteR3M). In Chapter 2, we have seen the importance of RRMs and how studying and characterizing them will get us a step closer in successfully designing the novel RRMs with target activity. We also learnt that another specific database, RRMdb [Nowacka et al., 2019], was developed to assess evolutionary relationships among RRM domain instances and the classification within the RRMdb was based on sequence similarity only.

Compared to this, the InteR3M database is developed by carefully inspecting experimental 3D structures. InteR3M database aims to provide sequence, structure, and the binding information available for each RRM domain instance. By collecting and organizing this information in InteR3M database, we aim to uncover the RNA binding code of RRM domains.

3.2 Scope & Requirements

We started with a very basic goal of collecting all the available information about RNA Recognition Motif (RRM) domain and integrate all the collected data to develop a comprehensive database of available RRM information.

3.2.1 Mission Statement

> The purpose of this database (InteR3M) is to organize, make available and maintain the available information about RNA Recognition Motif (RRM) domains and their interactions with nucleic acid partners. This includes their sequence, structure, biological function, and binding affinity.

The data collected and stored in the InteR3M database will help us and the interested scientific community to characterize the RNA binding code of RRM domain and to improve the modelling of 3D structures of RRM-RNA complexes.

3.2.2 Mission Objectives

The mission objectives consists of the general tasks related to the data maintained in the database and will help us in the database design procedure. We have compiled a set of objectives to achieve the goal of developing and maintaining the data in InteR3M database.

1. Keep track of new Pfam families from RRM clan.
2. Keep track of the proteins containing at least one or more RRM domains.
3. Provide access and maintain the list of all RRM domain instances.
4. Provide access and maintain the list of PDB entries having a structural instance of at least one RRM domain.

5. Provide access and maintain the list of experiments performed with RRM domain instances.
6. Provide access and maintain the list of molecular interactions between RRMs and nucleic acids.

3.2.3 Delineation of a correct set of RRM families

There are a number of generalist domain databases like Pfam, SCOP, CATH, InterPro, CDD and so on. All of these domain databases use different rationale for domain classification which can be either primarily sequence-based, primarily structure-based or integration-based (see Chapter 2 section 2.2).

We need to collect information from sequence-based domain databases to achieve the mission objectives 2 and 3. In this way, we will have information about the RRM domain instances even if they do not have experimentally solved 3D structures.

Pfam was selected from the generalist domain databases as it is well connected with other protein databases and widely used in the field. Pfam has been searched for the RRM families via keyword search. At the time of creation of the prototype for InteR3M database, the RRM-centric RRMdb database was online [Nowacka et al., 2019]. Thus, we also searched RRMdb as it was classifying RRM domain instances into 415 different families based on sequences and each instance of the family can be traced back to Pfam. Table 3.1 contains the list of the 42 Pfam families retrieved from Pfam and RRMdb.
The RRM_1 (PF00076) family has the highest number of domain instances and structures (experimentally determined) among all these 42 filtered Pfam families. RRM_1 family has 1,069 domain structures and no other Pfam family has more than 52 structures (Table 3.1). This shows the dominance of RRM_1 Pfam family over all these families from structural point of view.

Structural Inspection of RRM families

It is important to verify that the collected families from Pfam indeed contain RRM domain instances as defined by the typical RRM fold (See Chapter 2, section 2.3.1). One way to validate these families is the manual inspection of 3D structures from these families. Of the collected families, 16 families have no experimentally determined 3D structure. Thus, the 26 remaining families with 3D structures were processed through structural inspection to determine if the structures corresponding to these families contain the RRM fold (Table 3.2).

Many RRM domain instances from RRM_1 family are widely studied and well characterized from a structural point of view [Maris et al., 2005]. Thus, we selected two well-known RRM structural instances from the predominant PF00076 family as references, namely 1A9N_B (spliceosomal U2B″-U2A′ protein complex) and 2A3J_A (redesigned human U1A protein).

Based on the availability of structures we selected a few (1-3 unique) structures from each family for the manual structural inspection of these families. For each of the selected structure (query structure) from these families, we first checked how the structure-based domain databases, SCOP and CATH, classify these structures. This

Table 3.1: Pfam families retrieved from Pfam (v33.0) and RRMdb. One Protein (UniProt ID) and PDB entry may have multiple domain and structural instances, respectively.

Family Name	**Family Identifier**	**Number of Domain Instances**	**Number of Structural Instances**	**Clan Name**
Baculo_FP*	PF03258	224	0	None
BRAP2	PF07576	1,249	0	RRM
Calcipressin	PF04847	1,574	1	RRM
DbpA	PF03880	5,850	3	RRM
DUF1743	PF08489	334	10	RRM
DUF1866	PF08952	778	2	RRM
DUF4283*	PF14111	6,556	0	None
DUF4523	PF15023	94	0	RRM
GlcNAc-1_reg	PF18440	242	1	RRM
GUCT	PF08152	992	2	RRM
Limkain-b1†	PF11608	333	1	RRM
Nab6_mRNP_bdg*	PF10567	30	0	None
NCBP3*	PF10309	912	0	None
Nup35_RRM	PF05172	888	10	RRM
Nup35_RRM_2	PF14605	241	0	RRM
Peptidase_C48*	PF02902	14,660	51	Peptidase_CA
PHM7_cyt	PF14703	6,972	12	RRM
PNMA*	PF14893	1,063	0	GAG-polyprotein
PRE_C2HC*	PF07530	314	0	None
RL	PF17797	288	14	RRM
RNA_bind	PF08675	343	5	RRM
RRM	PF10378	408	0	None
RRM_1	PF00076	246,502	1,069	RRM
RRM_2	PF04059	2,110	0	RRM
RRM_3	PF08777	1,110	4	RRM
RRM_4	PF10598	1,368	30	None
RRM_5	PF13893	4,399	38	RRM
RRM_7	PF16367	2,003	6	RRM
RRM_8	PF11835	2,912	5	RRM
RRM_9	PF18444	124	6	RRM
RRM_DME	PF15628	685	0	None
RRM_occluded	PF16842	454	6	RRM
RRM_Rrp7	PF17799	769	1	RRM
SET_assoc	PF11767	514	0	RRM
Smg4_UPF3	PF03467	1,824	3	RRM
Spo7_2_N	PF15407	240	0	RRM
Sugar_tr*	PF00083	99,230	24	MFS
Tap-RNA_bind	PF09162	714	12	RRM
Transposase_22	PF02994	594	12	RRM
U1snRNP70_N	PF12220	1,302	10	RRM
XS	PF03468	1567	2	RRM
YlmH_RBD	PF17774	1,708	1	RRM

* These families were retrieved only from RRMdb.

† The Limkain-b1 family has been renamed to MARF1_RRM1.

will help to understand the topology in a better way. In addition, we aligned the query structure against the two reference structures using two different programs: PDBeFold and Kpax. The resulting aligned structures from both programs were manually inspected in PyMol, and the Pfam families were classified accordingly as true or false RRM families.

Demo 3.2.1 shows the demonstration for structural inspection of 1WHV_A from RNA_bind (PF08675) family. The details of all manual structural inspections can be found in Appendix A.2.

Demo 3.2.1: Demonstration of Structural Inspection†

The RNA_bind (PF08675) family contains 5 domain structural instances corresponding to 2 different domain instances. We performed the structural inspection of this family in following way for the 1WHV_A (430, 508) structural instance:

1WHV_A(430, 508):

- **SCOP classification***: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
RANGE 1whv:A        S H s S H S h
PDB 1a9n:B          S H s S H S h
PDB 2a3j:A          S H s S H S h
```

- **Kpax**: (For simplicity, we removed some fields from the result of Kpax)

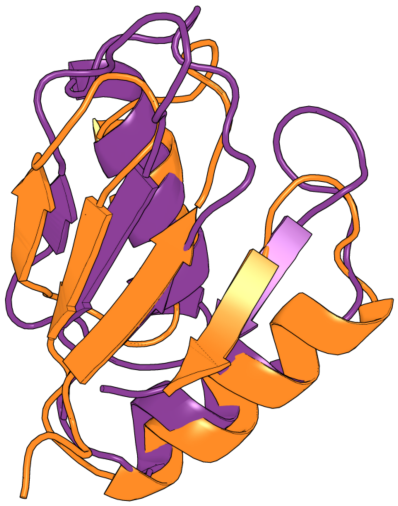
```
============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  Match
==== ======= ======= ======= ======= ======= ====  =====
  1   39.86   46.17  0.5014  0.6613  0.6740  1.88  1a9n_B
  2   37.17   44.60  0.4675  0.6465  0.6616  1.97  2a3j_A
============================================================
```

Both SCOP and CATH classification point to classes identical to the one obtained with RRM_1 (PF00076) in Pfam. In addition, the structures are aligned with respect to the secondary structures in PDBeFold (Figure 3.1). Finally, all the scores computed by Kpax structural alignment are consistent with a correct alignment, especially the M-Score which is greater than 0.60. Thus, this analysis allows to conclude that '1WHV_A' has the RRM fold and is a true RRM domain instance.
In the same way, we inspected two more structures (3CTR_A and 3D45_A) from this family before including this family in the list of true RRM families.

* TP=protein type, CL=protein class, CF=fold, SF=superfamily, FA=family
†NOTE: This demo shows a manual structural inspection for positive example of RRM family. Refer to Appendix A.1 for further explanation about SCOP and CATH classification codes and Kpax alignment scores.



(a) Structural alignment of 1WHV_A (orange) and 1A9N_B (green)



(b) Structural alignment of 1WHV_A (orange) and 2A3J_A (purple)

Figure 3.1: Structural inspection of 1WHV_A (orange) via alignment with a) 1A9N_B (green) and b) 2A3J_A (purple)

After this analysis, it became clear that not all the Pfam families from the RRM clan (CL0221) possess true RRM instances, matching the RRM definition. These Pfam families are DUF1743, U1snRNP70_N, PHM7cyt, RL, and GlcNac-1reg. In the end, we found only 19 true RRM Pfam families, 9 false RRM Pfam families (Table 3.2).

This set of 19 true RRM Pfam families was chosen as starting material for data collection for the database. This is one of the apriori choice we adopted for this database. The other choice was not to integrate predicted 3D structures obtained from AlphaFold but to limit our data to experimentally determined 3D structures from the PDB.

3.2.4 Use cases

A use case is a description of the ways in which a user interacts with the database to achieve a goal. Typically, a use case outlines the user's point of view and tells developers the needs and requirements of users. The use cases are important to design the database model. The use cases covering different applications for the interactions between RNA and RNA Recognition Motifs (RRMs) have been devised by questioning the partners of the RNAct project to help create the database model.

Below is the list of the use cases collected for the InteR3M database.

1. Retrieve sequences of RRM-containing proteins that have Pfam ID PF08777.
2. Retrieve RRM sequences that have Pfam ID PF00076 and 3D crystallography data of complexes with nucleic acids (NA).
3. Retrieve RRM sequences that have Pfam ID PF08777 and NMR data available

Table 3.2: Results of Curation of Pfam families for RRM domains with at least one PDB instance. These counts are from Pfam v33.0.

Family Identifier	**Family Name**	**Number of domain Structures (Number of corresponding distinct domain sequences)**	**Result of inspection**
PF00076	RRM_1	1,069 (247)	Pass
PF00083	Sugar_tr	24 (8)	Fail
PF02902	Peptidase_C48	51 (12)	Fail
PF02994	Transposase_22	12 (1)	Pass
PF03467	Smg4_UPF3	3 (2)	Pass
PF03468	XS	2 (1)	Pass
PF03880	DbpA	3 (2)	Pass
PF04847	Calcipressin	1 (1)	Pass
PF05172	Nup35_RRM	10 (5)	Pass
PF08152	GUCT	2 (2)	Pass
PF08489	DUF1743	10 (2)	Fail
PF08675	RNA_bind	5 (2)	Pass
PF08777	RRM_3	4 (2)	Pass
PF08952	DUF1866	2 (2)	Pass
PF09162	Tap-RNA_bind	12 (1)	Pass
PF10598	RRM_4	30 (3)	Fail
PF11608	Limkain-b1	1 (1)	Pass
PF11835	RRM_8	5 (2)	Pass
PF12220	U1snRNP70_N	10 (2)	Fail
PF13893	RRM_5	38 (8)	Pass
PF14703	PHM7_cyt	12 (4)	Fail
PF16367	RRM_7	6 (3)	Pass
PF16842	RRM_occluded	6 (1)	Pass
PF17774	YlmH_RBD	1 (1)	Pass
PF17797	RL	14 (2)	Fail
PF17799	RRM_Rrp7	1 (1)	Fail
PF18440	GlcNAc-1_reg	1 (1)	Fail
PF18444	RRM_9	6 (2)	Pass

(with or without NA).

4. Retrieve all NA sequences having dissociation constant lower than 10 µM.
5. Retrieve all NA sequences interacting with the same RRMs.
6. Retrieve all RRMs interacting with the NA sequence 'AAGCGCCAGAACU'.
7. Retrieve PDB IDs of RRMs that have Pfam ID PF00076 and NMR data of complexes with NA.
8. Retrieve all PDB IDs and Pfam IDs of RRM interacting with the same NA sequences.
9. Retrieve all complexes having domain RRM_1 or with Pfam ID PF03880.

10. Retrieve all the articles cited for domain PF16842 in relation with a PDB structure.
11. Retrieve all PDB IDs of RRM structures with crystallography data with less than 2 Å resolution.
12. Retrieve interacting residues between protein and NA from complexes containing Pfam ID PF18444.

The use case 1 suggests storing protein and domain information to meet these needs. One can simply store this information together but the protein information would be redundant when the protein has multiple RRM domains. Thus, it would be more efficient to store protein and domain information separately with a link (relation) between them. This results in a *Protein entity* linked with an *RRM_instance* entity, one instance of the former can be associated with one or more instances of the latter. The *RRM_instance* entity should also be linked to a *Pfam* entity as its reference family.

The use cases 2, 3, and 4 draw our attention towards the experiments performed with RRM domains and nucleic acids along with their resulting information. An experiment can be one of several types, including X-ray diffraction, solution NMR, ITC, filter binding assays, etc. At the very broad level, all of these experiments can be classified as either structural experiments or non-structural experiments depending on their resulting information. This leads us to introduce a *Ligand* entity (mostly of NA type) and an *Experiment* entity.

A structural experiment results in a 3D structure of protein(s) with or without an interacting ligand. One PDB entry is considered as one *structural experiment*. It might have one or more proteins with or without one or more ligands. A *non-structural experiment* does not result in a 3D structure. The goal of these experiments is to check the binding of protein and ligand. An instance of "Non-structural experiment" is usually an experiment performed to check the binding of single ligand with one protein containing one or more RRM domains.

Thus, based on these 3 use cases (use cases 2, 3, and 4) we can say at this stage that an *Experiment* entity can be either structural or non-structural. One instance of the *Experiment* entity uses one or more instances of the *Protein* entity (full or partial) containing one or more instances of *RRM instance* entity with or without any instance of *Ligand instance* entity. The *Experiment* entity is associated with either a *PDB structure* entity or a *Binding information* entity. The *PDB structure* entity represents entries from the PDB comprising the whole structure that may have one or more instances of an *RRM structure* entity with or without any instance of a *Ligand structure* entity.

Use cases 5-8 refer to various ways to query the data which can be implemented very easily using SQL technology. Use cases 9 and 10 lead us to add publication references and resolution values as attributes for the *PDB structure* entity. Finally use case 12 requires that contact information be structured in a separate entity (*Detailed contacts* entity) related to the *PDB structure* entity. This preliminary use case analysis will be refined when building the complete conceptual model of the database (see section 3.3.1).

3.3 Database Design

We designed the InteR3M database in three phases, conceptual, logical and physical database design, according the good practices in this field [Connolly and Begg, 2005].

3.3.1 Conceptual Database Design

The conceptual model has been designed for the InteR3M database based on the collected use cases (Figure 3.2). At this level, the design is entirely independent of implementation details like database management system (DBMS), programming or any hardware platforms.

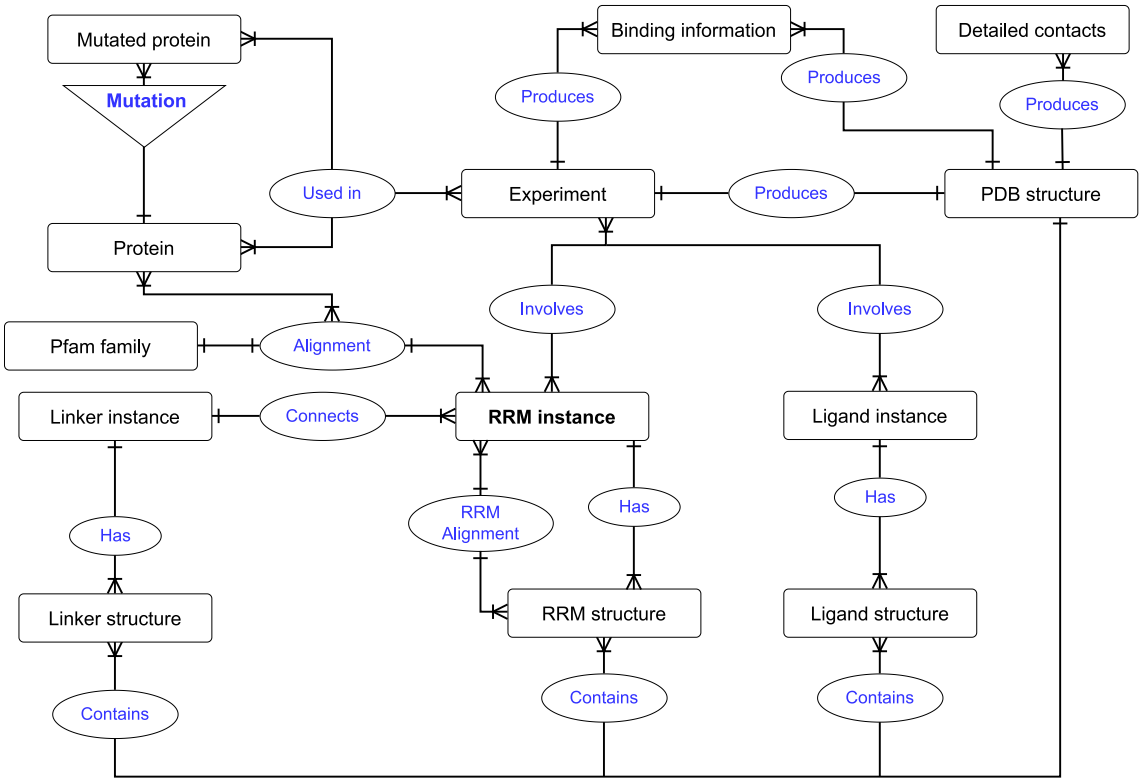


Figure 3.2: Conceptual model for InteR3M database

The *RRM instance* entity is central to the data model and is well connected with other entities from the data model. An RRM instance is a part of a protein, one protein may contain one or more RRM instances. When a protein contain two or more RRM instances, these RRM instances are connected by the linker instances. Ligand instances define the unique NA sequences that have been tested with the RRM instances in an experiment. An *Experiment* is an important entity as it carries the originality of this database. It is defined in relation with the *Protein* and *Ligand instance* entities. Experiments produce different information based on the type of experiment performed. In this schema, experiments produce mainly two types of information: structural information (PDB structure) and binding information. Experiments producing structural information are linked to the *PDB structure* entity that corresponds to entries from the PDB and structural information for RRM, ligand and linkers. For structural experiments, each entry in PDB with RRM is

considered as an individual experiment even if it has several proteins. For non-structural experiments, one experiment is performed with a single protein and a ligand to check the binding. The *Experiment* entity also keeps track of the mutations performed in the protein via the *Mutated protein* entity. The 3D structures of protein and nucleic acid complexes are further processed to extract interactions at the atomic level reported by the *Detailed contacts* entity.

The conceptual data model was tested and validated against the use cases collected. This conceptual model provides a base for the logical database design.

3.3.2 Logical Database Design

A preliminary logical model was developed on top of the conceptual model by considering constraints at the table level. This model was normalized to prevent any kind of anomalies in the database resulting in the final logical model for the InteR3M database (Figure 3.3). The logical model is composed of 19 tables and 106 attributes.
The details of all attributes and tables can also be found in the InteR3M data dictionary available on the website at https://inter3mdb.loria.fr/dictionary.

3.3.3 Physical Database Design

Physical database design is the final phase of database design process and it decides how the database is to be implemented. Field length specifications were defined for each field (also called an attribute or column) from all the tables. The views were created to merge/join data from multiple tables into a single view. The views were defined in a way to meet the user's needs without any information-retrieval issues. Views can be considered as a tool to support particular aspects of the implementation or application program, i.e., user interface in our case.

The *allpdbcontacts* view was defined to merge all atomic-level and residue-level interactions together by assigning the RRM and ligand instances and structures. The *contactsaligned* view was defined to align the contacts from *allpdbcontacts* against/with the RRM master alignment [Martinez et al., 2023].
Table 3.4 presents a complete list of tables and views created for the physical design of InteR3M database.

PostgreSQL automatically creates a unique index when a unique constraint or primary key is defined for a table. An index is a data structure in a database that helps to enhance the querying performance. Indexes provide a way for the database to look up the relevant rows directly based on the values in the indexed columns, rather than having to scan the entire table. Thus, we have 19 indexes for 19 tables in InteR3M database (Table 3.3).

3.4 Implementation

The InteR3M database implementation was carried out in two stages, the database implementation where we implemented the actual database in a Database Management System (DBMS) and the user interface implementation where we

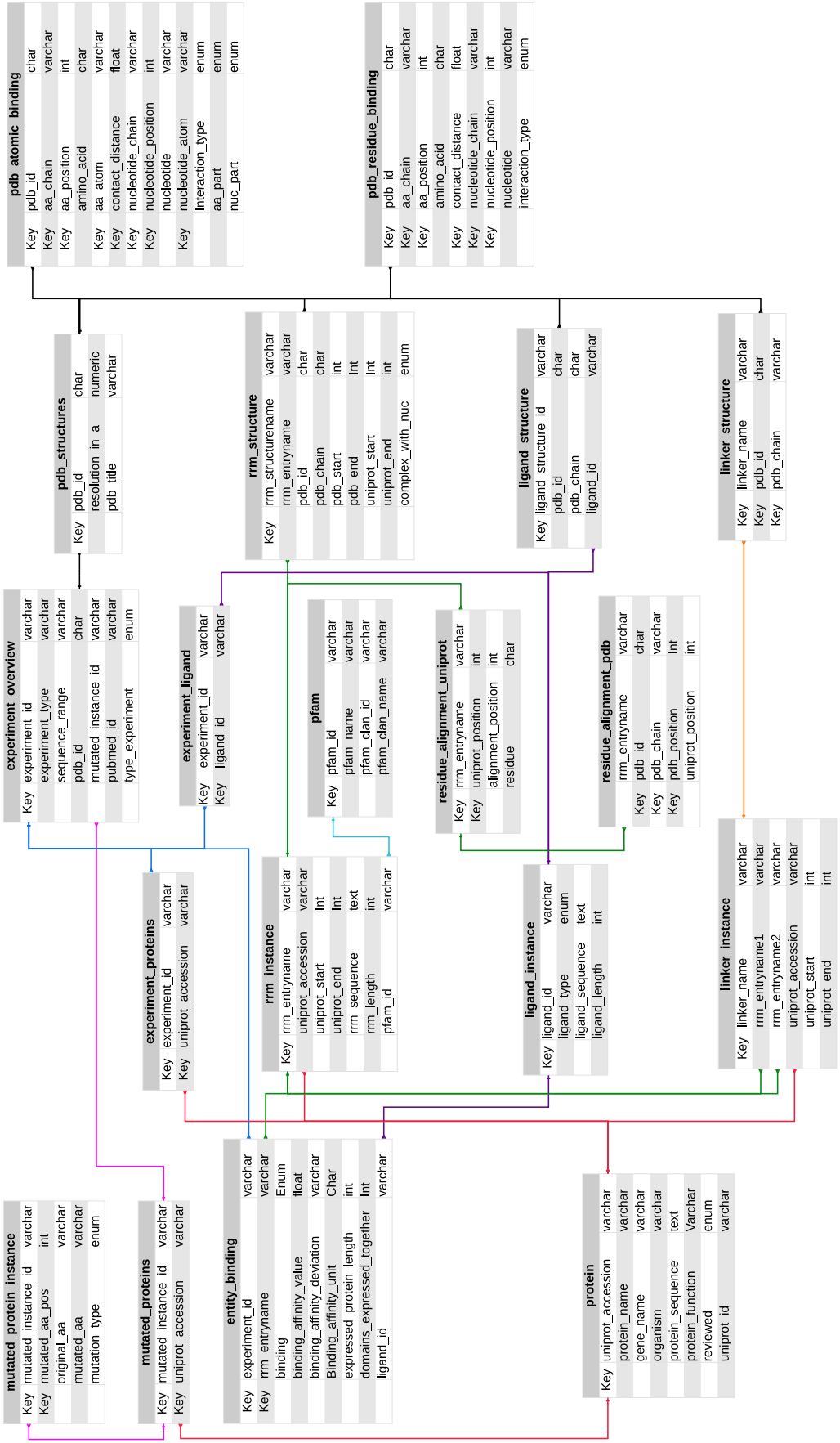


Figure 3.3: Logical model for InteR3M database. It is also available in 'About' section of InteR3M database at https://inter3mdb.loria.fr/about.

Table 3.3: List of indices in InteR3M database

Table	Index Name	Indexed Columns
entity_binding	entity_binding_pkey	experiment_id, ligand_id, rrm_entryname
experiment_ligands	experiment_ligands_pkey	experiment_id, ligand_id
experiment_overview	experiment_overview_pkey	experiment_id
experiment_proteins	experiment_proteins_pkey	experiment_id, uniprot_accession
ligand_instance	ligand_instance_pkey	ligand_id
ligand_structure	ligand_structure_pkey	ligand_structure_id
linker_instance	linker_instance_pkey	linker_name
linker_structure	linker_structure_pkey	pdb_id, linker_name, pdb_chain
mutated_protein_instance	mutated_protein_instance_pkey	mutated_aa_pos, mutated_instance_id
mutated_proteins	mutated_proteins_pkey	mutated_instance_id
pdb_atomic_binding	pdb_atomic_binding_pkey	aa_chain, aa_position, aa_atom, pdb_id, nucleotide_atom, contact_distance, nucleotide_chain, nucleotide_position
pdb_residue_binding	pdb_residue_binding_pkey	nucleotide_position, pdb_id, aa_chain, aa_position, contact_distance, nucleotide_chain
pdb_structure	pdb_structure_pkey	pdb_id
pfam	pfam_pkey	pfam_id
protein	protein_pkey	uniprot_accession
residue_alignment_pdb	residue_alignment_pdb_pkey	pdb_position, pdb_id, pdb_chain
residue_alignment_uniprot	residue_alignment_uniprot_pkey	uniprot_position, rrm_entryname
rrm_instance	rrm_instance_pkey	rrm_entryname
rrm_structure	rrm_structure_pkey	rrm_structurename

developed the application (user interface) that helps users to access data from database.

3.4.1 Data Collection

The overall process of data collection for InteR3M database is schematized in Figure 3.4. The primary data for domain information was collected from Pfam. The domain instances and structures were retrieved from the release files provided by Pfam database. Pfam stores domain instances using UniProt entry names that need to be mapped to UniProt accession identifiers to retrieve information about

Table 3.4: List of views in InteR3M database and their usage in database interface.

Name	Interface Usage
allpdbcontacts	Intermediate view for 'list of contacts' display
contactsaligned	List of contacts display
experiments_filter	Searching experiments from 'Multicriteria search'
experiment_per_protein	Experiment display, Protein display
ligands_per_experiment	Experiment display
notstructural	Experiment display, Ligand instance display, RRM instance display
structural	Experiment display, Ligand instance display, RRM instance display, RRM structure display
contacted_ligands	Ligand instance display, Protein display, RRM instance display
rrm	Protein display, RRM instance display
mutant_proteins	Mutated protein instance display
linker_insta_merged	Linker instance display
linker_structure_complex	Linker instance display
linkercontacts	Linker instance display

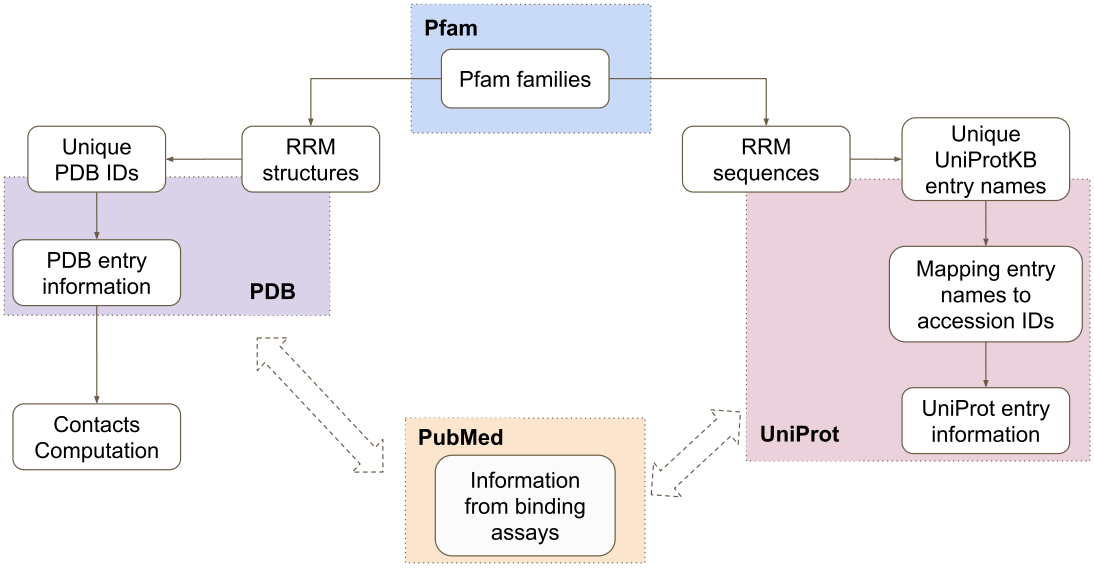


Figure 3.4: Workflow for data collection for InteR3M database. Starting Pfam families and PubMed references have been manually selected. Dashed arrows correspond to cross-referencing.

the corresponding proteins from UniProt. We used UniProt 'ID mapping' tool[1] for mapping UniProt entry names to UniProt accession identifiers. Then 'UniProt

[1]https://www.uniprot.org/id-mapping

website REST API'[2] was used to collect information about proteins containing RRM domains.
We also retrieved the structural instances of RRM domains from Pfam. Then, we used the PDB API to retrieve the experimental information using the PDB identifiers from the collected structural instances. The PDB API also provides information about the proteins and nucleic acids present in the given structure. With this information, we processed the PDB entries with RRMs having at least one nucleic acid to compute the contacts between RRM and nucleic acids.

Contacts computation

The overall workflow for contacts computation is schematized on Figure 3.5. The set of RRM-NA (RNA or DNA) complex structures were obtained from the Protein Data Bank (PDB). From each PDB entry, the chains containing RRM and NA were extracted and processed through an in-house script to retrieve the contacts between RRM and NA. An amino acid residue and a nucleotide were considered to interact if any atom from the residue and the nucleotide were less than 5.0 Å away from each other. This is a broadly used interaction definition to keep strong interactions such as hydrogen bonds or electrostatic interactions, while still accounting for hydrophobic interactions that can occur at distances of 3.8 – 5.0 Å. Each atomic level interaction is in the form of PDB ID, amino acid chain, amino acid, amino acid position, amino acid atom, nucleotide chain, nucleotide position, nucleotide and the distance between amino acid atom and nucleotide atom. We used x3DNA-DSSR [Colasanti et al., 2013] to determine the 'H-bonds' and 'stacking interactions' between RRM and NA chains. All atomic interactions were assigned one of the following interaction type if they met the respective conditions:

H-bond: If the atomic interaction is present in the list of H-bonds obtained from x3DNA-DSSR[3].

Ionic bond: If the interacting atoms are from the side chain of basic amino acid and phosphate of nucleotide respectively.

Van-der-waals: If neither of the above two conditions is met, the interaction is assigned as van-der-waals interaction.

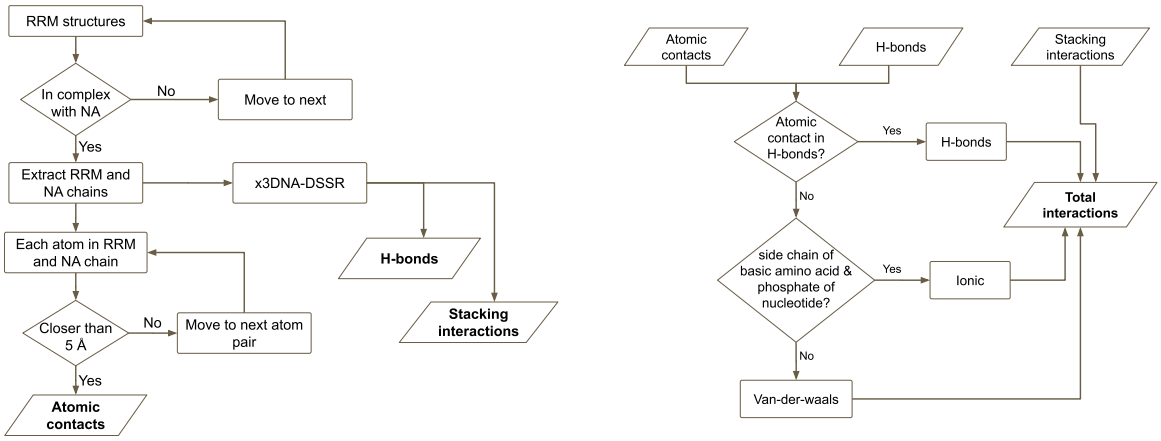
The RRM-nucleic acid binding information resulting from different binding assays was collected manually from literature.

In summary, InteR3M has 400,892 RRM domain instances from 256,266 unique proteins and 1,456 RRM structures from 727 unique PDB entries. Only 303 RRM domain instances have at least one structure in PDB. Contacts have been computed for the 656 RRM-NA complexes recorded in the database. The binding information was retrieved from 16 publications related to 34 distinct RRM instances.

3.4.2 Database Implementation

The InteR3M database is implemented in PostgreSQL (v15.1). PostgreSQL is a powerful and open source object-relational database management system

[2]https://www.uniprot.org/help/api_retrieve_entries
[3]http://home.x3dna.org/



(a) Workflow for the contacts computation. (b) Workflow for contact types determination.

Figure 3.5: Workflow for contacts computation and contact types determination. The bold font indicates the final output from each workflow.

(RDBMS) [Stonebraker et al., 1990]. The database schema (tables and views) was defined using the data definition language (DDL).

3.4.3 Implementation of User Interface

The user interface for InteR3M database was implemented in PHP language. We used twig as a template engine for PHP to design the web pages that will be displayed to the users. Twig is a fast, secure, and flexible modern template engine for PHP. The InteR3M user interface provides options to browse and search the InteR3M database. The interface provides two different search options, a simple search and a multi-criteria search. More details about these are given in section 3.5.

InteR3M database also provides the options to filter the contacts/interactions between RRM and nucleic acids based on several different options like amino acid residue name and nucleotide, or contact type.

3.4.4 Testing

Once the database was ready, we thoroughly tested it before making it publicly available. We carefully tested the search and browse functionalities of the database that helped us to improve the usability of InteR3M database. It was important to get the feedback from new users of the database that will make it more user-friendly and help us evaluate the database. Thus, we asked CAPSID and RNAct members to use the database and provide their feedback for testing purpose. The feedback we received mainly included the comments to add information for columns in search contacts section to better understand the meaning of each column. We also have been asked to make connections between ligand instances and structures so that users can easily hop between instances and structures of the ligand similar to RRMs.

3.5 Using InteR3M

InteR3M is built by considering real-world cases and ease to retrieve data. The Homepage contains a search bar (in top right corner) with the filtering options.

3.5.1 Search functionality

The users can search the InteR3M database for a given protein, RRM, or the 3D structure if they know their identifiers. The Search bar allows to enter a protein name or identifier from UniProt, an RRM name or identifier from Pfam or a 3D structure identifier from PDB. Each query will return a results page with a list of matching RRM instances and/or RRM structures. Each returned entry for RRM instances provides options to select and jump to RRM entryname or UniProt Accession. Each entry from returned results for RRM structures provides options to select and jump to experiment, protein, or RRM structure.
The Search bar is available at the top right of every page and is easily navigated by pressing the tab key.

3.5.2 RRM instance display

Once an RRM of interest is selected, the RRM instance display shows protein and domain information, a list of all available structural instances, a list of non-structural (binding) experiments retrieved from literature, and a list of ligands tested with this RRM instance in all (structural and non-structural) experiments. The list of structural instances provides information about interacting ligands if any and a link to list the atomic interactions between RRM and ligand from that structure.

3.5.3 Protein display

Upon selection of a particular protein, the Protein display presents brief information about the protein, a list of RRM domains, and all the experiments (structural and non-structural) performed for RRM domains from this protein along with the list of ligands included in the experiments. Each experiment from the list of experiments presents the information about experiment type and the source (PubMed identifier). The user can select the RRM domain instance, an experiment, and/or the ligand to learn further about a particular entry.

3.5.4 Ligand instance display

Upon selection of a particular ligand, the Ligand display presents brief information about the ligand followed by a list of experiments this ligand was involved in and a list of RRM domains with which the binding of ligand has been tested. The non-structural (binding) experiments are listed with the RRM domain with which the experiment was performed along with their binding affinity. The structural experiments are listed with the RRM domain with which the experiment was performed along with the option to visualize the interactions between them. At the bottom, the list of RRM domains indicates the RRM domains with their position in a protein and domain sequence.

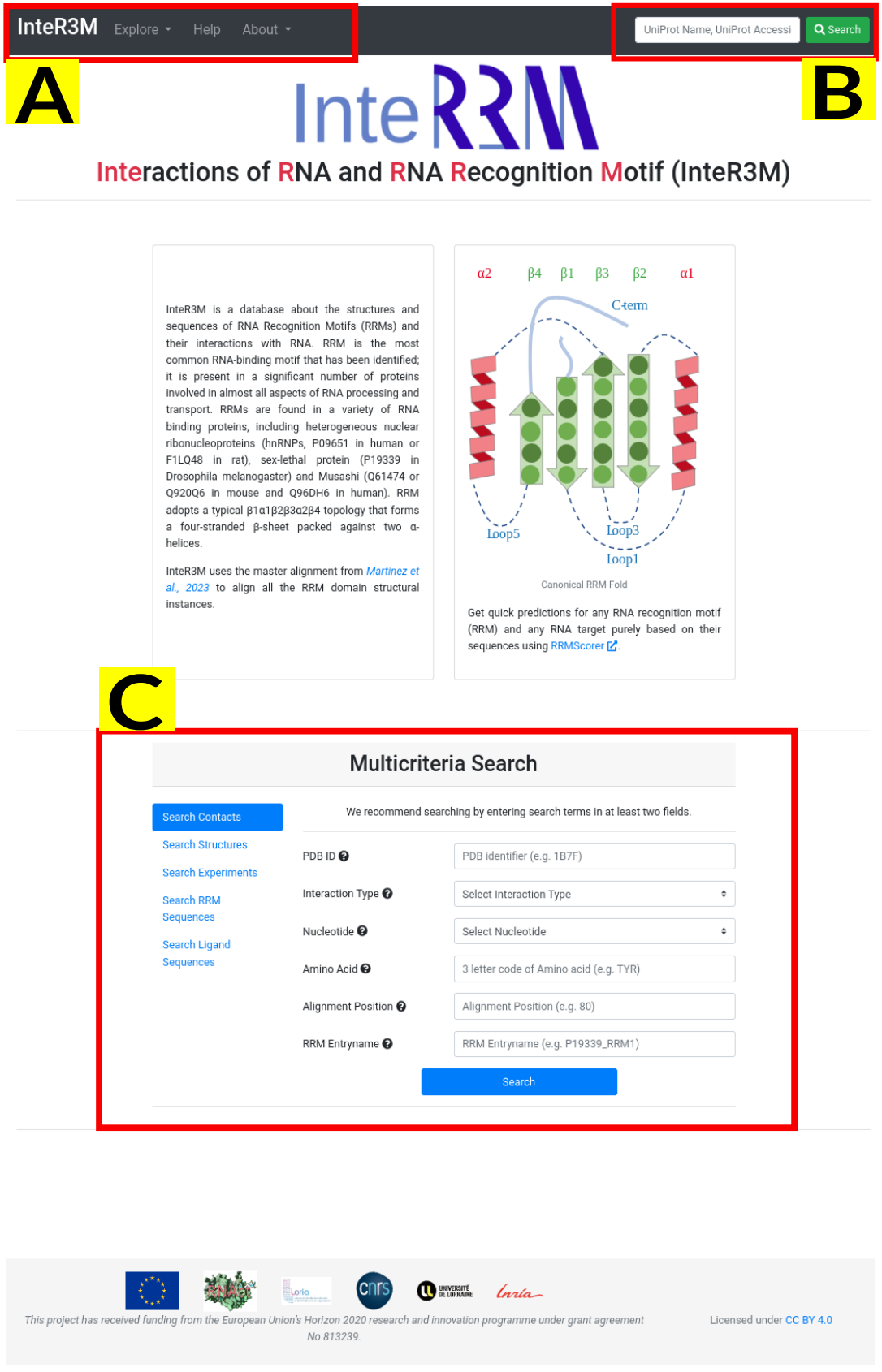


Figure 3.6: Homepage of InteR3M database. A) indicates the Navigation Bar, B) indicates the Search Bar, and C) indicates the Multicriteria Search.

3.5.5 Experiment display

The Experiment display provides brief information about the type and source of information, protein(s), RRMs and ligands involved in this experiment. The Experiment display also provides the information about mutations in the protein. From Experiment display, user can select protein, RRM, and/or ligand to get detailed information about a particular entry. For the complexes of RRM-ligand (NA) from structural experiments, the entries are listed with the link for the interactions between RRM and ligand, while the non-structural experiments are displayed with binding affinity information.

3.5.6 List-of-contacts display

The list-of-contacts display provides detailed contact information between an RRM and a ligand. From this display, user can jump to several displays like 'RRM instance', and 'Ligand instance' by a simple click. Whenever available, the list-of-contacts display also provides information about the amino acid residue position in the protein sequence ('UniProt Position') and in the RRM master alignment ('Alignment Position'). This display is available for showing particular contacts retrieved from the 'Search Contacts' option of multicriteria search, or directly from an RRM instance display or a Ligand instance display.

3.5.7 Multicriteria Search

The InteR3M database provides the functionalities for querying the database using 'Multicriteria Search' (Figure 3.6 C.). With this functionality, users can query for contacts between RRM and ligand, RRM structures, experiments performed with RRM domains, RRM domain instances and ligands.

The contacts (interactions) between RRM and ligand can be queried using one or more of these parameters: PDB identifier, interaction type, nucleotide, amino acid, and RRM entryname. The RRM structures can be queried using PDB identifier, experiment type, UniProt accession number, and/or PubMed identifier. The experiments can be queried based on PDB identifier, experiment type, ligand binding capacity, UniProt accession number, and/or PubMed identifier. Moreover, all these query variables can be combined thanks to the multicriteria search functionality. This answers the user needs described in the use cases (see Section 3.2.4). In the case of RRMs and ligands search, the user can use either the complete sequence or sequence motif to search for the RRM instances and ligand instances with the given sequence. For sequence search, it is recommended to use as long sequences as possible.
For each multicriteria search query, the user will land on the results page. Figure 3.7 shows the results for one 'Search Contacts' query using multicriteria search. Each result page from multicriteria search provides options for exporting the results, filtering the results and sorting the results in ascending or descending order based on a column. To filter the results, one can simply type the value on which the results will be filtered. For example, the user can filter the resulting contacts based on an interacting amino acid such as tyrosine (TYR) by typing "TYR" in the filter bar. The user can hide columns by clicking on a specific

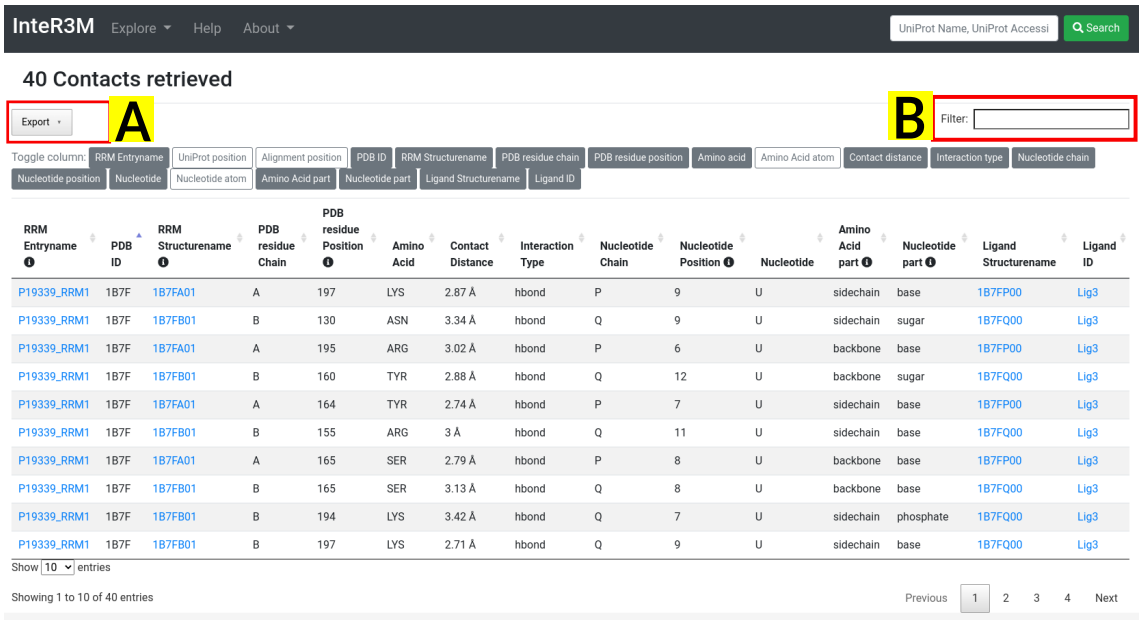


Figure 3.7: Results page for 'Search Contacts' using Multicriteria Search. We searched contacts using *Interaction Type* ('H-bond'), *Nucleotide* ('U'), and *RRM Entryname* ('P19339_RRM1'). A) represents Export options and B) represents Filter bar

column name in the 'Toggle column' section. For simplicity, we have hidden a few columns in Figure 3.7. The user can also sort the resulting contacts, by simply clicking on the column on which he/she wants to sort the results. The result page also provides options to export the results as either *CSV* file or *copy* to clipboard.

3.6 Strategies to update InteR3M database

The updates of InteR3M database can be from one of the two categories: minor updates and major updates. Major updates will be done with every release of Pfam database, i.e. generally every six months. Pfam release files will be checked for the addition or removal of any domain instances from one of the 19 RRM families present in InteR3M database. The newly added domain instances in each Pfam release will also be added to the InteR3M database along with their available 3D structures, while the obsolete domain instances (if any) from Pfam release will be moved to the obsolete section from InteR3M database.
For any new 3D structure of RRM-NA complex, the contacts will be computed using the same protocol (Figure 3.5). New families or families from RRM clan (CL0221) in Pfam with no structural coverage will be checked for availability of any experimental 3D structure. As soon as any structure is available from these families, the structures will be passed through the structural inspection protocol (Appendix A.2).
In addition, we will be computing one average structure with only backbone atoms from all RRM structural instances present in the InteR3M database. This average RRM structure can be considered as a representative structure for RRM domains, and thus can be used to compare with other structures.

Minor update will consist in adding binding information from literature and any changes made in the web interface. Minor releases of InteR3M database will be made regularly, for example, every three months.
The release history of InteR3M database can be accessed at https://inter3mdb.loria.fr/releases/.
All these updates will rely on existing scripts stored at https://gitlab.inria.fr/hdhondge/data_collection_inter3mdb.

3.7 Results

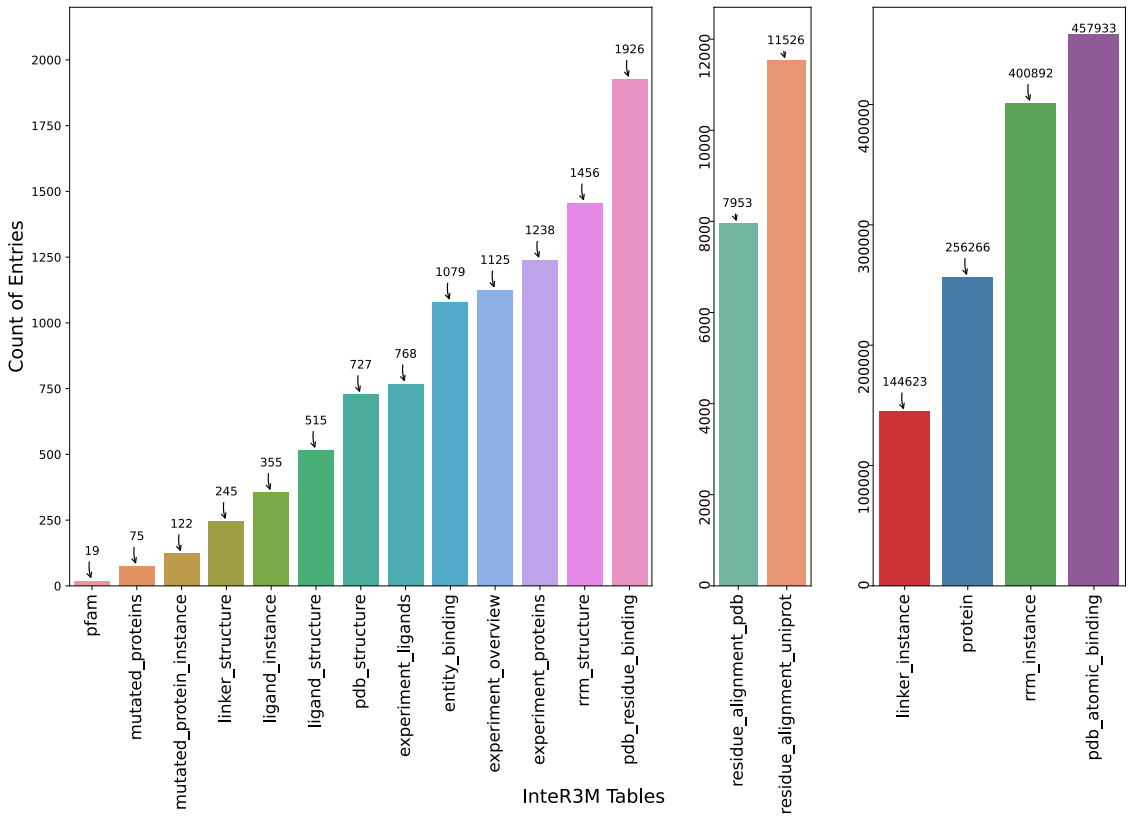


Figure 3.8: Count of entries per table InteR3M database

The InteR3M (v0.0.1) contains a total of 400,892 RRM domain instances from 256,266 unique proteins. All these RRM domain instances have been extracted from 19 different Pfam families. In average, the PDB[4] has 1 structure per 1209 protein entries from UniProtKB[5], i.e.

$$\frac{\text{UniProtKB entries}}{\text{PDB entries}} = \frac{246,440,937}{203,863} = 1208.855$$

In InteR3M, we have a more favourable structure to sequence ratio, i.e. 1 structure per 352 RRM containing proteins:

$$\frac{\text{RRM containing UniProtKB entries}}{\text{RRM containing PDB entries}} = \frac{256,266}{727} = 352.497$$

[4]Data collected on 19$^{\text{th}}$ of April 2023
[5]UniProtKB Release 2023_01

This suggests that the RRM-containing proteins have been studied more from a structural point of view compared to overall proteins.
The InteR3M database includes 459,859 interactions extracted from 656 RRM-NA complexes, with 1,926 residue-level stacking interactions and 457,933 atomic-level interactions (Figure 3.8). Proteins with multiple RRM domains have linkers, linking two RRM domains together. These linkers play an important role in the orientation of RRM domains. InteR3M has 144,623 linker instances from 95,288 different proteins.

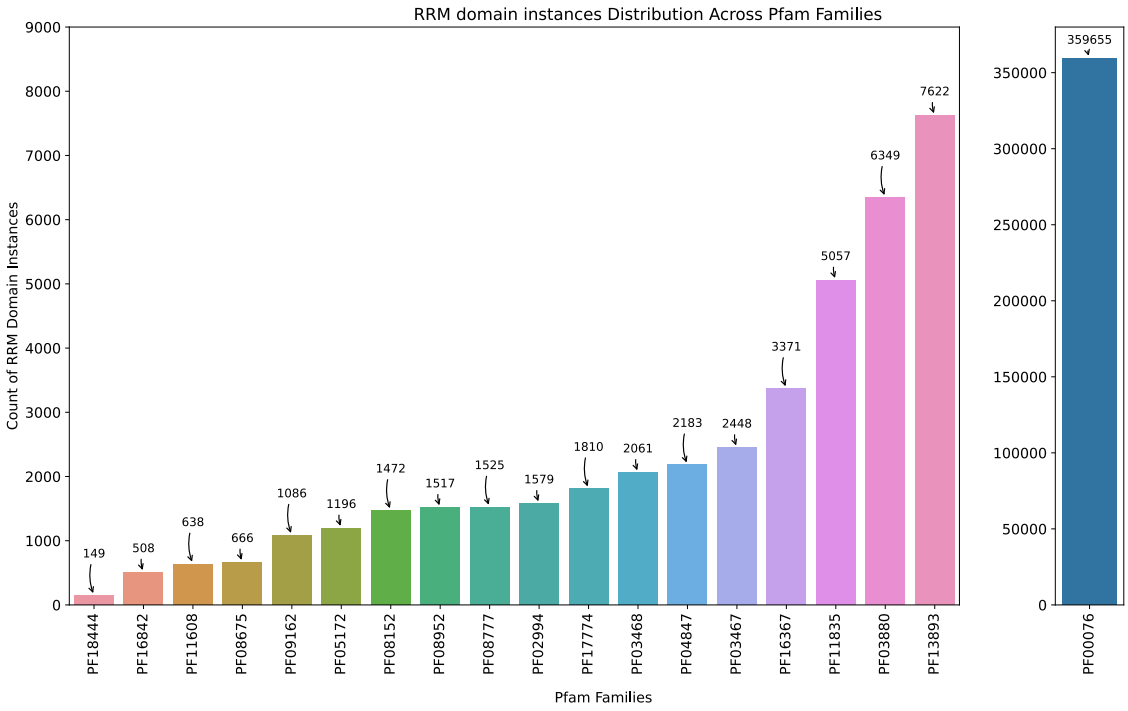


Figure 3.9: Distribution of RRM instances across Pfam families

Figure 3.9 shows the distribution of RRM instances across Pfam families. RRM_9 (PF18444) has the smallest number of RRM domain instances, i.e. 149, whereas RRM_1 (PF00076) is the predominant family among all comprising 359,655 RRM instances. Of all other families, none has more than 8,000 RRM instances.

Of these total (400,892) RRM instances, only 303 have their 3D structures solved in the PDB. These 303 RRM instances have a total of 1,456 structural instances from 727 unique PDB entries. The structural distribution of RRMs per Pfam family is shown in Figure 3.10. This clearly depicts the dominance of RRM_1 Pfam family in sequential and structural aspects of RRMs. RRM_5 (PF13893) is the second largest family with 7,622 RRM domain instances and 42 RRM structural instances. All other 17 Pfam families have less than 15 RRM structural instances each. Figure 3.10 also shows that there are more RRM structures in unbound form (orange bars, total unbound = 906) than the RRM structures in bound form (blue bars, total bound = 550) either with RNA or DNA. A total of 87 unique RRM instances have their structures in both bound and unbound states. The list of the RRM instances having both bound and unbound structures can be found in Appendix A.3.

The InteR3M database has information from 9 different types of experiments (including X-ray diffraction, NMR solution, Fluoroscence polarization, ITC etc.).

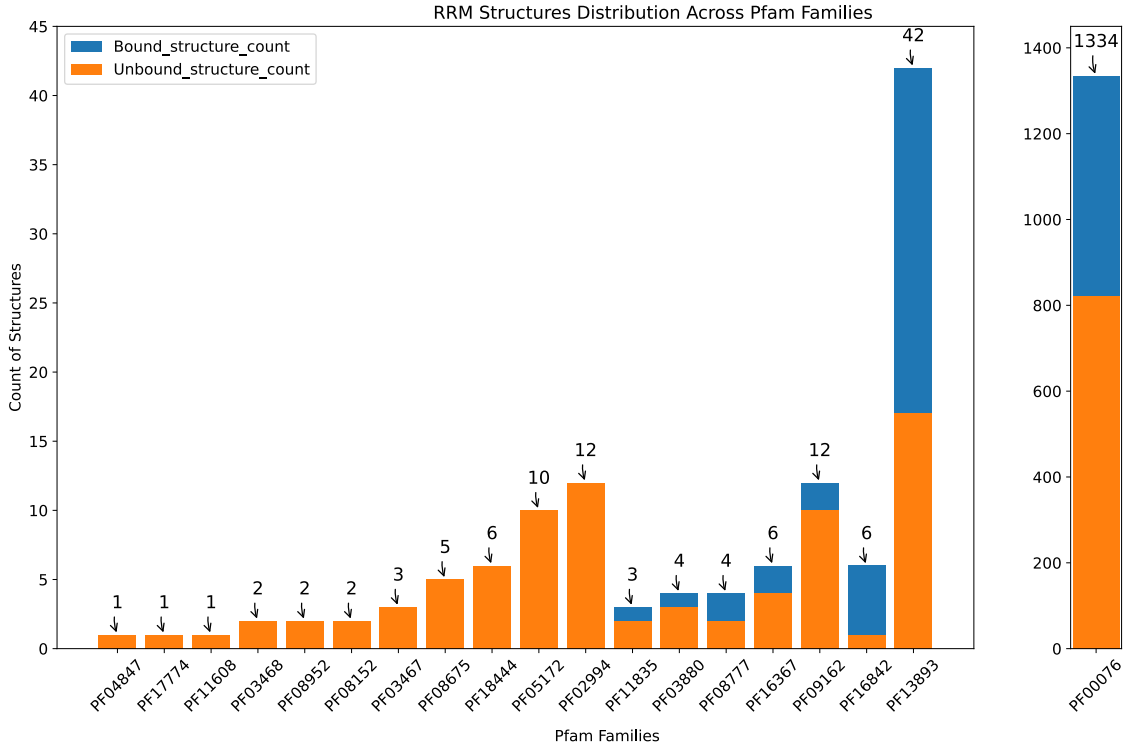


Figure 3.10: Distribution of RRM structures across Pfam families

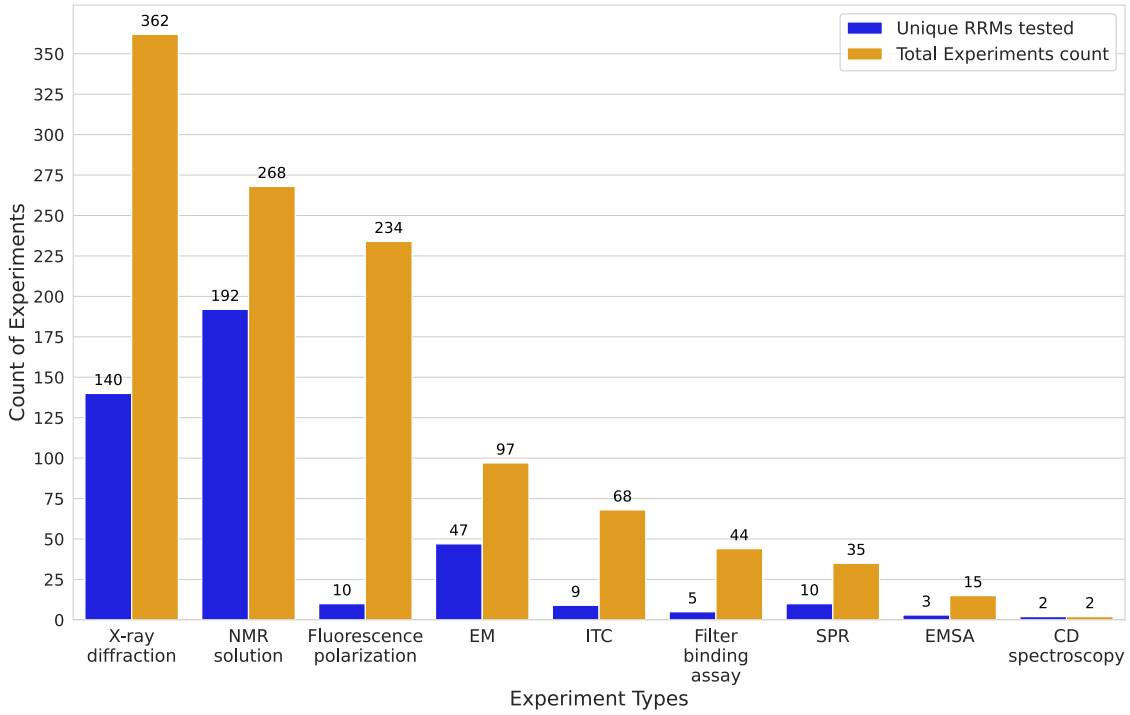


Figure 3.11: Distribution of experiments performed across each type of experiments. A single RRM instance might be used in more than one experiment of same or different types.

Figure 3.11 shows the distribution of experiment entries per unique RRM instances across different experiment types (techniques). The 'X-ray diffraction' technique

was used 362 times to determine the 3D structure of 140 unique RRM instances. The same RRM instance might be used in multiple experiments. The 'NMR solution' technique was used 268 times to determine the 3D structure of 192 unique RRM instances, making the NMR solution the technique used on the highest number of unique RRM instances. The 'Fluorescence polarization' technique has 234 experiments performed on 10 unique RRM instances.
Unlike non-structural experiments, structural experiments, i.e. X-ray diffraction, NMR solution, and EM, do not necessarily involve nucleic acid along with the RRM as their main goal is to determine the 3D structure of RRM containing proteins.

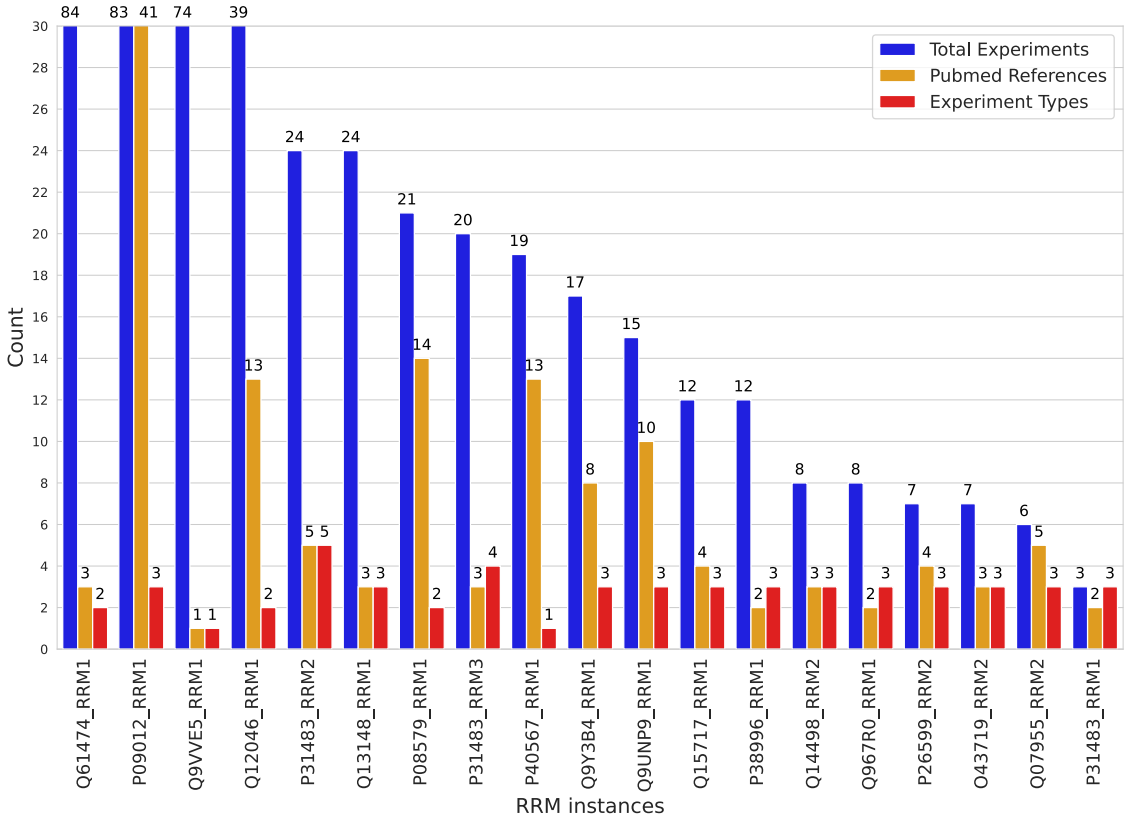


Figure 3.12: Distribution of Experiments performed for each RRM instance. All these RRM instances are from RRM_1 family (PF00076) except P40567_RRM1 that corresponds to RRM_5 (PF13893) family.
For simplicity, we have shown only RRM instances with multiple types of experiments, or with at least 50 total experiments, or with >10 PubMed references.

Figure 3.12 shows the most studied RRM instances from different perspectives, i.e. with the highest number of experiments, different PubMed references, and different experiment types. Currently, we have limited data for non-structural experiments as this step of collecting data from literature can not be automated. The RRM instance 'Q61474_RRM1' has the highest number of experiments (84) corresponding to 3 different PubMed references (studies). Each of these experiments can be classified into one of two types: 'Fluorescence polarization' or 'NMR solution'. This RRM instance corresponds to the first RRM domain from 'Musashi homolog 1' protein in Mouse.

The RRM instance with the highest number of publications (distinct studies) is the first RRM domain from 'U1 small nuclear ribonucleoprotein A' protein. A total of 83 experiments were performed for this RRM instance ('P09012_RRM1') that has been used in 41 different studies with 3 different types of experiments.
The RRM instance with the highest number of types of experiments is 'P31483_RRM2'. In InterR3M, there are 24 experiments entries related to this RRM instance from 5 different studies and 5 types of experiments. This RRM domain instance is from 'T-cell intracellular antigen-1' (TIA-1) protein in Human. Interestingly, all three RRM domain instances from the TIA-1 protein are present in our list of most studied RRM domain instances suggesting that all the three RRMs from this protein are well studied by different experiments and studies. The TIA-1 potein has dual regulatory role in transcriptional and post-transcriptional processes, shuttling between nucleus and cytoplasm [Kedersha et al., 2000, Zhang et al., 2005]. Within nucleus TIA-1 promotes the inclusion of exon 6 in FAS pre-mRNA resulting in an apoptotic form of FAS protein that is linked to autoimmune responses [Izquierdo et al., 2005]. In the cytoplasm, TIA-1 protein mediates translational repression of target mRNAs via binding to target RNA motifs (A/U rich sequence) present in their untranslated regions. Previous studies from Piecyk et al. [2000], Dixon et al. [2003] and Yu et al. [2003] have shown that TIA-1 protein also represses the expression of inflammatory mediators.

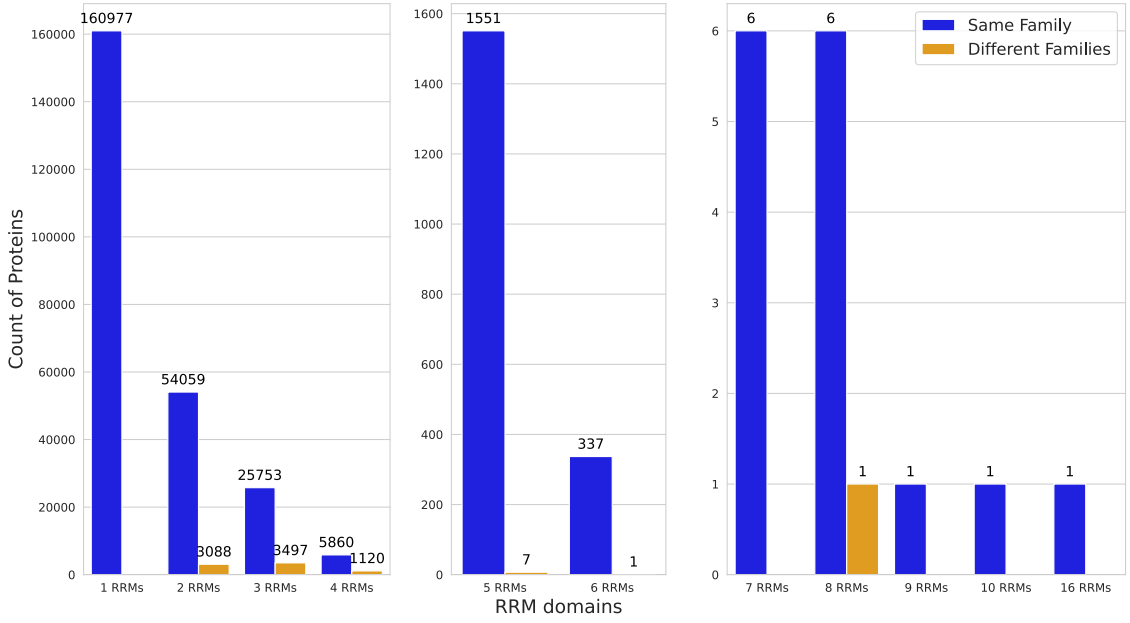
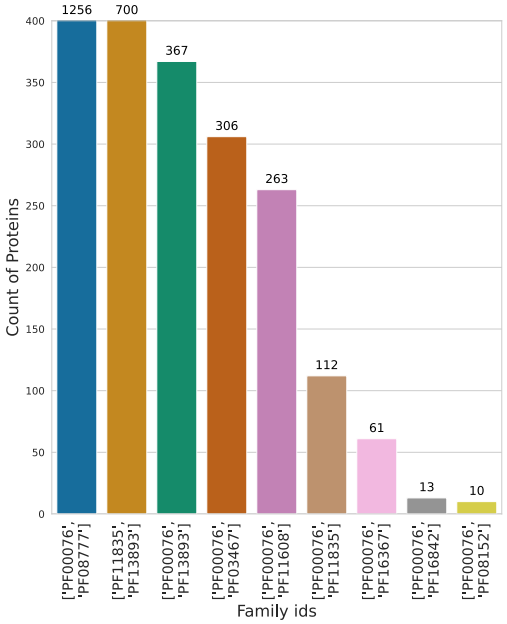


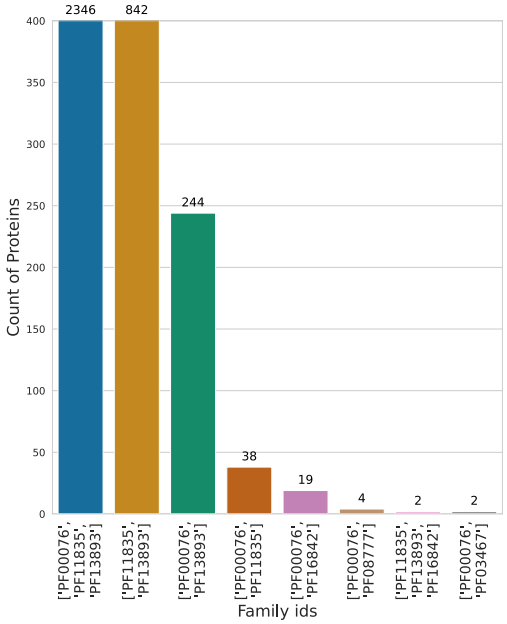
Figure 3.13: Distribution of single and multi-RRM proteins per RRM domain count

Figure 3.13 shows the distribution of proteins having single and multiple RRM domains. Majority of proteins contain a single RRM domains along with a few proteins (354) with more than 5 RRM domains. The "RPOLD domain-containing protein" (UniProt Accession: A0A2R6WJA9) has the highest number of RRM domains: 16. This protein has a 2143 amino acid long sequence and there is no experiment reported yet for any of the RRM domains from this protein. All 16 RRM domains from this protein correspond to RRM_1 Pfam family. From the figure 3.13, it is clear that RRM domains from some multi-RRM proteins (7714) correspond to different RRM families (yellow bars).

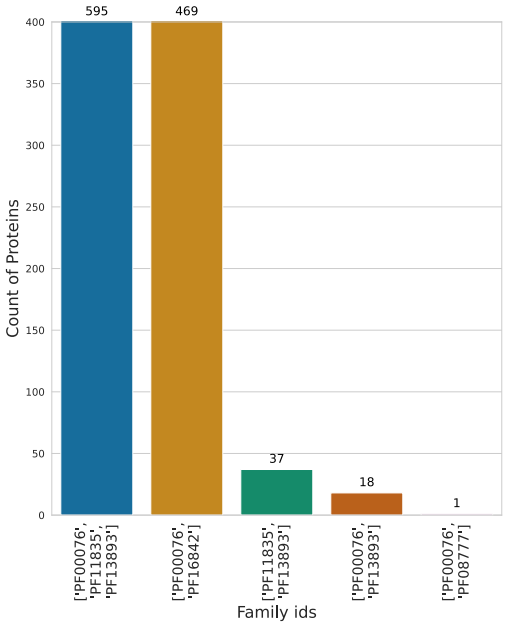
Although this number is a small fraction compared to proteins with multiple RRMs from the same family, a closer look at this might provide some insights into the evolutionary aspects of these RRM families.



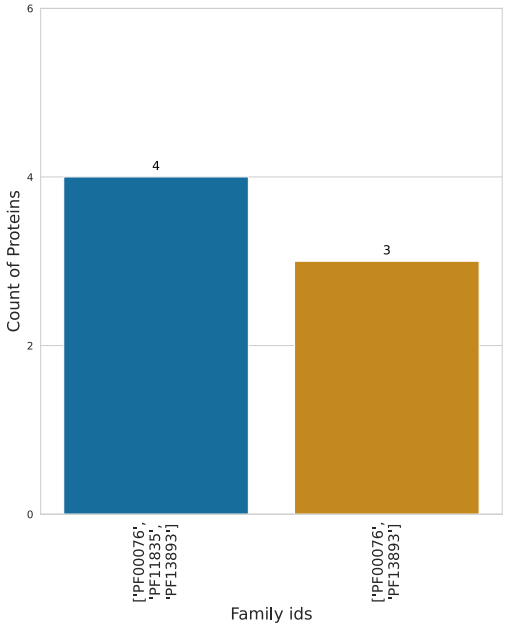
(a) Proteins with two RRM domains



(b) Proteins with three RRM domains



(c) Proteins with four RRM domains



(d) Proteins with five RRM domains

Figure 3.14: Distribution of proteins with multi-RRM domains from different Pfam families

Figure 3.14 shows the distribution of Pfam families among multi-RRM proteins. Each of the four subplot shows the Pfam families distribution across proteins with two, three, four and five RRM domains respectively. The two most abundant RRM families ('PF00076' and 'PF13893') from sequence and structural points of view can be found together within the same protein in all these categories. These two

families are together even in the proteins with 6 and 8 RRM domains. This indicates the closeness of these two families compared to rest of the RRM families. The phylogenetic analysis of all RRM instances can provide more details into this and help us to understand the evolution of RRM domains in order to know the functions and relations between multi-RRM proteins.

All the experiments in InteR3M have used 311 different nucleic acid sequences (ligand instances) to test their binding capacity with RRM domains. Of these 311 ligand instances, 190 have been used in structural experiments, 126 have been used in non-structural experiments like ITC, SPR, etc and 5 of them are used in both types of experiments.

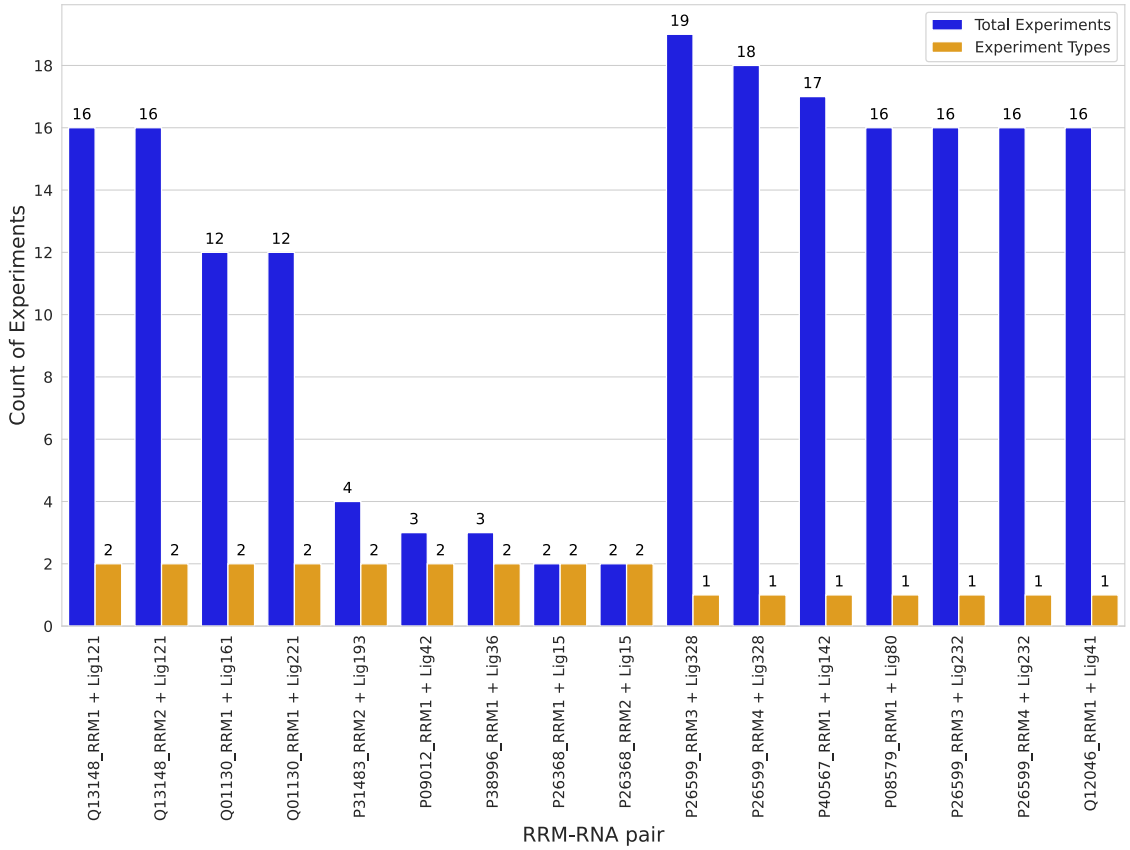


Figure 3.15: Most studied RRM-ligand pairs. For simplicity, we showed RRM-RNA pairs with multiple experiment types or with experiment count >15.

Figure 3.15 shows the distribution of experiment types and experiment count per RRM-ligand pair. This plot tells us about the most studied RRM-ligand pairs. The pairs of 'Q13148_RRM1/2' and 'Lig121' have been used in two different types of experiments and in a total of 16 experiments. InteR3M has a 3D structure for the complex of these pairs with PDB identifier '4BS2'. These pairs have been also tested with point mutations on protein using ITC experiments. Both of these RRM instances correspond to the protein named 'TAR DNA-binding protein 43' whereas the ligand has a GU-rich sequence 'GUGUGAAUGAAU'.
The RRM domain instances 'P26599_RRM3' and 'P26599_RRM4' have been used with the ligand 'Lig328' in 19 and 18 experiments, respectively. Both of these RRM instances, as well as several mutated versions, have been tested for RNA

binding capacity using 'filter binding assay'. Thus, InteR3M provides binding information resulting from checking the effect of point mutations in 'Polypyrimidine tract-binding protein 1' on the binding of the 185-nucleotides long RNA, referred to as 'Lig328'.

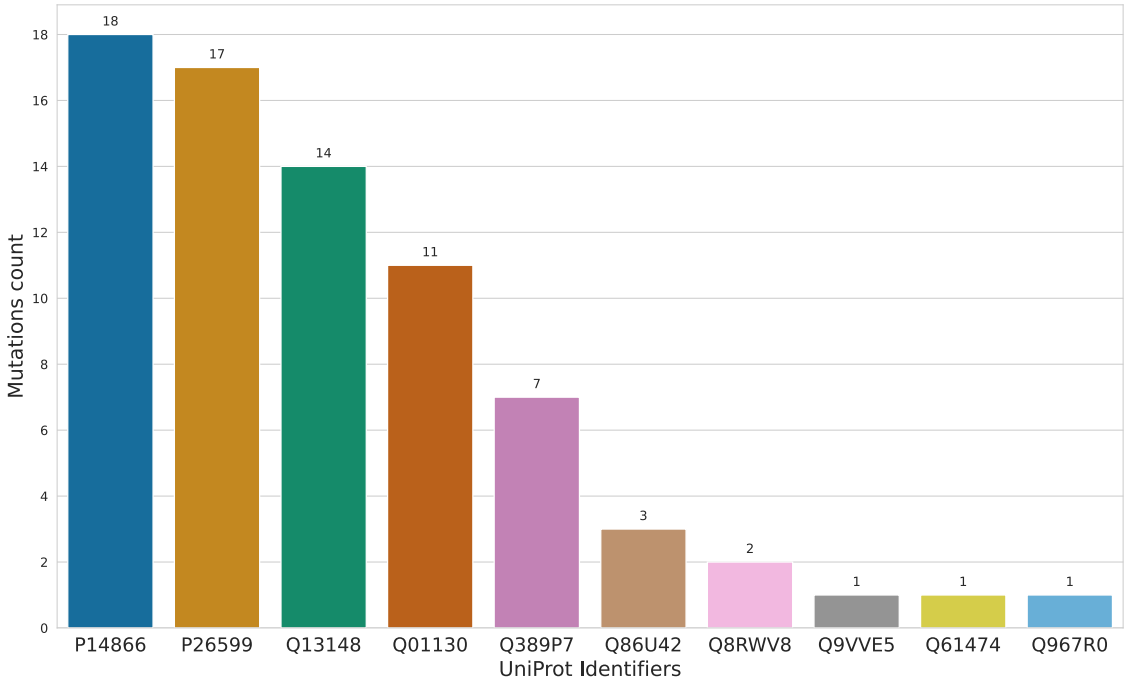


Figure 3.16: Number of mutations studied per protein containing RRM

InteR3M also stores the information about mutations performed on proteins to check their effect of these mutations on the RNA binding capacity. Figure 3.16 shows total count of mutations in proteins used to test the RNA binding. The 'Heterogeneous nuclear ribonucleoprotein L' protein with UniProt identifier 'P14866' has the highest number of mutations tested for RNA binding. This protein has 4 RRM domains, and mutation positions span the entire protein. These mutations were performed in this protein to check and identify the RNA binding surface of 'hnRNP L RRM1' and 'hnRNP L RRM23'.

In this example, one mutation refers to either point mutation or multiple mutations performed in a protein for the same experiment. These mutations can be either from or outside the RRM domain region.

The curation of information pertaining from various primary sources and its integration in a well designed domain-specific database such as InteR3M, have been essential in obtaining all these results. Such a database provides in-depth information about the targeted protein domain, which helps to better understand and characterize this domain.

3.8 Discussion & Conclusion

The domain-centric databases provide detailed information on specific domains of interest. Although there exist several generalist domain databases like Pfam, CATH,

CDD, InterPro etc. these databases provide only the very first basic information for domain-based analyses. Most of these generalist domain databases detect domains in an automated or semi-automated way. It is difficult to find all domain instances for a given domain type correctly using the same rationale. For example, there are five Pfam families that are reported as members of the RRM clan but do not have the RRM fold to be considered as RRM domain families.
Thus, it is important to study the protein domains in a domain-centric way and report newly generated information or any wrongly classified domains back to the generalist domain databases so that they can improve their classification.

Domain-centric databases can play important roles in developing the domain-centric approaches in different fields like functional studies of mammalian genes, characterization and identification of pathogenic viruses, and phylogenetic analysis [Phan et al., 2018]. One of the main advantages of domain-centric databases is that the domain instances they contain are manually curated and validated by developers of these domain-centric databases. Therefore, the users have more confidence on domain instances retrieved from domain-centric databases rather than generalist domain databases. Undoubtedly, the resulting domain instances depend on the rationale used to develop these databases and users should be aware of it before starting any downstream analysis.

We are making the code publicly available to facilitate either the creation of domain-centric databases similar to our InteR3M database or a few tasks from the whole process like contacts computation. The code used for data collection and update of the InteR3M database is publicly available and can be accessed at https://gitlab.inria.fr/hdhondge/data_collection_inter3mdb.

The InteR3M database is publicly available and can be accessed at https://inter3mdb.loria.fr/.

The major problems faced during the InteR3M database development were in the initial stages especially to delineate the correct set of RRM families. The choice of a generalist domain database as a primary source was very important and Pfam was selected based on collected use cases. Although not all Pfam families from the RRM clan have RRM fold, Pfam provided a good starting point to investigate further.

Currently, InteR3M have domain family information only from Pfam (sequence-based) database, but we would like to add the annotations from a structure-based classification like CATH or SCOP. The addition of information from structure-based domain classification would be helpful as it might add a few domain instances missed by Pfam and provide structural insights on domains from different Pfam families. This inspired us to develop a cross-mapping approach between Pfam (sequence-based) and CATH (structure-based) classifications presented in next chapter.

Chapter 4

CroMaSt: A workflow for assessing domain classification by cross-mapping of structural instances between protein domain databases

Summary

4.1 Introduction **65**

4.2 Approach **66**

4.3 Methods **70**

4.3.1 Selection of data sources 70

4.3.2 Retrieve the domain structural instances 70

4.3.3 Compare sets/lists of domain structural instances 71

4.3.4 Cross-mapping of the unique domain structural instances 71

4.3.5 Computation of average structures 72

4.3.6 Structural alignments 73

4.3.7 Implementation 74

4.4 Results **75**

4.5 Discussion **79**

4.6 Conclusion **80**

4.7 Future Perspectives **80**

Note: The work presented in this chapter is published in Bioinformatics Advances at https://doi.org/10.1093/bioadv/vbad081.

4.1 Introduction

Most proteins are composed of one or more domains that can be identified at the sequence or structural level. A protein domain is generally defined as a conserved structure identified by conserved residues in a multiple sequence alignment across different types of proteins often sharing similar functions. Moreover, it is generally assumed that the protein domains can fold independently from the rest of protein [Batey and Clarke, 2008]. Conceptually, a protein domain can also be considered as an abstract class whose definition/pattern is first induced from the analysis of certain occurrences of this domain (instances of the class) in real proteins. This definition/pattern is then used to discover new occurrences in new proteins. A domain can be associated with a specific, definite function in the proteins, making it the basic unit for understanding protein function and building synthetic proteins with given functions. There are several resources available that provide information about the protein domains and their classification [Wang et al., 2021]. The domain databases can be grouped into three categories depending on the rationale used for classification which can be either primarily sequence-based, primarily structure-based or integration-based (Section 2.2).

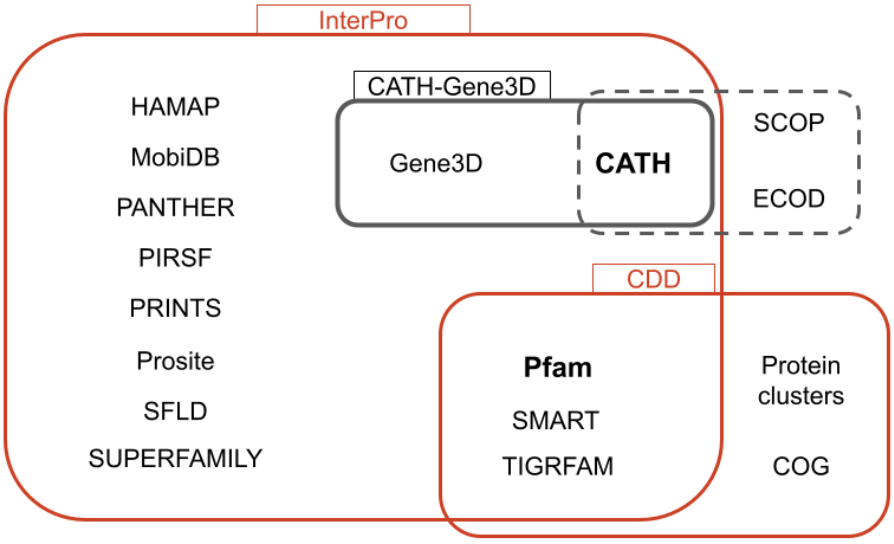


Figure 4.1: General landscape of domain databases. In red are the integrated domain databases (InterPro and CDD). CATH, SCOP and ECOD are structure-based domain databases (enclosed in dashed line) and the rest are sequence-based domain databases.

Figure 4.1 summarizes the general landscape of domain databases. In red are the integrated domain databases: InterPro [Blum et al., 2021] and CDD [Lu et al., 2020]. CATH [Sillitoe et al., 2021], SCOP [Andreeva et al., 2020] and ECOD [Cheng et al., 2015] are structure-based domain databases (framed with a dashed outline) and the rest are sequence-based domain databases. Pfam [Mistry et al., 2021] and CATH (in bold font) are used in this study.

All these domain databases constitute valuable and complementary sources of knowledge about domain families and can be used to investigate structure-function relationships in proteins. However, several problems arise when one begins to examine a particular type of domain across multiple domain databases. First, the domain families have different names in different databases and are not consistently mapped to each other. Second, domain families corresponding to the same type of domain in two different databases may not contain the same domain instances. Finally, for a given domain instance present in two different domain databases, the domain boundaries on the sequence (start and end residues) may be different.
Such difficulties are particularly deleterious for domain-centric investigation in the frame of synthetic biology and protein design. Indeed, such projects require a very precise knowledge about existing instances of a given type of domain associated with a specific function, in order to be able to engineer synthetic versions of this domain without losing the function. In practice, the exhaustive enumeration of all true domain instances of a certain type of domain is a complex problem and cannot be solved by querying a single domain database. To solve these issues, we propose here a generic iterative approach aiming to clarify domain definition by cross-mapping of domain structural instances between domain databases.

As a use case to develop and test our approach, we use the RNA Recognition Motif (RRM) domain, described in Chapter 2, section 2.3.1, with the goal to integrate all existing information about this domain, including all available experimental 3D structures.

Inconsistencies regarding RRM domain families appear with very simple searches inside a database (RRM Pfam families can be found outside the RRM clan) and between databases (16 Pfam families versus 2 CATH superfamilies). There are 16 Pfam families (13 from RRM clan) with RRM in their name. When the CATH database was searched using keyword 'RRM', several matching CATH domains were retrieved from and outside of the 'RRM' superfamilies (see query details in Appendix B.5.2).

Thus, different classification systems provide very different results for a given type of domain such as RRM. Although the RRM domain appeared well characterized in early studies, a great diversity exists today among domain families referring to RRMs. This raises the question of intra-domain allowed diversity: which domain instance is a "true RRM" and which one is not. At this stage, we decided to answer to this question from a 3D structure point of view and we designed a generic systematic approach: CroMaSt that classifies all structural instances of a given domain into 3 different categories (core, true and domain-like). We first describe our approach including our cross-mapping concept. The methods section then details the implementation of the CroMaSt generic workflow and is followed by the results section with RRM as a use case. Finally, we will discuss the possible usage of the CroMaSt workflow.

4.2 Approach

Our objective is to clarify the membership of structural domain instances to a given type of domain, often represented by a set of domain families in domain

databases. The CroMaSt workflow is designed in such a way that it uses the domain definitions from two well-known and widely used domain databases, Pfam (sequence-based) and CATH (structure-based). In our running example, we will consider the RRM domain type, which has been defined above, and is mostly represented by the CATH 3.30.70.330 superfamily (RRM domain) and the Pfam PF00076 domain family (RRM_1, RNA recognition motif). Two questions must be solved: (i) which other domain families in CATH or Pfam also represent the RRM domain type, and (ii) are there domain structural instances that are misclassified as RRMs in the RRM domain families.

Our generic approach assumes that there exists a consensus basic definition of the considered domain type, with a few validated structural instances (StIs), qualified as "true" StIs for this type of domain. Then, our first hypothesis is that it is possible to verify that a given domain StI belongs to the considered type of domain, by running structural alignment with true StIs. The result can be either manually inspected or automatically filtered using appropriate thresholds defined by the experts of this type of domain. However, this task can become tedious in view of the ever increasing number of StIs. Therefore, one needs to rely on existing domain classifications, in particular those with wide-coverage such as CATH and Pfam. Let CATH-rep1 and Pfam-rep1 be two most representative domain families in CATH and Pfam databases respectively. Our second hypothesis is that if a domain StI belongs to both CATH-rep1 and Pfam-rep1, then it is likely to be a true StI. If a domain StI is only present in CATH-rep1 (respectively in Pfam-rep1), it can be relevant to "cross-map" it to another domain family in Pfam database (respectively in CATH database), and to check all the StIs of this new domain family that may become a new representative family of the considered domain type.

Cross-mapping is the process of finding/locating an instance from one resource in another one. In our case, it refers to the process of finding a domain StI from Pfam in CATH or from CATH in Pfam.

CATH-rep1 to be cross-mapped

CATH-rep1 : '2DNL,A,1,3.30.70.330,427,515,Q8NE35,441,529'

This StI is in the format of: *'PDB_id, Chain_id, Domain_order_number, Family_id, PDB_start, PDB_end, UniProt_id, UniProt_start, UniProt_end'*.

For example, to cross-map the above domain StI (CATH-rep1) from CATH in Pfam database. I will search the CATH database until I find the match for this StI, in this case Pfam-rep2 from the CATH StIs box. Now we have successfully cross-mapped the CATH-rep1 in Pfam to Pfam-rep2. The domain family corresponding to this StI (Pfam-rep2) become a new possible candidate family of the considered domain type.

Pfam StIs

...

Pfam-rep1 : '2DIS,A,RRM_1,PF00076,11,78,A0AV96,153,220'
Pfam-rep2 : '2DNL,A,RRM_7,PF16367,426,515,Q8NE35,440,529'
Pfam-rep3 : '2MSS,A,RRM_1,PF00076,Q61474,111,180,111,180'
Pfam-rep4 : '4CQ1,F,RRM_5,PF13893,337,434,Q9UKA9,337,434'
Pfam-rep5 : '5X3Z,A,0,3.30.70.330,105,200,Q61474,109,200'

...

These StIs are in the format of: *'PDB_id, Chain_id, Domain_name, Family_id, PDB_start, PDB_end, UniProt_id, UniProt_start, UniProt_end'*.

At the end of the process, three categories of domain StIs are produced (Table 4.1): the "Core" category groups all domain StIs that are present in starting families and shared between two considered databases, the StIs from core category are used to build the 'core average structure', the "True" category groups all domain StIs that are cross-mapped to families distinct from starting families while displaying significant structural similarity with the core average structure, and the "Domain-like" category groups all domain StIs that could not be cross-mapped but display a significant structural similarity with core average structure.

More precisely our CroMaSt workflow is described in Fig. 4.2. Our approach starts with a single domain family (or a list of domain families) representing the domain type of interest in each of Pfam and CATH domain databases. We then filter all StIs from these families and compare the lists obtained from each domain database. The instances common to both Pfam and CATH families are included in the 'True domain' list and named the 'Core domain' set at the first iteration of the workflow. An average structure is computed for the set of core domains and named 'Average core structure'. The rest of the instances that are either specific to Pfam or specific to CATH are "cross-mapped" in the other database to fetch possible additional CATH or Pfam (respectively) families for the same type of domain. For all newly fetched families, an average structure is computed at the family level with cross-mapped StIs. After verifying that the structural alignment of the new family members with the average core structure exceeds a certain quality threshold, the newly found family is added to the list of families at the beginning of the workflow and a new iteration is started. At each iteration, a certain number of domain StIs specific to each database remain unmapped in the other database. For each unmapped domain instance, the average structure is computed at the sequence instance level, i.e. all StIs corresponding to the same protein sequence. These averaged structures are checked by structural alignment with average core structure and classified as 'Domain-like' structures when the alignment score exceeds a given threshold. This iterative procedure is followed until no other families are found. Domain StIs that do not fall in any category (*Core*, *True* or *Domain-like*) are labelled as 'Failed domains'.

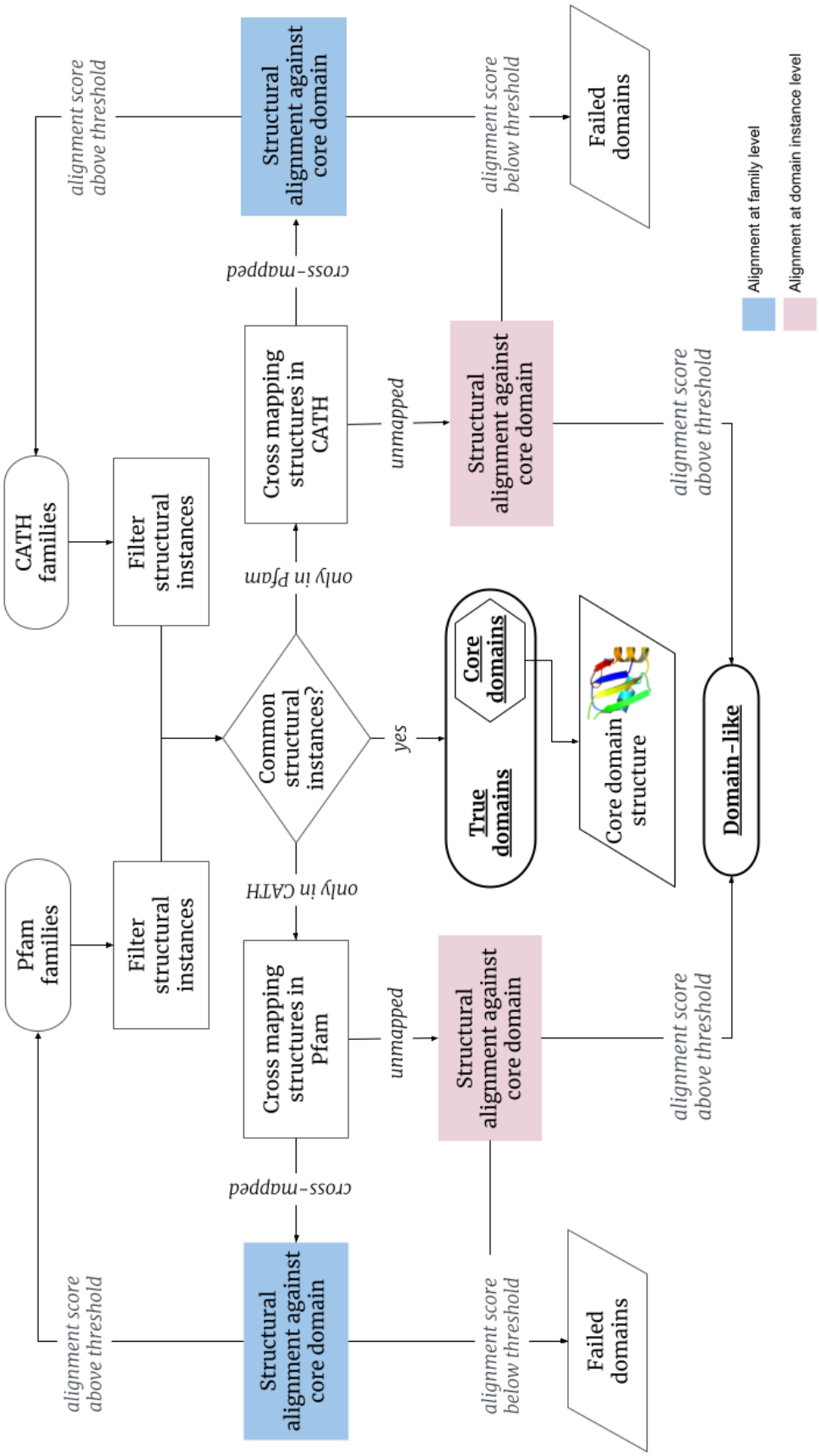


Figure 4.2: Conceptual model behind the CroMaSt workflow.

Table 4.1: Rationale behind classifying structural instances into categories.

Category	**Starting family member**	**Cross-mapped**	**Structurally well aligned**
Core	✓	✓	✓
True	-	✓	✓
Domain-like	-	-	✓

4.3 Methods

4.3.1 Selection of data sources

We selected two data sources to instantiate the CroMaSt approach, one sequence-based (Pfam) and one structure-based (CATH) domain databases. These two databases provide the raw files for each version of their update. The CroMaSt workflow uses these raw files. The workflow also uses Protein Data Bank (PDB) [Burley et al., 2021] and SIFTS (Structure integration with function, taxonomy and sequence) [Dana et al., 2019] resources for experimental 3D structures and residue-mapping between UniProt and PDB instances. More precisely, the files used by the workflow are as follows:

1. CATH domain description file - version 4.3.0 (Link)
2. File with Pfam-A matches for each PDB chain - version 33.0 (Link)
3. Obsolete PDB entries information (Link)
4. 3D coordinates of PDB structures
5. SIFTS mapping files

All above data files can be downloaded by running a tool provided with the workflow. This allows users to select the version of their choice for each source database.

4.3.2 Retrieve the domain structural instances

The workflow starts with at least one (super)family from each database. The structural instances (StIs) of a given domain refer to the parts of the PDB chains that contain an instance of this domain. This information can be extracted for given domain family IDs from Pfam and CATH databases but is not easily compared between the two sources as domain StIs retrieved from the Pfam database have start and end residues numbered according to UniProt, while these residues are numbered according to PDB when the domain StIs are retrieved from the CATH database. Therefore a residue mapping step is required to generate a unified representation that follows the form *'PDB_id, Chain_id, Domain_name or Domain_order_number, Family_id, PDB_start, PDB_end, UniProt_id, UniProt_start, UniProt_end'*. For example, a StI shared in Pfam and CATH has the following representation:

- '1CVJ, G, RRM_1, PF00076, 13, 83, P11940, 13, 83' in Pfam, and
- '1CVJ, G, 01, 3.30.70.330, 11, 87, P11940, 11, 87' in CATH.

This unified format facilitates identification of shared StIs between the two databases, taking into account the possible differences between start and end positions in the two data sources. Moreover, this format allows to save the start and end positions from the PDB chain in order to extract the 3D coordinates of the domain StIs for computing structural average or for structural alignment purpose.

The first step of the workflow consists of retrieving all StIs corresponding to given domain family IDs from each database. Domain length is used as a threshold to filter the domain StIs to avoid getting extremely short structures. Then, the start and end residues of all domain StIs are mapped to their corresponding PDB (for StIs deriving from Pfam) or UniProt (for StIs deriving from CATH) numbering. The SIFTS resource is used for this residue mapping. The workflow also keeps track of the domain StIs for which the residue mapping can not be done (obsolete or inconsistent entries). The complete step for retrieving the domain StIs is depicted in Fig. 4.3.

4.3.3 Compare sets/lists of domain structural instances

At this stage, the CroMaSt workflow compares two sets of domain StIs (one from Pfam, one from CATH) using the PDB and chain IDs and the start and end residue positions on the UniProt sequence. The workflow takes into account that the two databases can provide different lengths for the same domain instance, allowing for a difference between start (respectively end) residues for the same domain instance. By default, we use a maximal difference of 30 residues, but the user can modify this parameter according to their needs. This consideration of possible different lengths between various domain databases is important as it allows more than one source of information about a given domain to be covered.
This step of the workflow results in three different sets:

- Common domain StIs,
- Domain StIs unique to Pfam, and
- Domain StIs unique to CATH

The common domain StIs obtained at the first iteration of the workflow are the core domains and are used to compute the core structure of the domain (Step 4.3.5). The domain StIs unique to one database must be “cross-mapped” to the other database.

4.3.4 Cross-mapping of the unique domain structural instances

Each domain StI unique to one database (CATH or Pfam) is “cross-mapped” to the other database (Pfam or CATH, respectively). To do so, the PDB and chain IDs are used to query the domain description files downloaded from the data sources. When a hit is found, the start and end positions (from UniProt for Pfam database, and from PDB for CATH database) are checked with the same tolerated difference as in Step 4.3.3. If the cross-mapping is successful, the corresponding cross-mapped StI is created and its domain family becomes a new possible candidate for the ’true

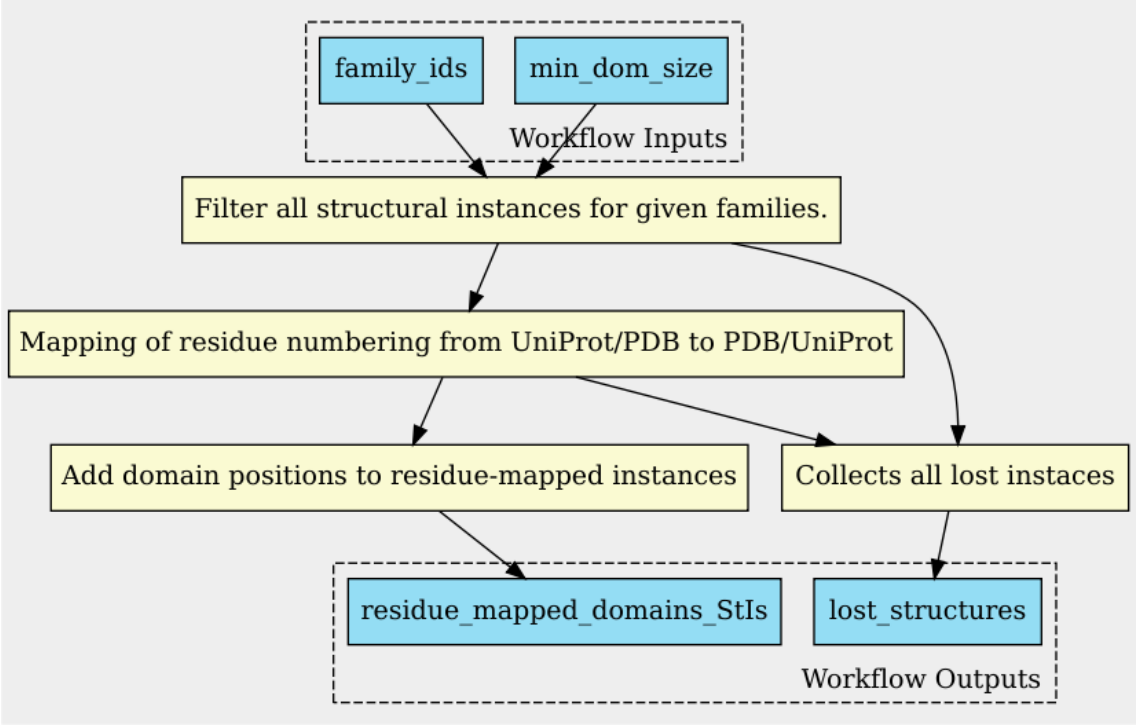


Figure 4.3: Process of filtering and residue-mapping structural domain instances. (Figure was generated locally using cwltool.)

domain' category. An average structure is computed at the family level from all StIs mapped to this domain family (see Step 3.5) and if the structural alignment test (see Step 3.6.1) is positive, this domain family is added to the set of domain families ready for the next iteration of the workflow. Whenever a new domain family is found and it undergo a structural alignment test (See Step 4.3.6) before being added to the set of domain families ready for next iteration of the workflow. and is stored for the next iteration. The corresponding cross-mapped StI is created. If the cross-mapping is not successful, the domain StI is stored as unmapped. In summary, this step provides three different sets of domain StIs as follows:

- Pfam StIs cross-mapped to new families in CATH
- CATH StIs cross-mapped to new families in Pfam
- Un-mapped StIs from both databases, these domain StIs are not classified or annotated in the other database

4.3.5 Computation of average structures

Family-level and instance-level average structure

The coordinates for average structure are computed using an in-house python script after aligning a set of 3D coordinates (extracted from PDB entries) using the Kpax program [Ritchie, 2016]. The resulting average structure only consists of the protein backbone without side-chains. When applied to a set of domain StIs, this computation could be biased towards the domain instances (defined by their UniProt sequence) having more StIs in the PDB. To avoid this, our workflow first

computes "instance-level" averages from all StIs corresponding to the same domain instance in UniProt. Then, these "instance-level" average structures are used to compute the "family-level" average structure as an 'average of averages'.

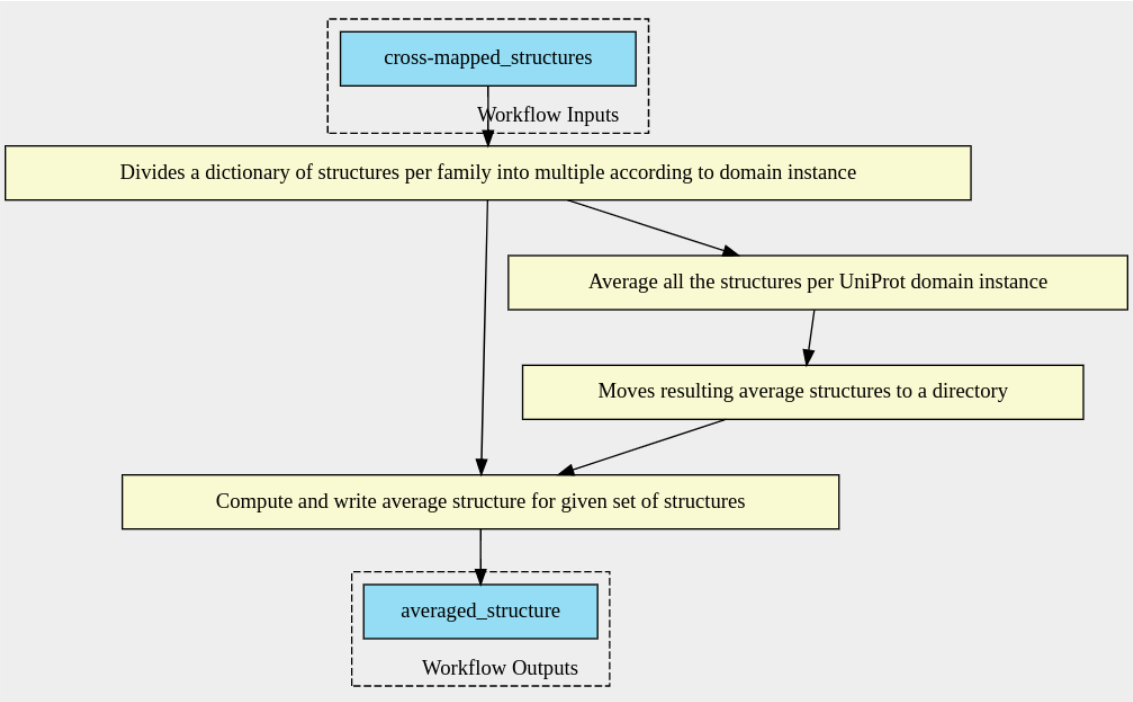


Figure 4.4: Steps to compute family-level average structure as the average of instance-level average structures (Figure was generated locally using cwltool).

Core average structure

Core domain StIs are the common domain StIs (from Step 4.3.3) retrieved during the first iteration. Core domain average structure is computed using these core domain StIs as an average of averages per domain instance as described in Step 4.3.5. It should be noted here that the core average structure of the domain is dependent on the families selected for the first iteration of the workflow.

Average structures for each cross-mapped families

An average structure is computed for each cross-mapped (newly found) family from all the unique domain StIs that are mapped to this family (Step 4.3.4). These average structures are computed at the "family-level" as described above (Step 4.3.5).

Average structures for un-mapped domain structural instances

Average structures are computed at the "instance level" for each unmapped domain instance produced at Step 4.3.4.

4.3.6 Structural alignments

The core average structure of the domain obtained at Step 4.3.5 plays a crucial role as a reference for domain membership in the structural alignments. The

workflow considers the core average structure as the prototype/template for the query domain. Thus, the core average structure is aligned against other structures to assess if the aligned structures are instances for the same domain type. For all structural alignments, the CroMaSt workflow uses Kpax as it provides a Gaussian-based Multiple Structural Alignments quality measure called 'M-score' [Ritchie, 2016], which circumvents the pitfalls of RMSD-based quality measures [Kufareva and Abagyan, 2011]. This step outputs a *csv* file with all the scores provided by Kpax for each target structure. By default, the CroMaSt workflow uses 'M-score' as the alignment score and 0.6 as the score threshold to evaluate the alignments. Users can choose to evaluate the alignments using any of the alignment score provided by Kpax (such as K-score, J-score, T-score and so on) with different thresholds.

Structural alignments for average structures of cross-mapped families

Each average structure for cross-mapped families from Pfam and CATH (Step 4.3.5) is aligned against the core average structure using Kpax. If the average structure for a cross-mapped family passes the given threshold (for alignment score), the family identifier is added to the list of families for next iteration of the workflow. In parallel, the StIs that were mapped to this new family are also kept for next iteration in order to get recognized as 'true domain' StIs when comparing lists from the two domain databases.

Structural alignments for un-mapped average structures

The un-mapped average structures (from Step 4.3.5) are also aligned in same manner like Step 4.3.6. The StIs corresponding to average structures passing the threshold (for given score) are included in the list of 'domain-like' StIs. The StIs corresponding to structures failing to pass the threshold, are considered as false positives and labeled as 'failed domains'.

4.3.7 Implementation

CroMaSt uses the common workflow language engine and the Conda package manager (to install the required dependencies). All the scripts are written in Python and wrapped in CWL [Crusoe et al., 2022] as cwltools. Although the developers of Common Workflow Language (CWL) are working on adding the functionality for loop from quite some time now, it does not support the loops yet. As the nature of the CroMaSt workflow is iterative, a new parameter file is created at the end of each iteration for next iteration. This step updates certain inputs from string types to file types along with updating the most important input parameters of family identifiers (from Step 4.3.6). The family identifiers are updated based on the alignment scores of average structures for cross-mapped families. FAIR principles were followed while developing workflow, which can be found on WorkflowHub [Goble et al., 2021] from doi: 10.48546/workflowhub.workflow.390.1. Details on how to use the CroMaSt workflow are provided on WorkflowHub and Git repository.

Table 4.2: Results from each step of CroMaSt, starting with Pfam family - RRM_1 (PF00076) and CATH superfamily - 3.30.70.330 (RRM (RNA Recognition Motif) domain).

Steps	**Iteration 1**		**Iteration 2**	
	Pfam	CATH	Pfam	CATH
Starting Families	1	1	14	0
StI filtered on domain length	1147	1527	96	80*
Obsolete and inconsistent entries	3	323	0	0
Residue-mapped StIs	1144	1204	96	-
Common StIs (**Core** & **True**)	886	886	80	80
Remaining StIs (not common)	258	318	16	0
Cross-mapped StIs	0	244	0	0
Properly aligned at family level	-	80	-	-
Not properly aligned at family level	-	164	-	-
Not cross-mapped StIs (unmapped)	258	74	16	0
Properly aligned at instance level (**Domain-like**)	255	74	15	-
Not properly aligned at instance level	3	0	1	-
Failed structures	3	164	1	0
New families found	14	0	0	0

*These StI entries are cross-mapped and properly aligned at the family level from the previous iteration.

4.4 Results

To run the CroMaSt workflow successfully requires:

1. At least one family identifier from each Pfam and CATH databases as input from user.

2. At least one common StI between the starting families from Pfam and CATH.

To demonstrate the capacity of the CroMaSt workflow to distinguish between true domain StIs and domain-like StIs and to list false positive and obsolete or inconsistent StIs, we apply CroMaSt to the RRM group of domain families.

We initiate the CroMaSt workflow with families PF00076 (RRM_1) from Pfam and superfamily 3.30.70.330 (RRM domain) from CATH. Table 4.2 shows the different results obtained at each step of the workflow. A total of 1147 domain StIs were filtered from RRM_1 Pfam family along with 3 inconsistent domain StIs, whereas 1527 domain StIs were filtered from CATH including 316 obsolete and 7 inconsistent domain StIs. Then, the 1144 and 1204 domain StIs from Pfam and CATH, respectively were residue-mapped. Out of all these residue-mapped domain StIs, 886 are shared between Pfam and CATH. Thus, 886 StIs constitute the core domain StIs, and are also included in the list of true domain StIs. Core average structure (Fig. 4.5 A.) for RRM domain was computed using these 886 StIs. From the remaining StIs (258 unique to Pfam and 318 unique to CATH), only 244 StIs from CATH were successfully cross-mapped to a total of 17 different Pfam families. Thus, average structures were computed for these 17 newly found Pfam

families using the cross-mapped StIs. After aligning these average structures against the core domain average structure, 14 of them passed the threshold allowing to include these families at the beginning of the next iteration. The remaining 3 families and their corresponding StIs (164) were considered as failed domain StIs. Thus, only 80 StIs from the 244 CATH StIs crossmapped to Pfam were kept for the next iteration.

The average structures were computed at the 'instance-level' for all un-mapped StIs from Pfam (258) and CATH (74). After the alignment of these average structures against the core domain average, only 3 StIs from Pfam failed to pass the threshold. Thus, all remaining StIs (255 from Pfam, and 74 from CATH) constitute the 'domain-like' StIs. In summary, the first iteration resulted in a total of 886 core domain StIs, 329 domain-like StIs, and 167 'failed domain' StIs, as well as 14 Pfam families and 80 CATH StIs ready for next iteration.

The second iteration started with the 14 Pfam families and 80 StIs from CATH. A total of 96 StIs were filtered from the 14 Pfam families with no inconsistent or obsolete entry. These two sets shared 80 StIs (true domain StIs) and the other 16 StIs from Pfam remained un-mapped in CATH. Nearly all of them (15/16) passed the alignment threshold leading to 15 domain-like StIs and 1 failed domain StI. Thus, at the end of the second iteration, no new family was found, hindering any further iterations of the workflow.

The CroMasSt workflow keeps track of all obsolete and inconsistent domain StIs which are detected mostly at the residue mapping step based on SIFTS. To illustrate that, we list here the inconsistent StIs encountered throughout the workflow (obsolete StIs are listed in Appendix B.5):

- "6DG0,B,RRM_1,PF00076,Q22039,230,294",
- "6DG0,A,RRM_1,PF00076,Q22039,230,294",
- "6PAI,D,RRM_1,PF00076,Q7Z3L0,95,165",
- "2KU7,A,00,3.30.70.330,1,140",
- "3DXB,F,02,3.30.70.330,452,556",
- "3DXB,G,02,3.30.70.330,452,556",
- "3DXB,D,02,3.30.70.330,452,556",
- "3DXB,B,02,3.30.70.330,452,556",
- "3DXB,A,02,3.30.70.330,452,556",
- "4V19,X,00,3.30.70.330,2,150"

The domain StIs with PDB ID '6DG0' (chains A and B) are associated with UniProt ID G5ECJ4 in PDB and with UniProt ID Q22039 in Pfam. Similarly, the domain StI present in PDB ID '6PAI' (chain D) is associated with UniProt ID Q14498 in PDB and with UniProt ID Q7Z3L0 in Pfam. Regarding the domain StI present in PDB ID '2KU7' (chain A), the start and end residues from the PDB entry are mapped in SIFTS to two different UniProt IDs: Q03164 and Q9UNP9. The situation is the same for all domain StIs in PDB ID '3DXB', the start and end residues map to

UniProt IDs, P0AA27 and Q9UHX1, respectively. Finally, the domain StI in PDB ID '4V19' (chain X) has no annotation for any UniProt entry in SIFTS.

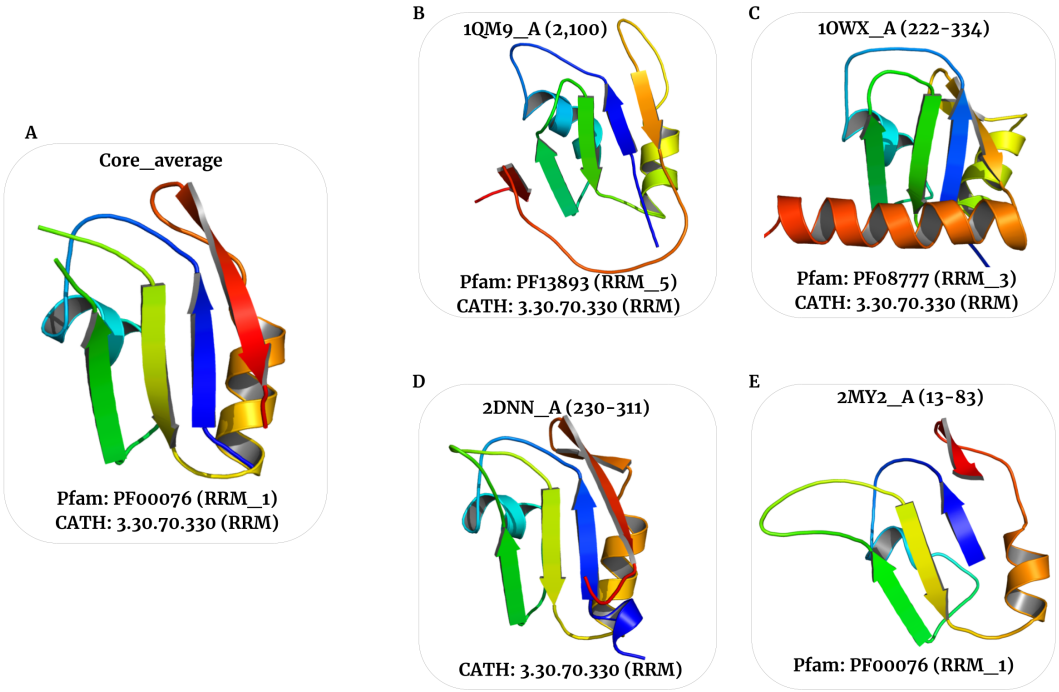


Figure 4.5: RRM domain structural instances A. Core average domain structure B. and C. True domain StIs unique to CATH at first iteration and successfully mapped to two different Pfam families. D. and E. Domain-like StIs from CATH and Pfam, respectively.

In summary, the CroMaSt workflow identified 966 true domain StIs among which 886 are core domain StIs, 344 domain-like StIs and 168 failed domain StIs. In terms of domain families, the CroMaSt workflow explored a total of 19 families (18 from Pfam and 1 from CATH, including the starting families) for the RRM type of domain, and 16 of them (15 from Pfam and 1 from CATH) qualified for the RRM domain type (Table 4.3). The structural alignment score for the other 3 families failed to pass the given threshold (Appendix B.1). Interestingly, CroMaSt detects in Pfam one DUF domain (Domains of Unknown Function), DUF1866, that can now be associated to an RNA binding function. In total, this run of CroMast workflow lasted approximately 54 minutes on a machine equipped with 8 2.40 GHz Intel(R) Xeon(R) Silver 4214R processor without any prior downloads of PDB and SIFTS entry files.

Figure 4.5 shows some of RRM domain StIs with the core average domain structure resulted from the CroMaSt workflow. Figure 4.5 B and C have some extensions after the RRM domain topology in the form of β sheet and α helix, respectively. These are the variations and extensions of the RRM domain studied previously. This is in good agreement with the variations described by Maris et al. [2005]. In addition, Fig. 4.5 D has a β sheet within the loop5 that is also found in many domain StIs from PF00076 (RRM_1) Pfam family and 3.30.70.330 (RRM) CATH superfamily.

One failed domain StI from each database is shown in Fig. 4.6. In both cases, one can observe that the topology of secondary structural elements is clearly different from the one of the average core structure.

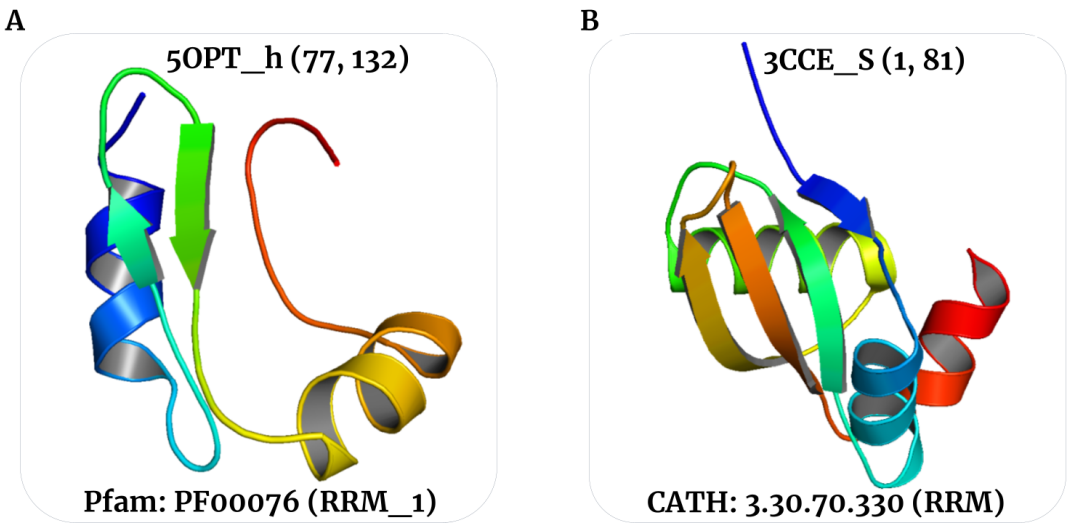


Figure 4.6: Failed domain StIs from Pfam PF00076 domain family (A) and from CATH superfamily 3.30.70.330 (B).

Table 4.3: List of 19 domain families explored by the CroMaSt workflow

Family IDs	**Family Name**	**Database**	**Passed alignment threshold?**
PF00076	RRM_1	Pfam	NA*
PF00276	Ribosomal_L23	Pfam	No
PF00511	PPV_E2_C	Pfam	No
PF01282	Ribosomal_S24e	Pfam	No
PF03467	Smg4_UPF3	Pfam	Yes
PF03880	DbpA	Pfam	Yes
PF04847	Calcipressin	Pfam	Yes
PF05172	Nup35_RRM	Pfam	Yes
PF08675	RNA_bind	Pfam	Yes
PF08777	RRM_3	Pfam	Yes
PF08952	DUF1866	Pfam	Yes
PF09162	Tap-RNA_bind	Pfam	Yes
PF11608	MARF1_RRM1	Pfam	Yes
PF11835	RRM_8	Pfam	Yes
PF13893	RRM_5	Pfam	Yes
PF16367	RRM_7	Pfam	Yes
PF16842	RRM_occluded	Pfam	Yes
PF17774	YlmH_RBD	Pfam	Yes
3.30.70.330	RRM (RNA recognition motif) domain	CATH	NA*

*Starting domain families from Pfam and CATH. NA: Not Applicable.

4.5 Discussion

The multiplicity of biological databases and the lack of systematic cross-references between them lead to important issues in data integration and consistency. Protein domain databases are no exceptions. Here, we show the way to increase interoperability between protein domain databases using a cross-mapping approach. Our CroMaSt method constitutes a systematic, reproducible and automated solution to retrieve domain StIs corresponding to a certain type of interest for the user. Along with this, CroMaSt also points out some irregularities in the databases.

It should be noted that the results returned by CroMaSt are determined by the starting domain families used as input. In fact, the set of core domain StIs is dependent on the starting domain families. This influences the computation of 'core domain average' structure and in turn all other results. For instance, we tested CroMaSt using PF13893 (RRM_5) and 3.30.70.330 (RRM domain) for Pfam family and CATH superfamily, respectively. Although CroMaSt explored the same families as described in our running example (Table 4.3), only 11 of them passed the threshold allowing them to be included in the next iteration. Thus, with these different starting domain families, CroMaSt returned a slightly different core average structure (RMSD-aligned: 0.96Å), 962 true domain StIs including 36 core domain StIs, 327 domain-like StIs and 183 failed domain StIs (see Appendix B.4). This strong dependency of CroMaSt on input domain families has at least two practical consequences. Firstly, it is recommended to start the workflow with the two most populated domain families in order to get the most exhaustive results. It is also not forbidden to start with more than one family from one or both databases. Second, this feature can be used to explore particular domain families within a given type of domain in order to characterize possible subtype average core structures.

The CroMaSt workflow can be easily applied to any structural domain different from RRMs by providing respective family identifiers to the workflow. Manual expertise of the domain of interest is useful (not necessary) to start the workflow with correct families and to inspect the structural alignments performed by the CroMaSt workflow. Ideally, the StIs from starting families should have a good quality structures with no chain breaks within the domain range. As mentioned above, it is recommended to initiate the workflow with the most populated families for domain of interest for better results, i.e. to get the core average structure that represents as many StIs as possible. The alignment threshold can be changed depending on the structural versatility of the domain type. If the user is unsure about the threshold, running the first iteration of the CroMaSt workflow with the default threshold is recommended. After the first iteration, all the structures are available for manual inspection, in order to decide the threshold according to the needs.

The CroMaSt workflow allows individual users to contribute to the Pfam and CATH protein domain database curation by (i) providing annotations to domain families lacking any annotation (e.g. the DUF domains in Pfam), (ii) detecting irregularities in domain annotations within databases, (iii) removing StIs wrongly assigned to a domain family and (iv) pointing to discrepancies occurring among databases. Thus, the CroMaSt workflow can bring a precious help in curating and

updating domain-specific databases. This way, users can take advantage of the strengths of both databases while trying to reduce the limitations of each. For example, the CATH database classifies structures based on the arrangement of secondary structures; but sometimes, two different domains might have the same topology, like the RRM domain and PPV_E2_C (PF00511) domain. The PPV_E2_C domain does not have any RNP sequence (Appendix B.1). We did check how other structure-based domain databases classify domain-like and failed domain StIs resulting from CroMaSt, and found that there are some domain instances missing in each of them (more details in Appendix B.2). We believe that this problem of inconsistencies can be solved by the cross-mapping approach. Currently, CroMaSt takes Pfam and CATH as sources for domain databases. However, these sources can be expanded to include other domain databases as well. The inclusion of different databases will facilitate interoperability between them. While developing the CroMaSt workflow, we considered only primary domain databases that do not take references from other domain databases during their classification procedure. As the number of source databases increases, the thresholds used (domain length difference, alignment score thresholds) in the workflow should not be too strict allowing domain definitions from other databases to be considered.

4.6 Conclusion

We built CroMaSt: "Cross-Mapper for Structural domains", a fully automated workflow that classifies all structural instances of a given domain into 3 different categories: core, true and domain-like. We show that CroMaSt can be used successfully to compute the prototype structure of a given type of domain, here the RRM domain. Thus, we used experimental 3D structures to clarify domain definition while addressing the inconsistencies within and between domain databases (Pfam and CATH). In addition, CroMaSt produces multiple structural alignments, that can provide new information about conserved and variable residues, loops or SSEs in the domain instances. This information could be readily used in protein design, for building synthetic proteins with the right domain-like properties.
We demonstrated the usage of CroMaSt workflow using 'M-score' from Kpax. In addition to the M-score, CroMaSt can use any other alignment score provided by Kpax like RMSD and T-score (TM-score defined by the TM-Align program).
The workflow is available from GitLab (https://gitlab.inria.fr/capsid.public_codes/CroMaSt) and WorkflowHub (doi: 10.48546/workflowhub.workflow.390.1).

4.7 Future Perspectives

This chapter has introduced the concept of "cross-mapping" domain StIs from two different protein domain classification systems, Pfam and CATH. Based on this idea of cross-mapping, we built the CroMaSt workflow to solve the inconsistency issues of domain StIs classification systems.

With the current approach of CroMaSt workflow, the time taken to execute workflow

is dependent on the number of considered domain StIs, both from starting families and from cross-mapped families at each iteration. There are two time consuming but very important steps in the CroMaSt workflow, residue-mapping between UniProt and PDB (Section 4.3.2), and computation of average structures (Section 4.3.5). Appendix B.5.3 provides a brief demonstration of 'residue mapping' process with an example. We will test running these steps on GPU to reduce the time taken by each of this step.

The CroMaSt workflow is designed to cross-map StIs between sequence-based Pfam database and structure-based CATH database. Thus, CroMaSt do not cover families without any StIs available (from sequential perspective). The structure-based domain databases like CATH and SCOP do not have any information of such families as there are no experimental structures available.
It is more likely that the domain databases like Pfam and CATH will integrate with the AlphaFold database providing representative structures for all families. Pfam have already started this initiative and from old website of Pfam, one could see the AlphaFold models for corresponding family.
CroMaSt workflow can be run with AlphaFold structures once the structure-based domain databases (CATH) have integrated information from AlphaFold database. The usage of AlphaFold models will give more power to CroMaSt workflow in terms of inclusion of the families with no structural coverage (from sequence-based classification) and addition of new families (from Structure-based classification).

We designed CroMaSt in a way to consider the fact that different databases have different policies and time period for updates. Thus, by allowing users to select the database versions we give complete control to the user. Inconsistencies in data might result from updating one database and not the other one. CroMaSt is designed to handle such inconsistencies by considering the inconsistent StI only from one database and then look for the structural similarity with core average structure. CroMaSt keeps track of the StIs that can not be processed and return the list of such StIs as one of the output.

Current version of CroMaSt supports cross-mapping of domain StIs from only two domain databases, Pfam (sequence-based) and CATH (structure-based). To open CroMaSt to any other source domain databases, we plan to create a unified format from release files of other domain databases. This will allow users to run CroMaSt with domain databases of their choice for cross mapping of StIs. The users can contribute to source domain databases in case of wrongly classified domain StIs or to assign classification for StIs without any classification.

Chapter 5

Data driven modeling of RRM-RNA complexes

Summary

5.1 Introduction **83**

5.2 Deciphering RRM-RNA Recognition Code **83**

5.2.1 General Approach 83

5.2.2 RRM Master alignment 83

5.2.3 Mapping Contacts onto the Alignment 85

5.2.4 Similarity among RRM-RNA complexes 87

5.2.5 Computation of scoring matrices 87

5.2.6 Discussion 88

5.3 Modeling 3D structures of RRM domains **89**

5.3.1 Methodology 89

5.3.2 Testing RRMpip 91

5.3.3 Results & Discussion 91

5.4 Modeling RRM-RNA complexes **93**

5.4.1 Anchored Docking: Definition and requirements 93

5.4.2 Extraction and Clustering of Anchoring Patterns 93

5.4.3 Resulting Anchoring patterns 95

5.4.4 Docking with Anchoring patterns 96

5.5 Evaluating RRM-RNA complexes **99**

5.5.1 Stability of an RRM-RNA complex 99

5.5.2 Free Energy Computation 110

5.1 Introduction

This chapter describes the usage of data from InteR3M database (Chapter 3) to decipher the RRM-RNA recognition code and to model 3D structures of RRM and RRM-RNA complexes followed by assessing the modelled 3D structures of RRM-RNA complexes.

5.2 Deciphering RRM-RNA Recognition Code

The work described in this section is a collaboration with the Bio2Byte group[1] from Vrije Universiteit Brussel (VUB). I did my secondment at Bio2Byte group from March 2021 to July 2021. During this period, I worked with another PhD student (Joel Roca-Martinez) from Bio2Byte group under the supervision of Prof. Wim Wranken. The main objective of this collaboration work is to decipher the RRM-RNA recognition code that will bring us a step closer to successfully design a RRM with desired RNA binding activity.

5.2.1 General Approach

The RRM domain is a well studied protein domain, even though a general recognition code between this motif and the RNA is not known [Auweter et al., 2006].

Figure 5.1 shows the data flow between different steps of this computational analysis and the green blocks represent my contributions in this collaboration work. We used the structural information about RRM, RNA and their interactions stored in InteR3M database to align RRM domains, RNAs, and compute RRM-RNA complex similarity among different RRM-RNA complex structures. Then we used an adaptation of GOR [Garnier et al., 1996] formula (Eq. 5.2.2) to compute the likelihood of a residue-nucleotide interaction at a specific position in RRM domain.
My major contributions in this work are in data curation and mapping of contacts on the alignment.

5.2.2 RRM Master alignment

The positions of interacting residues from RRM play an important role in the binding modes of RNA. They can be identified by aligning all the RRM structures having RNA binding information. Thus, we extracted all the available structural instances of RRM domains from InteR3M database, i.e. 1259 RRM structures. All the sequences corresponding to these RRM structures were extracted from InteR3M database and filtered with a sequence identity threshold of 99%. At the end, we retrieved 356 RRM sequences, and 314 entries were from a single Pfam family RRM_1 (PF00076) showing a strong bias towards this Pfam family.

[1] https://bio2byte.be/

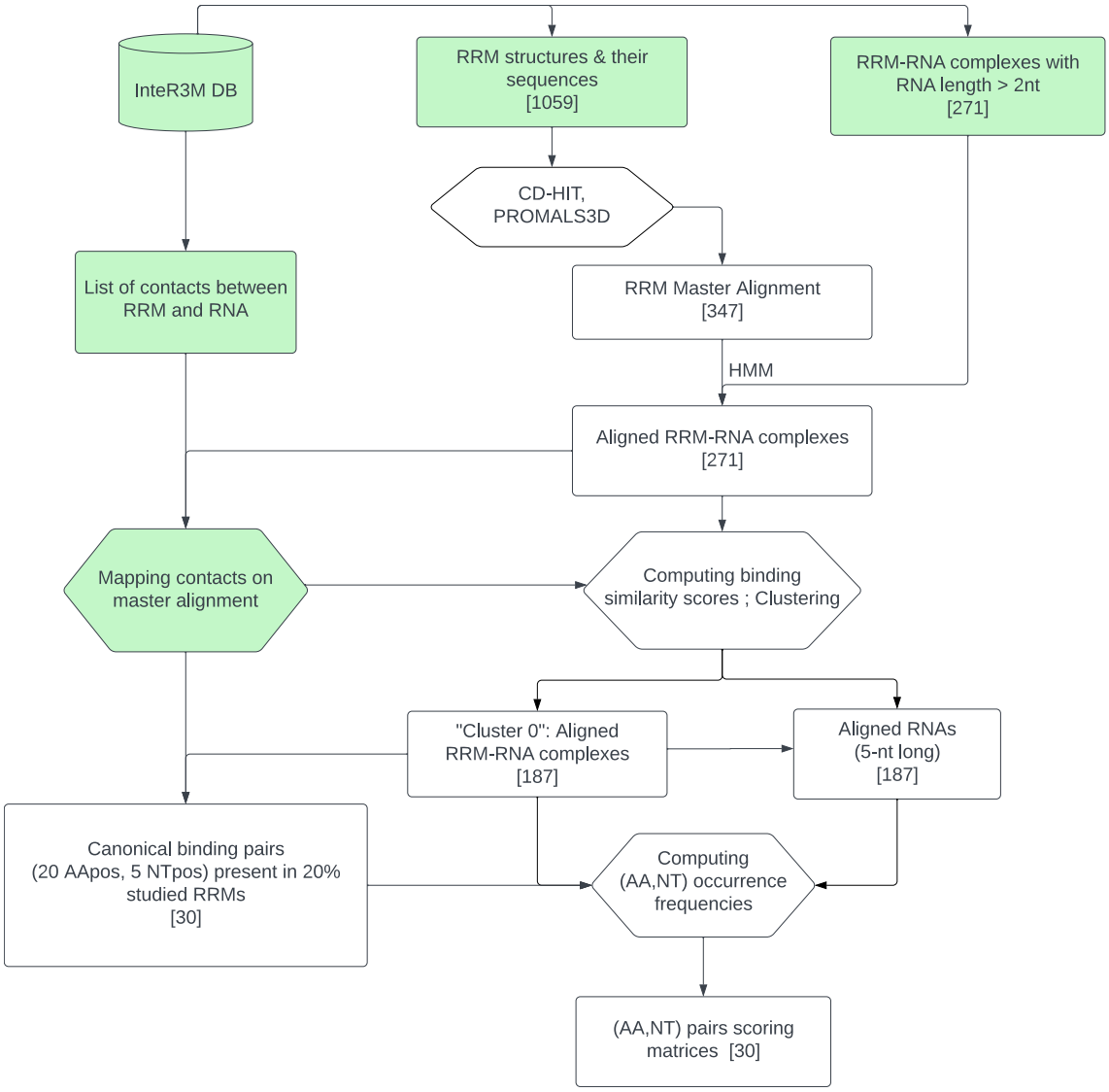


Figure 5.1: Data flow for computational analysis for deciphering RRM-RNA recognition code. The green blocks indicate my contributions in this work.

To overcome the bias, we used CD-HIT [Li and Godzik, 2006] to select 19 representative RRM sequences with a 30% sequence identity cut-off. PROMALS3D [Pei et al., 2008] uses sequence and structural information to generate high-quality multiple alignments. Thus, we used PROMALS3D to align these 356 RRM sequences. The structural information was used for the 19 representative RRMs and sequences were used for rest of the RRM domains.
After manually checking the resulting alignment, 9 sequences were removed from alignment because of unusually long β strands and/or α helices.

In this alignment, some extra residues were included at both the C- and N-terminal regions that might be relevant for RNA binding. All entries are identified by their UniProt code, RRM number, PDB Id. and chain, starting and ending positions of the RRM by PDB and UniProt, and UniProt starting and ending positions matching the sequence included in the file. The latter numbering helps to keep track of the extra added residues.

Extra information for RRM master alignment

After having the final alignment, it was a bit tricky to keep track of numbering of residues because all entries do not have the sequence only from PDB files. So we came up with a specific format for the headers in alignment and is as follows:
>UniProt-ID_RRM-no_PDB_chainid_PDBstart-PDBend_UniProtStart-UniProtEnd_UniProtStart'-UniProtEnd'
Where,

- PDBstart, PDBend are the start and end numbering for residues of RRM region in the PDB file/structure
- UniProtStart, UniProtEnd are the start and end numbering for residues of RRM region in the UniProt
- UniProtStart', UniProtEnd' are the start and end numbering for residues of RRM entry in the UniProt after adding 15 residues/linker at each terminal (if any)

Example:
>P19339_RRM2_1B7F_A_213-281_213-281_197-296

5.2.3 Mapping Contacts onto the Alignment

The detailed analysis of RRM-RNA contacts along the protein structure requires the mapping of interactions extracted from RRM-RNA complexes onto the RRM master alignment. This mapping can provide valuable insights into the structural and functional RNA-binding features of RRM domains and can help to design novel RRM with the desired RNA binding activity.

This mapping is not a straightforward task as the master alignment has some extra residues at the C- and N-terminal regions of RRM domains that may or not be present in the PDB entry. Therefore, we used SIFTS (Structure Integration with Function, Taxonomy and Sequence) resource for residue-level mapping between UniProt and PDB entries.

All the interactions between RRM and RNA were retrieved from InteR3M database for the 271 RRM-RNA complexes. Then using the SIFTS residue-level mapping these contacts were mapped onto the RRM master alignment by checking their corresponding residues from RRM domain entries.

The format of the final resulting file is as follows:

```
{
RRM_entry: {
  model_no: {
    alignment_position: {
      interacting_residue_pair: {
        ‘‘interactions’’: {list_of interactions},
        ‘‘stacking’’: boolean_value
        }
      }
```

```
    }
  }
}
```

Where,

- 'RRM_entry' represents the header of individual entry from the alignment
- 'model_no' represents the model number within the PDB entry (to differentiate models in NMR structure)
- 'alignment_position' represents the column position of the each residue from the alignment
- 'interacting_residue_pair' represents the interacting pair of amino acid and nucleotide with their position in PDB entry
- 'interactions' is the key with which we are storing all the interactions at atomic level
- 'list_of_interactions' is the list of all the interactions extracted for the *interacting_residue_pair*
- 'stacking' is the key we use to store information about stacking interactions in boolean form
- 'boolean_value' is True when there is stacking interaction between *interacting_residue_pair* and False when there is no stacking interaction

Example of a contact mapped onto the alignment

```
''P08579_RRM1_1A9N_B_12-83_9-80_0-95'': {
  ''0'': {
    ''73'': {
      ''6_I_10_C_Q'': {
        ''interactions'': {
            ''1'': [
              ''N4'',
              ''CD1'',
              ''4.27'',
              ''Base'',
              ''sidechain'',
              ''van-der-waals''
            ]
         },
        ''stacking'': false
        }
      }
    }
  }
```

5.2.4 Similarity among RRM-RNA complexes

All RNA sequences from RRM-RNA complexes were compared with each other, by sliding them with respect to each other and checking whether their nucleotides bind similar amino acid sequence positions in the RRM master alignment. The similarity score between two RRM-RNA complexes is computed using Eq. 5.2.1.

$$\text{Similarity Score} = \frac{\sum_{i}^{i=n} \frac{N_{\text{Matching positions}}}{N_{\text{Unique positions}}}}{N_{\text{Aligned nucleotides}}} \qquad \text{(Eq. 5.2.1)}$$

Where, $N_{\text{Matching positions}}$ is the number of matching positions between two RNAs, $N_{\text{Unique positions}}$ are the unique positions that those two nucleotides bind, and $N_{\text{Aligned positions}}$ is the aligned length of RNA.

The alignment (sliding window) with highest similarity score was retained for each RNA pair. With this a similarity matrix was constructed to identify the different binding modes.

We grouped the entries having a minimum score of 0.25 with at least 25% of the complexes in the cluster to select a homogeneous cluster. The first cluster (cluster 0) is composed of 187 entries and corresponds to the canonical binding mode of RRM.

RNA alignment

The method of similarity scores was used to align RNA sequences. To align all the 187 RNA sequences from cluster 0, we selected the medoid (3HHN_D), the entry with the highest similarity scores with respect to all other entries. All other 186 RNA sequences were aligned against the medoid to generate the RNA alignment. When comparing two RNAs, the sliding window position generating the highest score was considered as the best possible alignment for those RNA sequences.

5.2.5 Computation of scoring matrices

The RRM-RNA scoring method we have developed, RRMScorer, is an adaptation of the GOR method which was originally used for secondary structure prediction [Garnier et al., 1996, Kouza et al., 2017]. RRMScorer relies on the same information difference equation to calculate which nucleotide-residue contacts are preferred for specific amino acid positions in an RRM. RRMScorer uses Eq. 5.2.2 to compute the scores for each residue-nucleotide interacting position individually. The result is the sum of two terms; the first term computes the logarithm of the ratio between the number of times a nucleotide in position i (from the RNA alignment) has been observed interacting with an amino acid residue in position j (f_{N_i,R_j}) (from RRM master alignment), over the number of times that the nucleotide interacts with any other amino acid residue (f_{n-N_i,R_j}).

$$I(\Delta N_i; R_j) = log(\frac{f_{N_i,R_j}}{f_{n-N_i,R_j}}) + log(\frac{f_{n-N_i}}{f_{N_i}}) \qquad \text{(Eq. 5.2.2)}$$

Where, N_i is the nucleotide at position i, R_j is the residue at position j.

5.2.6 Discussion

We developed RRMScorer, a novel method to estimate RRM-RNA binding from sequence information only. Our method provides scores for the probability that a given RNA sequence binds to an RRM protein. We validated RRMScorer on both computational and experimental data. Our method is restricted to the canonical binding mode because of the limited data availability for other binding modes. From the protein structure side, the data availability is no longer a limitation after the AlphaFold Protein Structure Database release. Even though it does not solve the RNA recognition problem, current challenges purely based on protein structure, such as assessing the preferred RNA binding mode of an RRM, might be solved soon, although the current inability of such methods to cover dynamics and multiple conformations remains a bottleneck to be solved.

RRMScorer can be used to find good RNA candidates for a specific RRM domain on genomic scale studies that can be further coupled with computational methods for predicting the structure of the RRM-RNA complex.

5.3 Modeling 3D structures of RRM domains

There is a wide gap between the number of protein sequences and the number of protein structures available. UniProtKB release 2023_01 has 246,440,937 protein sequences, while there are only 202,292 experimentally determined protein structures available in the Protein Data Bank (PDB) as of 14-March-2023, corresponding to 61,463 distinct protein sequences. We collected 400,892 RRM (RNA Recognition Motif) sequences from UniProtKB (Swiss-Prot and TrEMBL) belonging to 19 different Pfam families and 1,456 RRM structures using 727 distinct entries from PDB, corresponding to 303 RRM instances. All the data about RRMs is stored in our InteR3M database. To bridge this gap between RRM sequences and RRM structures, comparative modeling could be one of the important tools. Three-dimensional structure determination using experimental methods like X-ray crystallography and NMR is a quite time-consuming and complex procedure. In contrast, comparative modeling gives acceptable results if a homologous template is available for the query sequence, in much less time and simpler ways.
The goal of this homology modeling pipeline for RRMs (RRMpip) is to model diverse structures of an RRM for a given sequence so that at least a few structures will be close to the RNA-bound form of the RRM.
Figure 5.2 depicts the workflow followed by RRMpip. This workflow is visualized using Common Workflow Language (CWL) [Amstutz et al., 2016].

This work was carried out in mid-2020, i.e. before the AlphaFold2 revolution. In this work, we developed the pipeline to model 3D structures of RRM domains using the state of the art methods available at the time. However, since the release of AlphaFold2, we have used the results from AlphaFold2 for modeling 3D structures of RRM domains. The use of AlphaFold2 has allowed us to generate 3D models of RRM domains with per-residue confidence score. We will be using the 3D structures from AlphaFold DB for any further analysis.

5.3.1 Methodology

Finding the best template

RRMpip needs a single input from the user, i.e. an amino acid sequence in *FASTA* format. The amino acid sequence provided by user was queried against the PDB database through HHblits [Remmert et al., 2012]. We limited the template hits to the RRM structures from InteR3M database only. At first, a single iteration of HHblits was performed and if there were no homologous RRM sequence found then further iterations were carried out to find the best templates.

Target-template sequence alignment

After template selection, target-template alignment was performed using *align2d()*[2] function from MODELLER. Although *align2d()* is based on a dynamic programming algorithm, it is different from standard sequence-sequence alignment methods because it takes into account structural information from the template when constructing an alignment.

[2]https://salilab.org/modeller/9v5/manual/node282.html

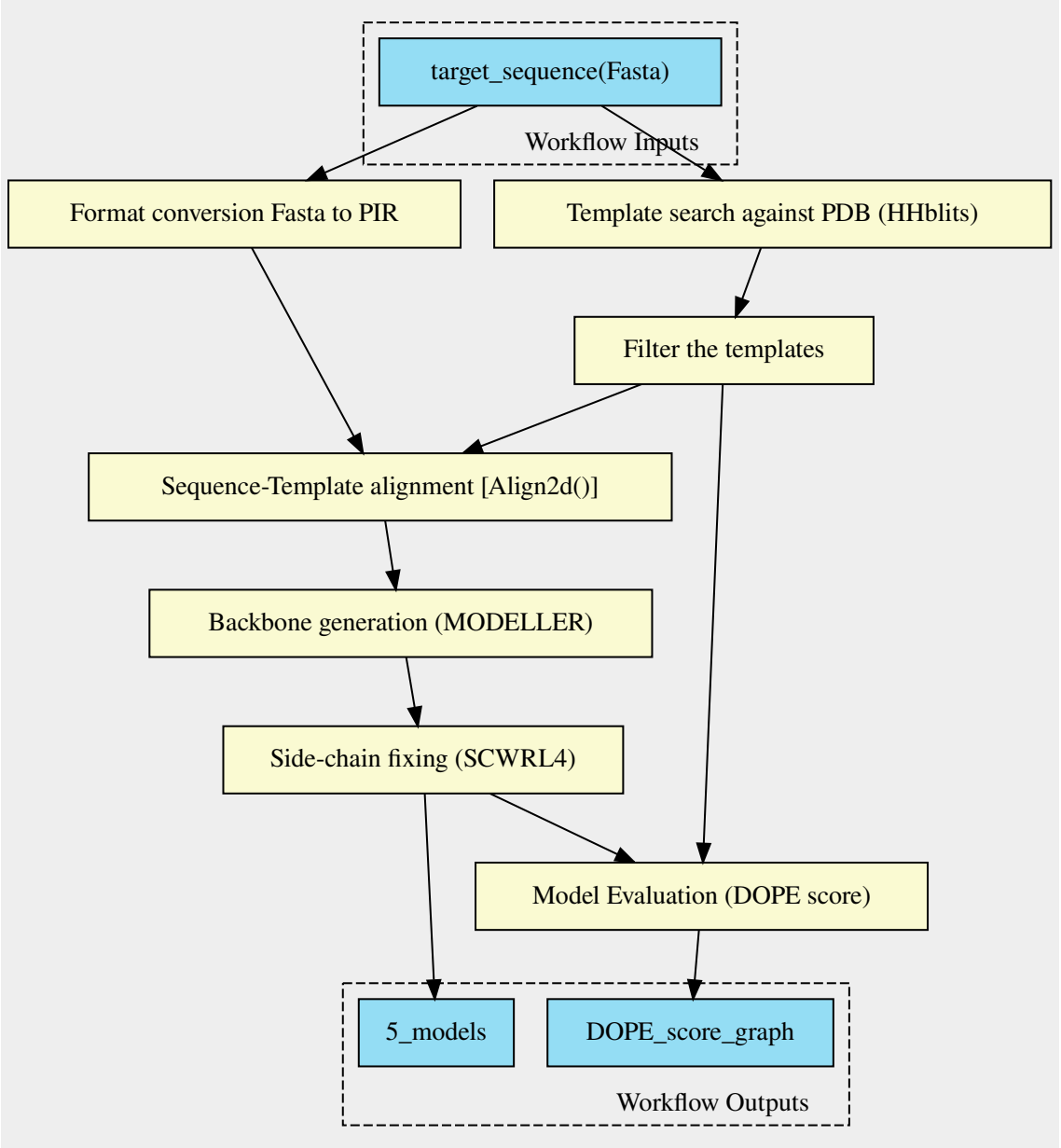


Figure 5.2: Workflow followed by the RRM Modeling Pipeline, RRMpip (visualized using CWL)

align2d() function takes two arguments, i.e., target sequence in PIR format and template structure, and returns the target-template alignment in PIR format.

Model building

The target-template alignment generated from *align2d()* was used to build models using MODELLER [Webb and Sali, 2016]. It takes two different arguments: target-template alignment in PIR format and template structure with chain ID, and returns five models (default). The number of models to be generated can be altered. The models generated by MODELLER were then processed with SCWRL4 [Krivov et al., 2009] for fixing the side-chain conformations.

Model evaluation

We used DOPE (Discrete Optimized Protein Energy) score [Shen and Sali, 2006] for evaluating the models. DOPE scores were calculated on a per-residue basis for each newly generated model and the template and then plotted for visualization.

The user needs to submit only a target sequence in fasta format and the pipeline will return five models generated by homology modeling, with the DOPE score. The pipeline also produces a graph to better understand the DOPE score for each part of the models according to the template.

5.3.2 Testing RRMpip

We used the sequences from experimentally determined structures of RRM domains for testing RRMpip. We randomly selected two RRM structures, one in the RNA-bound form and one in the unbound form.

- 1A9N_B from RRM_1 family (RNA-bound)
- 1D9A_A from RRM_1 family (unbound)

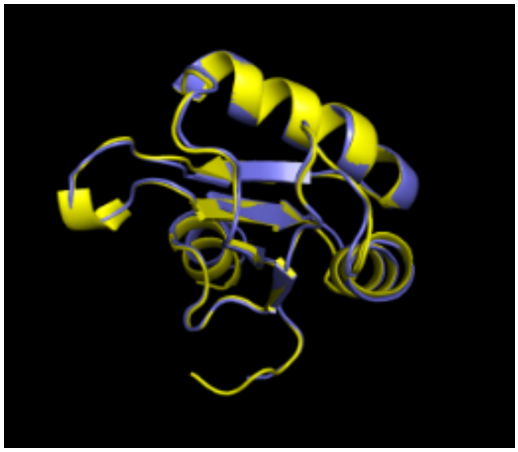
To avoid any bias, we ensured that no template would be from the same protein. The chosen templates to generate models for 1A9N_B and 1D9A_A have sequence identity of 75% (4PKD_B, RNA-bound) and 39% (1X5O_A, unbound), respectively. The bound/unbound state was not taken into account during the search for template. The 5 models of 1A9N_B have RMSD values ranging from 0.59 Å to 0.67 Å when aligned against the original structure, whereas the 5 models of 1D9A_A have RMSD values ranging from 1.83 Å to 2.02 Å.

The higher RMSD values for 1D9A_A models correspond to differences in the loops (Loop3 and 5), as visible on Figure 5.3b. Those higher differences in the models of 1A9N compared to the models of 1D9A could be explained by the bound state of both 1A9N and its template. The binding to RNA has a tendency to fix the interacting loops in a given position, while the unbound loops can have a larger diversity of positions.

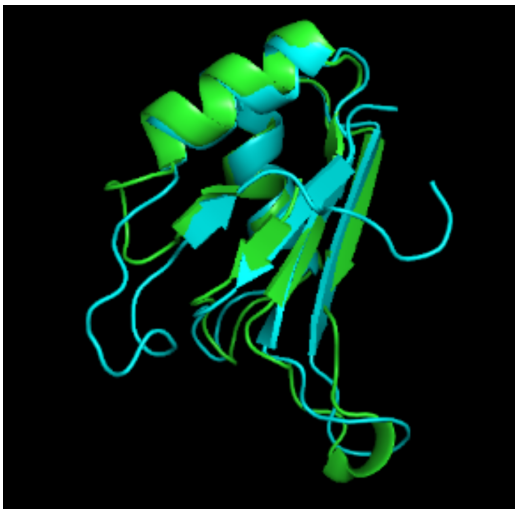
To summarize this testing of RRMpip, the structures generated by RRMpip (RRM modeling pipeline) are close to the experimentally determined structure. The DOPE score plots from the pipeline suggests that, when generating the DOPE score graph, only the aligned part of the template should be considered instead of the complete chain. Moreover, it reveals the need for adding a loop modeling step to the workflow.

5.3.3 Results & Discussion

RRMpip, a fully automated pipeline to model RRMs using a comparative modeling approach, was successfully built. RRMpip combines several tools to achieve the best results. As this is a preliminary version, there is much scope for improvement. In upcoming versions, loop modeling will be introduced (depending on homology). Loop modeling approaches can be used for correcting the folding of low-homology regions, with high accuracy for up to 30 residues. Loop modeling is one of the difficult parts of protein structure modeling. DaReUS-Loop, a



(a) Structure alignment of a generated model (yellow) and the experimentally determined (purple) structure (1A9N_B)



(b) Structure alignment a generated model (green) and the experimentally determined (cyan) structure (1D9A_A)

Figure 5.3: Structure alignment of one generated model against experimentally determined structures

data-based approach, could be an option for this purpose [Karami et al., 2019].
It is always better to have multiple criteria to assess or evaluate the models. There are several evaluation methods from different categories like MolProbity and Ramachandran plot from Physics-based methods, DOPE (already in use) and Qualitative Model Energy ANalysis (QMEAN) from knowledge-based methods, and DSSP from machine learning-based methods. All these methods can be used to evaluate the models generated from the RRMpip. Currently, only one criteria is being used to assess the models. It is assumed that 3D structure minimization is imperative prior to any computational analysis [Haddad et al., 2020]. But there might be various pre-processing steps required depending on the analysis workflow used afterwards (like coarse-grained docking using ATTRACT). Thus, we will try to make the output from this RRMpip pipeline interoperable with other computational analysis workflows. For future tests, the reliability of intra-Pfam and inter-Pfam RRM templates will be evaluated. We will also compare the results obtained for a given sequence using RNA-bound or unbound templates. Currently we expect that RNA-bound templates would give more reliable results to achieve the goal of modeling a bound-like structure.

5.4 Modeling RRM-RNA complexes

Initially, we planned to use the RRM 3D structures resulting from RRMpip for the modeling of RRM-RNA complexes but after the release of AlphaFold DB we shifted to directly use the 3D structures of RRM domains from AlphaFold DB. All the information about RRM domains and their interactions with nucleic acids from InteR3M database can be used to model structures of RRM-RNA complexes in a data-driven docking approach.

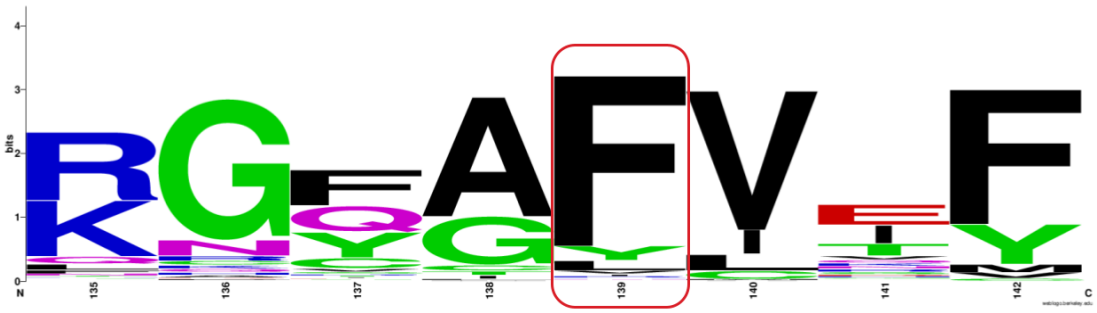
5.4.1 Anchored Docking: Definition and requirements

RRM domains have conserved residues in RNP1 and RNP2 regions that establish $\pi - \pi$ stacking interactions with the nucleotides. These stacking interactions can be used as anchors in the docking of RNA fragments. Using these stacking interaction, one nucleotide is anchored to a conserved aromatic residue at 5^{th} position in RNP1 and another nucleotide at 2^{nd} position in RNP2 (Figure 5.4). These two anchoring positions of nucleotides will guide the position of the next nucleotides in an iterative manner. This approach was first validated by de Beauchene et al. [2016] using fragment-based docking of tandem RRMs and RNA. In this study, the data was collected from the PDB in July 2015. Since then, there has been an increase in the number of RRM-RNA complex structures in the PDB. Moreover, there are more experimentally solved structures for single RRM domains in complex with RNA than for tandem RRM domains in complex with RNA. Thus, we focused on the docking of RNA onto a single RRM domain rather than on tandem RRM domains.

5.4.2 Extraction and Clustering of Anchoring Patterns

All the experimental structures of RRM-RNA complexes were extracted from the InteR3M database along with the contact information. There are a total of 496 structures of RRM-RNA complexes that have contacts between the atoms of RRM and RNA, without any filter for redundancies at the protein level (structures of the same RRM, bound to same or different RNA sequences). These 496 RRM-RNA complexes belong to 105 distinct RRM domain instances. As we are looking specifically at the stacking interactions from the RNP regions, to use them as anchors in the docking protocol, all these 496 structures were filtered for the stacking interactions at RNP1 position 5 (Figure 5.4a) and RNP2 position 2 (Figure 5.4b). A total of 257 structures of RRM-RNA complexes have the stacking interaction at either one or both positions. These 257 structural complexes are from 52 unique proteins and 72 unique RRM domains: 34 proteins with a single RRM domain, 16 proteins with two RRM domains, and 2 proteins with three RRM domains.

We want to get a few anchoring patterns (3D conformation and position of a stacking nucleotide relative to the RNP backbone) representative of all these 257 different structural complexes. They can be used to guide the docking by adding an energy penalty to the deviation of a given nucleotide of the fragment from the anchoring position. Our fragment-based docking method uses trinucleotide fragments, and the two RRM residues RNP1-5 and RNP2-2 usually bind two nucleotides that are at positions (i, i+1) or (i, i+2) in the RNA sequence. Thus, the docked fragments will



(a) RNP1 sequence motif



(b) RNP2 sequence motif

Figure 5.4: RNP sequence motifs from the alignment of RRM domains; the position of stacking amino acid is highlighted with a red outline.

The sequence logos were generated using WebLogo [Crooks et al., 2004].

use either one or a couple of anchoring pattern(s), one per stacking nucleotide in the fragment. As one docking will be run for each (couple of) representative anchoring pattern(s), we want as few representatives as possible to limit the computational time. But for the docking to succeed, one of the representatives must be close enough to the real target position, which is easier to achieve if we have many representatives. Clustering will be performed to get a few representative anchoring patterns. A trade-off must be found for the clustering cutoff to have a low number of representatives (loose clustering cutoff) but each pattern is close enough to its closest representative (tight clustering cutoff).

To get the representatives for all the stacking nucleotide conformations, we proceed with a new hierarchical agglomerative (HA) clustering approach called Radius, that produces the minimum number of clusters and their representatives such that each initial element is within a chosen distance from at least one of the final representatives [Moniot et al., 2022a]. The representatives are not among the initial elements but are the centroid of each cluster. For 3D structures of the same molecule, we use the root mean squared deviation (RMSD) as the clustering distance criteria.

As we want to cluster the stacking nucleotides after protein fitting, all the structural complexes were chopped by keeping only the amino acid residues from

RNP1 position 4-6 or RNP2 position 1-3 and the stacking nucleotides. The RRM master alignment was used to retrieve the positions of these amino acid residues. All the chopped structures were then divided into two different parts based on the position of the stacking interaction. Beta1 group consists of all chopped structures with only the stacking interaction on RNP2 retained, Beta3 group consists of all chopped structures with only the stacking interaction on RNP1 retained. The structures having stacking interactions at both positions were also classified into Beta1 and Beta3 with the stacked nucleotide at the respective position.

All the structures from each group were superimposed on the amino acids backbone. The reference structure used for the superposition was RRM1 from 1B7F chain A (protein sex-lethal). All the nucleotides were extracted from the superimposed structures and converted into the ATTRACT coarse-grain representation. To cluster all nucleotides together requires the same number of pseudo-atoms, therefore the extra bead (coarse-grain pseudo-atom) created from the N7 atom of purines was removed, resulting in the same number of beads for purines and pyrimidines. Its position can be recreated exactly from the position of the 3 other base beads.

The atomic coordinates of all the nucleotides in the coarse-grain model from each group were extracted and stored in a numpy matrix, input for the HA clustering method. The Radius HA clustering method groups together the given structures based on the threshold RMSD given by the user, and returns the following:

- A list of all members (nucleotide structures in our case) in each cluster,
- Prototypes (representative structures) for each cluster,
- The radius of each cluster, suggesting the variability within each cluster.

5.4.3 Resulting Anchoring patterns

We tried clustering the coarse-grain coordinates of the stacking nucleotides fitted on the reference RNPs at different RMSD thresholds for both Beta1 and Beta3 groups. The following table summarizes the results from clustering of the anchoring patterns:

Table 5.1: Results from clustering of the anchoring patterns

Group	**Clustering Threshold**	**Number of Clusters**	**Number of Singletons***	**Max distance of each bead to prototype** [GP1,GS1,GS2,GX1,GX2,GX3]
Beta1	3.0 Å	8	3	[5.1, 3.6, 3.7, 3.1, 5.1, 4.1]
Beta3	3.0 Å	6	1	[4.8, 3.9, 4.4, 3.4, 4.4, 3.8]
Beta1	3.5 Å	6	2	[6.3, 4.7, 3.9, 3.7, 5.0, 4.9]
Beta3	3.5 Å	4	0	[4.8, 3.9, 4.4, 3.7, 4.4, 4.9]

* Singleton: A cluster is said to be singleton if it contains only structure(s) from the same instance of RRM domain
X in the table above represents the base and GX1, GX2, and GX3 are beads for the atoms of the base for the corresponding nucleotide

The maximum distance of each bead to its prototype in the clusters gives us

information about which bead positions are the most variable. Currently, during docking we are using the same maximal distance for the positional restraints of each bead of a nucleotide toward the corresponding bead in the prototype, but different maximal distances could be applied for the different beads depending on their position variability. This would result in a more accurate position of each bead instead of an overall position of the nucleotide. For example, in most cases from Table 1, the bead (GP1) representing the phosphate group has the highest positional variability, so this bead would have slightly looser positional restraints or a lower energy penalty. Yet given the relatively small number of clustered structures, this could lead to over-fitting.

Figures 5.5 show the positions of each bead along with its prototype for one cluster from each group.

We took the clusters from 3.5 Å and 3.0 Å for beta1 and beta3 groups. These RMSD thresholds were chosen such as to minimize the radius of the clusters (which determines the precision of the docking restraints that can be used) while obtaining a small number of singletons. The singletons were excluded from docking as they are not representative enough. Using them would increase the number of wrong poses while barely increasing the chance to obtain hits. Finally, we obtained 4 and 5 prototypes for Beta1 and Beta3, respectively.

Following are the excluded RRM domain instances (Singletons):

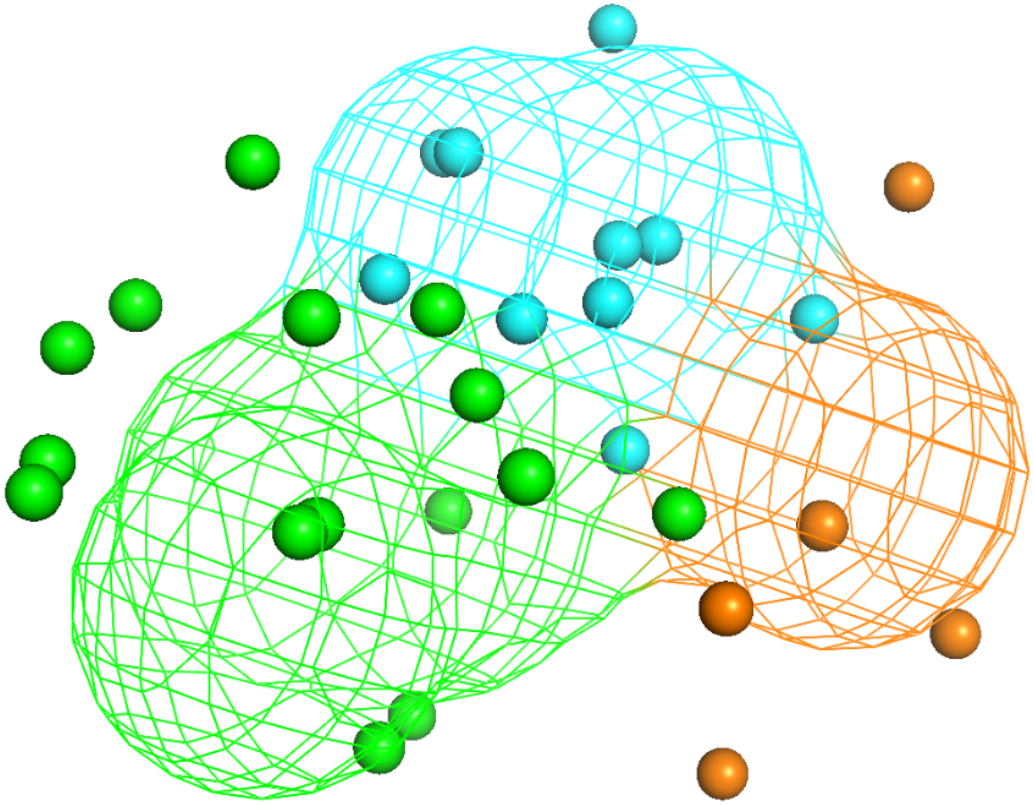
- P25299_RRM1 (2 structures in Beta1 group and 1 structure in Beta3)
- Q4DY32_RRM2 (1 structure in Beta1 group)

The position of the stacking nucleotide is dependent on the position of the side-chain of the amino acid involved in stacking. As a consequence, by looking at the side-chain positions in the clustered structures, it appears that most positions are specific to one cluster, while only a few positions appear in several clusters Thus the most suitable prototype for docking on a specific target protein could be predicted from the bound position of the side-chain of the amino acid, if this position could be inferred from the sequence or an unbound structure of the protein. As an example, Figure 5.6 shows two prototypes along with the stacking amino acid from their respective cluster.

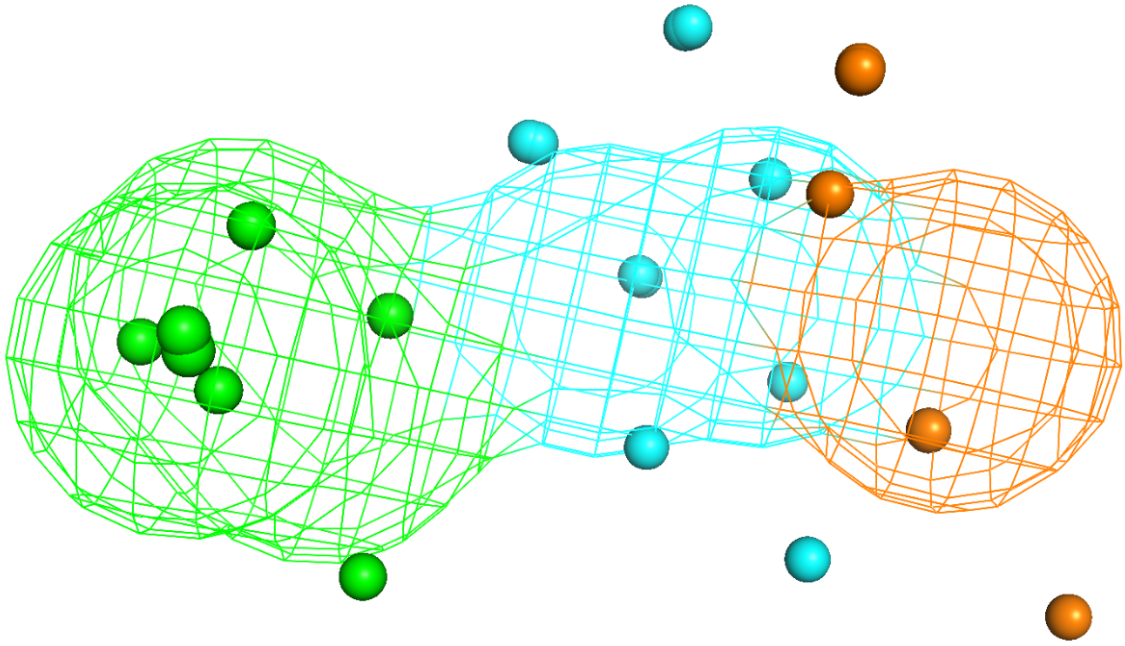
5.4.4 Docking with Anchoring patterns

This part was done by another PhD student ESR4 from our lab (Anna Kravchenko). The docking pipeline was developed to automatically build RRM-RNA models using only the sequences and the identification of stacking nucleotides as input from the user. This RRM-RNA docking (RRM-RNA dock) pipeline uses AlphaFold DB to obtain a 3D structures of RRM domain and the anchoring prototypes (from Section 5.4.3) for anchored docking. RRM-RNA dock is still in development and currently can dock only tri-nucleotide fragments containing the anchoring nucleotides. The RRM-RNA dock pipeline is available at https://github.com/AnnaKravchenko/RRM-RNA-dock.

The resulting 3D structure of RRM-RNA complex from RRM-RNA dock can be assessed and further studied for interactions between RRM and RNA using



(a) Prototype and members for cluster 5 from Beta3 group at 3.0Å



(b) Prototype and members for cluster 3 from Beta1 group at 3.5Å

Figure 5.5: Prototype and members for one cluster each from Beta1 and Beta3 group. Prototype for the cluster is shown as mesh and all the members from this cluster are shown as spheres; Orange color is for phosphate beads, cyan color is for sugar atoms and green color is for base beads.

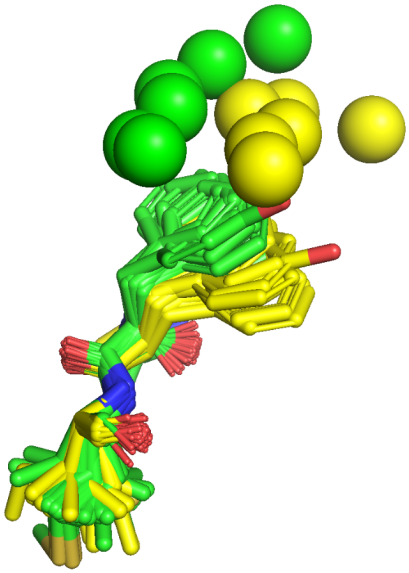


Figure 5.6: The prototypes from clusters 1 (green) and 4 (yellow) of the Beta3 group along with the amino-acid residues used for fitting. Although there is some overlap in the position of the side-chain of the stacking amino acid residue from the 2 clusters, they are still mostly distinct from each other

molecular dynamic simulations.

5.5 Evaluating RRM-RNA complexes

The goal of the RNAct project is to design an RRM with the desired RNA binding activity. For this, we need to be able to distinguish, between two RRMs, which one binds with highest affinity to the given RNA sequence. This is why we are interested in developing a molecular dynamic (MD) simulation protocol capable of detecting the effect of RRM point mutations on RNA binding.

Binding free energies can be computed using molecular dynamic simulations. The methods to compute absolute binding free energies can be used to compare the binding of ligands irrespective of chemical similarity [Feng et al., 2022].
The computation of relative binding free energy results in the difference between the binding free energies of two ligands by computing the change in free energy. In alchemical transformation, this change in free energy is computed by transforming one ligand (beginning state) to the other (end state). Thus, the methods to compute the relative binding free energies are well suited for molecules that are chemically similar. Thus, computation of relative binding free energy is most suitable for comparing or ranking the binding activity of chemically similar ligands [Baumann et al., 2021].
Absolute binding free energy computations require more sampling because of the larger transformations (from initial state to end state) involved compared to relative binding free energy. This makes the computation of absolute binding free energy computationally expensive [Cournia et al., 2020].
Our goal is to find the effect of RRM point mutations on RNA binding and computation of relative binding free energy is more relevant in this case. Thus, we focused only on the computation of relative binding free energies. This section focuses on establishing a protocol for the computation of relative binding free energies for RRM-RNA complexes.

Before proceeding with the computation of binding free energy, we wanted to make sure that MD simulations can be used to replicate stability and instability for a RRM-RNA pair known to be stable (binding) and less stable (not binding), respectively. If MD simulations can differentiate between stable and less stable complex, we can use this MD protocol as a step to filter for good/bad RRM, in terms of binding to the desired RNA before a finer filter is designed by computing binding energy.

5.5.1 Stability of an RRM-RNA complex

I started with checking the stability of an RRM-RNA complex by running a standard MD simulation. Musashi homolog 1 (UniProt ID: Q61474) is one of the proteins of interest in the RNAct[3] project. RNAct aims to modify the RRM1 domain from Musashi1 to change its RNA specificity for regulation of the fatty acid pathway in *E. Coli*. So, I used the Musashi1 RRM1-RNA complex (PDB ID: 2RS2) for this task.

[3] https://rnact.eu/

System preparation

The RRM domain from Musashi (from Mouse) was taken from PDB (PDB ID: 2RS2). This NMR solution structure of wild-type RRM1 from 'RNA-binding protein Musashi homolog 1' is in complex with RNA having sequence 'GUAGU'.
We built the system for this complex in explicit solvent using TIP3P water model inside a truncated octahedral box with a space of 30 Å around the complex in each direction. This ensures that all atoms in the starting structure of the complex will be no less than 30 Å from the edge of the water box. The Na^+ and Cl^- ions were added at a concentration of 0.15 M NaCl salt. The AMBER force fields (OL3 for RNA and ff19SB for Protein) were used to describe molecular interactions. All this was done in tleap program from AmberTools. The system consists of 82,610 atoms.

Simulation Protocol

The MD simulation was performed using NAMD 3.0. MD simulations were run at 310 K with a time step of 1 fs. Particle mesh ewald (PME) was used to compute the long-range electrostatic interactions. The SHAKE algorithm was used to constrain the covalent bonds involving hydrogen atoms. A cutoff distance of 9 Å was used for non-bonded interactions. The interactions between atoms that are further apart than this cutoff distance were ignored. Langevin dynamics was used to maintain the constant temperature and pressure (1 atm) [4].
Once set up, the system was minimized for 20 ps using a conjugate gradient and line search algorithm. The system was slowly heated up from 0 to 310 K over a period of 310 ps. Then the system was equilibrated with NVT and NPT ensembles, for 500 ps each. Finally, the production run was performed for 300 ns under NPT ensemble.

Analysis of the trajectory for wild type

After the production run, the complete system was wrapped and unwrapped using PBCTools plugin[5] in VMD [Humphrey et al., 1996]. As we are interested in the stability of RRM-RNA complex and their inter-molecular interactions, we have removed water molecules along with salt (Na^+, Cl^-) ions and hydrogen atoms to save space and time for further analysis.

The first frame of the trajectory was used as reference to superimpose all other frames and compute the RMSF values per residue and RMSD values for each frame in this system.

RMSF analysis indicates the flexibility of an individual residue in the simulated system. Firstly, the RMSF has been computed per atom and then averaged per residue. The terminal residues with higher RMSF values than the rest of the complex can be removed to avoid bias when computing the RMSD.

[4]Detailed setting information for using AMBER force field in NAMD can be found here: https://ambermd.org/namd/namd_amber.html

[5]https://www.ks.uiuc.edu/Research/vmd/plugins/pbctools/

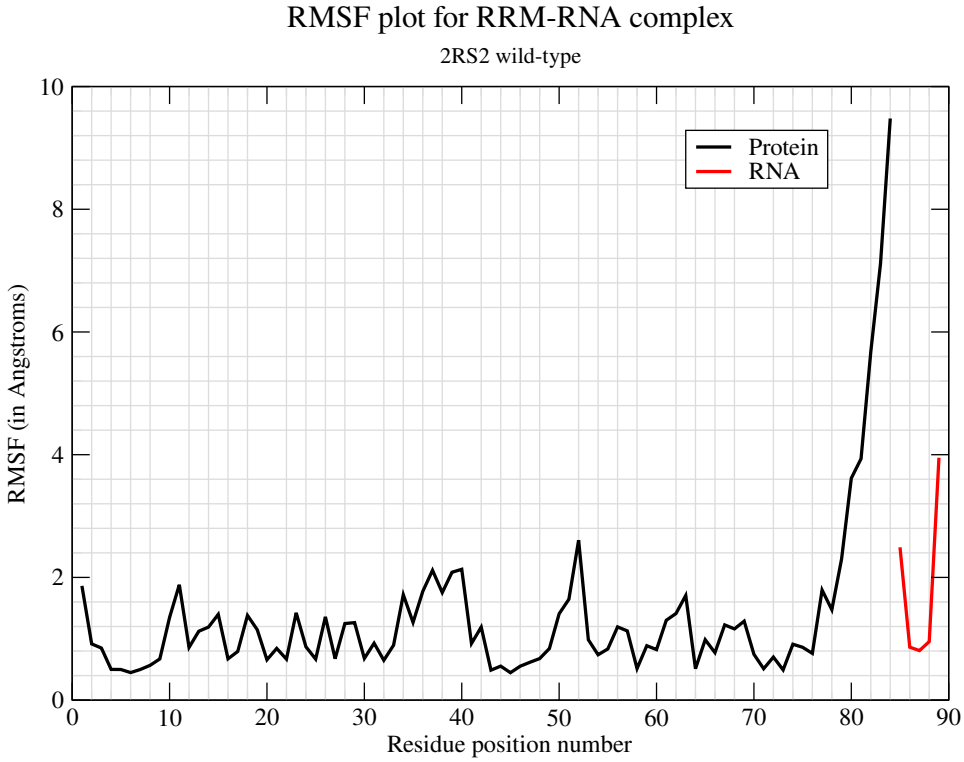


Figure 5.7: RMSF plot for the Musashi RRM1-RNA ('GUAGU') complex

The RMSF shows a low conformational flexibility (<2.7 Å) for the protein, except the C-terminal residues 82-84 having RMSF values >5 Å. These residues were removed from the protein for further analysis. The RMSF values <1 Å for the middle three nucleotides (83-85) correspond to very small fluctuations from their positions, whereas the terminal nucleotides (G82 and U86) have RMSF values >2.4 Å, revealing more fluctuations.

RMSD analysis: The RNA shows more deviation compared to the protein and complex. As the RNA is very short compared to the protein in this complex, it contributes very little to the complex RMSD compared to the protein contribution (Figure 5.8). Overall, this complex is quite stable as the RMSD keeps moving around 3 Å after the first few frames (ns).

Convergence analysis: This is useful to check the convergence of the simulation, i.e., if no new conformations have been explored by the last part of the trajectory. One of the ways to check the convergence of a simulation is clustering the simulation trajectory and see if there are any new clusters for the end part of the trajectory. Thus, to perform clustering of the simulation trajectory, we divided the complete trajectory of 300 ns into three parts, each comprising 100 frames: 0-100 ns, 101-200 ns, and 201-300 ns. Then we performed "star-shaped clustering" on these three parts of the trajectory with different thresholds to obtain different number of clusters using an in-house script. Table 5.2 provides the number of resulting clusters at

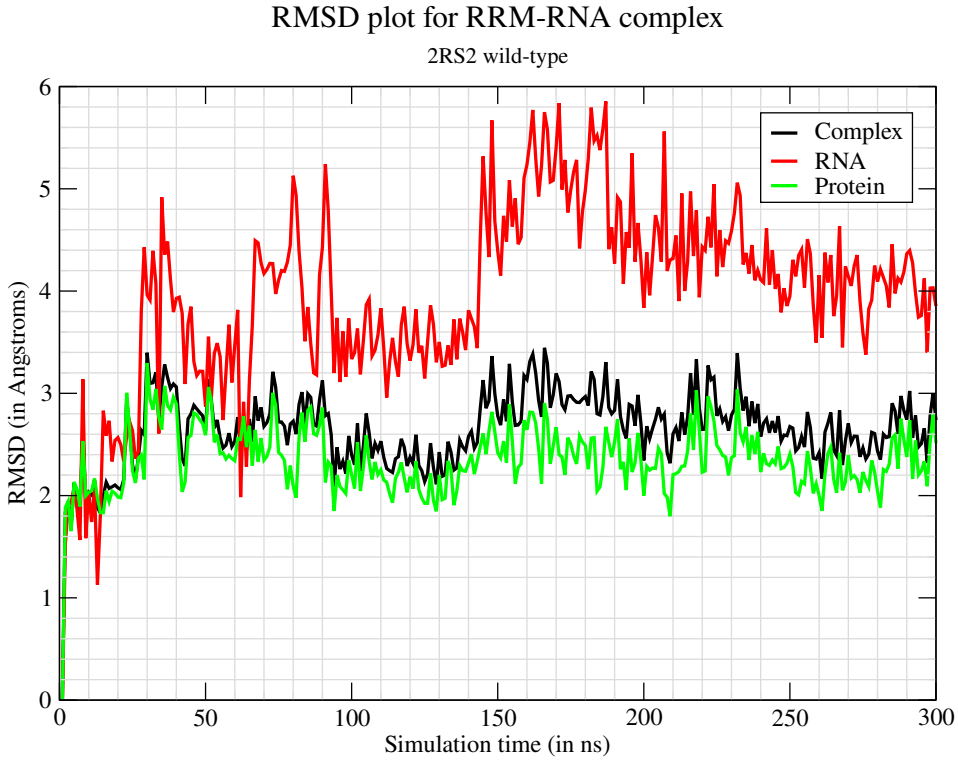


Figure 5.8: RMSD plot for the Musashi1 RRM1-RNA ('GUAGU') complex

different thresholds.
A high threshold of RMSD will result into a small number of clusters composed of frames with higher RMSD (significantly different). This will make it difficult to track the conformational changes that could be detected at lower RMSD threshold such as movement of side-chain from a residue. A low threshold of RMSD will result into a large number of clusters comprised of frames with lower RMSD (significantly similar). The analysis of too many clusters becomes very tedious and makes it difficult to track the conformational changes because multiple clusters might have similar conformations with a very little deviation. Thus, it is important to perform clustering at different threshold to find the optimal number of clusters and number of members per cluster.

Table 5.2: Results from clustering of the trajectory for Musashi1 RRM1-RNA complex

Clustering Threshold	**Number of Clusters**
3.0 Å RMSD	2
2.7 Å RMSD	5
2.5 Å RMSD	6
2.3 Å RMSD	9

Each cluster represents a conformation of the simulated system. Figure 5.9 shows the clustering results for 2.7 Å threshold. The simulation explored a total of 5 different

conformations. All three parts of the simulation have explored 4 conformations while the last part from 201 - 300 ns has not explored the 5^{th} conformation (cluster). As there is no new conformations explored in the last part of the simulation, the simulation has converged.

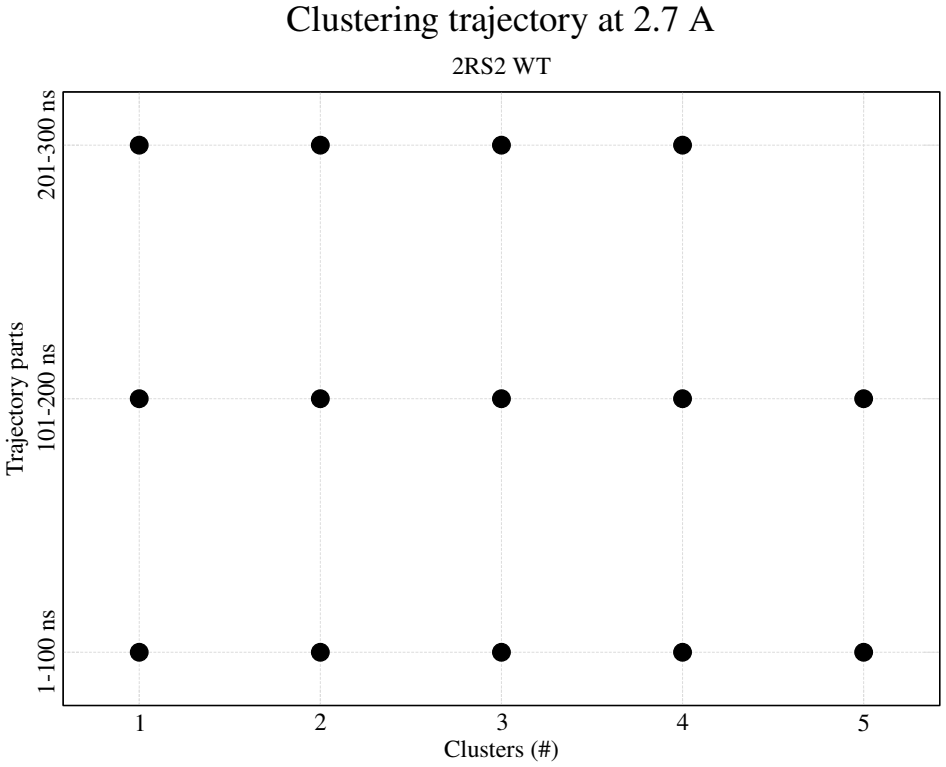


Figure 5.9: Convergence plot for the Musashi1 RRM1-RNA complex

Contacts analysis : We want to check if we can detect a low stability (known in vitro) by a loss of contacts in the MD simulation. A contacts analysis can be done by comparing native and non-native contacts. Native contacts are the contacts present in a protein's native (natural) state. Non-native contacts are the contacts absent in a protein's native state.

CPPTRAJ [Roe and Cheatham III, 2013] provides functionality to compute native contacts. We wrote a script to take advantage of 'nativecontacts' functionality from CPPTRAJ and visualize these contacts in a heatmap. The resulting heatmap contains three different colors: green for conserved native contacts, blue for non-native contacts and red for lost native contacts. The intensity of the color shows the percentage of contacts for that particular section.

In this analysis, we considered the first 50 ns (frames) as reference for the native contacts, to allow the complex to relax and to capture all the possible contacts. We have only considered the contacts found in at least 10% of the frames. No native contacts are lost in the simulation, but some contacts are formed after the 50 ns of the simulation, i.e., shown in blue (Figure 5.10).

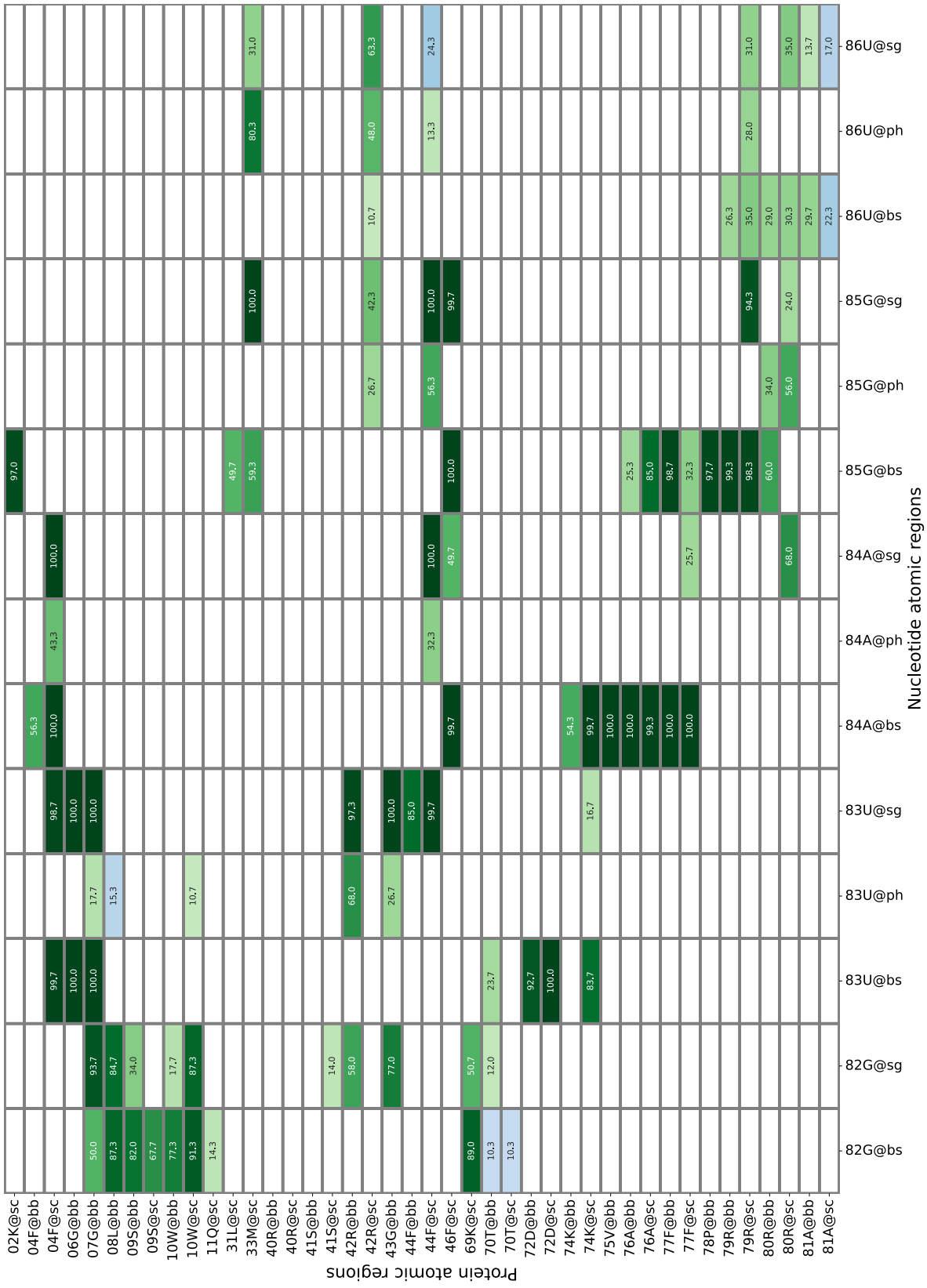


Figure 5.10: Tracking of native and non-native contacts over time from the Musashi1 RRM1-RNA ('GUAGU') complex. The green color represents native contacts, blue represents non-native contacts, and red color represents lost native contacts. The intensity of color shows how many frames have that respective contact.

Thus, the MD simulations were able to replicate the stability of Musashi1 RRM1-RNA (GUAGU) complex and the contacts between RRM and RNA from the simulation are in agreement with Ohyama et al. [2012] (Table 5.3).

Table 5.3: Unique stacking interactions for recognition of target RNA by Musashi1 RRM1 from the study of Ohyama et al. [2012]

AA with actual position	**Nucleotide**	**AA with position from our system**	**Location of AA**
Phe23	3A	Phe4	RNP2
Trp29	1G	Trp10	First loop
Phe65	4G	Phe46	RNP1
Phe96	3A	Phe77	Immediately after β4

The nucleotides numbered 83-85 in our system ('UAG' motif from RNA sequence) have some contacts that are present in all the frames of simulation. This suggests 'UAG' as the binding motif for this RRM domain.
We want to check if we can predict the non-binding of other RNA sequences and to test this we changed the RNA sequence from 'GUAGU' to 'GCCCU' in the same RRM-RNA complex.

The mutations in the RNA sequence were performed using the 'mutate_bases'[6] utility program from x3DNA-DSSR tool. The modified RNA sequence does not have the binding motif (UAG) so it may not form the interactions required for the stability of RRM-RNA complex, resulting in weak binding or dissociation of the complex. The system was prepared, simulated and analysed in same way as the previous one.

Analysis of trajectory for complex with mutRNA

Figure 5.11 shows the RMSF per residue from the simulated system of Musashi1 RRM1-mutRNA ('GCCCU') complex. The C-terminal residues (82-84) with RMSF values > 5.0 Å were removed for further analysis.

Figure 5.12 showing the RMSD of the system suggests that the complex is stable, as its RMSD stays around 3.0 Å. There is a sudden drop in RMSD of the RNA, after 36^{th} frame, that arises mainly from the fluctuation of $3'$ nucleotides.

Table 5.4 shows the number of clusters obtained from clustering of Musashi1 RRM1-mutRNA complex at different thresholds.

Table 5.4: Results from clustering of the trajectory for Musashi1 RRM1-RNA complex

Clustering Threshold	**Number of Clusters**
3.0 Å	2
2.7 Å	4
2.5 Å	8

[6] http://forum.x3dna.org/general-discussions/mutate_bases/

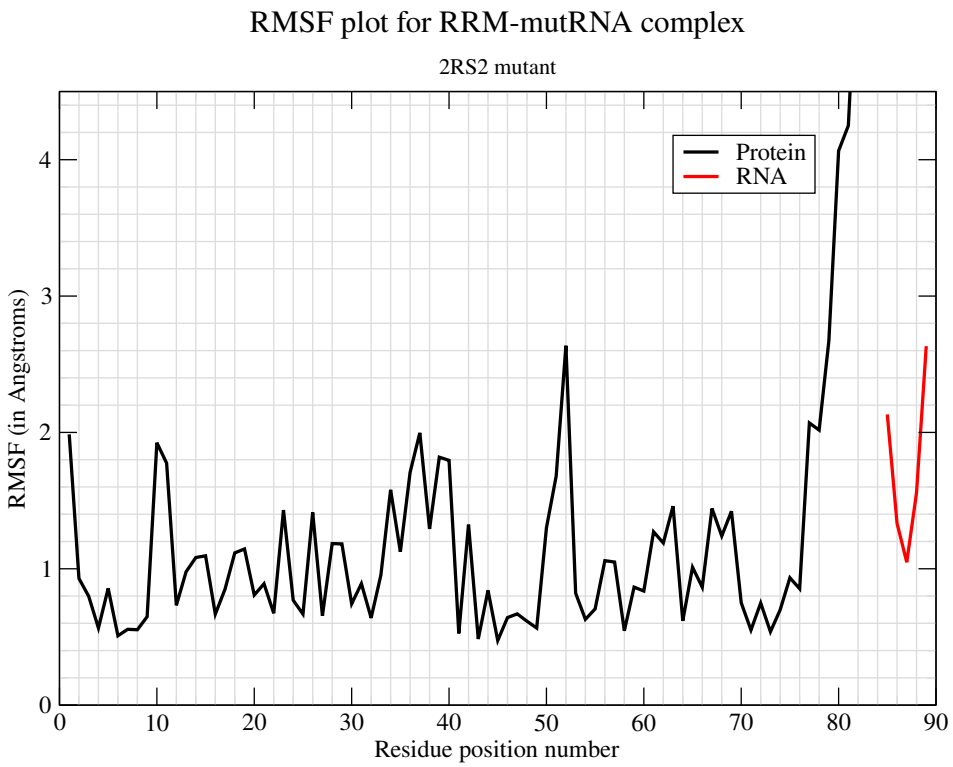


Figure 5.11: RMSF plot for the Musashi RRM1-RNA ('GCCCU') complex

We chose the clusters at 2.7 Å as it has optimal number of clusters and number of members per clusters (>5). Figure 5.13 shows the clusters from the three parts of the trajectory at 2.7 Å. All three parts have three clusters in common and the fourth cluster is present only in the first and last parts of the trajectory. Thus, the simulation has converged and no new conformation was explored by the last part of the trajectory.
Figure 5.14 shows the intermolecular contacts between Musashi1 RRM1 and RNA from the simulation. Most native contacts are conserved in the RRM-mutRNA complex, like Phe4-C2, Phe76-C3. The results for this are not exactly as we were expecting , but still there are some contacts missing between loop3 (between β2 and β3) and the last nucleotide.

Comparison of Wild Type and Mutant complex

We wanted to compare both trajectories (wild type and mutant) to see if there are any conformations that are unique to only one trajectory. Those unique conformations from one trajectory might provide more insights into the conformational differences. We clustered the two trajectories together. For clustering by RMSD, we need the same number of atoms in both trajectories. Thus, we kept only the common atoms from the RNA of both complexes and removed the atoms that are different. For pyrimidine (C) - purine (A or G) pair, we kept only the sugar-phosphate backbone of the RNA by removing all atoms from the nucleobase. After removing all these atoms, both systems have 731

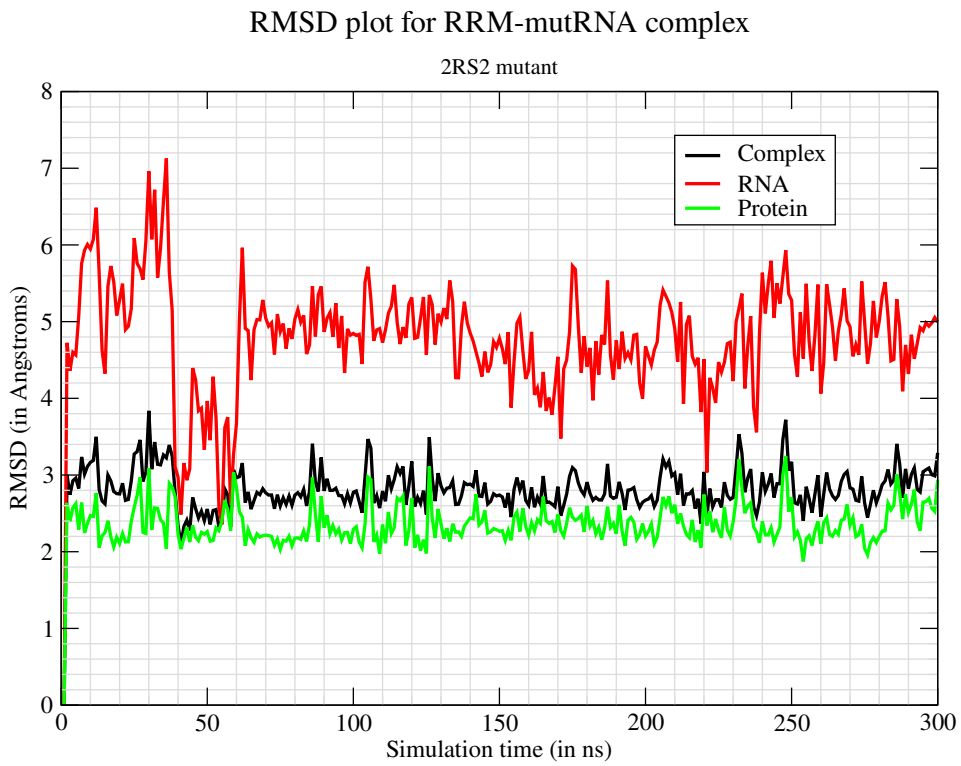


Figure 5.12: RMSD plot for the Musashi RRM1-RNA ('GCCCU') complex

atoms.

We performed 'star-shaped clustering' on the merged trajectories. Table 5.5 shows the number of resulting clusters with different thresholds. Out of all the 7 clusters formed at 2.7 Å threshold, 4 clusters are shared by both trajectories while 1 and 2 clusters are unique to the mutant and wild type complex, respectively.

Table 5.5: Results from clustering of the trajectory for wild type and mutant complex of Musashi1 RRM1-RNA

Clustering Threshold	**Number of Clusters**	**Common Clusters to both Trajectories**
3.0 Å	4	3
2.7 Å	7	4
2.5 Å	12	3

After visualizing the frames (members) from clusters that are unique to one of the trajectories, visually significant differences are in the conformation of the C-terminal part of the protein and in the last nucleotide.

Unlike our hypothesis, the change in binding motif from RNA sequence has not resulted in a dissociation of the RRM-RNA complex during 300 ns MD simulation. The classical MD simulations are probably not able to distinguish between strong and weak binders, in our case RNA sequences. This MD protocol might not be very

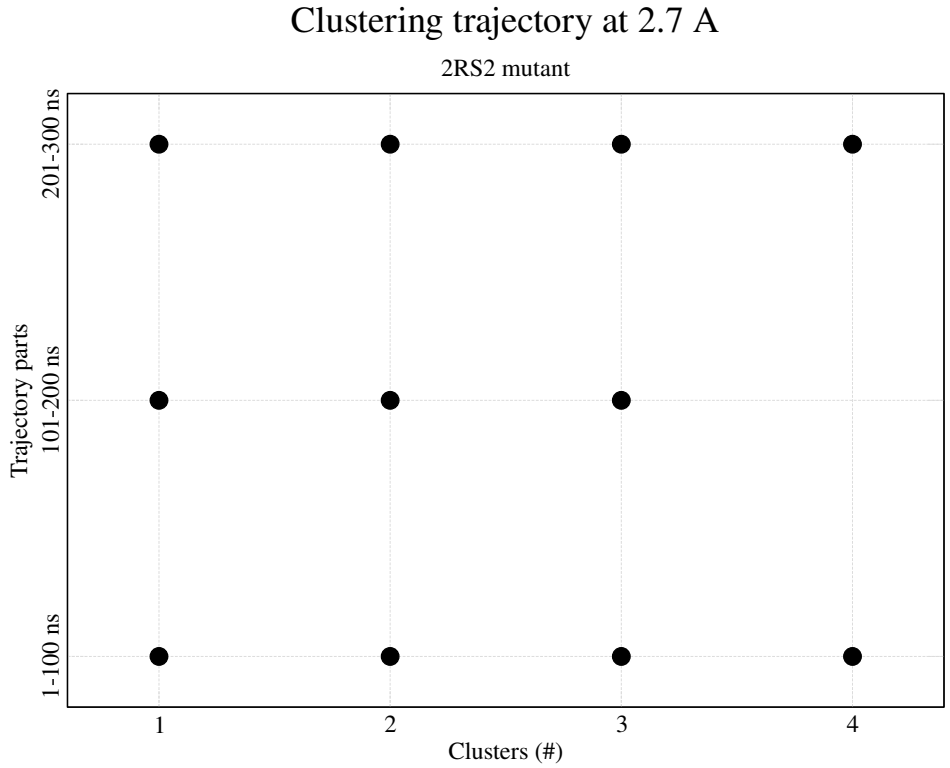


Figure 5.13: Convergence plot for the Musashi1 RRM1-mutRNA ('GCCCU') complex

helpful to use as a filtering step for RRM-RNA complexes resulting from docking approaches like RRM-RNA dock.

We also performed MD simulation for both the wild type and mutant complexes with a simulated annealing protocol. We increased the temperature of the system by 50 K after every 100 ns. But, even with the simulated annealing protocol we were not able to distinguish between strongly and weakly bound RRM-RNA complexes.

We can further investigate these two trajectories for the differences at the binding interface. To do this, we will perform clustering only for the residues at the binding interface and nucleotides. We can also employ steered MD protocol, i.e. constant velocity pulling[7] to check which RNA sequence binds stronger to the RRM domain.

[7]https://www.ks.uiuc.edu/Training/Tutorials/namd/namd-tutorial-html/node18.html

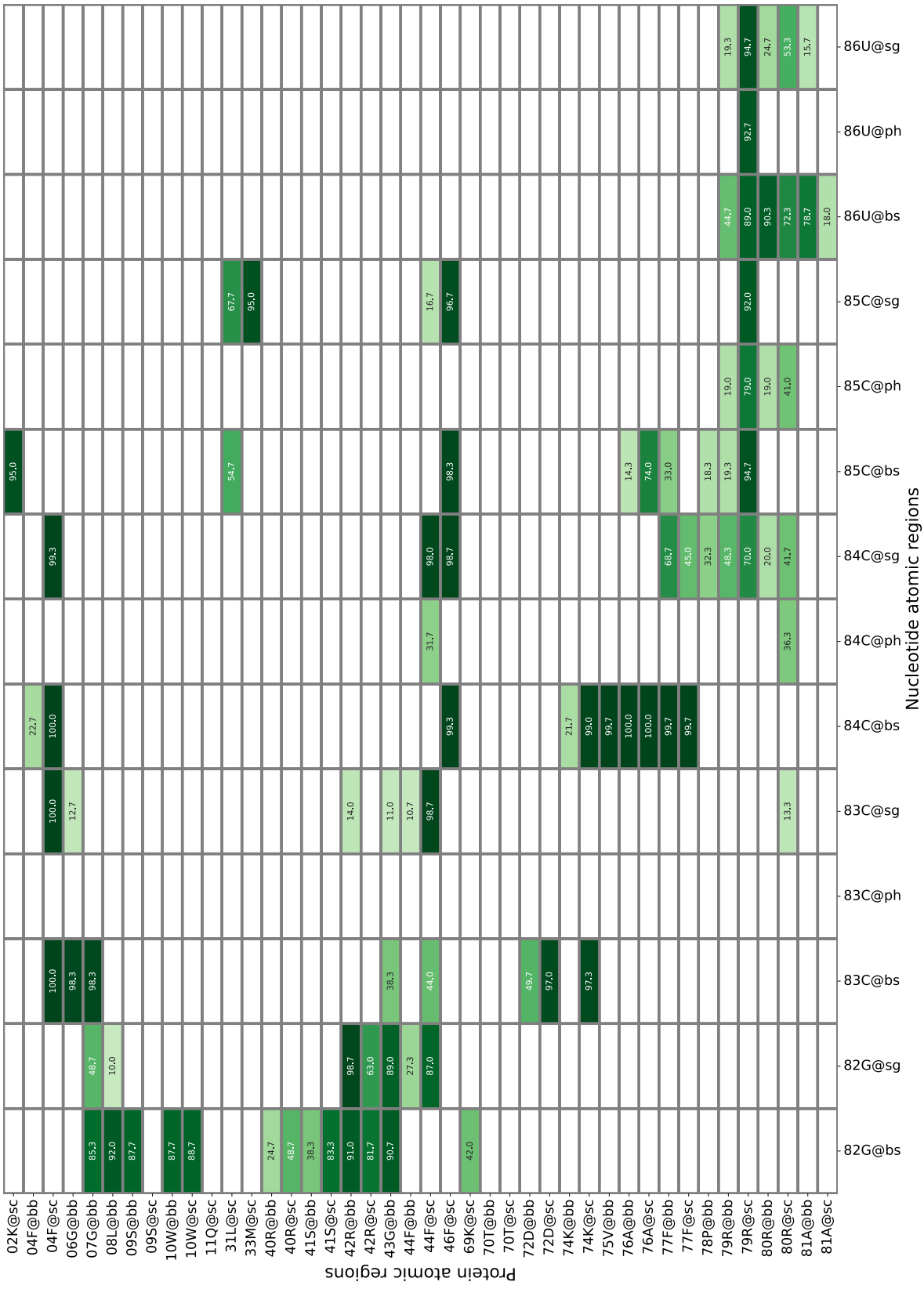


Figure 5.14: Tracking of native and non-native contacts over time from the Musashi1 RRM1-mutRNA ('GCCCU') complex. The green color represents native contacts, blue represents non-native contacts, and red color represents lost native contacts. The intensity of color shows how many frames have that respective contact.

5.5.2 Free Energy Computation

The computation of free energy will help us to check the effect of point mutations in RRM domain on the RNA-binding activity. To establish and check the protocol for RRM-RNA complexes, we first performed MD simulations with complexes having binding affinity information available. From InteR3M database, we already have the information about RRM-RNA complexes with experimental data for binding affinities.
The RNA binding affinities of RRM domain from SRSF2 protein with several point mutations have been quantified in vitro [Daubner et al., 2012, Phelan et al., 2012, Kim et al., 2015]. Kim et al. [2015] showed that the proline mutations at 95^{th} position changes the RNA binding specificity of SRSF2 RRM. The mutagenesis studies from Phelan et al. [2012] demonstrated that the residues from loop3 region (in between β2 and β3) R47, D48 and K52 are responsible for mediating the RNA binding. Daubner et al. [2012] determined the 3D structures of SRSF2 RRM, SRSF2 RRM-RNA (UCCAGU), and SRSF2 RRM-RNA (UGGAGU). Most interacting residues from the SRSF2 RRM-RNA complexes were substituted by alanine residue in order to evaluate the importance of each residue involved in RNA binding activity.

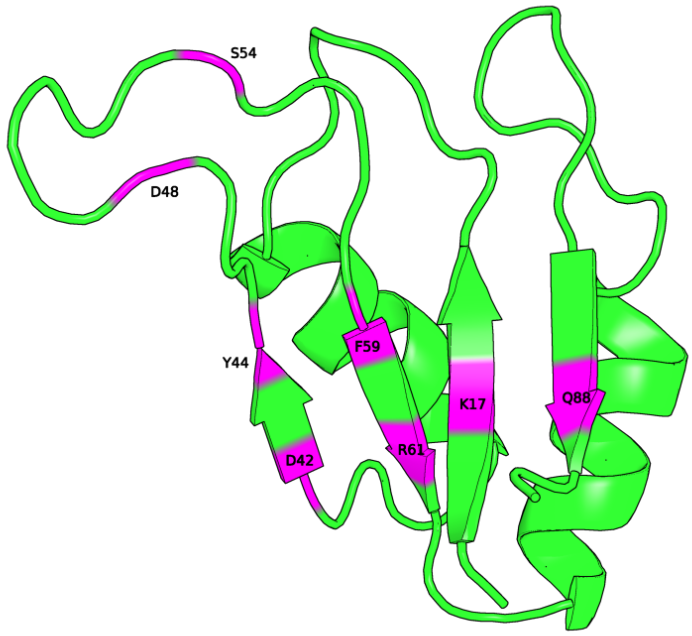


Figure 5.15: Position of all point mutations studied by Daubner et al. [2012] in SRSF2 RRM. Residues with magenta color represent the mutated residues.

We decided to select the study from Daubner et al. [2012] for testing the protocol for computing binding free energies of RRM-RNA complexes, as it covers the point mutations from different regions of RRM domain. Figure 5.15 shows the positions of all point mutations (from Daubner et al. [2012] study) on RRM domain.

System preparation

All the systems with dual topology were prepared using VMD and its plugins. We used 3D structure of SRSF2 RRM-RNA from PDB ID: 2LEB. The primary topology files were generated with the 'autopsf plugin'[8] from VMD using CHARMM36 force field. The dual topology for the wild type and mutant states were created using the 'mutator plugin'[9] from VMD. After this, the explicit solvent was added using TIP3P model inside a cubic box with a space of at least 30 Å around the RRM-RNA complex in each direction. Na^+ and Cl^- ions were added in the system to make the salt concentration of 0.15 M.
For each point mutation, two systems were prepared in the same manner, bound (RRM-RNA complex) and unbound (RRM). For each prepared system, we also created a file with the perturbation flags for concerned atoms. The format of this file is similar to the '.pdb' file with flag of '**-1**' for the disappearing atoms (in our case wild type residue), '**1**' for appearing atoms (in our case mutated residue), and '**0**' for rest of the atoms.

Simulation protocol

All MD simulations were performed using NAMD 3.0 at 310 K with a time step of 1 fs. Langevin dynamics was used to maintain the constant temperature and pressure (1 atm). Particle mesh ewald (PME) was used to compute the long-range electrostatic interactions. A cutoff distance of 12 Å was used for non-bonded interactions. The SHAKE algorithm was used to constrain the covalent bonds involving hydrogen atoms. We started simulations with the minimization step for 20 ps (20,000 steps) using conjugate gradient and line search algorithm. After minimization, the system was equilibrated with NVT and NPT ensembles, respectively for 500 ps each. Finally, bidirectional (forward and backward) alchemical tranformations were performed with $\lambda = 0.0625$. Thus, each alchemical transformation will take place in $\frac{1}{\lambda} = 16$ steps. Figure 5.16 shows the thermodynamic cycle we used to compute the relative binding free energy ($\Delta\Delta G$) with point mutations on RRM (receptor) while keeping the RNA (ligand) same.

We computed the relative binding free energy for a set of point mutations on RRM that can be found in Table 5.6.

Analysis

We used 'ParseFEP plugin'[10] from VMD for analysis of free-energy perturbation calculations [Liu et al., 2012]. ParseFEP computes the free-energy difference and provides an estimate of the statistical error based on the output files from FEP simulation. It combines the results of the forward and the backward simulations in the form of the simple-overlap sampling (SOS) estimator, or the Bennett acceptance-ratio (BAR) estimator [Bennett, 1976] of the free energy.

[8] https://www.ks.uiuc.edu/Research/vmd/plugins/autopsf/
[9] https://www.ks.uiuc.edu/Research/vmd/plugins/mutator/
[10] https://www.ks.uiuc.edu/Research/vmd/plugins/parsefep/

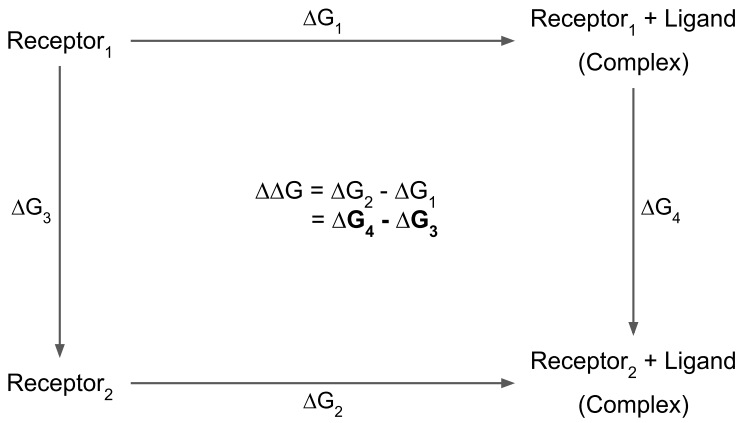
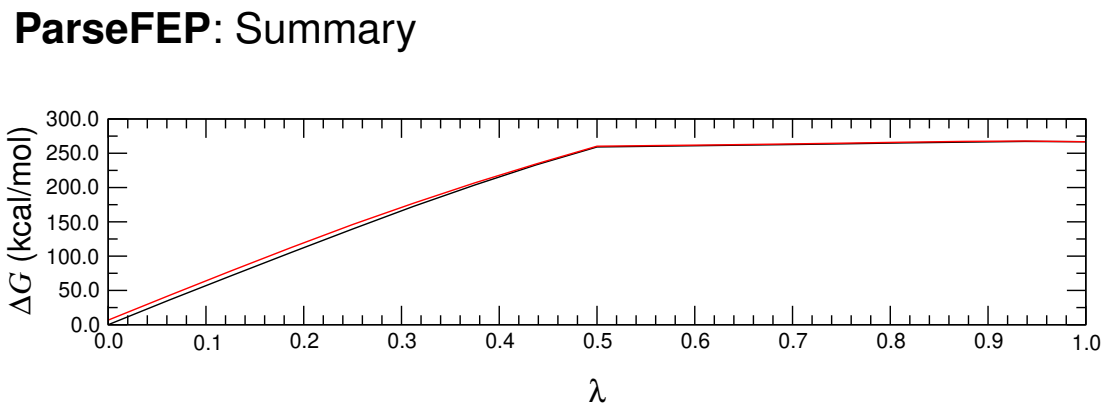


Figure 5.16: A thermodynamic cycle describing the binding of two Receptors, (Receptor_1 and Receptor_2) to a Ligand. The relative free energy of binding can be calculated from either the physical (ΔG_2 – ΔG_1) or alchemical (ΔG_4 – ΔG_3) legs of the cycle.

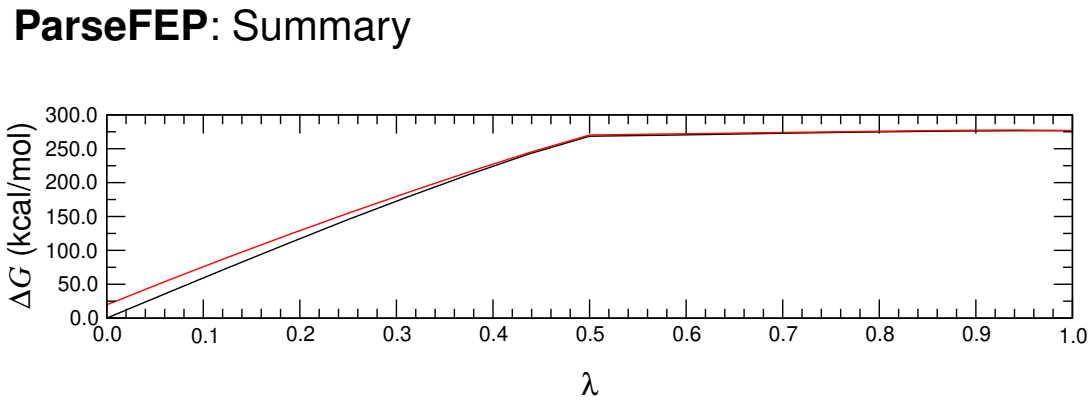
Table 5.6: Binding affinities of SRSF2 RRM mutants with RNA ('UCCAGU'), experimental values are from the study of Daubner et al. [2012] and computational values are computed in this thesis.

Point mutation performed	**Experimental**		**Computational**
	K_d (in µM)	**Affinity decrease (in folds)**	**$\Delta\Delta G$ (in kcal mol^{-1})**
SRSF2 RRM (aa 1-101) + 5′-UCCAGU-3′			
Wild type	0.27 ± 0.02		
K17A	2.38 ± 0.38	9	−−
D42A	4.09 ± 0.58	15	1.48 ± 0.02
Y44A	>5	>20	−−
D48A	0.83	3	-1.42 ±0.58
S54A	0.37	1	1.322 ± 0.09
F59A	Unfolded	−−	4.08 ± 0.20
R61A	>5	>20	2.83 ± 0.42
Q88A	0.35	1	0.202 ±0.11

R61A point mutation : The Arg61 resides on $\beta 3$ within RNP1 sequence motif. Figure 5.17 shows the overall free energy profile for R61A mutation in unbound and bound form of SRSF2 RRM domain.
The total free energy change (ΔG_3) of R61A mutation in unbound form of SRSF2 RRM is 263.44 kcal mol^{-1} with a total error of 0.250 kcal mol^{-1}. Whereas, in the bound form the free energy change (ΔG_4) for R61A point mutation is 266.27 kcal mol^{-1} with a total error of 0.334 kcal mol^{-1}.



(a) Free energy change for Arg61 to Ala mutation in SRSF2 RRM domain (unbound form)



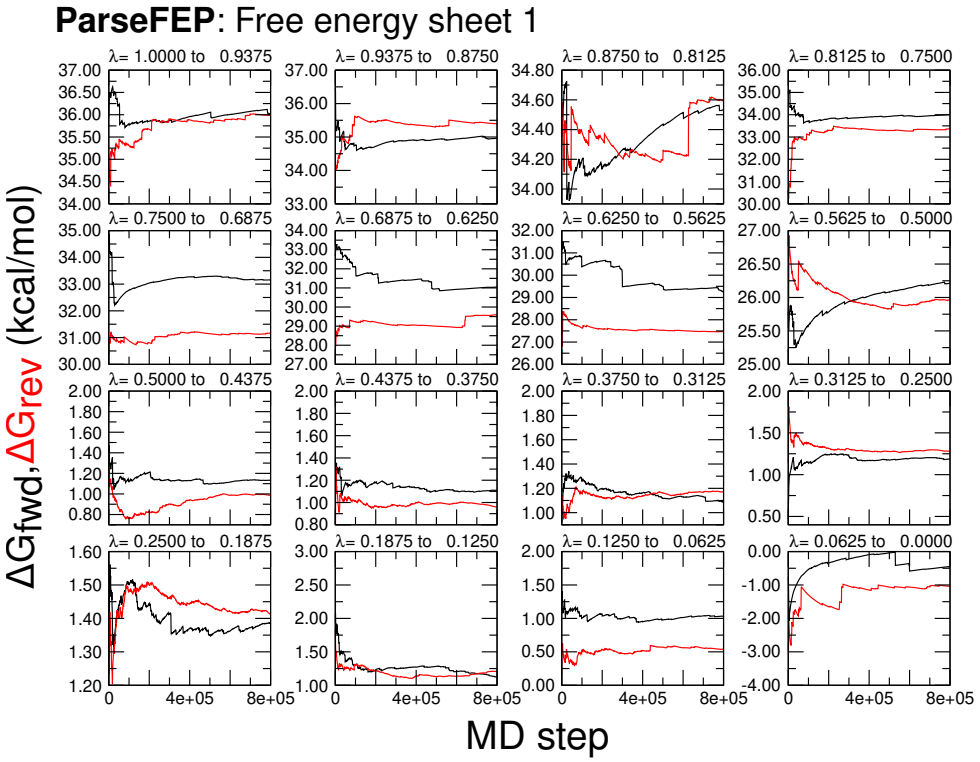
(b) Free energy change for Arg61 to Ala mutation in SRSF2 RRM domain (bound form)

Figure 5.17: Free energy change for Arginine (position 61) to Alanine mutation in SRSF2 RRM domain. Black line indicates the forward transformation, i.e., from Arg to Ala and red line indicates the backward transformation from Ala to Arg.

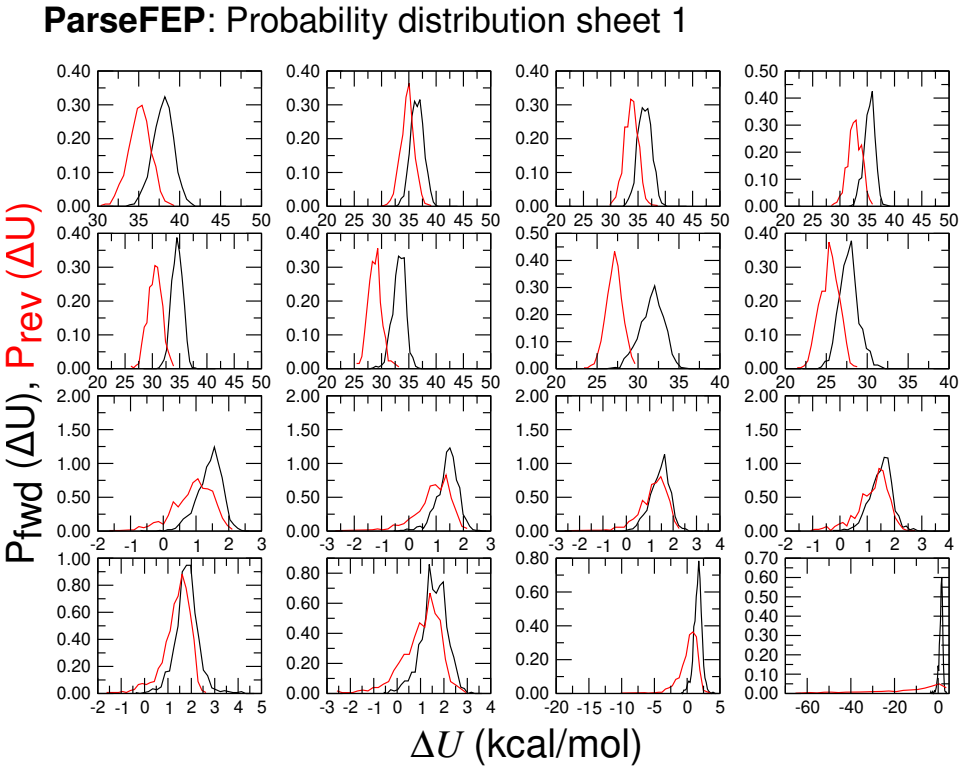
Overlapping free energy profiles for forward and backward transformations is a sign for the convergence of the free energy calculation. This is further broken down per window (λ) size in Figure 5.18a and Figure 5.19a for unbound and bound form of RRM, respectively.
The convergence of the free energy calculation can be assessed by monitoring the time-evolution of $\Delta G(\lambda)$ for every individual window (λ-state) and the overlap of configurational ensembles embodied in their density of states, $P_0[\Delta U(x)]$ and $P_1[\Delta U(x)]$, where $\Delta U(x) = U_1(x) - U_0(x)$ denotes the difference in potential energy between the target and the reference states (Figure 5.18 and Figure 5.19).

To compute the relative binding free energy for R61A point mutation we will use the equation $\Delta\Delta G = \Delta G_4 - \Delta G_3$ (see Figure 5.16).

$$\begin{aligned}\Delta\Delta G &= (266.27 \pm 0.334\,\mathrm{kcal\,mol^{-1}}) - (263.44 \pm 0.250\,\mathrm{kcal\,mol^{-1}}) \\ &= (266.27 - 263.44) \pm (\sqrt{0.334^2 + 0.250^2})\mathrm{kcal\,mol^{-1}} \\ &= 2.83 \pm 0.417\,\mathrm{kcal\,mol^{-1}}\end{aligned}$$

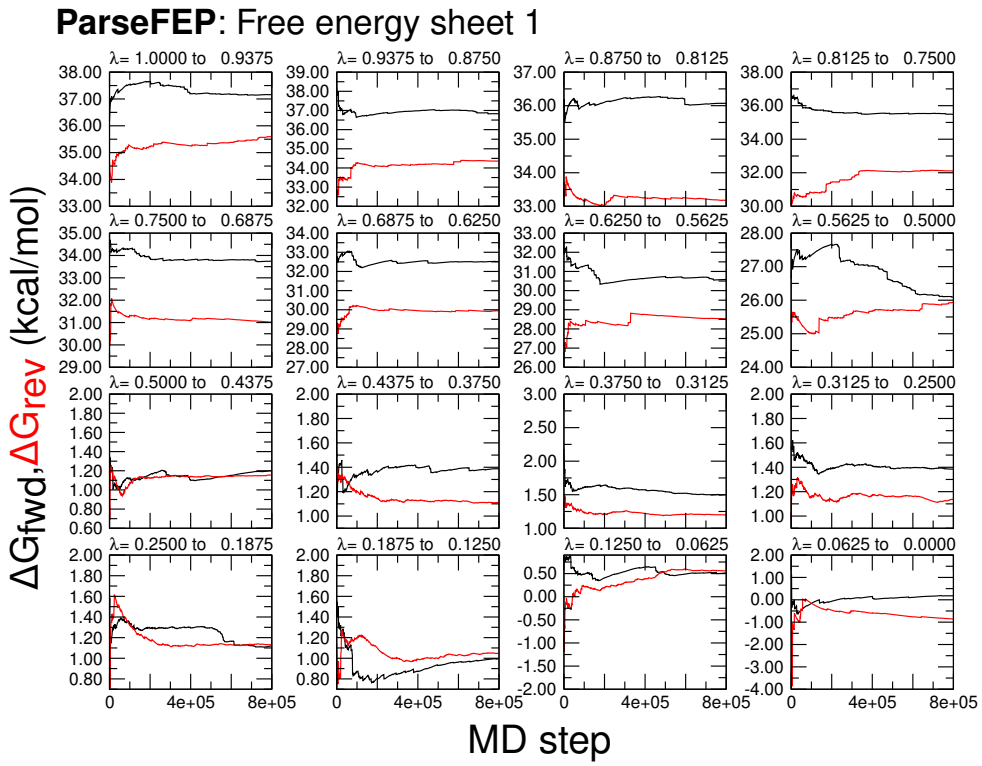


(a) Time evolution of free energy differences for each λ window (unbound RRM)

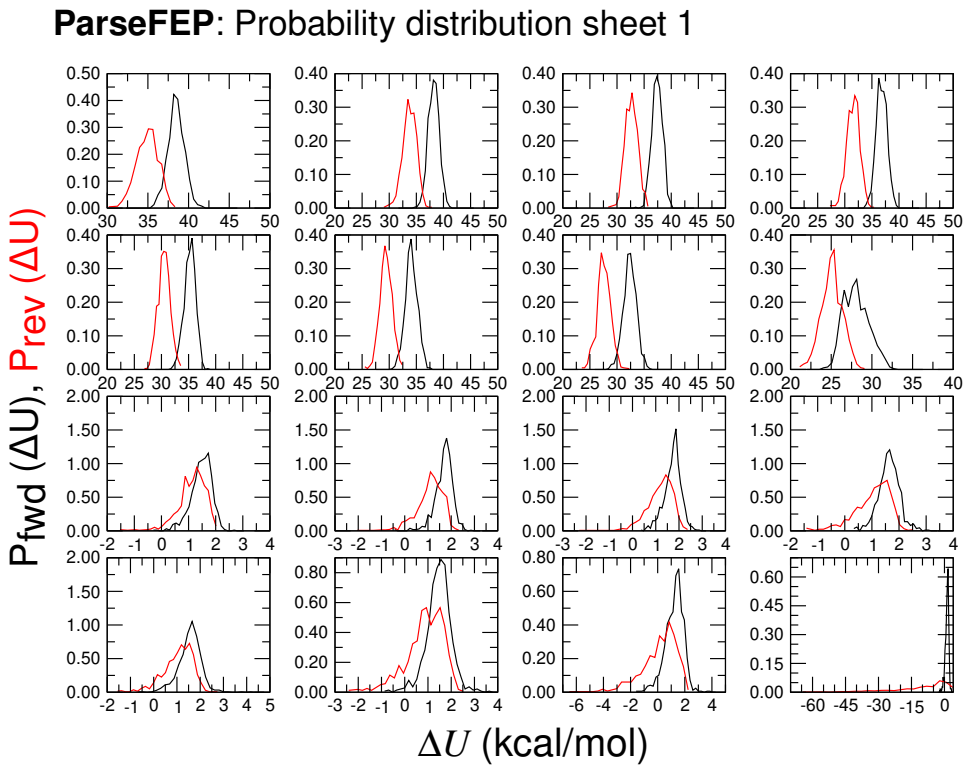


(b) Probability distribution plots for backward and forward transformations (unbound RRM)

Figure 5.18: Output plots from the soft-core potential calculation, of ΔG, and $P_0[\Delta U]$ and $P_1[\Delta U]$ generated by ParseFEP (for unbound RRM). Each subplot corresponds to a λ sampling window.



(a) Time evolution of free energy differences for each λ window (RRM-RNA complex)



(b) Probability distribution plots for backward and forward transformations (RRM-RNA complex)

Figure 5.19: Output plots from the soft-core potential calculation, of ΔG, and $P_0[\Delta U]$ and $P_1[\Delta U]$ generated by ParseFEP (for bound RRM-RNA complex). Each subplot corresponds to a λ sampling window.

Thus, the relative binding free energy for R61A point mutation in SRSF2 RRM-RNA complex is $2.83 \pm 0.417\,\mathrm{kcal\,mol^{-1}}$.

In the same way, we computed relative binding free energies for other point mutations from SRSF2 RRM and the results are in Table 5.6.

The higher the binding free energy, the lower the binding affinity and the less stable the complex. The lower the binding free energy, the higher the binding affinity and the more stable the complex.
Thus, the point mutation R61A in SRSF2 RRM is less stable relative to wild type SRSF2 RRM-RNA complex. In the study from Daubner et al. [2012], the point mutation R61A resulted in a decrease in affinity by 20-fold.

The wild type RRM-RNA complex is favourable compared to all point mutations performed in our study (FEP simulations) except the point mutation D48A. The D48 (Aspartate) amino acid resides in the loop3 between β2 and β3 (See Figure 5.15). The relative binding free energy computed from FEP simulations for this point mutation (D48A) is -1.414 $\pm$ 0.32 $\mathrm{kcal\,mol^{-1}}$ (see Appendix C.2). In contrast, in the study from Daubner et al. [2012] this point mutation (D48A) resulted in a decrease in affinity by 3-fold.

This could be because this point mutation needs more sampling time or shorter window size (λ-step). Because of the lack of time we were not able to check the effect of window size of the convergence of calculations. But there are studies showing the impact of window size and sampling time on the relative binding free energy calculations using FEP protocol [Guest et al., 2022].

To proceed with the relative binding free energy computations with FEP protocol, we need to first benchmark FEP protocol for RRM-RNA system with different sampling time and window size, like the study from Guest et al. [2022]. This would help us to better understand the impact of sampling time and window size on the accuracy and error rate of FEP calculations.

After benchmarking the FEP protocol, we can test it on other point mutations either on the same system using studies from Phelan et al. [2012], Kim et al. [2015] or on any other RRM-RNA system.

Chapter 6

Conclusions & Perspectives

Summary

6.1	**Summary of the main contributions**	**118**
6.2	**InteR3M Database**	**118**
6.3	**CroMaSt Workflow**	**119**
6.4	**RRMScorer**	**120**
6.5	**RRM-RNA Dock**	**120**
6.6	**Evaluating 3D structures of RRM-RNA complexes**	**121**
6.7	**Future Directions**	**121**

6.1 Summary of the main contributions

In this thesis, I have explored the diversity among RRM domains to better understand their binding characteristics with the aim of contributing to the design of RRM-containing proteins with desired binding activity. I carefully curated a set of Pfam families having the RRM domains via structural inspection of family members. This made it clear that not all the families from RRM clan of Pfam have RRM domains.

The main contributions of the thesis are as follows:

1. I have developed a database for Interactions of RNA and RNA Recognition Motif (InteR3M), available at https://inter3mdb.loria.fr/.
2. I have developed a workflow for assessing domain classification by cross-mapping of structural instances between protein domain databases (CroMaSt) and assessing structural alignment of unmapped instances with an RRM structural prototype, available at https://workflowhub.eu/workflows/390.
3. I have contributed to the computational scoring method for estimating the binding between an RRM and a ssRNA (RRMScorer), available at https://bio2byte.be/rrmscorer/.
4. I have contributed to integrate information from InteR3M database and RRMScorer predictions to model the 3D structures of RRM-ssRNA complex (RRM-RNA dock), available at https://github.com/AnnaKravchenko/RRM-RNA-dock.
5. I have tested different molecular dynamics (MD) simulation protocols to evaluate the 3D models of RRM-ssRNA complexes.

The contributions 3 and 4 from this list were performed in the frame of collaborations with RNAct partners: Joel Roca-Martinez and Anna Kravchenko, respectively, within the RNAct project.

6.2 InteR3M Database

The foundation for InteR3M database is laid by the careful delineation of a correct set of RRM families from Pfam database (Section 3.2.3). InteR3M database integrates the domain information from Pfam, protein information from UniProtKB, structural information from PDB, and experimental binding (K_d values) information from literature. The current version of InteR3M database contains 400,892 RRM domain instances, of which only 303 RRMs have at least one experimentally solved 3D structure in the PDB. A total of 459,859 interactions are stored from 656 RRM-RNA complexes. InteR3M database provides an easy to use interface for querying the interactions, experiments, structures, and sequences of RRM domains.

The InteR3M database will be updated with every new release of Pfam database, and the binding information from literature will be added regularly (Section 3.6). We are planning to add the CroMaSt annotations (core, true, false) and provide a

prototype structure for the RRM domain. In addition, we are considering to add information from CATH-Gene3D database to include RRM domains captured only by CATH.

In the near future, we will consider integrating supplementary binding information from other RNA-binding domain resources such as RBPDB as well as evolutionary information from RRMdb. We will also investigate other generalist domain databases, such as ECOD or CDD, for the RRM domain instances that can not be captured by Pfam. In this way the InteR3M database can play the role of a hub for RRM domain information.
The code used for data collection is publicly available, so the same method can be followed to create the database for any other RNA/DNA-binding domain. The code is available at https://gitlab.inria.fr/hdhondge/data_collection_inter3mdb.

6.3 CroMaSt Workflow

The "Cross-Mapper of Structural domains" (CroMaSt) workflow is a fully automated workflow that classifies all structural instances (StIs) of a given domain into 3 different categories: core, true and domain-like. This workflow provides an easy way to curate a list of domain instances for a given domain type from two different domain classification systems, Pfam (mostly sequence-based) and CATH (structure-based). In addition, CroMaSt computes an average prototype structure for the given domain type and also detects wrongly classified domain instances thanks to structural alignment. Current version of CroMaSt was built to take advantage of both, sequence-based (Pfam) and structure-based (CATH), classification systems. It can be used to look for structurally related families across protein domain classification databases.

Currently, CroMaSt can be run with only two source databases, Pfam and CATH. To open CroMaSt to any other source domain databases, we plan to create a unified format from release files of other domain databases. This will allow users to use CroMaSt with domain databases of their choice for cross mapping of StIs. Moreover, users can contribute to source domain databases in case of wrongly classified domain StIs or to assign classification for StIs lacking any classification.

The current version of CroMaSt uses only experimentally determined 3D structures because these are directly available from source databases. CroMaSt workflow can be used with AlphaFold structures once the structure-based domain databases (CATH) have integrated information from AlphaFold database. The usage of AlphaFold models will give more power to CroMaSt workflow in terms of inclusion of the families hitherto devoid of any experimental structure (from sequence-based classification) and addition of new families (from structure-based classification).

We believe that CroMaSt will be a valuable tool for developers of protein domain data resources like Pfam and CATH. CroMaSt can be used to improve the quality and knowledge of domains in these domain data resources. For example, to learn more about any DUF (Domain of Unknown Function) families from Pfam.

Moreover, CroMaSt can be useful for people interested in evolutionary relationships of protein domains. One can use the ECOD (Evolutionary Classification of Protein Domains) [Cheng et al., 2015] as a source database for cross-mapping of domain structural instances.

6.4 RRMScorer

RRMScorer provides scores for the binding probability of any RRM-RNA pair asssuming they use the canonical binding mode. RRMScorer was validated on both computational and experimental data. The key tasks of this work are the generation of RRM master alignment and the mapping of contacts onto the alignment. These tasks led us to understand the role of each residue positions and identify different binding modes of RRM domains. For this work, we focused only on the canonical binding mode of RRMs, therefore the other binding modes remain to be investigated. However, for the moment, the number of instances for these less frequent binding modes is too low to compute statistically reliable scores.

The RRM-RNA interaction data used to develop the RRMScorer method is retrieved from the InteR3M database. With the addition of new RRM-RNA complexes in InteR3M database, RRMScorer will be updated accordingly. Hopefully, the various updates of InteR3M database will provide more 3D structures of RRM-RNA complexes with non-canonical binding modes. This will be useful to analyze and decipher the RRM recognition code for other binding modes. RRMScorer can score an RRM-RNA complex while tracking the individual scores of each residue-nucleotide pair. This information can be useful to rationally design new RRMs. The RRMScorer can be coupled with state-of-the-art methods used to predict the 3D structure of the RNA-RRM complex, such as RoseTTAFoldNA [Baek et al., 2022] and ATTRACT.
Finally, this methodology can be used to decipher the recognition code for other RNA/DNA-binding domains, assuming there exists sufficient structural information on 3D complexes of these domains with RNA or DNA.

6.5 RRM-RNA Dock

We employed anchored docking protocol for RRM-RNA docking with the ATTRACT docking program. The interactions between RRM and RNA were used to identify anchoring patterns, i.e. prototypes of 3D atomic positions (relative to the protein backbone) of a nucleotide stacked on a conserved aromatic amino acid. We used all the experimentally determined 3D structures of RRM-RNA complexes to extract the stacking interactions. Then, we used the representatives for all these stacking interactions to model the 3D structures of RRM-RNA complexes. The current version of 'RRM-RNA dock' can be used to dock only the tri-nucleotide fragment containing the anchored nucleotide onto the RRM domain.

The current version of 'RRM-RNA dock' uses the same constraint distance for each bead. The use of different constraint distances depending on the variability of the beads position within the cluster might help to get better results.

The 'RRM-RNA dock' pipeline is still in development as discussed in section 5.4.4. Another PhD student (Anna Kravchenko) from the RNAct project in our lab is working on it to integrate the RRMScorer predictions.

6.6 Evaluating 3D structures of RRM-RNA complexes

We tested the standard MD protocol to distinguish between strongly bound and weakly bound RRM-RNA complexes using MSI1 RRM1-RNA complex. But this protocol was not able to successfully distinguish between these two systems.
We then used free energy perturbation (FEP) protocol to compute the relative binding free energy of RRM-RNA complexes. We tested this protocol with point mutations on SRSF2 RRM (from Daubner et al. [2012]), but the results are not completely in agreement with the study from Daubner et al. [2012].
Due to the lack of time, we could not test other MD protocols or the same protocol with different settings.

To develop an efficient FEP protocol on such a system formed by an RRM-RNA complex, a benchmarking study can be performed. This will help to understand the effect of sampling time and window size on the accuracy and error rate of FEP calculation [Guest et al., 2022].
A few other MD protocols can also be tested to evaluate the 3D structures of RRM-RNA complexes like constant velocity pulling and accelerated molecular dynamics [Pawnikar et al., 2022].

6.7 Future Directions

This thesis presents a set of computational tools that can be used in different studies. The evolutionary aspects of a protein domain are useful to understand the function of newly discovered proteins and reveal co-evolving interactions at molecular level [P Bagowski et al., 2010, Basu et al., 2009]. The curated data from InteR3M database can be used to extend the evolutionary analysis of RRM domains initiated recently [Nowacka et al., 2019]. In particular, this will help to understand the promiscuity of different RRM families. Promiscuity refers to the presence of a domain in combination with many other domains. This can be extended to all proteins containing at least one RRM domain resulting in all possible multi-domain architectures of RRM domains. These multi-domain architecture can provide some valuable insights into the ability of RRM to interact with a wide range of molecular partners (RNA, DNA, and proteins). In addition, it can reveal if any other neighboring domains have an impact on the function of RRM domains leading to either a deviant domain or a novel function. Forslund and Sonnhammer [2012] reviewed several studies to dissect the evolution of domain architectures and provides an overview on different scenarios.

The study from Oliveira et al. [2017], which focuses on RNA-binding proteins (RBPs) in *Trypanosoma cruzi* and *Saccharomyces cerevisiae*, provides interesting insights on domain architecture among RBPs. This study also states that in some

cases, the function of a protein of interest is conserved despite the phylogenetic distance. The phylogenetic analysis of RRM-containing proteins in plants performed by Gomez-Porras et al. [2011] revealed that the RRM from plants and cyanobacteria do not have a common origin. Thus, it would be interesting to perform a phylogenetic analysis of RRM-containing proteins to understand and learn about common origin of RRMs and how they evolved to preserve their functions. It would also provide insights on the multiple divergence steps of RRM domain evolution, that lead to their extremely large diversity.

The RRM domain plays an important role in several key biological processes including post-transcriptional gene regulation, formation of amyloid-like aggregates [Berchowitz et al., 2015], and abnormal cell proliferation [Chen et al., 2019]. The 'HuR (ELAV1) protein' with three RRM domains is an established regulator of post-transcriptional gene regulation in humans. The HuR protein is overexpressed and over-active, i.e. with an increased subcellular localization within cytoplasm, in most cancers [Schultz et al., 2020]. HuR is a promising target for cancer therapies [Blanco et al., 2016]. HuR nucleocytoplasmic shuttling sequence (HNS) and RRM3 play critical role in its cytoplasmic localization [Doller et al., 2010]. The study from Grammatikakis et al. [2017] summarizes the impact of each post-translational modification on HuR localization and function. We believe that the computational tools developed during this thesis will be very useful in such cases. For example, the RRMScorer can be used to find RNA candidates that can bind to HuR (or individual RRMs) with relatively higher score than mRNA known as targets. The structure of HuR with these RNA candidates can be modelled using 'RRM-RNA dock'. These 3D models can be studied to compare inter-molecular interactions between HuR and RNA candidates. Currently, there is no MD protocol for alchemical transformations of RNA but other MD protocols can be used to compute the relative binding free energies for the different RNA sequences to filter out and select the best candidates. Then, the selected candidates can be tested in-vitro to inhibit the HuR.

In addition to HuR, RRM-containing proteins are involved in many other diseases. For example, Tar DNA-binding protein 43 (TDP-43) with 2 RRM domains is a key player in Amyotrophic Lateral Sclerosis (ALS). Elevated expression of Musashi (MSI1 and MSI2) proteins has been observed in several tumors from different organs [Fox et al., 2016, Kudinov et al., 2017] and some chronic diseases like chronic myelogenous leukemia (CML), acute myelogenous leukemia (AML), and acute lymphoblastic leukemia (ALL) [Kharas et al., 2010]. These RRM-containing proteins are well-established therapeutic targets and our tools can help in guiding the inhibition of these over-expressed proteins.

Controlled management of multi-drug resistant pathogens is both critical and challenging [Fournier et al., 2006]. Recently, Ciani et al. [2022] uncovered the existence of an RRM-containing protein in *Acinetobacter baumannii* (gram-negative multi-drug resistant pathogen) that binds to AU-rich regions. Despite the fact that the specific role of this protein in *A. baumannii* is not known yet, its RNA-binding properties could be targeted, in view of both functional studies and novel therapeutic strategies.

Application domains for RNA-binding protein design are very diverse but

computational skills required to support such protein design projects are also very diverse. The contributions reported in this thesis illustrate this diversity which ranges from sequence and structure analyses to MD simulations, including database creation, workflow development and docking algorithms. Nowadays, all these skills should be brought together in the pluridisciplinary task forces addressing the challenging issues raised by protein design.

Appendix A

Structural Inspection

A.1 Classification codes and Alignment scores

A.1.1 SCOP classification

The SCOP database contains classification for non-redundant protein domains. A representative is selected based on its sequence (UniProtKB) and structure (PDB) and used for manual SCOP classification.
SCOP has six levels of classification:

- **Family** groups closely related proteins with a clear evidence for their evolutionary origin.
- **Superfamily** brings together more distantly related protein domains.
- **Fold** groups superfamilies on the basis of the global structural features shared by the majority of their members.
- **IUPR** (Intrinsically Unstructured Protein Region) organises superfamilies of proteins or protein regions that do not adopt globular folded structure.
- **Classes** bring together folds and IUPRs with different secondary structural content.
- **Protein type** groups folds and IUPRs into four groups: soluble, membrane, fibrous and intrinsically disordered.

SCOP uses following abbreviations to denote the classification levels: TP=protein type, CL=protein class, CF=fold, SF=superfamily, FA=family

A.1.2 CATH classification

The name CATH derives from the initials of the top four levels of the classification - (C)lass, (A)rchitecture, (T)opology and (H)omologous Superfamily.

- **Class** refers to the secondary structure content (e.g. mainly-alpha, mainly-beta, mixed alpha/beta or 'few secondary structures').

Level	CATH Code	Description
C	3	Alpha Beta
A	3.30	2-Layer Sandwich
T	3.30.70	Alpha-Beta Plaits
H	3.30.70.330	RRM (RNA recognition motif) domain

Figure A.1: CATH classification levels for RNA Recognition Motif domain

- **Architecture** refers to the general arrangement of the secondary structures irrespective of connectivity between them (e.g. alpha/beta sandwich).
- **Topology**, also known as the 'fold' level, takes into account the connectivity of secondary structures in the chain.
- **Homologous Superfamily** refers to domains that are believed to be related by a common ancestor.

Figure A.1 shows the levels of RRM domain classification in CATH.

A.1.3 Kpax alignment scores

K-Score: K-score (Kpax Score) is calculated using

$$K_{A,B} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \mu_{i,j} K_{i,j}$$

where,

- $\mu_{i,j}$ is 1 when residue i of the first protein is aligned to residue j of the second protein, and zero otherwise.
- $K_{i,j}$ is the similarity score for residues i and j[1].

Despite being a global structural similarity score, it is worth noting that this penalty-free score is 'pose-invariant' in that it does not depend on the orientations of the given proteins, and that for two perfectly matching backbones it will be numerically equal to the number of aligned residues.

G-Score: Global alignment score calculated using:

$$G = \sum_{i,j} \mu_{i,j} G_{i,j}$$

where,

- $\mu_{i,j}$ is 1 when residue i of the first protein is aligned to residue j of the second protein, and zero otherwise.

[1]For details refer to Ritchie et al. [2012]

- $G_{i,j}$ is Gaussian overlap between residue i and j.

J-Score: Normalised K-score.

M-Score: M-Score as a measure of MSA quality; novel atomic Gaussian based MSA scoring function, which circumvents the number/RMSD trade-off problem.

$$M = \frac{\sum_{j=1}^{C} max(C_j, 1) - C}{(T - L)}$$

where,

- C represents the total number of columns and C_j the count of residues in the j^{th} column.
- T represents total number of residues for a given sequence.
- L represents the longest amino acid chain aligned.
- The max function ensures that at least one unit is subtracted for each column and thus deals with columns that contain poorly superposed C_α atoms.

TM-Score: TM-Score defined by the program TM-align.

RMSD: The root mean squared deviation of two aligned structures.

N/*: The total count of aligned residues in an alignment.

I/@: The total count of identical residues in an alignment.

P/!: The identity percentage of residues in an alignment.

Len: The length of the target structure.

Seg: The count of continuous segments in the target structure.

TP*: This value shows whether the retrieved structure belongs to the same CATH or SCOP family.

Match (Family)*: This shows the CATH or SCOP classification code of the matching database structure.

*: These values are useful when searching CATH or SCOP databases. In our case, we are not interested in these values as we are not searching these databases.

A.2 Structural Inspection of Pfam families

Transposase_22 (PF02994):

2YKO_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4007201
- **CATH classification**: Superfamily 3.30.70.1820 (L1 transposable element, RRM domain)

- **PDBeFold**: (Secondary structure alignment)

```
PDB 2yko:A  S h S S H s S -
PDB 1a9n:B  S h S S H - S h
PDB 2a3j:A  S h S S H - S h
```

- **Kpax**:

```
=========================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   40.05   42.12  0.4828  0.5857  0.6386  2.69   68  76
  2   40.02   41.26  0.4825  0.6014  0.6440  2.45   69  71
=========================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
10  14.7   80   1  +1  1a9n_B[0.0.0.0]
10  14.5   80   1  +1  2a3j_A[0.0.0.0]
====================================
```

2LDY_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4007201
- **CATH classification**: Superfamily 3.30.70.1820 (L1 transposable element, RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 2ldy:A  S h h S S H S h
PDB 1a9n:B  S h - S S H S h
PDB 2a3j:A  S h - S S H S h
```

- **Kpax**:

```
=========================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   43.95   42.11  0.4938  0.6151  0.6686  2.74   74  76
  2   43.07   39.70  0.4840  0.5901  0.6460  2.76   73  76
=========================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
10  13.5   80   1  +1  1a9n_B[0.0.0.0]
11  15.1   80   1  +1  2a3j_A[0.0.0.0]
====================================
```

CATH and SCOP have classification for both of the structural instances point to a family named 'L1 transposable element RRM domain-like'. In addition

both of these query structures aligned nicely with PDBeFold and Kpax. Thus, we include this family into a list of 'true RRM families'.

RRM_3(PF08777):

1OWX_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 1owx:A   S H S S H S h
PDB 1a9n:B   S H S S H S h
PDB 2a3j:A   S H S S H S h
```

- **Kpax**:

```
=========================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD N/*  D/$
==== ======= ======= ======= ======= ======= ==== ===  ===
  1   37.08   40.50  0.4065  0.6042  0.6463  2.58  71  68
  2   35.39   40.01  0.3880  0.6051  0.6454  2.49  71  69
=========================================================
=====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
11   15.5  80   1    +1 2a3j_A[0.0.0.0]
10   14.1  80   1    +1 1a9n_B[0.0.0.0]
=====================================
```

5KNW_A:

- **SCOP classification**: NO ENTRY
- **CATH classification**: NO ENTRY
- **PDBeFold**: (Secondary structure alignment)

```
PDB 5knw:A   S H S S H S h
PDB 1a9n:B   S H S S H S h
PDB 2a3j:A   S H S S H S h
```

- **Kpax**:

```
=========================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   38.00   48.99  0.4270  0.6894  0.7068  1.95   69  71
```

```
  2   38.88   48.72  0.4369  0.6784  0.6945  1.91   67  70
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
 12  17.4  80   1   +1  2a3j_A[0.0.0.0]
 4    6.0  80   1   +1  1a9n_B[0.0.0.0]
====================================
```

CATH and SCOP have classification for only the first structural instance (1OWX_A) and not the other one. Both of these databases classified the first structure into RRM domain family. In addition both of these query structures aligned nicely with PDBeFold and Kpax. Thus, we include this family into a list of 'true RRM families'.

XS (PF03468):

4E8U_A:

- **SCOP classification**: No Entry
- **CATH classification**: Superfamily 3.30.70.2890 (Not yet named)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 4e8u:A   s S H s s H h S h
PDB 1a9n:B   - S H s s H - S h
PDB 2a3j:A   - S H s s H - S h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   40.92   48.33  0.4304  0.6822  0.7057  2.06   70  74
  2   39.32   46.37  0.4135  0.6684  0.6897  2.20   70  72
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
10  14.3   80   1   +1  1a9n_B[0.0.0.0]
 9  12.9   80   1   +1  2a3j_A[0.0.0.0]
====================================
```

For the same reasons as above, this family was included in the list of 'true RRM families'.

DbpA (PF03880):

2G0C_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4004284
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 2g0c:A  S H S S H S -
PDB 1a9n:B  S H S S H S h
PDB 2a3j:A  S H S S H S h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   27.72   44.76  0.3758  0.7293  0.7292  1.95   61  65
  2   27.79   42.92  0.3767  0.7167  0.7128  1.91   61  62
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
 6   9.8  80   1   +1 1a9n_B[0.0.0.0]
 5   8.2  80   1   +1 2a3j_A[0.0.0.0]
====================================
```

5B88_A:

- **SCOP classification**: NO ENTRY
- **CATH classification**: Superfamily 3.30.70.3360 (not yet named)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 5b88:A  S H - S H - -
PDB 1a9n:B  S H s S H s h
PDB 2a3j:A  S H s S H s h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   28.58   31.32  0.3766  0.5344  0.5809  2.76   63   60
  2   28.49   30.72  0.3754  0.5076  0.5603  2.78   60   64
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
 4   6.3  80    1  +1 2a3j_A[0.0.0.0]
 6  10.0  80    1  +1 1a9n_B[0.0.0.0]
```

```
==================================
```

For the same reasons as above, this family was included in the list of 'true RRM families'.

Nup35_RRM (PF05172):

3P3D_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 3p3d:A  S H h S H S h
PDB 1a9n:B  S H h S H S h
PDB 2a3j:A  S H h S H S h
```

- **Kpax**:

```
=========================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD
     N/*  D/$  I/@  P/!  Len  Seg  TP  Match[Family]
==== ======= ======= ======= ======= ======= ====
     ===  ===  ===  ===  ===  ===  ==  =============
  1   48.25   50.89  0.5784  0.7092  0.7209  1.87
       69  68 12  17.4  80   1   +1 1a9n_B[0.0.0.0]
  2   39.92   45.74  0.4785  0.6608  0.6812  2.12
       69  67 13  18.8  80   1   +1 2a3j_A[0.0.0.0]
=========================================================
```

1WWH_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 1wwh:A  S H S S H S h
PDB 1a9n:B  S H S S H S h
PDB 2a3j:A  S H S S H S h
```

- **Kpax**:

```
=========================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   48.75   52.10  0.6056  0.7283  0.7400  1.87   71   71
```

```
  2   45.45   48.93  0.5646  0.6974  0.7123  1.92   70   70
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
 9  12.7  80    1  +1 1a9n_B[0.0.0.0]
 10 14.3  80    1  +1 2a3j_A[0.0.0.0]
====================================
```

5UAZ_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 5uaz:A  S H h S H s S h
PDB 1a9n:B  S H h S H s S h
PDB 2a3j:A  S H h S H s S h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   49.80   51.91  0.5805  0.7204  0.7256  1.66   68   69
  2   41.52   46.00  0.4840  0.6604  0.6818  2.17   69   68
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
15  22.1  80   1   +1 1a9n_B[0.0.0.0]
 9  13.0  80   1   +1 2a3j_A[0.0.0.0]
====================================
```

For the same reasons as above, this family was included in the list of 'true RRM families'.

RNA_bind (PF08675):

3CTR_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 3ctr:A  S H S s h h -
PDB 1a9n:B  S H S s h s h
PDB 2a3j:A  S H S s h s h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   28.61   31.13  0.3823  0.4962  0.4982  1.95   43   63
  2   27.29   30.19  0.3646  0.4883  0.4895  1.95   43   61
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
 10  23.3  80   1  +1  1a9n_B[0.0.0.0]
 9   20.9  80   1  +1  2a3j_A[0.0.0.0]
====================================
```

1WHV_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 1whv:A  S H s S H S h
PDB 1a9n:B  S H s S H S h
PDB 2a3j:A  S H s S H S h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD N/* D/$
==== ======= ======= ======= ======= ======= ==== === ===
  1   39.86   46.17  0.5014  0.6613  0.6740  1.88  65  64
  2   37.17   44.60  0.4675  0.6465  0.6616  1.97  65  66
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
14  21.5  80   1  +1  1a9n_B[0.0.0.0]
13  20.0  80   1  +1  2a3j_A[0.0.0.0]
====================================
```

3D45_A:

- **SCOP classification**: refer 1whv

- **CATH classification**: Superfamily 3.30.420.10 (Ribonuclease H-like superfamily/Ribonuclease H)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 3d45:A   S H S S H - -
PDB 1a9n:B   S H S S H s h
PDB 2a3j:A   S H S S H s h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   38.77   47.78  0.5039  0.7126  0.7253  1.87  65   63
  2   36.99   45.41  0.4808  0.6993  0.7011  1.87  64   65
=============================================================
=====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
 14  21.5  80   1   +1 1a9n_[0.0.0.0]
 12  18.8  80   1   +1 2a3j_A[0.0.0.0]
=====================================
```

For the same reasons as above, this family was included in the list of 'true RRM families'.

Tap-RNA_bind (PF09162):

1KOH_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4001295
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 1koh:A    S H - - H S -
PDB 1a9n:B    S H s s H S h
PDB 2a3j:A    S H s s H S h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   36.05   39.68  0.4506  0.5881  0.6251  2.35   67  70
  2   35.18   38.03  0.4398  0.5739  0.6201  2.54   69  70
=============================================================
=====================================
```

```
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  ==============
7  10.4   80   1   +1  1a9n_B[0.0.0.0]
8  11.6   80   1   +1  2a3j_A[0.0.0.0]
=====================================
```

1FT8_A:

- **SCOP classification**: No SCOP2 classification is available for 1ft8 A explicitly. This entry is represented by the following domains:
  - 8025950 1KOH A
  - 8025952 1KOH A
  - 8038329 1KOH A
  - 8038331 1KOH A
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 1ft8:A  S h H S s H S -
PDB 1a9n:B  S - H S s H S h
PDB 2a3j:A  S - H S s H S h
```

- **Kpax**:

```
=========================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   36.56   37.51  0.4570  0.5696  0.6176  2.51   69  70
  2   38.35   37.05  0.4794  0.5624  0.6082  2.54   68  71
=========================================================
=====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  ============
 8  11.6  80    1  +1  2a3j_A[0.0.0.0]
 7  10.3  80    1  +1  1a9n_B[0.0.0.0]
=====================================
```

CATH classified both of these structures as RRM domains. SCOP classified these structures into 'Non-canonical RBD domain' family that do not point/map to any Pfam family. Although the secondary structure alignment from PDBeFold, does not seem to be aligned very well, but the structures are aligned nicely with Kpax. Thus, we include this family into a list of 'true RRM families'.

RRM_8 (PF11835):

2E5I_A:

- **SCOP classification**: No Entry
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 2e5i:A    S H S S H S -
PDB 1a9n:B    S H S S H S h
PDB 2a3j:A    S H S S H S h
```

- **Kpax**:

```
============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   47.00   55.71  0.5912  0.7776  0.7902  1.83   74  72
  2   45.34   54.43  0.5704  0.7579  0.7823  2.05   75  72
============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
18  24.3   80   1  +1  1a9n_B[0.0.0.0]
16  21.3   80   1  +1  2a3j_A[0.0.0.0]
====================================
```

2MQM_A:

- **SCOP classification**: No Entry
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 2mqm:A   h S H S S H S -
PDB 1a9n:B   - S H S S H S h
PDB 2a3j:A   - S H S S H S h
```

- **Kpax**:

```
============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   46.55   58.39  0.5893  0.8124  0.8246  1.79   75  74
  2   42.68   55.66  0.5403  0.7813  0.7971  1.91   74  74
============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
19  25.3   80   1  +1  1a9n_B[0.0.0.0]
17  23.0   80   1  +1  2a3j_A[0.0.0.0]
====================================
```

For the same reasons as above, this family was included in the list of 'true RRM families'.

RRM_5 (PF13893):

6EXN_Y:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 6exn:Y  S H S S H S -
PDB 1a9n:B  S H S S H S h
PDB 2a3j:A  S H S S H S h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   64.67   74.31  0.8134  0.9685  0.9630  0.71   79  79
  2   48.47   63.11  0.6097  0.8616  0.8563  1.24   75  76
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
25  31.6   80   1  +1  1a9n_B[0.0.0.0]
18  24.0   80   1  +1  2a3j_A[0.0.0.0]
====================================
```

2ADC_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 2adc:A  h S H S S H S s
PDB 1a9n:B  - S H S S H S h
PDB 2a3j:A  - S H S S H S h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
```

```
  1   47.36   60.17  0.6114  0.8715  0.8676  1.42   74  74
  2   48.87   59.39  0.6309  0.8619  0.8603  1.49   74  74
=========================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  ============
20  27.0   80   1  +1  1a9n_B[0.0.0.0]
20  27.0   80   1  +1  2a3j_A[0.0.0.0]
====================================
```

For the same reasons as above, this family was included in the list of 'true RRM families'.

RRM_7 (PF16367):

2MKK_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 2mkk:A   S H s h s H s S -
PDB 1a9n:B   S H s - s H - S h
PDB 2a3j:A   S H s - s H - S h
```

- **Kpax**:

```
=========================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   32.31   41.35  0.3746  0.5912  0.6313  2.53   67   67
  2   33.79   38.68  0.3917  0.5566  0.6052  2.68   66   73
=========================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family
===  ===  ===  ===  ==  ============
 9  13.4  80    1  +1 1a9n_B[0.0.0.0]
 9  13.6  80    1  +1 2a3j_A[0.0.0.0]
====================================
```

2MKJ_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 2mkj:A   s H S S H s S -
PDB 1a9n:B   s H S S H - S h
PDB 2a3j:A   s H S S H - S h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   35.69   40.40  0.4160  0.6075  0.6410  2.34   69   73
  2   32.75   36.55  0.3817  0.5695  0.6217  2.74   72   72
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
10  14.5  80   1   +1 1a9n_B[0.0.0.0]
12  16.7  80   1   +1 2a3j_A[0.0.0.0]
====================================
```

For the same reasons as above, this family was included in the list of 'true RRM families'.

RRM_occluded (PF16842):

4N0T_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 4n0t:A   h S H S S H S h
PDB 1a9n:B   - S H S S H S h
PDB 2a3j:A   - S H S S H S h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
 1    51.10   59.10  0.6554  0.8295  0.8272  1.40   70  74
 2    46.39   54.75  0.5950  0.7956  0.8042  1.85   73  72
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
 8  11.4   80   1  +1  1a9n_B[0.0.0.0]
15  20.5   80   1  +1  2a3j_A[0.0.0.0]
```

```
==================================
```

6ASO_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2000014, SF=3000110, FA=4000236
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 6ASO:A   S H S S H S h
PDB 1a9n:B   S H S S H S h
PDB 2a3j:A   S H S S H S h
```

- **Kpax**:

```
============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   50.15   58.14  0.6431  0.8217  0.8190  1.48   70  74
  2   45.02   53.52  0.5774  0.7840  0.7937  1.93   73  72
============================================================
==================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
 8  11.4   80   1  +1  1a9n_B[0.0.0.0]
15  20.5   80   1  +1  2a3j_A[0.0.0.0]
==================================
```

For the same reasons as above, this family was included in the list of 'true RRM families'.

YlmH_RBD (PF17774):

2FPH_X:

- **SCOP classification**: NO ENTRY
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 2fph:X   h S H S S H S h
PDB 1a9n:B   - S H S S H S h
PDB 2a3j:A   - S H S S H S h
```

- **Kpax**:

```
============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
```

```
  1   34.94   42.56  0.4237  0.6252  0.6396  2.05   65  69
  2   32.82   40.17  0.3979  0.6016  0.6248  2.26   66  67
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
8   12.3  80   1   +1 1a9n_B[0.0.0.0]
12  18.2  80   1   +1 2a3j_A[0.0.0.0]
====================================
```

For the same reasons as above, this family was included in the list of 'true RRM families'.

RRM_9 (PF18444):

4WPM_A:

- **SCOP classification**: No Entry
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 4wpm:A  h H s S H S -
PDB 1a9n:B  s H s S H S h
PDB 2a3j:A  s H s S H S h
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   44.48   51.28  0.5394  0.7186  0.7410  2.09   73  71
  2   39.26   48.35  0.4761  0.6901  0.7180  2.23   73  71
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
13  27.8   80   1  +1  1a9n_B[0.0.0.0]
10  13.7   80   1  +1  2a3j_A[0.0.0.0]
====================================
```

4WWU_A:

- **SCOP classification**: No Entry
- **CATH classification**: No classification
- **PDBeFold**: (Secondary structure alignment)

```
PDB 4wwu:A  S H - S H S -
PDB 1a9n:B  S H s S H S h
```

```
PDB 2a3j:A   S H s S H S h
```

- **Kpax**:

```
=========================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   45.89   48.67  0.6089  0.7810  0.7757  1.80   68  68
  2   39.70   45.76  0.5268  0.7552  0.7490  1.92   68  67
=========================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
 9  13.2   80   1  +1  1a9n_B[0.0.0.0]
10  14.7   80   1  +1  2a3j_A[0.0.0.0]
====================================
```

For the same reasons as above, this family was included in the list of 'true RRM families'.

Peptidase_C48 (PF02902):

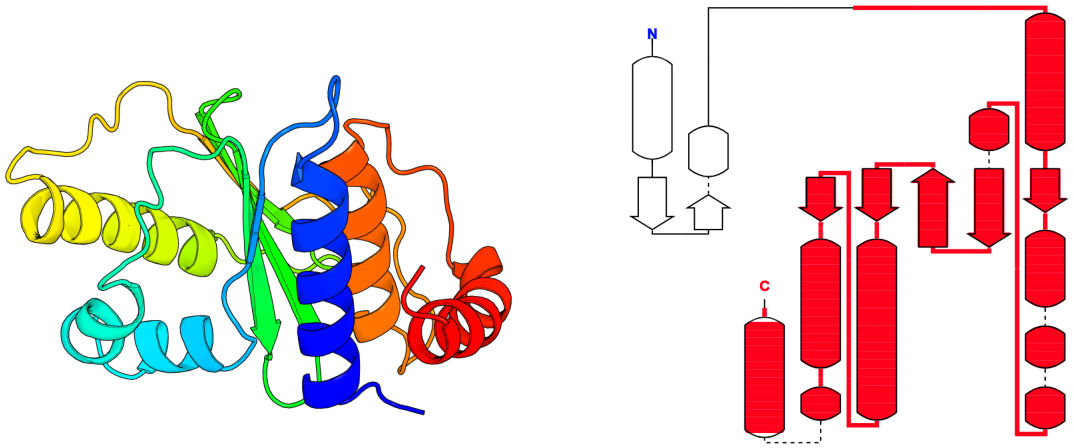
2IY1_A:

- **SCOP classification***: TP=1, CL=1000003, CF=2001107, SF=3001808, FA=4000883
- **CATH classification**: Superfamily 3.40.395.10 (Adenoviral Proteinase; Chain A)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 2iy1:A   h s s h H s h h h S S s h s h h
PDB 1a9n:B   - - - s H - - - - S S h s h - -
PDB 2a3j:A   - - - s H - - - - S S h s h - -
```

- **Kpax**:

```
=========================================================
Rank  K-Score G-Score J-Score M-Score T-Score  RMSD  N/* D/$
====  ======= ======= ======= ======= =======  ====  === ===
  1    41.67   21.79  0.3483  0.3217  0.3621   2.91  42  69
  2    38.18   20.73  0.3190  0.2966  0.3184   2.59  34  65
=========================================================
======================================
I/@  P/!  Len  Seg  TP   Match [Family]
===  ===  ===  ===  ==  ================
 3   7.1   80   1   +1  1a9n_B[0.0.0.0]
 4  11.8   80   1   +1  2a3j_A[0.0.0.0]
======================================
```



(a) 3D model of 2IY1_A from the PDB (visualised with PyMol)

(b) Secondary structure arrangement in 2IY1_A, the domain part is highlighted in red.

Figure A.2: The structural visualization of 2IY1_A (464, 642).

2HKP_A:

- **SCOP classification**: TP=1, CL=1000003, CF=2001107, SF=3001808, FA=4000883
- **CATH classification**: Superfamily 3.30.310.130 (TATA-Binding Protein; Ubiquitin-related)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 2hkp:A  h s s h h s h h h S s s H S h h - -
PDB 1a9n:B  - - - - - - - - - S - - H S s h s h
PDB 2a3j:A  - - - - - - - - - S - - H S s h s h
```

- **Kpax**:

```
=========================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/* D/$
==== ======= ======= ======= ======= ======= ====  === ===
 1   41.76   22.76   0.3550  0.3294  0.3801  3.15  45  67
 2   39.83   21.77   0.3386  0.3255  0.3626  2.91  42  64
=========================================================
=====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  === ==============
 4   8.9   80   1   +1  2a3j_A[0.0.0.0]
 6  14.3   80   1   +1  1a9n_B[0.0.0.0]
=====================================
```

Both of the above structures don't have RRM fold (topology). Figure A.2 shows the 3D structure of 2IY1_A and the arrangement of secondary structure, visualized using 'PDB Topology Viewer Plugin'. The arrangement secondary

structural elements is very different in this structure compared to the RRM domain. In addition, SCOP and CATH classified these structures into non-RRM families. Finally, the Kpax scores are low, especially the M-Score which is lower than 0.35. Thus, we include this family into a list of 'false RRM families'.

Sugar_tr (PF00083):

4LDS_A:

- **SCOP classification**: No Entry
- **CATH classification**: Superfamily 1.20.1250.20
- **PDBeFold**: (Secondary structure alignment)

```
PDB 4lds:A   h h H h h h h h h h h h h h h h
PDB 1a9n:B   - s H s s h s h - - - - - - - -
PDB 2a3j:A   - s H s s h s h - - - - - - - -
```

- **Kpax**:

```
=============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   32.08   19.84  0.1752  0.2981  0.3387  2.90   40  72
  2   33.21   19.49  0.1814  0.2829  0.3179  3.42   38  80
=============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
 3   7.5   80    1  +1  1a9n_B[0.0.0.0]
 4  10.5   80    1  +1  2a3j_A[0.0.0.0]
====================================
```

For the same reasons as above, this family was included in the list of 'false RRM families'.

RRM_Rrp7 (PF17799):

4M5D_B:

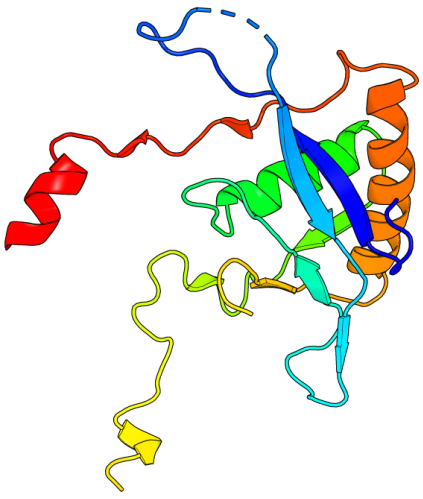
- **SCOP classification**: TP=1, CL=1000003, CF=2001463, SF=3000110, FA=4007617
- **CATH classification**: Superfamily 3.30.70.330 (RRM domain)
- **PDBeFold**: (Secondary structure alignment)

```
PDB 4m5d:B      s s h S h S h S h h
PDB 1a9n:B      - - - S h S - S h s
PDB 2a3j:A      - - - S h S - S h s
```

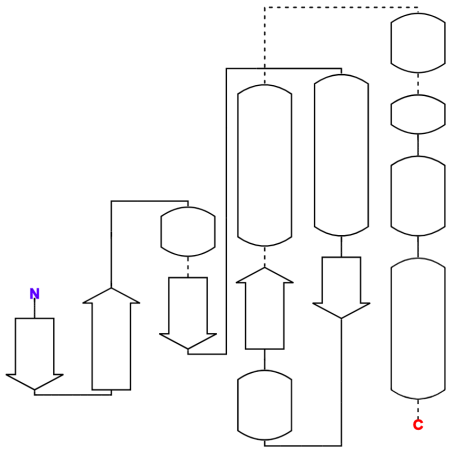
- **Kpax**:

```
============================================================
Rank K-Score G-Score J-Score M-Score T-Score RMSD  N/*  D/$
==== ======= ======= ======= ======= ======= ====  ===  ===
  1   39.66   37.89  0.3645  0.5529  0.5799  2.38   61  79
  2   38.43   36.29  0.3532  0.5353  0.5752  2.58   63  79
============================================================
====================================
I/@  P/!  Len  Seg  TP  Match[Family]
===  ===  ===  ===  ==  =============
 9  14.8   80    1  +1  1a9n_B[0.0.0.0]
 8  12.7   80    1  +1  2a3j_A[0.0.0.0]
====================================
```

This structure (4M5D_B) does not have a continuous fragment in PDB (Figure A.3a). Figure A.3b shows the secondary structure arrangement in the 4M5D_B, which is different compared to the RRM fold. The structural alignment are okay, but these are because of the wrongly aligned secondary structures. The $\beta 2$ sheet from 4M5D_B is aligned onto the $\beta 4$ sheet of the 1A9N_B and 2A3J_A. Lin et al. [2013] identified this N-terminal domain of 4M5D_B as a Deviant RRM Domain.
The available structure (4M5D_B) from this family does not have the RRM fold, so this family was included in the list of 'false RRM families'.



(a) 3D model of 4M5D_B from the PDB (visualised with PyMol)



(b) Secondary structure arrangement in 4M5D_B.

Figure A.3: The structural visualization of 4M5D_B (8, 173).

A.3 List of RRM instances having bound and unbound structures

1. F1LQ48_RRM1
2. F1LQ48_RRM2
3. F1LQ48_RRM3
4. G5ECJ4_RRM1
5. O00425_RRM1
6. O00425_RRM2
7. O43719_RRM2
8. O45189_RRM1
9. O75821_RRM1
10. O95319_RRM3
11. P05455_RRM1
12. P06103_RRM1
13. P07910_RRM1
14. P08199_RRM1
15. P08199_RRM2
16. P08579_RRM1
17. P08621_RRM1
18. P09012_RRM1
19. P09651_RRM1
20. P09651_RRM2
21. P11940_RRM2
22. P19339_RRM1
23. P19339_RRM2
24. P22626_RRM1
25. P23588_RRM1
26. P25299_RRM1
27. P26368_RRM1
28. P26368_RRM2
29. P26599_RRM1
30. P26599_RRM2
31. P26599_RRM3
32. P26599_RRM4
33. P29558_RRM1
34. P31483_RRM2
35. P35637_RRM1
36. P38159_RRM1
37. P38996_RRM1
38. P40565_RRM1
39. P42305_RRM1
40. P43332_RRM1
41. P49960_RRM1
42. P49960_RRM2
43. P49960_RRM3
44. P49960_RRM4
45. P52597_RRM1
46. P52597_RRM2
47. P53617_RRM1
48. P53927_RRM1
49. P62995_RRM1
50. P84103_RRM1
51. Q00916_RRM1
52. Q01130_RRM1
53. Q07955_RRM1
54. Q07955_RRM2
55. Q12046_RRM1
56. Q13148_RRM1
57. Q13148_RRM2
58. Q14103_RRM2
59. Q15717_RRM1
60. Q15717_RRM2
61. Q15717_RRM3
62. Q16630_RRM1
63. Q17RY0_RRM1
64. Q389P7_RRM1
65. Q4G0J3_RRM2
66. Q60900_RRM1
67. Q60900_RRM2
68. Q61474_RRM1
69. Q61474_RRM2
70. Q64368_RRM1
71. Q8I3T5_RRM1
72. Q8IUH3_RRM1
73. Q8IUH3_RRM2
74. Q92879_RRM1
75. Q92879_RRM2
76. Q92879_RRM3
77. Q93062_RRM1
78. Q99181_RRM1
79. Q99181_RRM2
80. Q9BZB8_RRM1
81. Q9H0Z9_RRM1
82. Q9NW64_RRM1
83. Q9NWB1_RRM1
84. Q9UBU9_RRM1
85. Q9UHX1_RRM1
86. Q9UNP9_RRM1
87. Q9Y388_RRM1

Appendix B

CroMaSt Data

B.1 Average structures at family level

All the cross-mapped structural instances (StIs) are first averaged into domain instances (UniProt domain) followed by averaging all the domain instances together, resulting in the average structure at family level (See Methods section). All these average structures at family level are aligned against the core average domain structure using Kpax, and based on the alignment score the families are either added at the beginning of next iteration or discarded. Table B.1 shows the excerpt from the alignment result file.

The three families failing to pass the given threshold (Mscore < 0.6) are Ribosomal_L23 (PF00276), PPV_E2_C (PF00511), and Ribosomal_S24e (PF01282). The average structures for these three families are shown in Figure B.1. The topology (order of secondary structure elements) for all these structures are mentioned below -

1. Core average domain - $\beta1$ $\alpha1$ $\beta2$ $\beta3$ $\alpha2$ $\beta4$
2. Ribosomal_L23 (PF00276) - $\alpha1$ $\beta1$ $\alpha2$ $\beta2$ $\beta3$
3. PPV_E2_C (PF00511) - $\beta1$ $\alpha1$ $\beta2$ $\beta3$ $\alpha2$ $\beta4$
4. Ribosomal_S24e (PF01282) - $\beta1$ $\beta2$ $\alpha1$ $\beta3$ $\beta4$

The topology of 'core average domain' for RRM domain and 'PPV_E2_C' are similar, thus to confirm we visualized the structural alignment of PPV_E2_C with core average domain (Figure B.2). RNP regions are highlighted (RNP1: Green, RNP2: Blue) in the sequence alignment (Figure B.2 C.) showing the difference between these sequences (structures). The sequence from RNP regions of 'PPV_E2_C' (PF00511_avgStruct_core_avgStruct.pdb) does not match with the RNP sequence from core average structure (core_avgStruct_query.pdb), moreover 'PPV_E2_C' structure lacks aromatic residues in this region, that can form stacking interactions with nucleotides.

The structural instances used to compute the averaged structure for PPV_E2_C (PF00511) are listed below. All of these StIs are originally from CATH cross-mapped to PPV_E2_C (PF00511) Pfam family.

Table B.1: Structural alignment results of averaged StIs at family level against core average domain structure

Query In	**Target In**	**Mscore**	**Ncover**	**Naligned**	**RMSD-aligned (in Å)**
core_avgStruct	PF13893_avgStruct	0.9313	77	74	0.96
core_avgStruct	PF11835_avgStruct	0.8604	81	73	1.43
core_avgStruct	PF16367_avgStruct	0.7995	97	72	1.66
core_avgStruct	PF04847_avgStruct	0.7979	84	70	1.52
core_avgStruct	PF05172_avgStruct	0.7875	76	69	1.51
core_avgStruct	PF16842_avgStruct	0.7852	82	72	1.87
core_avgStruct	PF11608_avgStruct	0.7845	78	68	1.72
core_avgStruct	PF03467_avgStruct	0.7758	97	75	1.98
core_avgStruct	PF08952_avgStruct	0.7620	90	69	1.77
core_avgStruct	PF03880_avgStruct	0.7555	83	67	1.62
core_avgStruct	PF08675_avgStruct	0.7262	109	65	1.62
core_avgStruct	PF17774_avgStruct	0.6756	85	66	1.99
core_avgStruct	PF09162_avgStruct	0.6526	82	69	2.31
core_avgStruct	PF08777_avgStruct	0.6186	88	69	2.47
core_avgStruct	PF00276_avgStruct	0.5557	97	51	1.93
core_avgStruct	PF00511_avgStruct	0.5362	88	64	2.70
core_avgStruct	PF01282_avgStruct	0.4440	70	43	1.93

*The complete file for these alignment results can be found in the Results archive.

1. “3MI7,X,0,3.30.70.330,284,365,domain_1,P03120,284,365”,
2. “2Q79,A,0,3.30.70.330,283,362,domain_1,P03120,285,364”,
3. “1BY9,A,0,3.30.70.330,283,362,domain_1,P03120,285,364”,
4. “1R8P,A,0,3.30.70.330,1,81,domain_1,P03120,286,365”,
5. “1R8P,B,0,3.30.70.330,1,81,domain_1,P03120,286,365”,
6. “1ZZF,A,0,3.30.70.330,1,81,domain_1,P03120,286,365”,
7. “1ZZF,B,0,3.30.70.330,1,81,domain_1,P03120,286,365”,
8. “1DBD,A,0,3.30.70.330,1,100,domain_1,P03122,311,410”,
9. “1DBD,B,0,3.30.70.330,1,100,domain_1,P03122,311,410”,
10. “6BUS,3,0,3.30.70.330,316,410,domain_1,P03122,316,410”,
11. “6BUS,1,0,3.30.70.330,319,410,domain_1,P03122,319,410”,
12. “6BUS,2,0,3.30.70.330,319,410,domain_1,P03122,319,410”,
13. “6BUS,4,0,3.30.70.330,319,410,domain_1,P03122,319,410”,
14. “2BOP,A,0,3.30.70.330,326,410,domain_1,P03122,326,410”,
15. “1JJH,B,0,3.30.70.330,325,410,domain_1,P03122,326,410”,

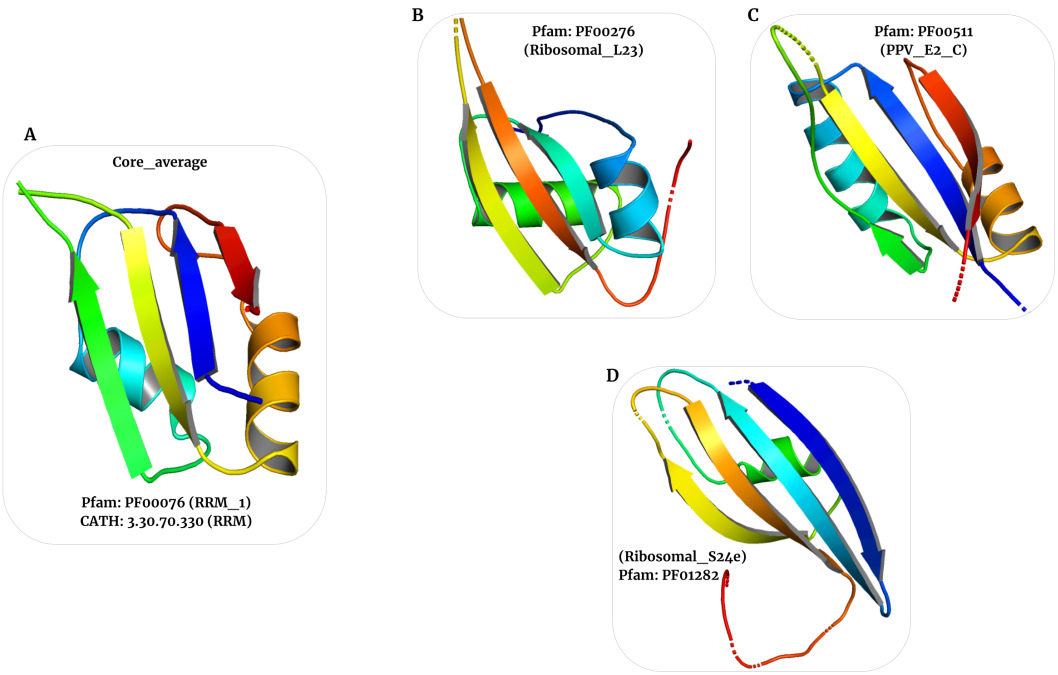


Figure B.1: A) Core average domain structure B) Average structure at family level for Ribosomal_L23 (PF00276) Pfam family C) Average structure at family level for PPV_E2_C (PF00511) Pfam family D) Average structure at family level for Ribosomal_S24e (PF01282) Pfam family.

16. “1JJH,A,0,3.30.70.330,325,410,domain_1,P03122,326,410”,
17. “1JJH,C,0,3.30.70.330,326,410,domain_1,P03122,326,410”,
18. “1JJ4,B,0,3.30.70.330,285,364,domain_1,P06790,286,364”,
19. “1JJ4,A,0,3.30.70.330,286,364,domain_1,P06790,286,364”,
20. “1F9F,D,0,3.30.70.330,284,365,domain_1,P06790,287,365”,
21. “1F9F,B,0,3.30.70.330,284,365,domain_1,P06790,287,365”,
22. “1F9F,A,0,3.30.70.330,285,365,domain_1,P06790,287,365”,
23. “1F9F,C,0,3.30.70.330,284,365,domain_1,P06790,287,365”,
24. “1A7G,E,0,3.30.70.330,291,372,domain_1,P17383,291,372”,
25. “1DHM,A,0,3.30.70.330,1,83,domain_1,P17383,291,372”,
26. “1DHM,B,0,3.30.70.330,1,83,domain_1,P17383,291,372”,
27. “1R8H,A,0,3.30.70.330,281,366,domain_1,Q84294,282,368”,
28. “1R8H,B,0,3.30.70.330,281,366,domain_1,Q84294,282,368”,
29. “1R8H,D,0,3.30.70.330,281,366,domain_1,Q84294,282,368”,
30. “1R8H,E,0,3.30.70.330,281,366,domain_1,Q84294,282,368”,

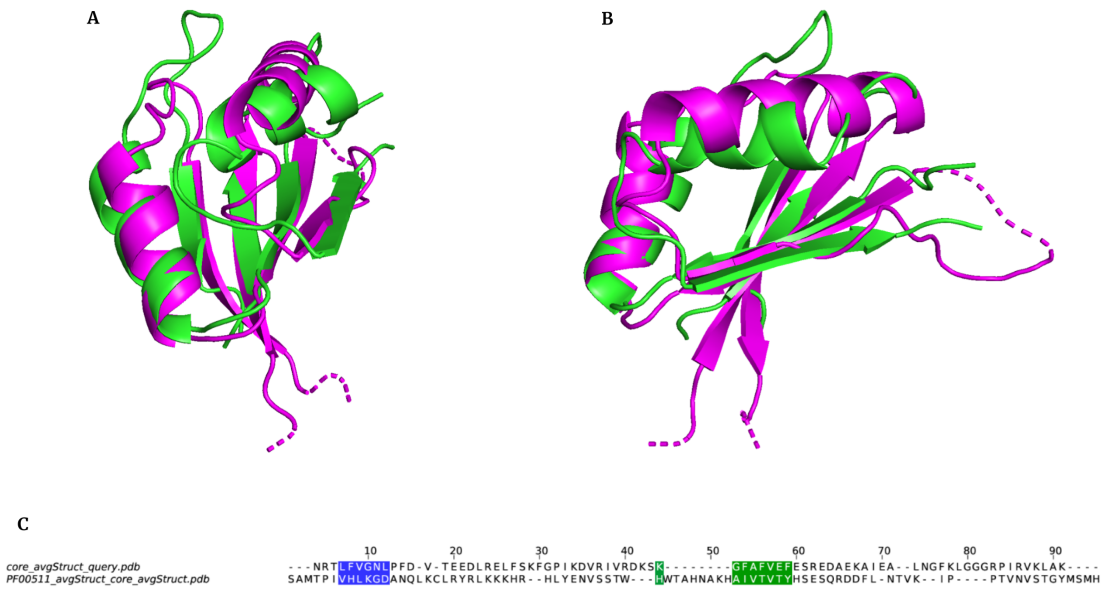


Figure B.2: Alignment of PPV_E2_C (PF00511) against Core average domain A,B) Two different views for the structural alignment C) Fasta alignment resulted from structural alignment; Blue highlighted region represents RNP2 and Green highlighted region represents RNP1.

31. "1R8H,C,0,3.30.70.330,281,366,domain_1,Q84294,282,368",
32. "1R8H,F,0,3.30.70.330,281,366,domain_1,Q84294,282,368",
33. "2AYE,B,0,3.30.70.330,281,366,domain_1,Q84294,282,368",
34. "2AYE,C,0,3.30.70.330,281,366,domain_1,Q84294,282,368",
35. "2AYE,D,0,3.30.70.330,281,366,domain_1,Q84294,282,368",
36. "2AYE,E,0,3.30.70.330,281,366,domain_1,Q84294,282,368",
37. "2AYG,A,0,3.30.70.330,281,366,domain_1,Q84294,282,368",
38. "2AYG,B,0,3.30.70.330,281,366,domain_1,Q84294,282,368",
39. "2AYB,A,0,3.30.70.330,281,366,domain_1,Q84294,282,368",
40. "2AYB,B,0,3.30.70.330,281,366,domain_1,Q84294,282,368",
41. "2AYE,A,0,3.30.70.330,282,366,domain_1,Q84294,283,368",
42. "2AYE,F,0,3.30.70.330,282,366,domain_1,Q84294,283,368"

All the StIs used to compute the average structure at family level can be found in the Results archive.

B.2 Comparison of CroMaSt results with structure-based domain databases

CroMaSt uses the cross-mapping approach for individual StI between Pfam and CATH. Although the results from CroMaSt covers both sequential (from Pfam) and structural (from CATH) features, we wanted to compare the results with other structure-based classifications, i.e., ECOD and SCOP. This comparison can be done at different levels (family and instance-level), but it is an extensive and time-consuming procedure. Thus, we randomly selected a StI from each Pfam family (with at least 1 StI in list of true domain StIs) and cross-mapped these StIs to families in ECOD and SCOP. Table B.2 shows StIs from Pfam and the cross-mapped families in ECOD and SCOP, respectively.

SCOP does not have a family exclusively named as 'RRM' or 'RNA Recognition Motif', but the 'Canonical RBD' family of SCOP can be cross-referenced to the 'PF00076' (RRM_1) of Pfam. The 'Canonical RBD' family classifies under the superfamily 'RNA-binding domain, RBD' in SCOP. All of the SCOP families from Table B.2 are classified under the superfamily 'RNA-binding domain, RBD'.
In ECOD, the family (F) level classification for domains is primarily based on Pfam, domains having significant sequence similarity. Thus, the family naming convention is similar to the Pfam. All the ECOD families from Table B.2 are classified under the same topology (T) - 'RNA-binding domain, RBD'.
In CATH, all these StIs from Table B.2 are from superfamily 3.30.70.330 (RRM domain).

Table B.2 contains only StIs from the list of true domain StIs from CroMaSt. Three of these StIs [2E5I_A (208, 293), 2AD9_A (49, 146), 2FPH_X (82, 166)], do not have any classification in the SCOP database. The classification for other StIs in SCOP and all the StIs in ECOD, is in good agreement with CroMast results (Pfam and CATH).

B.3 RRM clan in Pfam

The RRM clan from Pfam contains families that are related to the RNA recognition motif domains and are thought to be evolutionarily related. This clan contains 33 families and the total number of domains in the clan is 433471. Table B.3 lists all the families from RRM clan.

B.4 Starting CroMaSt with different Pfam family

We also started CroMaSt with different family to check the effect of starting families on the results. Table B.4 shows the summarized results when the CroMaSt was started with RRM_5 (PF13893) Pfam family and 3.30.70.330 (RRM (RNA Recognition Motif) domain) CATH superfamily.

Table B.2: Cross-mapping of representative structures from Pfam families to ECOD and SCOP databases

Pfam Family Name (Family Ids)	**Representative StI**	**Family in ECOD**	**Family in SCOP**
RRM_1 (PF00076)	1B7F_A (127, 197)	RRM_1_2 (EF21352)	Canonical RBD
Smg4_UPF3 (PF03467)	1UW4_A (50, 140)	Smg4_UPF3 (EF12679)	Smg-4/UPF3-like
DbpA (PF03880)	2G0C_A (405, 476)	DbpA (EF02236)	DbpA RNA-binding domain-like
Calcipressin (PF04847)	1WEY_A (15, 98)	Calcipressin (EF01259)	Canonical RBD
Nup35_RRM (PF05172)	3P3D_A (266, 363)	Nup35_RRM (EF10021)	Canonical RBD
RNA_bind (PF08675)	3CTR_A (445, 514)	RNA_bind (EF12083)	Canonical RBD
RRM_3 (PF08777)	1OWX_A (231, 334)	RRM_3 (EF12206)	Canonical RBD
DUF1866 (PF08952)	2DNR_A (893, 970)	DUF1866 (EF03296)	Canonical RBD
Tap-RNA_bind (PF09162)	1FO1_A (123, 191)	Tap-RNA_bind (EF13157)	Non-canonical RBD domain*
MARF1_RRM1 (PF11608)	2DIU_A (8, 90)	Limkain-b1 (EF08767)	Limkain b1 domain-like
RRM_8 (PF11835)	2E5I_A (208, 293)	RRM_1_6 (EF24069)	No results
RRM_5 (PF13893)	2AD9_A (49, 146)	RRM_1_6 (EF24069)	No results
RRM_7 (PF16367)	2DNL_A (426, 515)	RRM_1_3 (EF21353)	Canonical RBD
RRM_occluded (PF16842)	2L9W_A (311, 393)	RRM_occluded (EF20754)	Occluded RRM-like*
YlmH_RBD (PF17774)	2FPH_X (82, 166)	YlmH_2nd (EF14865)	No results

*No SCOP2 classification is available for given PDB ID explicitly.
Note: Representative StIs are randomly chosen from the list of true domains for each of the listed Pfam family.

Table B.3: All the Pfam families from RRM clan

Sr. No.	Family Ids	Family Name	Explored by CroMaSt?
1	PF00076	RRM_1	NA*
2	PF02994	Transposase_22	No
3	PF03467	Smg4_UPF3	No
4	PF03468	XS	No
5	PF03880	DbpA	Yes
6	PF04059	RRM_2	No
7	PF04847	Calcipressin	Yes
8	PF05172	Nup35_RRM	No
9	PF07576	BRAP2	No
10	PF08152	GUCT	No
11	PF08489	DUF1743	No
12	PF08675	RNA_bind	Yes
13	PF08777	RRM_3	Yes
14	PF08952	DUF1866	Yes
15	PF09162	Tap-RNA_bind	Yes
16	PF11608	MARF1_RRM1	Yes
17	PF11767	SET_assoc	No
18	PF11835	RRM_8	Yes
29	PF12220	U1snRNP70_N	No
20	PF13893	RRM_5	Yes
21	PF14605	Nup35_RRM_2	No
22	PF14703	PHM7_cyt	No
23	PF15023	DUF4523	No
24	PF15407	Spo7_2_N	No
25	PF16367	RRM_7	Yes
26	PF16842	RRM_occluded	Yes
27	PF17774	YlmH_RBD	Yes
28	PF17799	RRM_Rrp7	No
29	PF17797	RL	No
30	PF18440	GlcNAc-1_reg	No
31	PF18444	RRM_9	No
32	PF18528	Ret2_MD	No
33	PF19977	xRRM	No

*Starting Pfam family for CroMaSt workflow

Table B.4: Results from each step of CroMaSt, starting with Pfam family - RRM_5 (PF13893) and CATH superfamily - 3.30.70.330 (RRM (RNA Recognition Motif) domain)

Steps	Iteration 1		Iteration 2	
	Pfam	CATH	Pfam	CATH
Starting Families	1	1	11	0
StI filtered on domain length	39	1527	1194	926*
Obsolete and inconsistent entries	0	323	3	0
Residue-mapped StIs	39	1204	1191	-
Common StIs (**Core** & **True**)	36	36	926	926
Remaining StIs (not common)	3	1168	265	0
Cross-mapped StIs	0	1094	0	0
Properly aligned at family level	-	926	-	-
Not properly aligned at family level	-	168	-	-
Not cross-mapped StIs (unmapped)	3	74	265	0
Properly aligned at instance level (**Domain-like**)	3	67	257	-
Not properly aligned at instance level	0	7	8	-
Failed structures	0	175	8	0
New families found	11	0	0	0

*These StI entries are cross-mapped and properly aligned at the family level from the previous iteration.

B.5 List of obsolete and inconsistent structural instances

There are a total of 326 StIs from Pfam and CATH considered as obsolete and inconsistent entries by CroMaSt. Below is the complete list of all these 326 StIs: First 10 are inconsistent entries and rest are obsolete entries (in Protein Data Bank). First 3 are from Pfam and rest are from CATH.

Format of these structural instance entries differ slightly for Pfam and CATH.

- Pfam inconsistent and obsolete entries:
  "PDB_id,Chain,Fam_name,Fam_id,UNP_id,UNP_start,UNP_end"
- CATH inconsistent entries:
  " PDB_id,Chain,Domain_position,Fam_id,PDB_start,PDB_end"
- CATH obsolete entries:
  " PDB_idChainDomain_position,Fam_id,PDB_start,PDB_end"

B.5.1 Inconsistent structural instances

1. "6DG0,B,RRM_1,PF00076,Q22039,230,294",
2. "6DG0,A,RRM_1,PF00076,Q22039,230,294",
3. "6PAI,D,RRM_1,PF00076,Q7Z3L0,95,165",
4. "2KU7,A,00,3.30.70.330,1,140",
5. "3DXB,F,02,3.30.70.330,452,556",
6. "3DXB,G,02,3.30.70.330,452,556",
7. "3DXB,D,02,3.30.70.330,452,556",
8. "3DXB,B,02,3.30.70.330,452,556",
9. "3DXB,A,02,3.30.70.330,452,556",
10. "4V19,X,00,3.30.70.330,2,150"

The list of 316 **Obsolete structural instances** can be found here in a file at https://github.com/HrishiDhondge/Data_files.

B.5.2 Inconsistencies in Pfam and CATH

Inconsistencies in and between domain databases regarding RRM domain families appear with very simple searches. Inconsistencies are at the level of the databases, inside a database (ex : RRM families outside the RRM clan) and between databases (16 Pfam families versus 2 CATH superfamilies). Firstly, there are 16 Pfam families with RRM in their name: MARF1_RRM1 (PF11608), Nup35_RRM (PF05172), Nup35_RRM_2 (PF14605), RRM (PF10378), RRM_1 (PF00076), RRM_2 (PF04059), RRM_3 (PF08777), RRM_4 (PF10598), RRM_5 (PF13893), RRM_7 (PF16367), RRM_8 (PF11835), RRM_9 (PF18444), RRM_DME (PF15628), RRM_occluded (PF16842), RRM_Rrp7 (PF17799), xRRM (PF19977). In addition, Pfam also has an RRM clan (CL0221) with a total of 33 Pfam families including only 13 of the above-mentioned families with RRM in their name (See suppl Table S3). In particular, the RRM clan does not include RRM (PF10378), RRM_4 (PF10598) and RRM_DME (PF15628). These 3 Pfam families do not belong to any clan in Pfam.

In CATH, 2 superfamilies have RRM keyword in their name: 3.30.70.330 RRM (RNA recognition motif) domain, 3.30.70.1820 L1 transposable element, RRM domain. When the CATH database was searched using keyword 'RRM', several matching CATH domains were retrieved apart from these two CATH superfamilies. Some of those hits are as follows: 3m4xA01 from Superfamily 3.30.70.3130[1], 3m4xA02 from Superfamily 3.40.50.150, 3m4xA03 from Superfamily 2.30.130.60, 3dxbA01 from Superfamily 3.40.30.10.

B.5.3 Demonstration of Residue-mapping

Let's try to understand residue-mapping for one StI and why it takes more time: We have a CATH StI retrieved directly from CATH release files: **'1WF2,A,0,3.30.70.330,1,98'** in the format of *'PDB_id, Chain_id, Domain_order_number, Family_id, PDB_start, PDB_end.*

Before trying to find this StI in Pfam, we need the corresponding residue positions from UniProt. Thus, we will use SIFTS resource to find the start and end residue positions of this StI in UniProt.
At first, we start by trying to find the 1st and 98th residues from 'A' chain of '1WF2' structure in '1wf2.xml' file downloaded from SIFTS.

```
−<residue dbSource="PDBe" dbCoordSys="PDBe" dbResNum="1" dbResName="GLY">
    <crossRefDb dbSource="PDB" dbCoordSys="PDBresnum" dbAccessionId="1wf2" dbResNum="1" dbResName="GLY" dbChainId="A"/>
    <crossRefDb dbSource="CATH" dbCoordSys="PDBresnum" dbAccessionId="3.30.70.330" dbResNum="1" dbResName="GLY" dbChainId="A"/>
    <crossRefDb dbSource="SCOP" dbCoordSys="PDBresnum" dbAccessionId="114573" dbResNum="1" dbResName="GLY" dbChainId="A"/>
    <residueDetail dbSource="PDBe" property="codeSecondaryStructure">T</residueDetail>
    <residueDetail dbSource="PDBe" property="nameSecondaryStructure">loop</residueDetail>
    <residueDetail dbSource="PDBe" property="Annotation">Cloning artifact</residueDetail>
  </residue>
```

Figure B.3: Snapshot of SIFTS file for 1st residue from '1WF2' PDB structure with 'A' chain.

[1]merged with 3.30.70.1170: Sun protein; domain 3

The SIFTS file ('1wf2.xml') don't have UniProt entry information for this residue and PDBe annotation states that this residue is part of 'cloning artifact' (Figure B.3). This suggests this residue is not the part of actual protein and rather an artifact in this experiment. We encounter the same issue with the end residue (98) of this StI. Thus, we increment the starting residue position by 1 and decrement the end residue position by 1. We repeat this until we find the corresponding residue positions from UniProt.
At the end, we find the corresponding positions from UniProt resulting in residue-mapped StI ready for cross-mapping with Pfam StIs. Below is the residue mapped StI: **'1WF2,A,0,3.30.70.330,1,98,P07910,8,92'** in the format of *'PDB_id, Chain_id, Domain_order_number, Family_id, PDB_start, PDB_end, UniProt_id, UniProt_start, UniProt_end'*.

Appendix C

Binding Free Energy Computations

C.1 D42A point mutation

The Asp42 resides on $\beta 2$ of RRM domain (see Figure 5.15). Figure C.1 shows the overall free energy profile for D42A mutation in unbound and bound form of SRSF2 RRM domain.

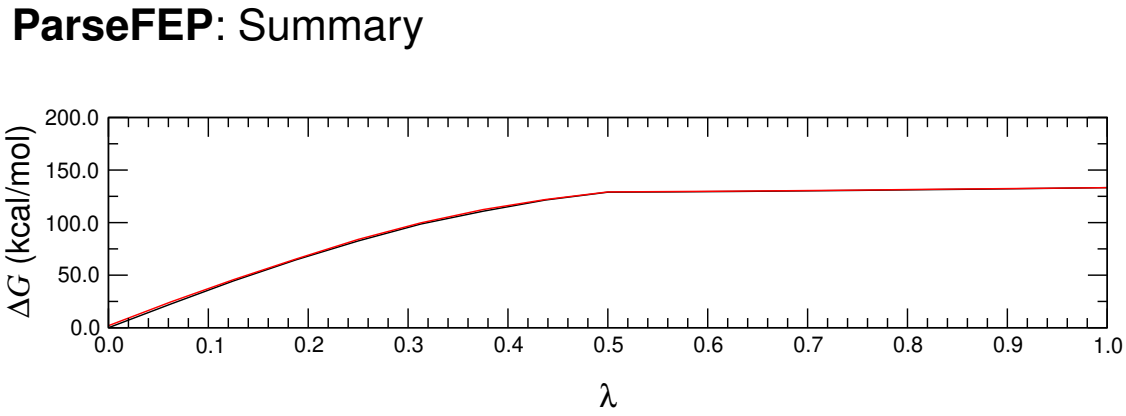
The total free energy change (ΔG_3) of D42A mutation in unbound form of SRSF2 RRM is $132.47\,\mathrm{kcal\,mol^{-1}}$ with a total error of $0.194\,\mathrm{kcal\,mol^{-1}}$. Whereas, in the bound form the free energy change (ΔG_4) for D42A point mutation is $133.95\,\mathrm{kcal\,mol^{-1}}$ with a total error of $0.196\,\mathrm{kcal\,mol^{-1}}$.

Figure C.2 and Figure C.3 show the time evolution of free energy differences for each of the λ window and the probability distribution plots (for backward and forward transformations) for unbound and bound form of SRSF2 RRM domain, respectively.

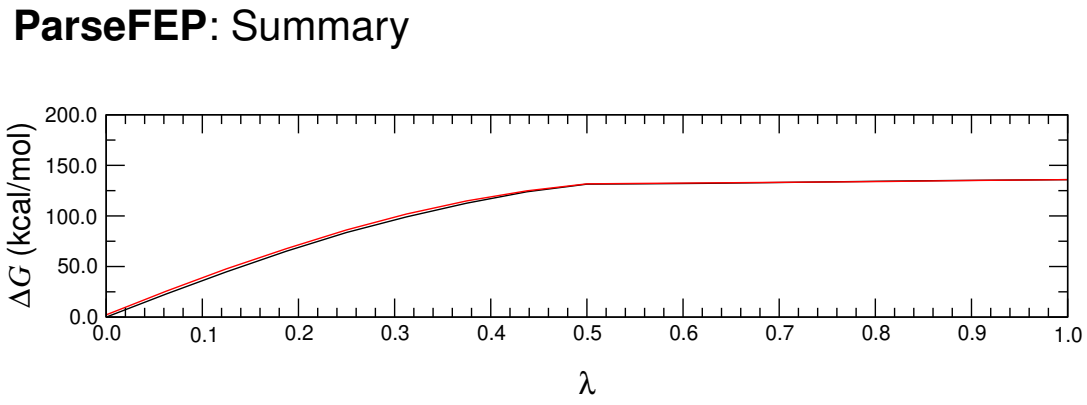
To compute the relative binding free energy for D42A point mutation we will use the equation $\Delta\Delta G = \Delta G_4 - \Delta G_3$ (see Figure 5.16).

$$\begin{aligned}\Delta\Delta G &= (133.95 \pm 0.196\,\mathrm{kcal\,mol^{-1}}) - (132.47 \pm 0.194\,\mathrm{kcal\,mol^{-1}}) \\ &= (133.95 - 132.47) \pm (\sqrt{0.196^2 + 0.194^2})\mathrm{kcal\,mol^{-1}} \\ &= 1.48 \pm 0.276\,\mathrm{kcal\,mol^{-1}}\end{aligned}$$

The relative binding free energy for D42A point mutation in SRSF2 RRM-RNA complex is $1.48 \pm 0.276\,\mathrm{kcal\,mol^{-1}}$.

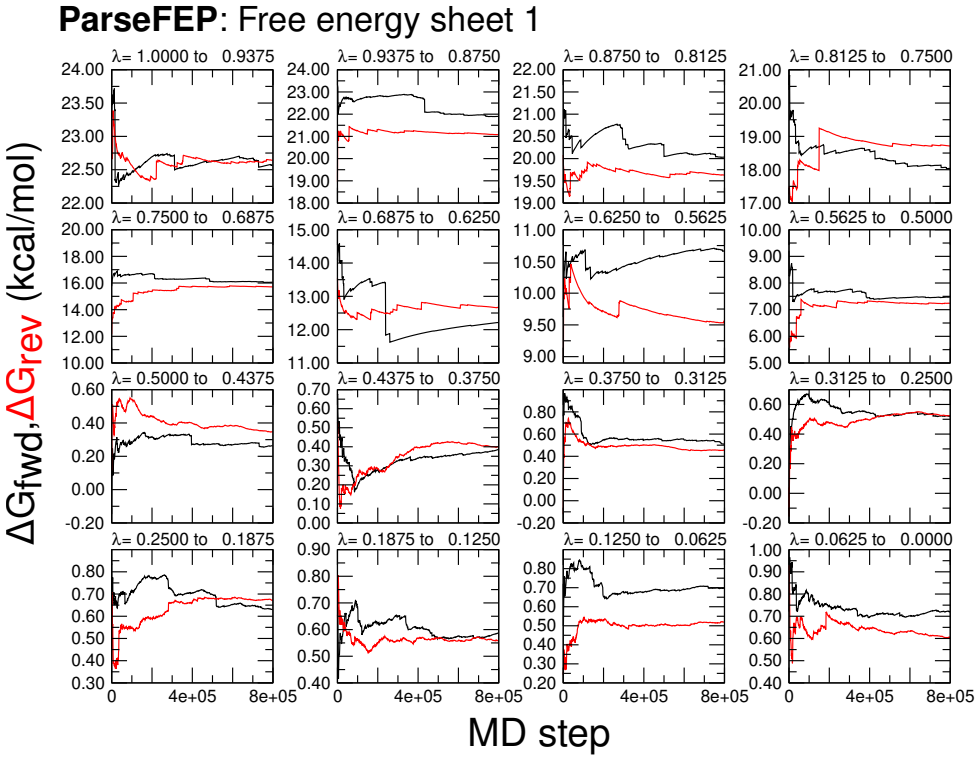


(a) Free energy change for Asp42 to Ala mutation in SRSF2 RRM domain (unbound form)

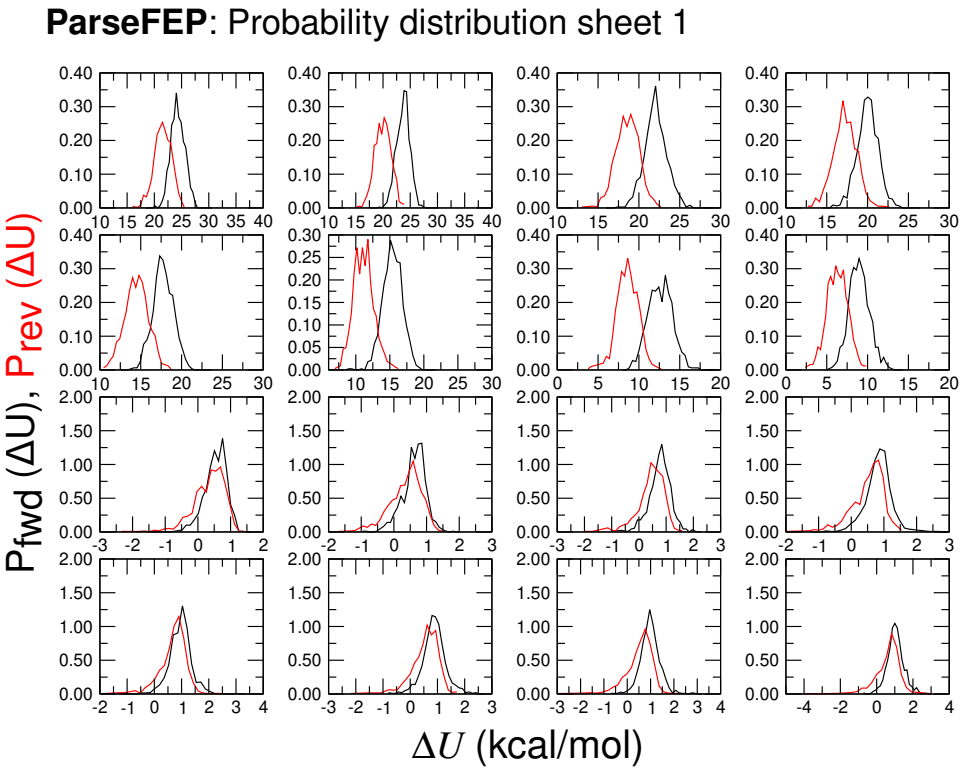


(b) Free energy change for Asp42 to Ala mutation in SRSF2 RRM domain (bound form)

Figure C.1: Free energy change for Aspartate (position 42) to Alanine mutation in SRSF2 RRM domain. Black line indicates the forward transformation, i.e., from Asp to Ala and red line indicates the backward transformation from Ala to Asp.

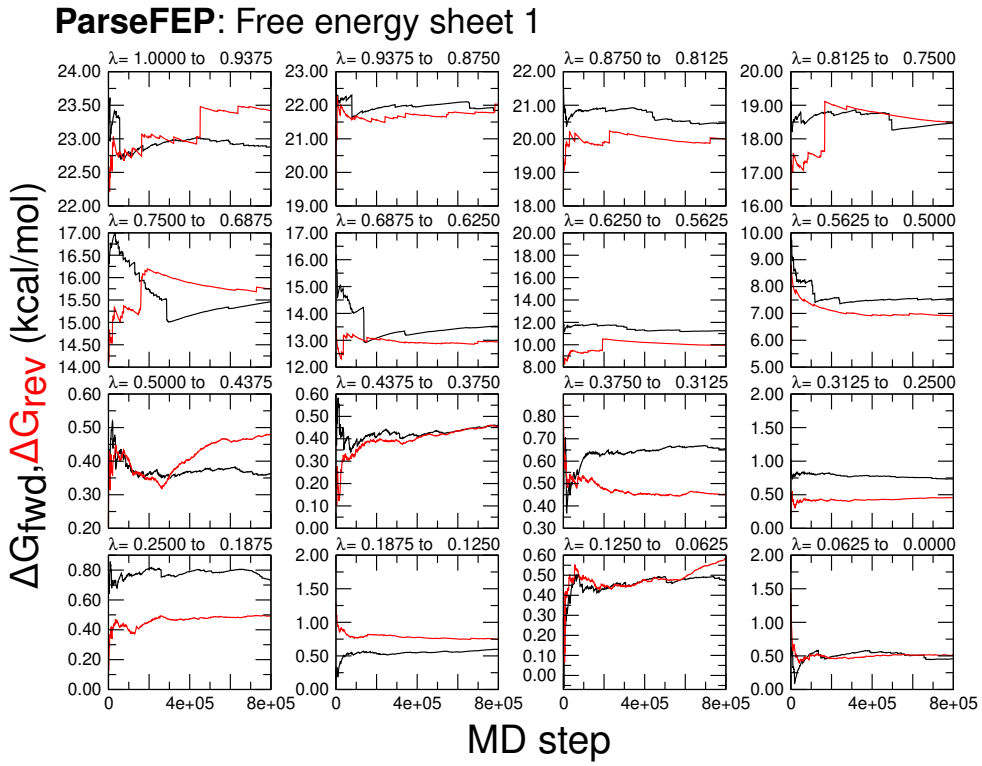


(a) Time evolution of free energy differences for each λ window

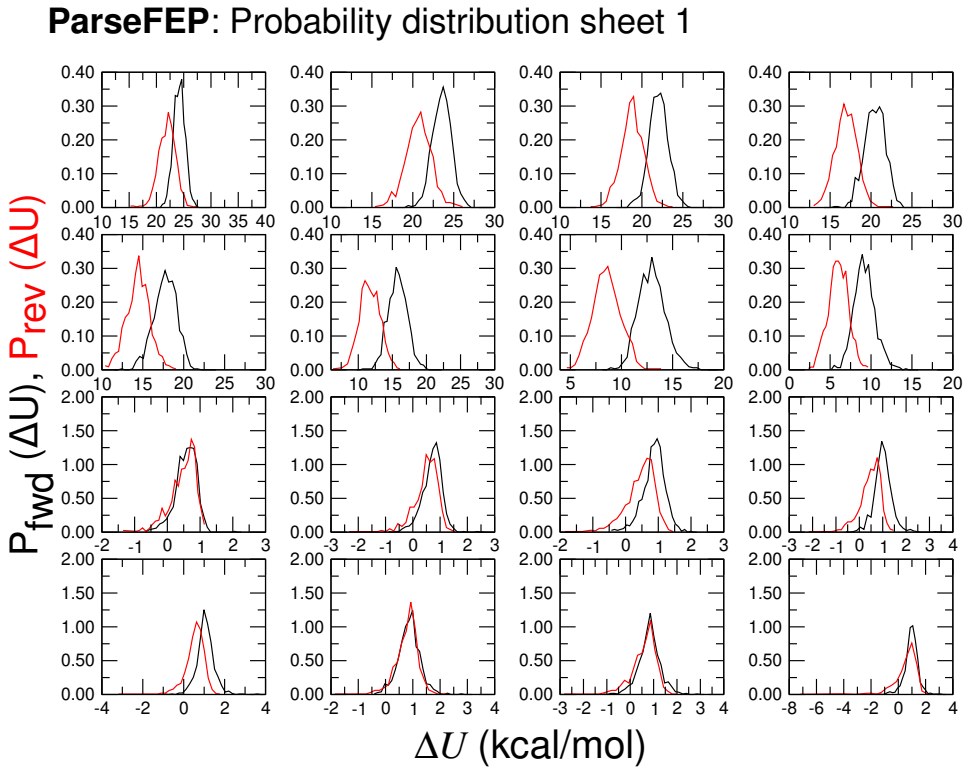


(b) Probability distribution plots for backward and forward transformations

Figure C.2: Output plots from the soft-core potenial calculation, of ΔG, and $P_0[\Delta U]$ and $P_1[\Delta U]$ generated by ParseFEP for alchemical transformation of D42A in SRSF2 RRM (unbound). Each subplot corresponds to a λ sampling window.



(a) Time evolution of free energy differences for each λ window



(b) Probability distribution plots for backward and forward transformations

Figure C.3: Output plots from the soft-core potenial calculation, of ΔG, and $P_0[\Delta U]$ and $P_1[\Delta U]$ generated by ParseFEP for alchemical transformation of D42A in SRSF2 RRM (bound form). Each subplot corresponds to a λ sampling window.

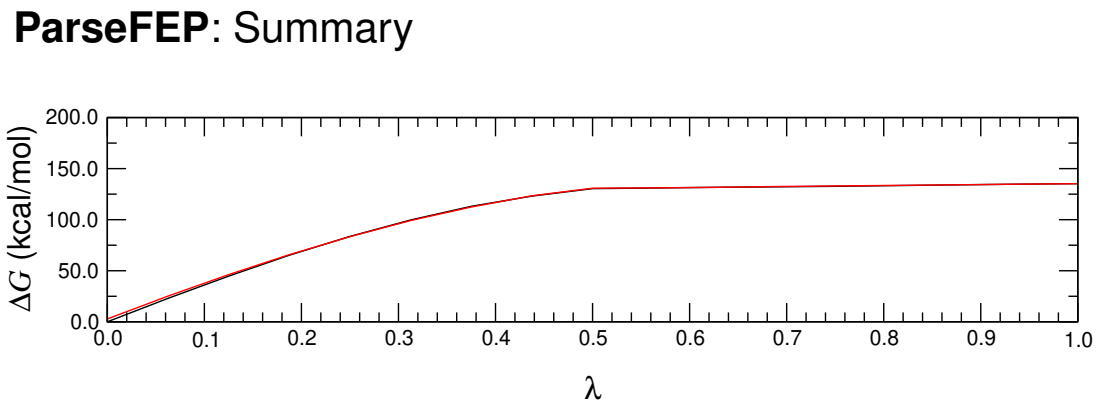
C.2 D48A point mutation

The Asp48 resides in loop3 of RRM domain (see Figure 5.15). Figure C.4 shows the overall free energy profile for D48A mutation in unbound and bound form of SRSF2 RRM domain.
The total free energy change (ΔG_3) of D48A mutation in unbound form of SRSF2 RRM is $133.36\,\mathrm{kcal\,mol^{-1}}$ with a total error of $0.342\,\mathrm{kcal\,mol^{-1}}$. Whereas, in the bound form the free energy change (ΔG_4) for D42A point mutation is $131.94\,\mathrm{kcal\,mol^{-1}}$ with a total error of $0.480\,\mathrm{kcal\,mol^{-1}}$.
Figure C.5 and Figure C.6 show the time evolution of free energy differences for each of the λ window and the probability distribution plots (for backward and forward transformations) for unbound and bound form of SRSF2 RRM domain, respectively.

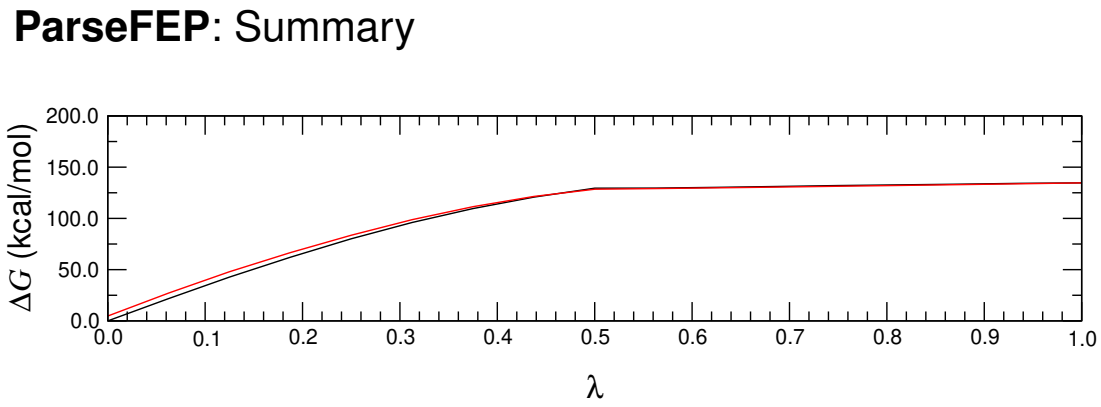
To compute the relative binding free energy for D42A point mutation we will use the equation $\Delta\Delta G = \Delta G_4 - \Delta G_3$ (see Figure 5.16).

$$\begin{aligned}\Delta\Delta G &= (131.94 \pm 0.480\,\mathrm{kcal\,mol^{-1}}) - (133.36 \pm 0.342\,\mathrm{kcal\,mol^{-1}}) \\ &= (131.94 - 133.36) \pm (\sqrt{0.480^2 + 0.342^2})\mathrm{kcal\,mol^{-1}} \\ &= -1.42 \pm 0.589\,\mathrm{kcal\,mol^{-1}}\end{aligned}$$

The relative binding free energy for D42A point mutation in SRSF2 RRM-RNA complex is $-1.42 \pm 0.589\,\mathrm{kcal\,mol^{-1}}$.

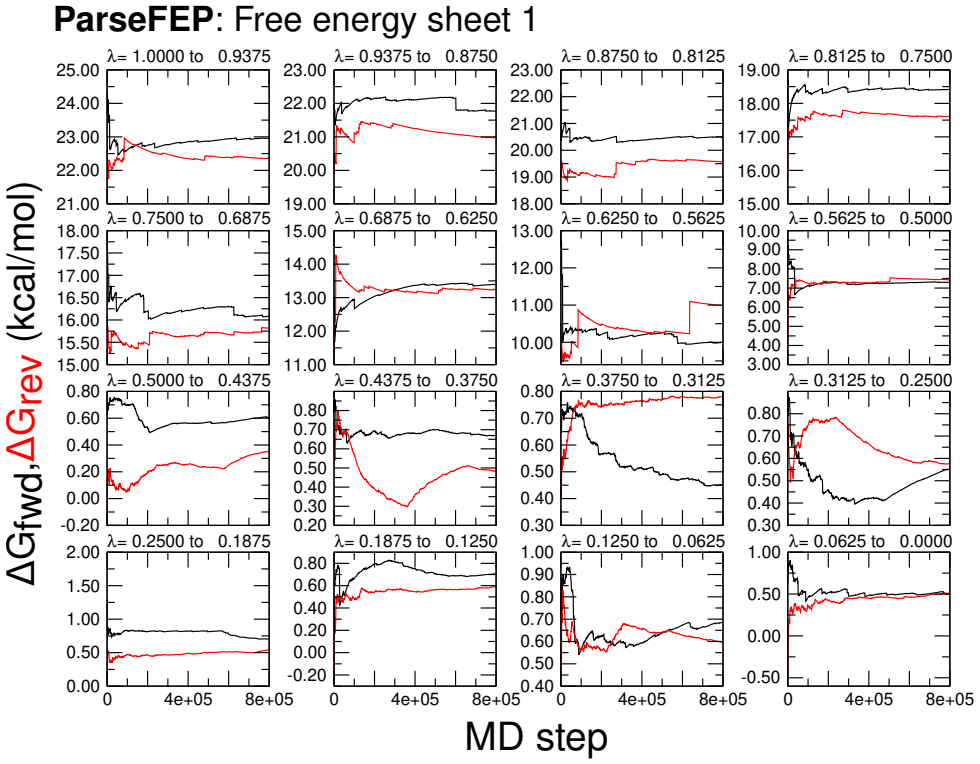


(a) Free energy change for Asp48 to Ala mutation in SRSF2 RRM domain (unbound form)

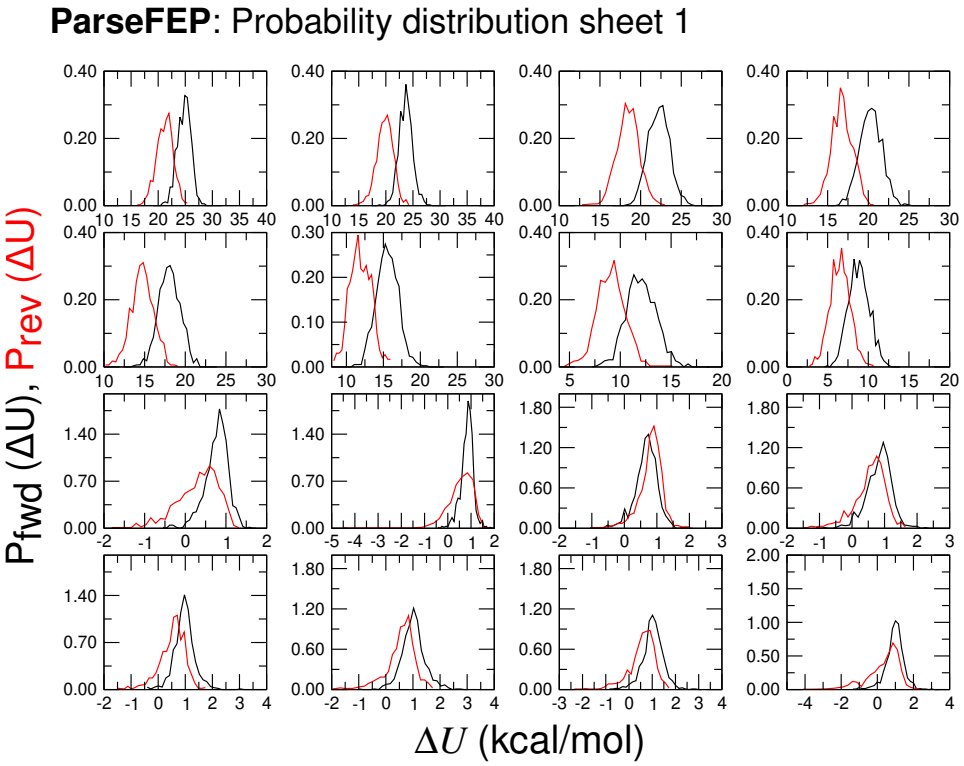


(b) Free energy change for Asp48 to Ala mutation in SRSF2 RRM domain (bound form)

Figure C.4: Free energy change for Aspartate (position 48) to Alanine mutation in SRSF2 RRM domain. Black line indicates the forward transformation, i.e., from Asp to Ala and red line indicates the backward transformation from Ala to Asp.

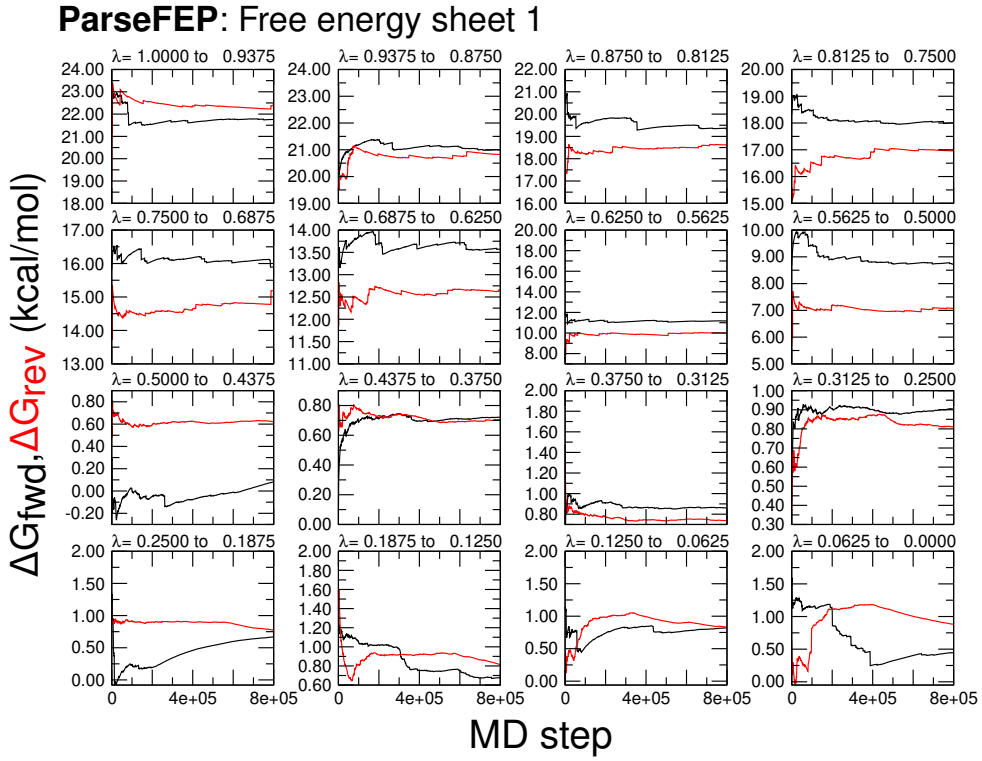


(a) Time evolution of free energy differences for each λ window

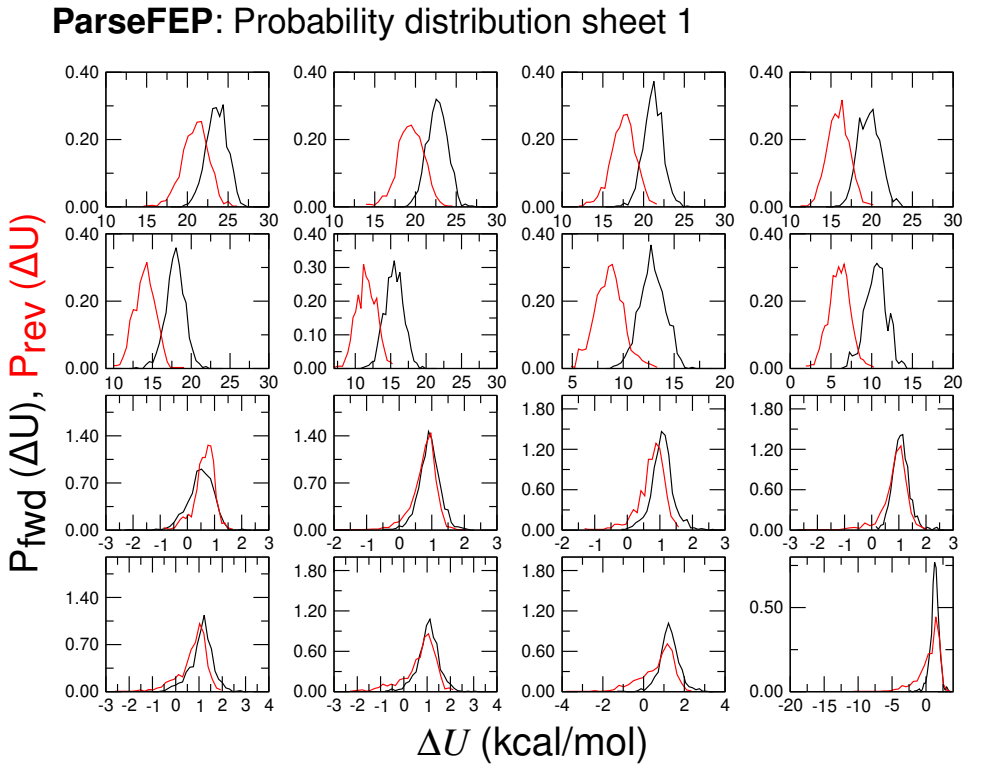


(b) Probability distribution plots for backward and forward transformations

Figure C.5: Output plots from the soft-core potenial calculation, of ΔG, and $P_0[\Delta U]$ and $P_1[\Delta U]$ generated by ParseFEP for alchemical transformation of D48A in SRSF2 RRM (unbound form). Each subplot corresponds to a λ sampling window.



(a) Time evolution of free energy differences for each λ window



(b) Probability distribution plots for backward and forward transformations

Figure C.6: Output plots from the soft-core potenial calculation, of ΔG, and $P_0[\Delta U]$ and $P_1[\Delta U]$ generated by ParseFEP for alchemical transformation of D48A in SRSF2 RRM (bound form). Each subplot corresponds to a λ sampling window.

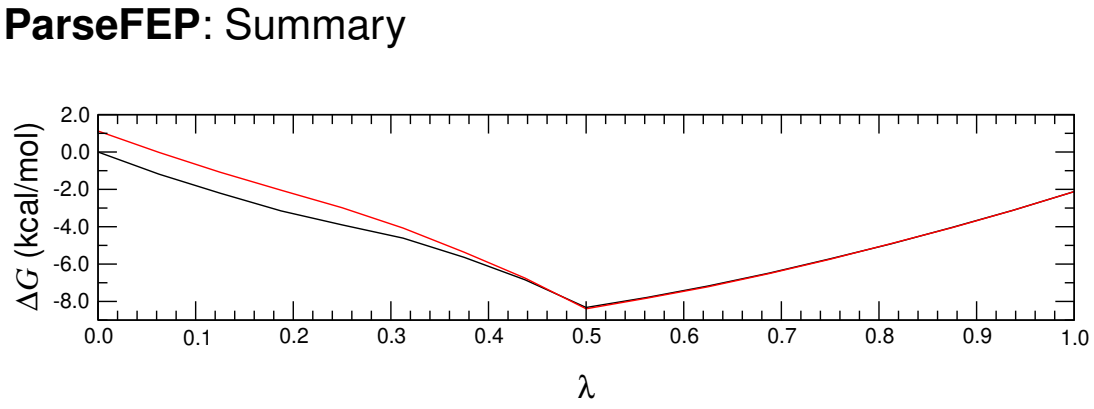
C.3 S54A point mutation

The Ser54 resides in loop3 of RRM domain (see Figure 5.15). Figure C.7 shows the overall free energy profile for S54A mutation in unbound and bound form of SRSF2 RRM domain.
The total free energy change (ΔG_3) of S54A mutation in unbound form of SRSF2 RRM is $-2.457\,\mathrm{kcal\,mol^{-1}}$ with a total error of $0.064\,\mathrm{kcal\,mol^{-1}}$. Whereas, in the bound form the free energy change (ΔG_4) for S54A point mutation is $-1.135\,\mathrm{kcal\,mol^{-1}}$ with a total error of $0.075\,\mathrm{kcal\,mol^{-1}}$.
Figure C.8 and Figure C.9 show the time evolution of free energy differences for each of the λ window and the probability distribution plots (for backward and forward transformations) for unbound and bound form of SRSF2 RRM domain, respectively.

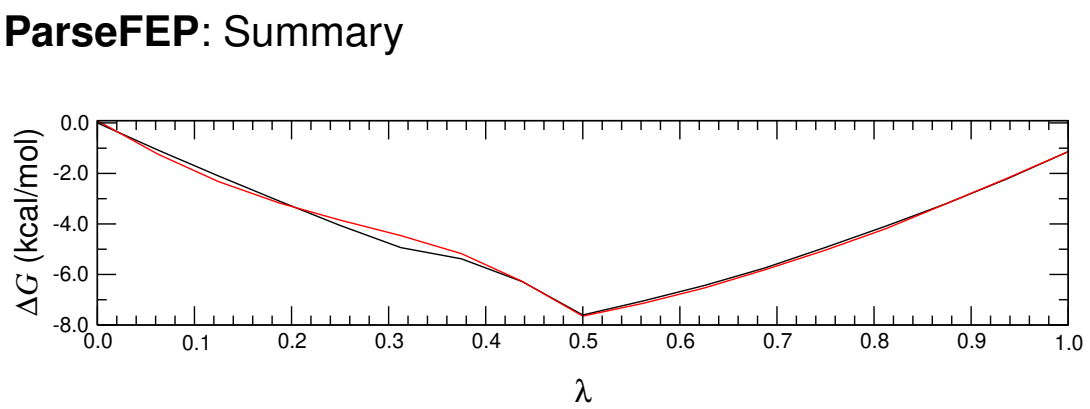
To compute the relative binding free energy for S54A point mutation we will use the equation $\Delta\Delta G = \Delta G_4 - \Delta G_3$ (see Figure 5.16).

$$\begin{aligned}\Delta\Delta G &= (-1.135 \pm 0.075\,\mathrm{kcal\,mol^{-1}}) - (-2.457 \pm 0.064\,\mathrm{kcal\,mol^{-1}}) \\ &= (-1.135 - (-2.457)) \pm (\sqrt{0.075^2 + 0.064^2})\mathrm{kcal\,mol^{-1}} \\ &= 1.322 \pm 0.099\,\mathrm{kcal\,mol^{-1}}\end{aligned}$$

The relative binding free energy for S54A point mutation in SRSF2 RRM-RNA complex is $1.322 \pm 0.099\,\mathrm{kcal\,mol^{-1}}$.

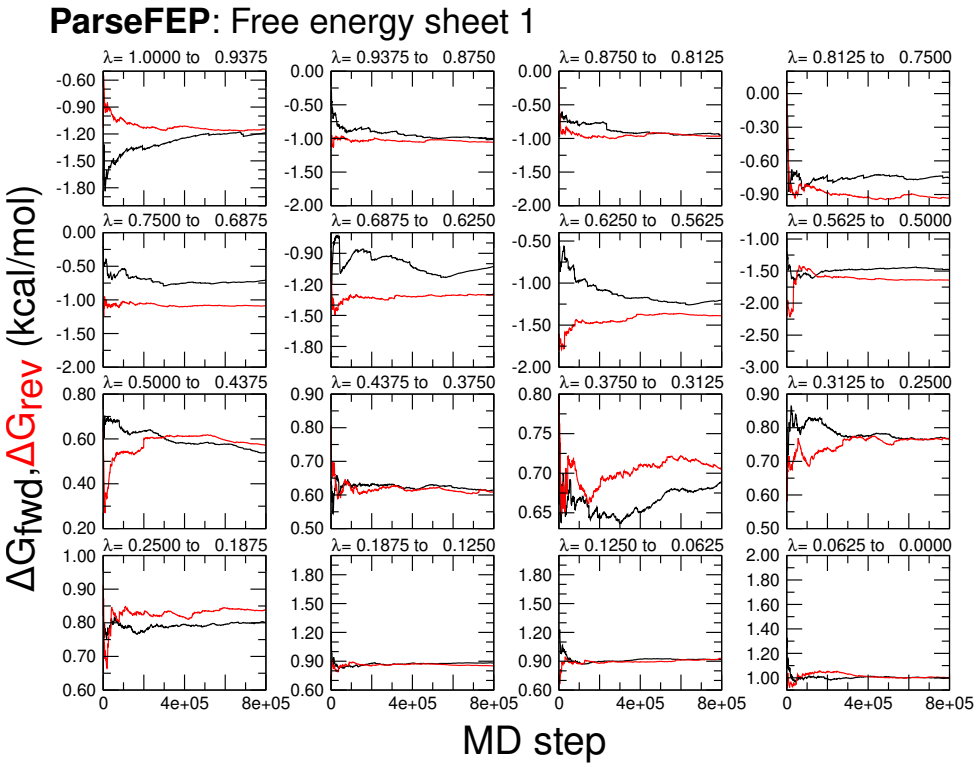


(a) Free energy change for Ser54 to Ala mutation in SRSF2 RRM domain (unbound form)

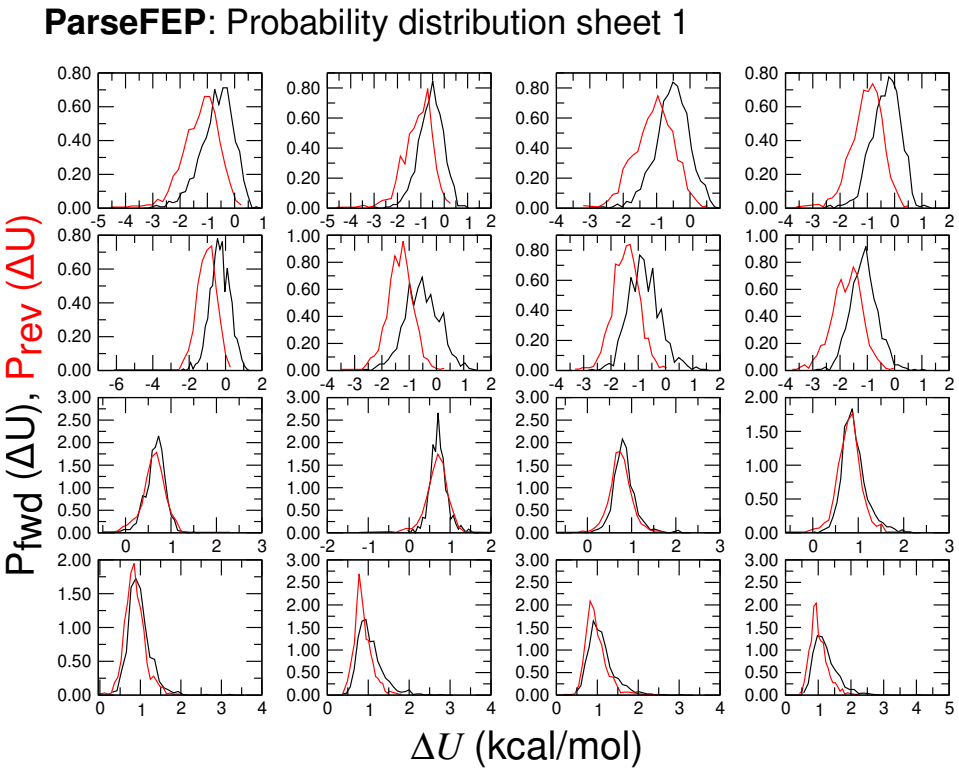


(b) Free energy change for Ser54 to Ala mutation in SRSF2 RRM domain (bound form)

Figure C.7: Free energy change for Serine (position 54) to Alanine mutation in SRSF2 RRM domain. Black line indicates the forward transformation, i.e., from Ser to Ala and red line indicates the backward transformation from Ala to Ser.

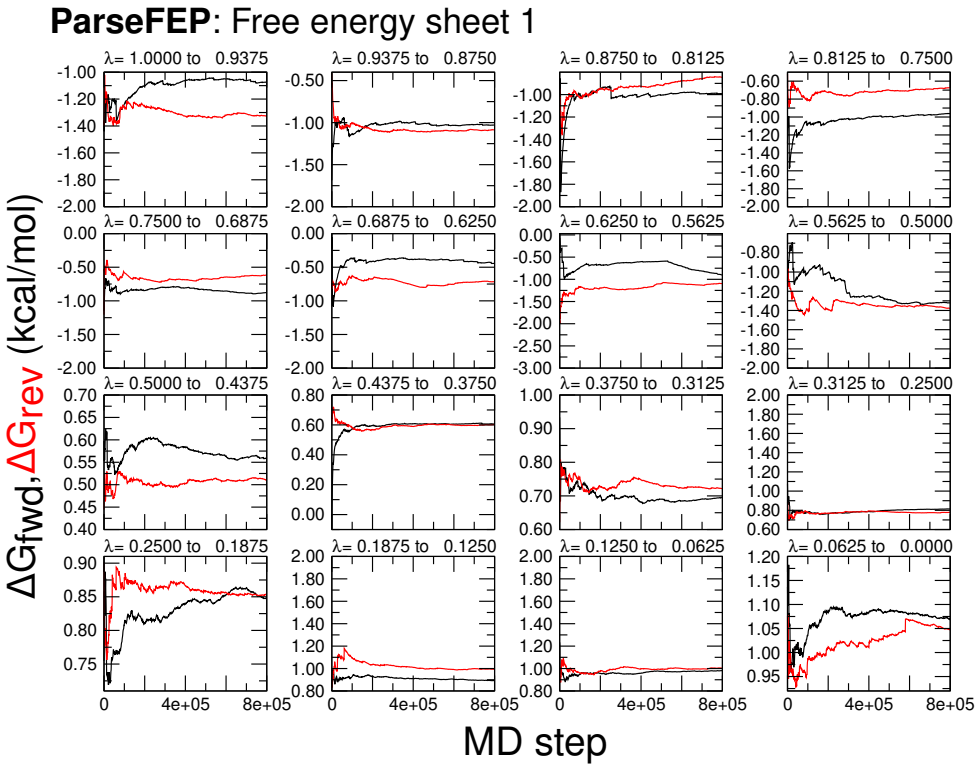


(a) Time evolution of free energy differences for each λ window

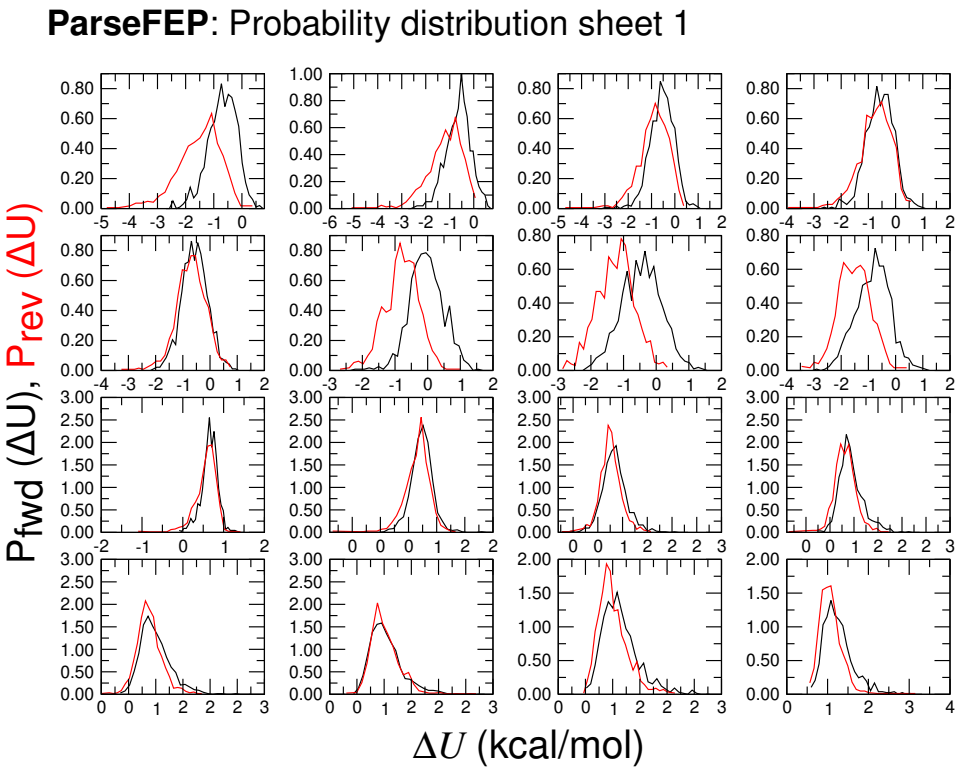


(b) Probability distribution plots for backward and forward transformations

Figure C.8: Output plots from the soft-core potenial calculation, of ΔG, and $P_0[\Delta U]$ and $P_1[\Delta U]$ generated by ParseFEP for alchemical transformation of S54A in SRSF2 RRM (unbound form). Each subplot corresponds to a λ sampling window.



(a) Time evolution of free energy differences for each λ window



(b) Probability distribution plots for backward and forward transformations

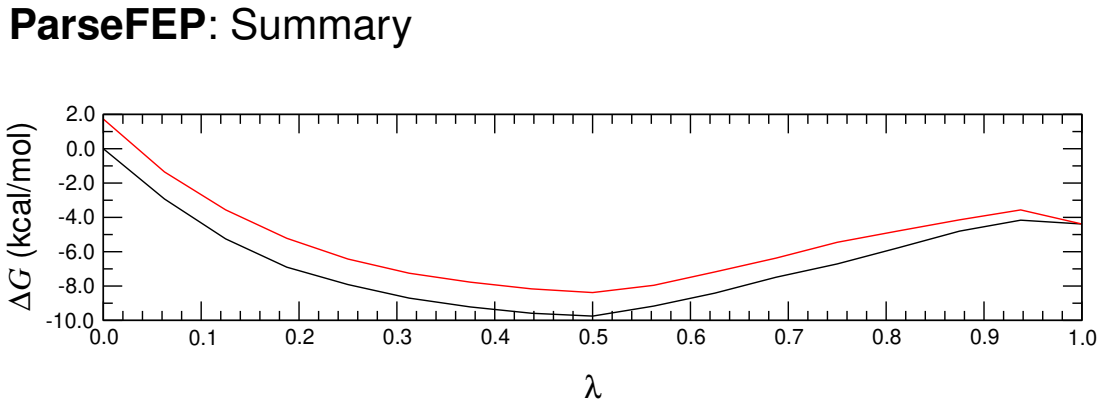
Figure C.9: Output plots from the soft-core potenial calculation, of ΔG, and $P_0[\Delta U]$ and $P_1[\Delta U]$ generated by ParseFEP for alchemical transformation of S54A in SRSF2 RRM (bound form). Each subplot corresponds to a λ sampling window.

C.4 F59A point mutation

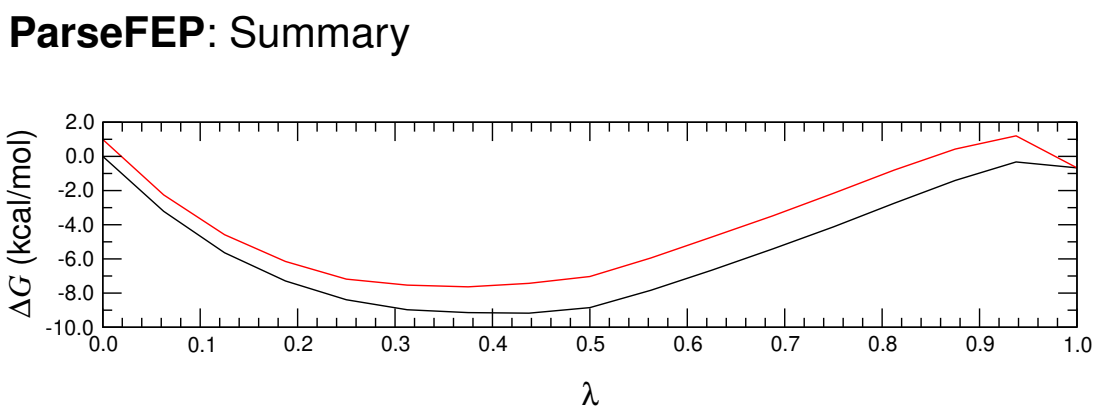
The Phe59 resides in $\beta 3$ of RRM domain (see Figure 5.15). Figure C.10 shows the overall free energy profile for F59A mutation in unbound and bound form of SRSF2 RRM domain.
Figure C.11 and Figure C.12 show the time evolution of free energy differences for each of the λ window and the probability distribution plots (for backward and forward transformations) for unbound and bound form of SRSF2 RRM domain, respectively.
The total free energy change (ΔG_3) of F59A mutation in unbound form of SRSF2 RRM is $-5.42\,\mathrm{kcal\,mol^{-1}}$ with a total error of $0.132\,\mathrm{kcal\,mol^{-1}}$. Whereas, in the bound form the free energy change (ΔG_4) for F59A point mutation is $-1.34\,\mathrm{kcal\,mol^{-1}}$ with a total error of $0.158\,\mathrm{kcal\,mol^{-1}}$.
To compute the relative binding free energy for F59A point mutation we will use the equation $\Delta\Delta G = \Delta G_4 - \Delta G_3$ (see Figure 5.16).

$$\begin{aligned}\Delta\Delta G &= (-1.34 \pm 0.158\,\mathrm{kcal\,mol^{-1}}) - (-5.42 \pm 0.132\,\mathrm{kcal\,mol^{-1}}) \\ &= (-1.34 - (-5.42)) \pm (\sqrt{0.158^2 + 0.132^2})\mathrm{kcal\,mol^{-1}} \\ &= 4.08 \pm 0.206\,\mathrm{kcal\,mol^{-1}}\end{aligned}$$

The relative binding free energy for D42A point mutation in SRSF2 RRM-RNA complex is $4.08 \pm 0.206\,\mathrm{kcal\,mol^{-1}}$.

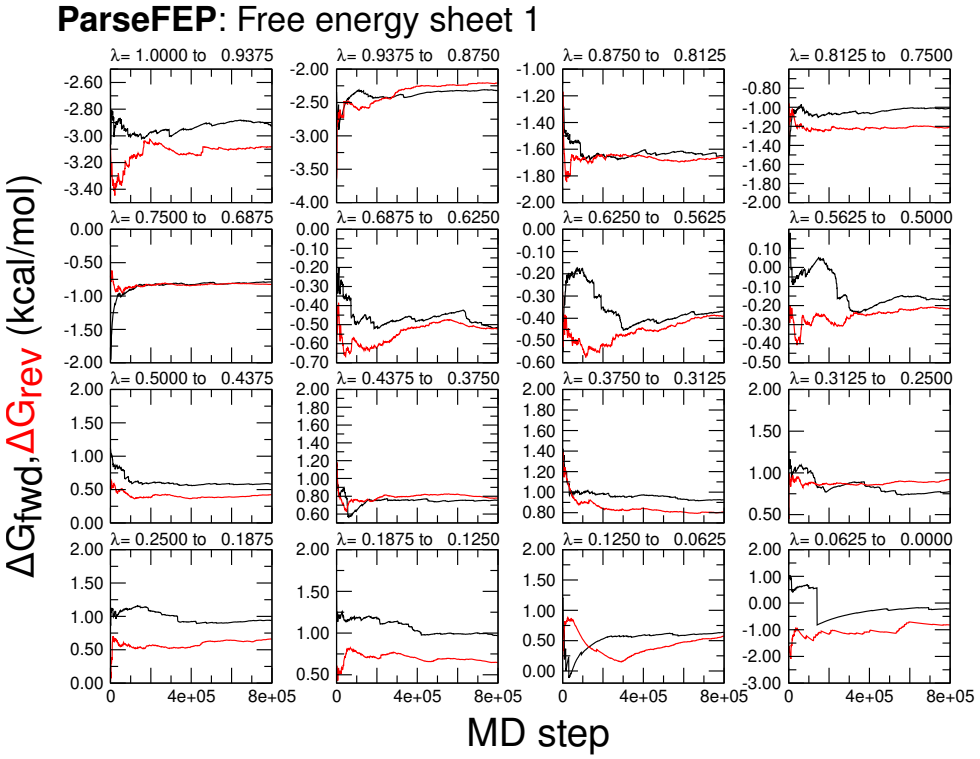


(a) Free energy change for Phe59 to Ala mutation in SRSF2 RRM domain (unbound form)

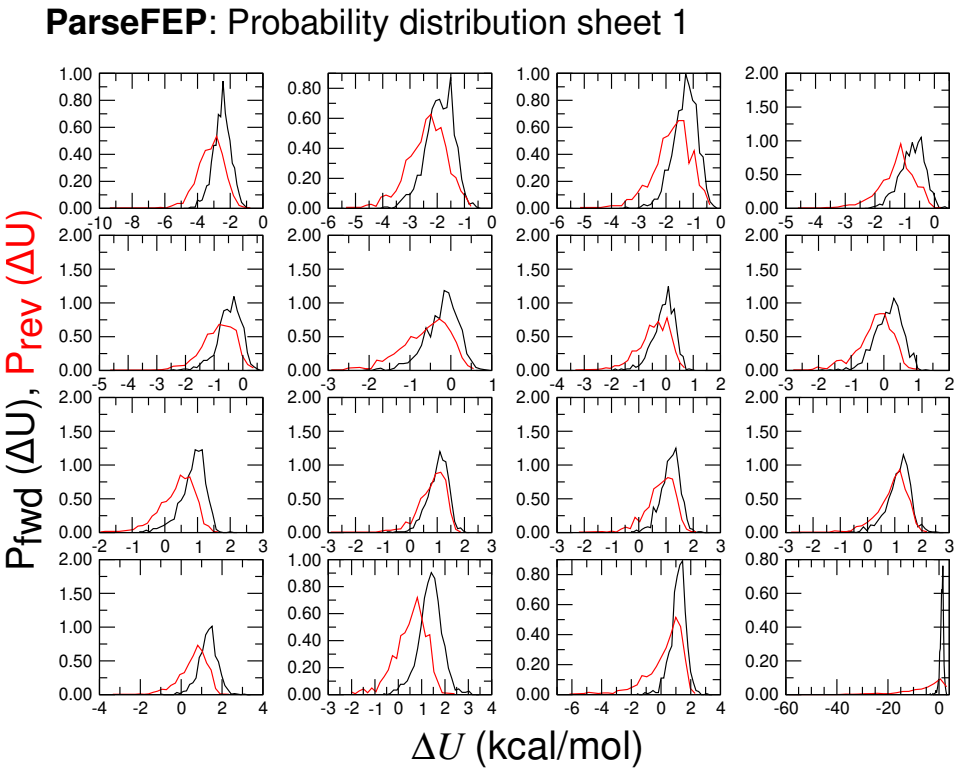


(b) Free energy change for Phe59 to Ala mutation in SRSF2 RRM domain (bound form)

Figure C.10: Free energy change for Arginine (position 61) to Alanine mutation in SRSF2 RRM domain. Black line indicates the forward transformation, i.e., from Arg to Ala and red line indicates the backward transformation from Ala to Arg.

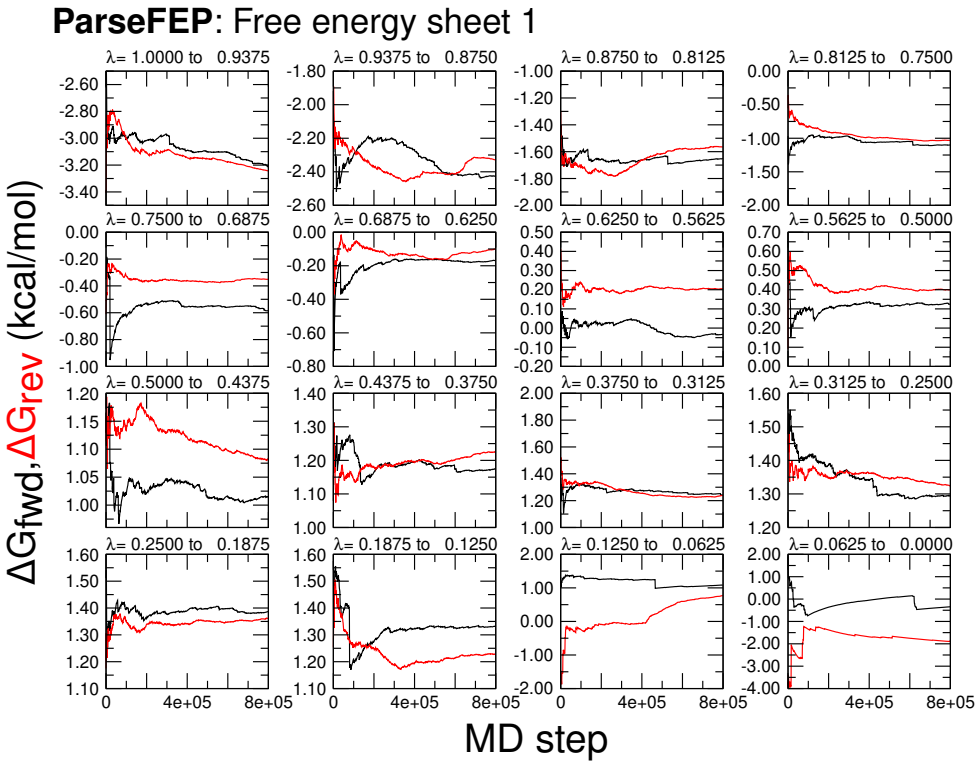


(a) Time evolution of free energy differences for each λ window

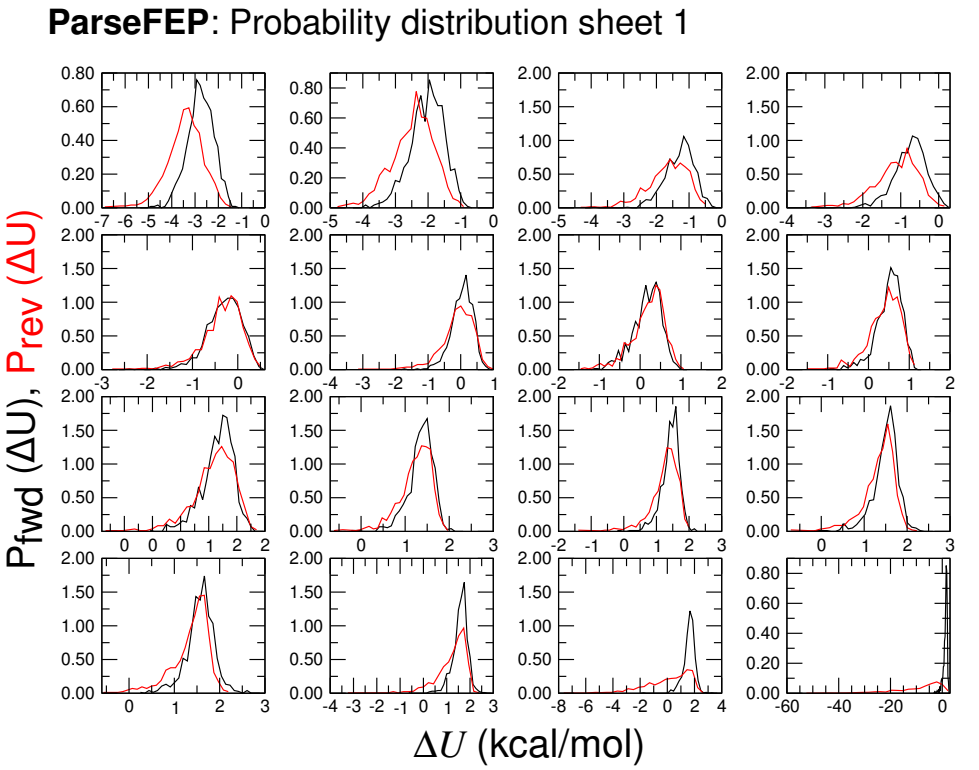


(b) Probability distribution plots for backward and forward transformations

Figure C.11: Output plots from the soft-core potenial calculation, of ΔG, and $P_0[\Delta U]$ and $P_1[\Delta U]$ generated by ParseFEP for alchemical transformation of F59A in SRSF2 RRM (unbound form). Each subplot corresponds to a λ sampling window.



(a) Time evolution of free energy differences for each λ window



(b) Probability distribution plots for backward and forward transformations

Figure C.12: Output plots from the soft-core potenial calculation, of ΔG, and $P_0[\Delta U]$ and $P_1[\Delta U]$ generated by ParseFEP for alchemical transformation of F59A in SRSF2 RRM (bound form). Each subplot corresponds to a λ sampling window.

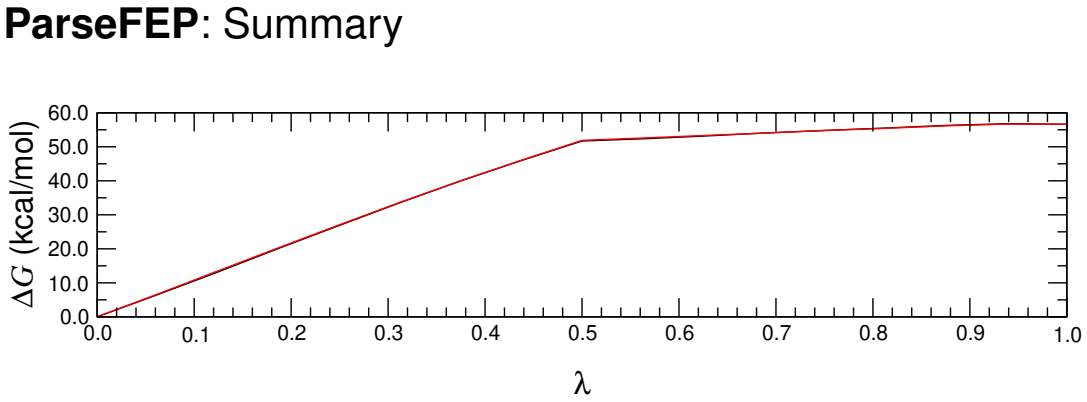
C.5 Q88A point mutation

The Gln88 resides in $\beta 4$ of RRM domain (see Figure 5.15). Figure C.13 shows the overall free energy profile for Q88A mutation in unbound and bound form of SRSF2 RRM domain.
The total free energy change (ΔG_3) of Q88A mutation in unbound form of SRSF2 RRM is $56.528\,\mathrm{kcal\,mol^{-1}}$ with a total error of $0.0847\,\mathrm{kcal\,mol^{-1}}$. Whereas, in the bound form the free energy change (ΔG_4) for Q88A point mutation is $56.730\,\mathrm{kcal\,mol^{-1}}$ with a total error of $0.0800\,\mathrm{kcal\,mol^{-1}}$.
Figure C.14 and Figure C.15 show the time evolution of free energy differences for each of the λ window and the probability distribution plots (for backward and forward transformations) for unbound and bound form of SRSF2 RRM domain, respectively.

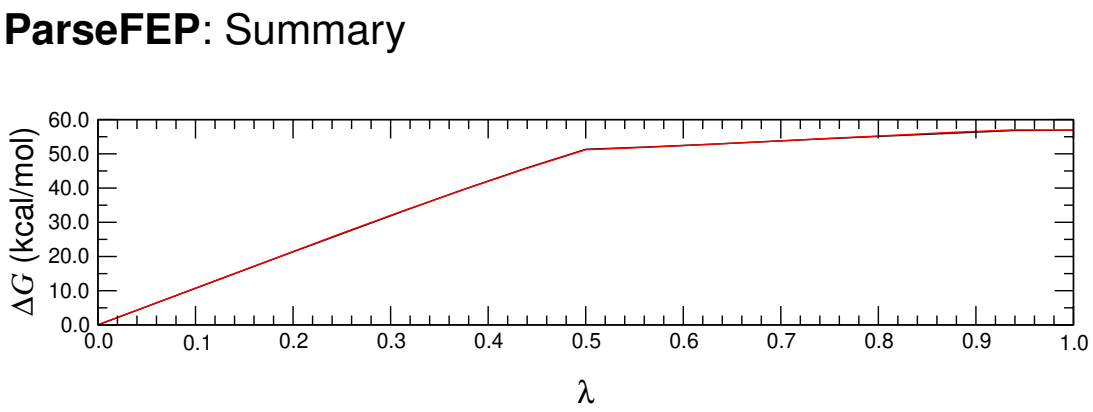
To compute the relative binding free energy for Q88A point mutation we will use the equation $\Delta\Delta G = \Delta G_4 - \Delta G_3$ (see Figure 5.16).

$$\begin{aligned}\Delta\Delta G &= (56.730 \pm 0.080\,\mathrm{kcal\,mol^{-1}}) - (56.528 \pm 0.084\,\mathrm{kcal\,mol^{-1}}) \\ &= (56.730 - 56.528) \pm (\sqrt{0.080^2 + 0.084^2})\mathrm{kcal\,mol^{-1}} \\ &= 0.202 \pm 0.116\,\mathrm{kcal\,mol^{-1}}\end{aligned}$$

The relative binding free energy for Q88A point mutation in SRSF2 RRM-RNA complex is $0.202 \pm 0.116\,\mathrm{kcal\,mol^{-1}}$.

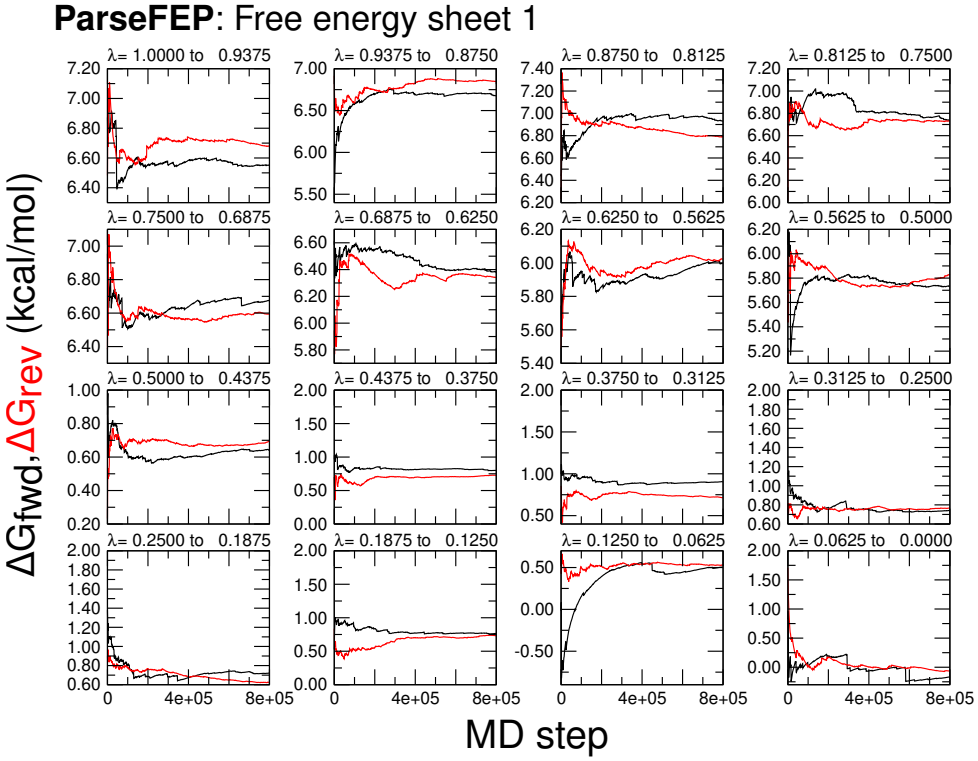


(a) Free energy change for Gln88 to Ala mutation in SRSF2 RRM domain (unbound form)

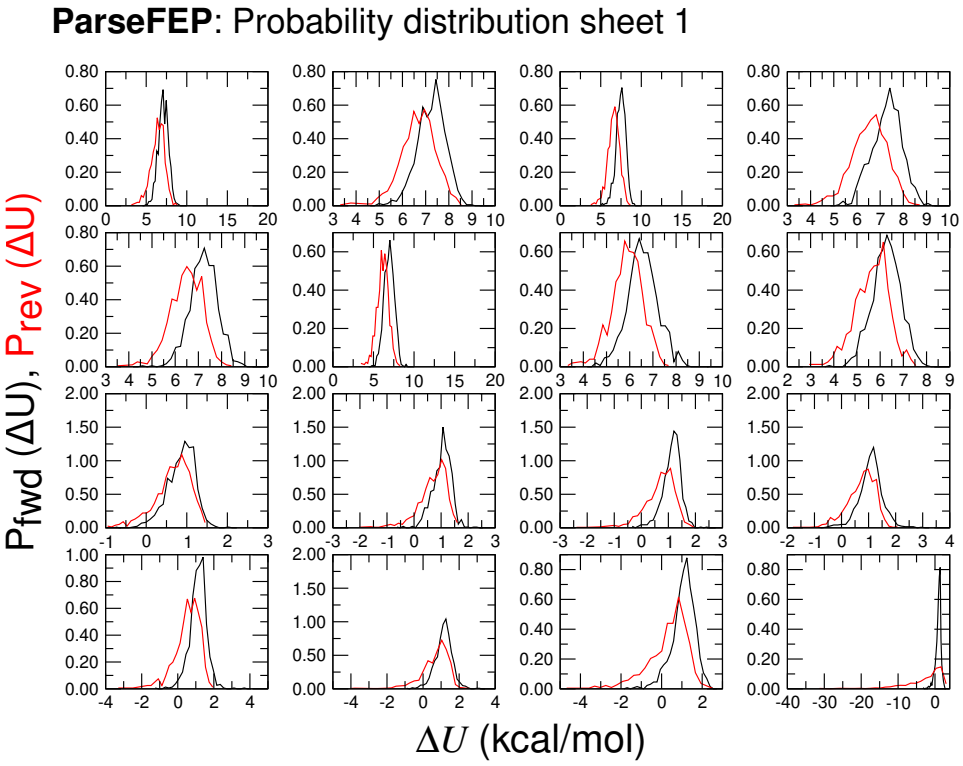


(b) Free energy change for Gln88 to Ala mutation in SRSF2 RRM domain (bound form)

Figure C.13: Free energy change for Arginine (position 61) to Alanine mutation in SRSF2 RRM domain. Black line indicates the forward transformation, i.e., from Gln to Ala and red line indicates the backward transformation from Ala to Gln.

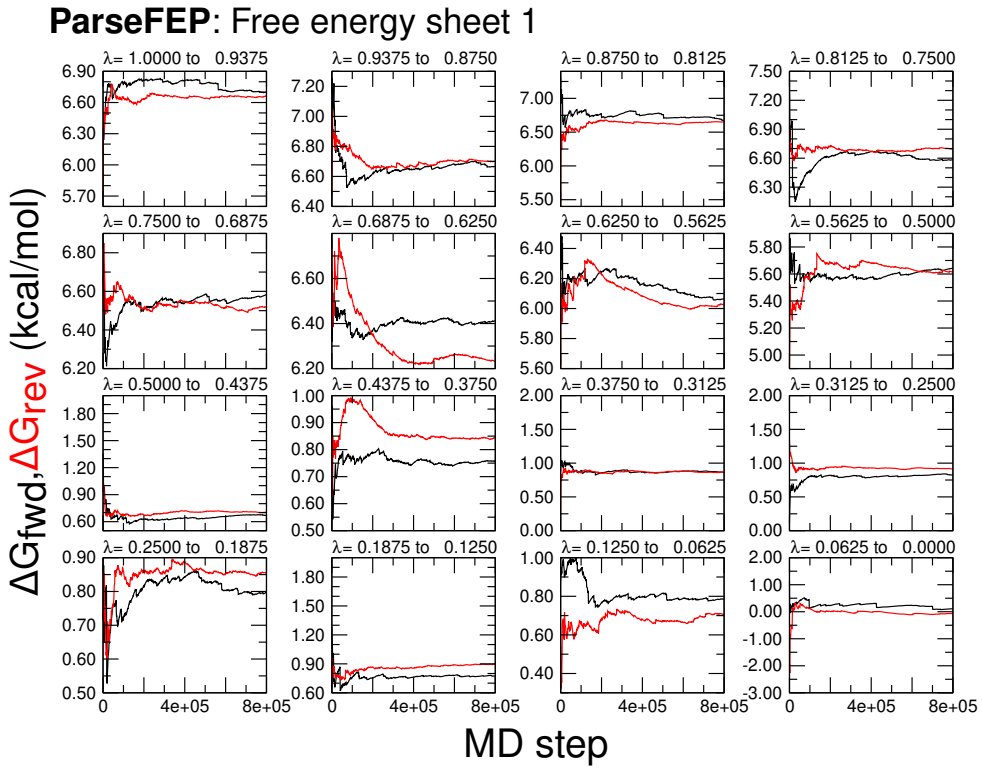


(a) Time evolution of free energy differences for each λ window

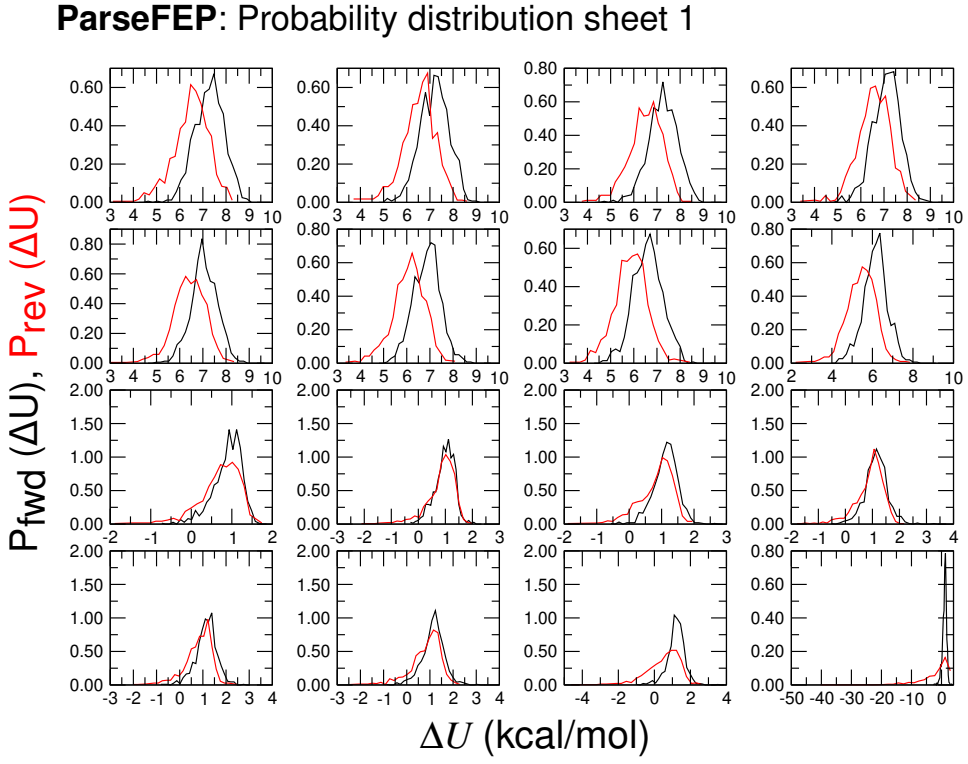


(b) Probability distribution plots for backward and forward transformations

Figure C.14: Output plots from the soft-core potenial calculation, of ΔG, and $P_0[\Delta U]$ and $P_1[\Delta U]$ generated by ParseFEP for alchemical transformation of Q88A in SRSF2 RRM (unbound form). Each subplot corresponds to a λ sampling window.



(a) Time evolution of free energy differences for each λ window



(b) Probability distribution plots for backward and forward transformations

Figure C.15: Output plots from the soft-core potenial calculation, of ΔG, and $P_0[\Delta U]$ and $P_1[\Delta U]$ generated by ParseFEP for alchemical transformation of Q88A in SRSF2 RRM (bound form). Each subplot corresponds to a λ sampling window.

Bibliography

F. H.-T. Allain, C. C. Gubser, P. W. Howe, K. Nagai, D. Neuhaus, and G. Varani. Specificity of ribonucleoprotein interaction determined by rna folding during complex formation. *Nature*, 380(6575):646–650, 1996.

P. Amstutz, M. Crusoe, N. Tijanić, B. Chapman, J. Chilton, M. Heuer, A. Kartashov, D. Leehr, H. Ménager, M. Nedeljkovich, et al. Common workflow language, v1. 0. specification, common workflow language working group. *Peter Amstutz MRC, Nebojša Tijanic, editor*, 2016.

A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, and A. G. Murzin. Scop2 prototype: a new approach to protein structure mining. *Nucleic acids research*, 42(D1):D310–D314, 2014.

A. Andreeva, E. Kulesha, J. Gough, and A. G. Murzin. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res*, 48(D1):D376–D382, 01 2020.

D. R. Armstrong, J. M. Berrisford, M. J. Conroy, A. Gutmanas, S. Anyango, P. Choudhary, A. R. Clark, J. M. Dana, M. Deshpande, R. Dunlop, et al. Pdbe: improved findability of macromolecular structure data in the pdb. *Nucleic acids research*, 48(D1):D335–D343, 2020.

S. D. Auweter, F. C. Oberstrass, and F. H.-T. Allain. Sequence-specific binding of single-stranded rna: is there a code for recognition? *Nucleic acids research*, 34 (17):4943–4959, 2006.

M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

M. Baek, R. McHugh, I. Anishchenko, D. Baker, and F. DiMaio. Accurate prediction of nucleic acid and protein-nucleic acid complexes using rosettafoldna. *bioRxiv*, pages 2022–09, 2022.

M. K. Basu, E. Poliakov, and I. B. Rogozin. Domain mobility in proteins: functional and evolutionary implications. *Briefings in bioinformatics*, 10(3):205–216, 2009.

S. Basu, S. Alagar, and R. P. Bahadur. Unusual rna binding of fus rrm studied by molecular dynamics simulation and enhanced sampling method. *Biophysical Journal*, 120(9):1765–1776, 2021.

A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-

Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. The pfam protein families database. *Nucleic acids research*, 30(1):276–280, 2002.

S. Batey and J. Clarke. The folding pathway of a single domain in a multidomain protein is not affected by its neighbouring domain. *Journal of molecular biology*, 378(2):297–301, 2008.

W. J. Bauer, J. Heath, J. L. Jenkins, and C. L. Kielkopf. Three rna recognition motifs participate in rna recognition and structural organization by the pro-apoptotic factor tia-1. *Journal of molecular biology*, 415(4):727–740, 2012.

H. M. Baumann, V. Gapsys, B. L. de Groot, and D. L. Mobley. Challenges encountered applying equilibrium and nonequilibrium binding free energy calculations. *The Journal of Physical Chemistry B*, 125(17):4241–4261, 2021.

G.-J. Bekker, M. Yokochi, H. Suzuki, Y. Ikegawa, T. Iwata, T. Kudou, K. Yura, T. Fujiwara, T. Kawabata, and G. Kurisu. Protein data bank japan: Celebrating our 20th anniversary during a global pandemic as the asian hub of three dimensional macromolecular structural data. *Protein Science*, 31(1):173–186, 2022.

M. Bellucci, F. Agostini, M. Masin, and G. G. Tartaglia. Predicting protein associations with long noncoding rnas. *Nature methods*, 8(6):444–445, 2011.

C. H. Bennett. Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.

L. E. Berchowitz, G. Kabachinski, M. R. Walker, T. M. Carlile, W. V. Gilbert, T. U. Schwartz, and A. Amon. Regulated formation of an amyloid-like translational repressor governs gametogenesis. *Cell*, 163(2):406–418, 2015.

H. Berman, K. Henrick, H. Nakamura, and J. L. Markley. The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. *Nucleic acids research*, 35(suppl_1):D301–D303, 2007.

E. Birney, S. Kumar, and A. R. Krainer. Analysis of the rna-recognition motif and rs and rgg domains: conservation in metazoan pre-mrna splicing factors. *Nucleic acids research*, 21(25):5803–5816, 1993.

G. Blackshields, F. Sievers, W. Shi, A. Wilm, and D. G. Higgins. Sequence embedding for fast construction of guide trees for multiple sequence alignment. *Algorithms for Molecular Biology*, 5:1–11, 2010.

B. D. Blakeley and B. R. McNaughton. Synthetic rna recognition motifs that selectively recognize hiv-1 trans-activation response element hairpin rna. *ACS chemical biology*, 9(6):1320–1329, 2014.

F. Blanco, M. Jimbo, J. Wulfkuhle, I. Gallagher, J. Deng, L. Enyenihi, N. Meisner-Kober, E. Londin, I. Rigoutsos, J. Sawicki, et al. The mrna-binding protein hur promotes hypoxia-induced chemoresistance through posttranscriptional regulation of the proto-oncogene pim1 in pancreatic cancer cells. *Oncogene*, 35 (19):2529–2541, 2016.

M. Blum, H. Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, L. Richardson, G. A. Salazar,

L. Williams, P. Bork, A. Bridge, J. Gough, D. H. Haft, I. Letunic, A. Marchler-Bauer, H. Mi, D. A. Natale, M. Necci, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman, and R. D. Finn. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*, 49(D1):D344–D354, 01 2021.

K. J. Brayer and D. J. Segal. Keep your fingers off my dna: protein–protein interactions mediated by c2h2 zinc finger domains. *Cell biochemistry and biophysics*, 50:111–131, 2008.

S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, L. Chen, G. V. Crichlow, C. H. Christie, K. Dalenberg, L. Di Costanzo, J. M. Duarte, et al. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, 49(D1): D437–D451, 2021.

J. Carey. [8] gel retardation. In *Methods in enzymology*, volume 208, pages 103–117. Elsevier, 1991.

Y. Chen, F. Yang, L. Zubovic, T. Pavelitz, W. Yang, K. Godin, M. Walker, S. Zheng, P. Macchi, and G. Varani. Targeted inhibition of oncogenic mir-21 maturation with designed rna-binding proteins. *Nature chemical biology*, 12(9):717–723, 2016.

Z.-H. Chen, Y.-J. Jing, J.-B. Yu, Z.-S. Jin, Z. Li, T.-T. He, and X.-Z. Su. Esrp1 induces cervical cancer cell g1-phase arrest via regulating cyclin a2 mrna stability. *International Journal of Molecular Sciences*, 20(15):3705, 2019.

H. Cheng, Y. Liao, R. D. Schaeffer, and N. V. Grishin. Manual classification strategies in the ecod database. *Proteins: Structure, Function, and Bioinformatics*, 83(7):1238–1251, 2015.

C. Ciani, A. Pérez-Ràfols, I. Bonomo, M. Micaelli, A. Esposito, C. Zucal, R. Belli, V. G. D'Agostino, I. Bianconi, V. Calderone, et al. Identification and characterization of an rrm-containing, rna binding protein in acinetobacter baumannii. *Biomolecules*, 12(7):922, 2022.

A. Cléry and F. Allain. From structure to function of rna binding domains. *RNA binding proteins*, pages 137–58, 2012.

A. V. Colasanti, X.-J. Lu, and W. K. Olson. Analyzing and building nucleic acid structures with 3dna. *Journal of visualized experiments: JoVE*, (74), 2013.

I. Coluzza. Computational protein design: a review. *Journal of Physics: Condensed Matter*, 29(14):143001, 2017.

T. M. Connolly and C. E. Begg. *Database systems: a practical approach to design, implementation, and management*. Pearson Education, 2005.

Z. Cournia, B. K. Allen, T. Beuming, D. A. Pearlman, B. K. Radak, and W. Sherman. Rigorous free energy simulations in virtual screening. *Journal of chemical information and modeling*, 60(9):4153–4169, 2020.

F. Cozzolino, I. Iacobucci, V. Monaco, and M. Monti. Protein–dna/rna interactions:

an overview of investigation methods in the-omics era. *Journal of Proteome Research*, 20(6):3018–3030, 2021.

F. H. Crick. The origin of the genetic code. *Journal of molecular biology*, 38(3): 367–379, 1968.

G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. Weblogo: a sequence logo generator. *Genome research*, 14(6):1188–1190, 2004.

M. R. Crusoe, S. Abeln, A. Iosup, P. Amstutz, J. Chilton, N. Tijanić, H. Ménager, S. Soiland-Reyes, B. Gavrilović, C. Goble, et al. Methods included: Standardizing computational reuse and portability with the common workflow language. *Communications of the ACM*, 65(6):54–63, 2022.

J. M. Dana, A. Gutmanas, N. Tyagi, G. Qi, C. O'Donovan, M. Martin, and S. Velankar. Sifts: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic acids research*, 47(D1):D482–D489, 2019.

G. M. Daubner, A. Cléry, S. Jayne, J. Stevenin, and F. H.-T. Allain. A syn–anti conformational difference allows srsf2 to recognize guanines and cytosines equally well. *The EMBO journal*, 31(1):162–174, 2012.

K. Daze and F. Hof. Molecular interaction and recognition. *Encyclopedia of physical organic chemistry*, pages 1–51, 2016.

I. C. de Beauchene, S. J. de Vries, and M. Zacharias. Fragment-based modelling of single stranded rna bound to rna recognition motif containing proteins. *Nucleic acids research*, 44(10):4565–4580, 2016.

D. A. Dixon, G. C. Balch, N. Kedersha, P. Anderson, G. A. Zimmerman, R. D. Beauchamp, and S. M. Prescott. Regulation of cyclooxygenase-2 expression by the translational silencer tia-1. *The Journal of experimental medicine*, 198(3): 475–481, 2003.

A. Doller, K. Schlepckow, H. Schwalbe, J. Pfeilschifter, and W. Eberhardt. Tandem phosphorylation of serines 221 and 318 by protein kinase cδ coordinates mrna binding and nucleocytoplasmic shuttling of hur. *Molecular and cellular biology*, 30(6):1397–1410, 2010.

D. A. Dougherty. The cation- π interaction. *Accounts of chemical research*, 46(4): 885–893, 2013.

X. Du, Y. Li, Y.-L. Xia, S.-M. Ai, J. Liang, P. Sang, X.-L. Ji, and S.-Q. Liu. Insights into protein–ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences*, 17(2):144, 2016.

R. C. Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, et al. Protein complex prediction with alphafold-multimer. *BioRxiv*, pages 2021–10, 2021.

M. Feng, G. Heinzelmann, and M. K. Gilson. Absolute binding free energy calculations improve enrichment of actives in virtual compound screening. *Scientific Reports*, 12(1):13640, 2022.

R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, et al. Pfam: the protein families database. *Nucleic acids research*, 42(D1):D222–D230, 2014.

S. Forli, R. Huey, M. E. Pique, M. F. Sanner, D. S. Goodsell, and A. J. Olson. Computational protein–ligand docking and virtual drug screening with the autodock suite. *Nature protocols*, 11(5):905–919, 2016.

K. Forslund and E. L. Sonnhammer. Evolution of protein domain architectures. *Evolutionary Genomics: Statistical and Computational Methods, Volume 2*, pages 187–216, 2012.

P. E. Fournier, H. Richet, and R. A. Weinstein. The epidemiology and control of acinetobacter baumannii in health care facilities. *Clinical infectious diseases*, 42 (5):692–699, 2006.

R. G. Fox, N. K. Lytle, D. V. Jaquish, F. D. Park, T. Ito, J. Bajaj, C. S. Koechlein, B. Zimdahl, M. Yano, J. L. Kopp, et al. Image-based detection and targeting of therapy resistance in pancreatic adenocarcinoma. *Nature*, 534(7607):407–411, 2016.

V. Gapsys and B. L. de Groot. Alchemical free energy calculations for nucleotide mutations in protein–dna complexes. *Journal of Chemical Theory and Computation*, 13(12):6275–6289, 2017.

J. Garnier, J.-F. Gibrat, and B. Robson. [32] gor method for predicting protein secondary structure from amino acid sequence. In *Methods in enzymology*, volume 266, pages 540–553. Elsevier, 1996.

A. W. Ghoorah, M.-D. Devignes, M. Smaïl-Tabbone, and D. W. Ritchie. Protein docking using case-based reasoning. *Proteins: Structure, Function, and Bioinformatics*, 81(12):2150–2158, 2013.

M. K. Gilson and H.-X. Zhou. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.*, 36:21–42, 2007.

C. Goble, S. Soiland-Reyes, F. Bacall, S. Owen, A. Williams, I. Eguinoa, B. Droesbeke, S. Leo, L. Pireddu, L. Rodríguez-Navas, et al. Implementing fair digital objects in the eosc-life workflow collaboratory. *Zenodo*, 2021.

J. Gomez-Porras, M. Lewinski, D. M. Riaño-Pachón, and D. Staiger. Molecular evolution of rrm-containing proteins and glycine-rich rna-binding proteins in plants. *Nature Precedings*, pages 1–1, 2011.

I. Grammatikakis, K. Abdelmohsen, and M. Gorospe. Posttranslational control of hur function. *Wiley Interdisciplinary Reviews: RNA*, 8(1):e1372, 2017.

A. M. Griffin, H. G. Griffin, and D. G. Higgins. Clustal v: multiple alignment of dna and protein sequences. *Computer Analysis of Sequence Data: Part II*, pages 307–318, 1994.

E. E. Guest, L. F. Cervantes, S. D. Pickett, C. L. Brooks III, and J. D. Hirst. Alchemical free energy methods applied to complexes of the first bromodomain of brd4. *Journal of Chemical Information and Modeling*, 62(6):1458–1470, 2022.

Y. Haddad, V. Adam, and Z. Heger. Ten quick tips for homology modeling of high-resolution protein 3d structures. *PLoS computational biology*, 16(4):e1007449, 2020.

C. Hardin, T. V. Pogorelov, and Z. Luthey-Schulten. Ab initio protein structure prediction. *Current opinion in structural biology*, 12(2):176–181, 2002.

L. M. Hellman and M. G. Fried. Electrophoretic mobility shift assay (emsa) for detecting protein–nucleic acid interactions. *Nature protocols*, 2(8):1849–1861, 2007.

J. C. Hoch, K. Baskaran, H. Burr, J. Chin, H. R. Eghbalnia, T. Fujiwara, M. R. Gryk, T. Iwata, C. Kojima, G. Kurisu, et al. Biological magnetic resonance data bank. *Nucleic Acids Research*, 51(D1):D368–D376, 2023.

C. E. Holt and E. M. Schuman. The central dogma decentralized: New perspectives on rna function and local translation in neurons. *Neuron*, 80(3):648–657, 2013. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2013.10.036. URL `https://www.sciencedirect.com/science/article/pii/S0896627313009884`.

P. Hosseinzadeh, G. Bhardwaj, V. K. Mulligan, M. D. Shortridge, T. W. Craven, F. Pardo-Avila, S. A. Rettie, D. E. Kim, D.-A. Silva, Y. M. Ibrahim, et al. Comprehensive computational design of ordered peptide macrocycles. *Science*, 358(6369):1461–1466, 2017.

J. Huang and A. D. MacKerell Jr. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. *Journal of computational chemistry*, 34(25):2135–2145, 2013.

X. Huang and W. Miller. A time-efficient, linear-space local similarity algorithm. *Advances in applied mathematics*, 12(3):337–357, 1991.

W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

J. M. Izquierdo, N. Majós, S. Bonnal, C. Martínez, R. Castelo, R. Guigó, D. Bilbao, and J. Valcárcel. Regulation of fas alternative splicing by antagonistic effects of tia-1 and ptb on exon definition. *Molecular cell*, 19(4):475–484, 2005.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Y. Karami, J. Rey, G. Postic, S. Murail, P. Tufféry, and S. J. De Vries. Dareus-loop: a web server to model multiple loops in homology models. *Nucleic Acids Research*, 47(W1):W423–W428, 2019.

J. Kästner and W. Thiel. Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method:"umbrella integration". *The Journal of chemical physics*, 123(14):144104, 2005.

N. Kedersha, M. R. Cho, W. Li, P. W. Yacono, S. Chen, N. Gilks, D. E. Golan, and P. Anderson. Dynamic shuttling of tia-1 accompanies the recruitment of mrna to mammalian stress granules. *The Journal of cell biology*, 151(6):1257–1268, 2000.

M. G. Kharas, C. J. Lengner, F. Al-Shahrour, L. Bullinger, B. Ball, S. Zaidi, K. Morgan, W. Tam, M. Paktinat, R. Okabe, et al. Musashi-2 regulates normal hematopoiesis and promotes aggressive myeloid leukemia. *Nature medicine*, 16 (8):903–908, 2010.

E. Kim, J. O. Ilagan, Y. Liang, G. M. Daubner, S. C.-W. Lee, A. Ramakrishnan, Y. Li, Y. R. Chung, J.-B. Micol, M. E. Murphy, et al. Srsf2 mutations contribute to myelodysplasia by mutant-specific effects on exon recognition. *Cancer cell*, 27 (5):617–630, 2015.

D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, 3(11):935–949, 2004.

P. Kollman. Free energy calculations: applications to chemical and biochemical phenomena. *Chemical reviews*, 93(7):2395–2417, 1993.

P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of chemical research*, 33(12):889–897, 2000.

B. L. Kormos, Y. Benitex, A. M. Baranger, and D. L. Beveridge. Affinity and specificity of protein u1a-rna complex formation based on an additive component free energy model. *Journal of molecular biology*, 371(5):1405–1419, 2007.

M. Kouza, E. Faraggi, A. Kolinski, and A. Kloczkowski. The gor method of protein secondary structure prediction and its application as a protein aggregation prediction tool. *Prediction of protein secondary structure*, pages 7–24, 2017.

M. Krepl, A. Cléry, M. Blatter, F. H. Allain, and J. Sponer. Synergy between nmr measurements and md simulations of protein/rna complexes: application to the rrms, the most common rna recognition motifs. *Nucleic acids research*, 44(13): 6452–6470, 2016.

E. Krissinel and K. Henrick. Multiple alignment of protein structures in three dimensions. In *Computational Life Sciences: First International Symposium, CompLife 2005, Konstanz, Germany, September 25-27, 2005. Proceedings 1*, pages 67–78. Springer, 2005.

G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack Jr. Improved prediction of protein side-chain conformations with scwrl4. *Proteins: Structure, Function, and Bioinformatics*, 77(4):778–795, 2009.

A. E. Kudinov, J. Karanicolas, E. A. Golemis, and Y. Boumber. Musashi rna-binding proteins as cancer drivers and novel therapeutic targetsmusashi proteins as cancer drivers and therapeutic targets. *Clinical Cancer Research*, 23(9):2143–2153, 2017.

I. Kufareva and R. Abagyan. Methods of protein structure comparison. In *Homology modeling*, pages 231–257. Springer, 2011.

B. Kuhlman and P. Bradley. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20(11):681–697, 2019.

R. Kumari, R. Kumar, O. S. D. D. Consortium, and A. Lynn. g_mmpbsa— a gromacs tool for high-throughput mm-pbsa calculations. *Journal of chemical information and modeling*, 54(7):1951–1962, 2014.

J. E. Ladbury and B. Z. Chowdhry. Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions. *Chemistry & biology*, 3(10):791–801, 1996.

B. Lang, A. Armaos, and G. G. Tartaglia. Rnact: Protein–rna interaction predictions for model organisms with supporting experimental data. *Nucleic acids research*, 47(D1):D601–D606, 2019.

E. Lewars. Computational chemistry. *Introduction to the theory and applications of molecular and quantum mechanics*, 318, 2011.

D.-W. Li and R. Brüschweiler. Protocol to make protein nmr structures amenable to stable long time scale molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 10(4):1781–1787, 2014.

W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

W. Li, R. D. Schaeffer, Z. Otwinowski, and N. V. Grishin. Estimation of uncertainties in the global distance test (gdt_ts) for casp models. *PloS one*, 11(5): e0154786, 2016.

J. Lin, J. Lu, Y. Feng, M. Sun, and K. Ye. An rna-binding complex involved in ribosome biogenesis contains a protein with homology to trna cca-adding enzyme. *PLoS biology*, 11(10):e1001669, 2013.

P. Liu, F. Dehez, W. Cai, and C. Chipot. A toolkit for the analysis of free-energy perturbation calculations. *Journal of chemical theory and computation*, 8(8):2606–2616, 2012.

S. Lu, J. Wang, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, M. Yang, D. Zhang, C. Zheng, C. J. Lanczycki, and A. Marchler-Bauer. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res*, 48(D1):D265–D268, 01 2020.

P. J. Lukavsky, D. Daujotyte, J. R. Tollervey, J. Ule, C. Stuani, E. Buratti, F. E. Baralle, F. F. Damberger, and F. H. Allain. Molecular basis of ug-rich rna recognition by the human splicing factor tdp-43. *Nature structural & molecular biology*, 20(12):1443–1449, 2013.

F. Madeira, M. Pearce, A. R. Tivey, P. Basutkar, J. Lee, O. Edbali, N. Madhusoodanan, A. Kolesnikov, and R. Lopez. Search and sequence analysis tools services from embl-ebi in 2022. *Nucleic acids research*, 50(W1):W276–W279, 2022.

C. Maris, C. Dominguez, and F. H.-T. Allain. The rna recognition motif, a plastic

rna-binding platform to regulate post-transcriptional gene expression. *The FEBS journal*, 272(9):2118–2131, 2005.

M. A. Marti-Renom, M. Madhusudhan, and A. Sali. Alignment of protein sequences by their profiles. *Protein Science*, 13(4):1071–1087, 2004.

J. R. Martinez, H. B. Dhondge, M. Sattler, and W. Vranken. Deciphering the rrm-rna recognition code: A computational analysis. *PLoS Computational Biology*, 2023.

A. May and M. Zacharias. Accounting for global protein deformability during protein–protein and protein–ligand docking. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1754(1-2):225–231, 2005.

X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–157, 2011.

B. R. Miller III, T. D. McGee Jr, J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg. Mmpbsa. py: an efficient program for end-state free energy calculations. *Journal of chemical theory and computation*, 8(9):3314–3321, 2012.

J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Res*, 49(D1):D412–D419, 01 2021.

N. J. Moerke. Fluorescence polarization (fp) assays for monitoring peptide-protein or nucleic acid-protein binding. *Current protocols in chemical biology*, 1(1):1–15, 2009.

A. Moniot, I. Chauvot de Beauchêne, and Y. Guermeur. Inferring ε-nets of finite sets in a rkhs. In *Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization: Dedicated to the Memory of Teuvo Kohonen/Proceedings of the 14th International Workshop, WSOM+ 2022, Prague, Czechia, July 6-7, 2022*, pages 53–62. Springer, 2022a.

A. Moniot, Y. Guermeur, S. J. de Vries, and I. Chauvot de Beauchene. Protnaff: protein-bound nucleic acid filters and fragment libraries. *Bioinformatics*, 38(16): 3911–3917, 2022b.

S. J. Nolan, J. C. Shiels, J. B. Tuite, K. L. Cecere, and A. M. Baranger. Recognition of an essential adenine at a protein- rna interface: comparison of the contributions of hydrogen bonds and a stacking interaction. *Journal of the American Chemical Society*, 121(38):8951–8952, 1999.

C. Notredame, D. G. Higgins, and J. Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1): 205–217, 2000.

M. Nowacka, P. Boccaletto, E. Jankowska, T. Jarzynka, J. M. Bujnicki, and S. Dunin-Horkawicz. Rrmdb—an evolutionary-oriented database of rna recognition motif sequences. *Database*, 2019, 2019.

R. Nussinov. How can plos computational biology help the biological sciences? *PLOS Computational Biology*, 9(10):e1003262, 2013.

T. Ohyama, T. Nagata, K. Tsuda, N. Kobayashi, T. Imai, H. Okano, T. Yamazaki, and M. Katahira. Structure of musashi1 in a complex with target rna: the role of aromatic stacking interactions. *Nucleic acids research*, 40(7):3218–3231, 2012.

C. Oliveira, H. Faoro, L. R. Alves, and S. Goldenberg. Rna-binding proteins and their role in the regulation of gene expression in trypanosoma cruzi and saccharomyces cerevisiae. *Genetics and molecular biology*, 40:22–30, 2017.

C. P Bagowski, W. Bruins, and A. JW te Velthuis. The nature of protein domain evolution: shaping the interaction network. *Current genomics*, 11(5):368–376, 2010.

S. G. Patching. Surface plasmon resonance spectroscopy for characterisation of membrane protein–ligand interactions and its potential for drug discovery. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1838(1):43–55, 2014.

S. Pawnikar, A. Bhattarai, J. Wang, and Y. Miao. Binding analysis using accelerated molecular dynamics simulations and future perspectives. *Advances and Applications in Bioinformatics and Chemistry*, pages 1–19, 2022.

J. Pei, B.-H. Kim, and N. V. Grishin. Promals3d: a tool for multiple protein sequence and structure alignments. *Nucleic acids research*, 36(7):2295–2300, 2008.

X. Periole and S.-J. Marrink. The martini coarse-grained force field. *Biomolecular simulations: methods and protocols*, pages 533–565, 2013.

M. V. Phan, T. Ngo Tri, P. Hong Anh, S. Baker, P. Kellam, and M. Cotten. Identification and characterization of coronaviridae genomes from vietnamese bats and rats based on conserved protein domains. *Virus evolution*, 4(2):vey035, 2018.

M. M. Phelan, B. T. Goult, J. C. Clayton, G. M. Hautbergue, S. A. Wilson, and L.-Y. Lian. The structure and selectivity of the sr protein srsf2 rrm domain with rna. *Nucleic acids research*, 40(7):3232–3244, 2012.

M. Piecyk, S. Wax, A. R. Beck, N. Kedersha, M. Gupta, B. Maritim, S. Chen, C. Gueydan, V. Kruys, M. Streuli, et al. Tia-1 is a translational silencer that selectively regulates the expression of tnf-α. *The EMBO journal*, 19(15):4154–4163, 2000.

B. G. Pierce, K. Wiehe, H. Hwang, B.-H. Kim, T. Vreven, and Z. Weng. Zdock server: interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics*, 30(12):1771–1773, 2014.

S. Piotr, B. Ranjit, and Z. Martin. Protein-dna docking with a coarse-grained force field. *BMC Bioinformatics*, 2012.

M. Ramanathan, D. F. Porter, and P. A. Khavari. Methods to study rna–protein interactions. *Nature methods*, 16(3):225–234, 2019.

D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, et al. A compendium of rna-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, 2013.

M. Remmert, A. Biegert, A. Hauser, and J. Söding. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 9(2):173–175, 2012.

D. W. Ritchie. Calculating and scoring high quality multiple flexible protein structure alignments. *Bioinformatics*, 32(17):2650–2658, 2016.

D. W. Ritchie, A. W. Ghoorah, L. Mavridis, and V. Venkatraman. Fast protein structure alignment using gaussian overlap scoring of backbone peptide fragment similarity. *Bioinformatics*, 28(24):3274–3281, 2012.

D. R. Roe and T. E. Cheatham III. Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data. *Journal of chemical theory and computation*, 9(7):3084–3095, 2013.

A. M. Rossi and C. W. Taylor. Analysis of protein-ligand interactions by fluorescence polarization. *Nature protocols*, 6(3):365–387, 2011.

S. P. Ryder, C. C. Clingman, L. M. Deveau, and F. Massi. Allosteric inhibition of a stem cell rna-binding protein by an intermediary metabolite, 2012.

C. W. Schultz, R. Preet, T. Dhir, D. A. Dixon, and J. R. Brody. Understanding and targeting the disease-related rna binding protein human antigen r (hur). *Wiley Interdisciplinary Reviews: RNA*, 11(3):e1581, 2020.

P. Setny and M. Zacharias. A coarse-grained force field for protein–rna docking. *Nucleic acids research*, 39(21):9118–9129, 2011.

M.-y. Shen and A. Sali. Statistical potential for assessment and prediction of protein structures. *Protein science*, 15(11):2507–2524, 2006.

C. R. Shotwell, J. D. Cleary, and J. A. Berglund. The potential of engineered eukaryotic rna-binding proteins as molecular tools and therapeutics. *Wiley Interdisciplinary Reviews: RNA*, 11(1):e1573, 2020.

F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1):539, 2011.

I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. M. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S. D. Lam, K. Berka, I. H. Varekova, R. Svobodova, J. Lees, and C. A. Orengo. CATH: increased structural coverage of functional space. *Nucleic Acids Res*, 49(D1): D266–D273, 01 2021.

M. O. Sinnokrot, E. F. Valeev, and C. D. Sherrill. Estimates of the ab initio limit for π- π interactions: The benzene dimer. *Journal of the American Chemical Society*, 124(36):10887–10893, 2002.

J. Söding. Protein homology detection by hmm–hmm comparison. *Bioinformatics*, 21(7):951–960, 2005.

M. Souaille and B. Roux. Extension to the weighted histogram analysis method:

combining umbrella sampling with free energy calculations. *Computer physics communications*, 135(1):40–57, 2001.

V. K. Srivastava and R. Yadav. Chapter 9 - Isothermal titration calorimetry. In G. Misra, editor, *Data Processing Handbook for Complex Biological Data Sources*, pages 125–137. Academic Press, 2019. ISBN 978-0-12-816548-5. doi: https://doi.org/10.1016/B978-0-12-816548-5.00009-5. URL `https://www.sciencedirect.com/science/article/pii/B9780128165485000095`.

M. Stonebraker, L. A. Rowe, and M. Hirohama. The implementation of postgres. *IEEE transactions on knowledge and data engineering*, 2(1):125–142, 1990.

The-UniProt-Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023.

J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680, 1994.

J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. The clustal_x windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic acids research*, 25(24):4876–4882, 1997.

S. Vajda, D. R. Hall, and D. Kozakov. Sampling and scoring: A marriage made in heaven. *Proteins: Structure, Function, and Bioinformatics*, 81(11):1874–1884, 2013.

B. Vallat, C. Madrid-Aliste, and A. Fiser. Modularity of protein folds as a tool for template-free modeling of structures. *PLoS computational biology*, 11(8):e1004419, 2015.

R. Valverde, L. Edwards, and L. Regan. Structure and function of kh domains. *The FEBS journal*, 275(11):2712–2726, 2008.

E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, et al. Robust transcriptome-wide discovery of rna-binding protein binding sites with enhanced clip (eclip). *Nature methods*, 13(6):508–514, 2016.

G. Van Zundert, J. Rodrigues, M. Trellet, C. Schmitz, P. Kastritis, E. Karaca, A. Melquiond, M. van Dijk, S. De Vries, and A. Bonvin. The haddock2. 2 web server: user-friendly integrative modeling of biomolecular complexes. *Journal of molecular biology*, 428(4):720–725, 2016.

M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

I. Wang, J. Hennig, P. K. A. Jagtap, M. Sonntag, J. Valcárcel, and M. Sattler. Structure, dynamics and rna binding of the multi-domain splicing factor tia-1. *Nucleic acids research*, 42(9):5949–5966, 2014.

X. Wang, J. McLachlan, P. D. Zamore, and T. M. T. Hall. Modular recognition of rna by a human pumilio-homology domain. *Cell*, 110(4):501–512, 2002.

Y. Wang, H. Zhang, H. Zhong, and Z. Xue. Protein domain identification methods and online resources. *Computational and Structural Biotechnology Journal*, 19:1145–1153, 2021. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2021.01.041. URL `https://www.sciencedirect.com/science/article/pii/S2001037021000453`.

T. A. Wassenaar and A. E. Mark. The effect of box shape on the dynamic properties of proteins simulated under periodic boundary conditions. *Journal of computational chemistry*, 27(3):316–325, 2006.

B. Webb and A. Sali. Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, 54(1):5–6, 2016.

W. Xie, H. Zhu, M. Zhao, L. Wang, S. Li, C. Zhao, Y. Zhou, B. Zhu, X. Jiang, W. Liu, et al. Crucial roles of different rna-binding hnrnp proteins in stem cells. *International Journal of Biological Sciences*, 17(3):807, 2021.

Y. Ye and A. Godzik. Fatcat: a web server for flexible structure comparison and structure similarity searching. *Nucleic acids research*, 32(suppl_2):W582–W585, 2004.

W. You, Z. Tang, and C.-E. A. Chang. Potential mean force from umbrella sampling simulations: What can we learn and what is missed? *Journal of chemical theory and computation*, 15(4):2433–2443, 2019.

F. M. Ytreberg, R. H. Swendsen, and D. M. Zuckerman. Comparison of free energy methods for molecular systems. *The Journal of chemical physics*, 125(18):184114, 2006.

Q. Yu, S. J. Cok, C. Zeng, and A. R. Morrison. Translational repression of human matrix metalloproteinases-13 by an alternatively spliced form of t-cell-restricted intracellular antigen-related protein (tiar). *Journal of Biological Chemistry*, 278 (3):1579–1584, 2003.

M. Zacharias. Protein–protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Science*, 12(6):1271–1282, 2003.

C. Zardecki, S. Dutta, D. S. Goodsell, R. Lowe, M. Voigt, and S. K. Burley. Pdb-101: Educational resources supporting molecular explorations through biology and medicine. *Protein Science*, 31(1):129–140, 2022.

A. Zemla. Lga: a method for finding 3d similarities in protein structures. *Nucleic acids research*, 31(13):3370–3374, 2003.

T. Zhang, N. Delestienne, G. Huez, V. Kruys, and C. Gueydan. Identification of the sequence determinants mediating the nucleo-cytoplasmic shuttling of tiar and tia-1 rna-binding proteins. *Journal of cell science*, 118(23):5453–5463, 2005.

Y. Zhang and M. F. Sanner. Autodock crankpep: combining folding and docking to predict protein–peptide complexes. *Bioinformatics*, 35(24):5121–5127, 2019.

Y. Zhang and J. Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

W. Zhao, S. Zhang, Y. Zhu, X. Xi, P. Bao, Z. Ma, T. H. Kapral, S. Chen, B. Zagrovic, Y. T. Yang, et al. Postar3: an updated platform for exploring post-transcriptional regulation coordinated by rna-binding proteins. *Nucleic Acids Research*, 50(D1): D287–D294, 2022.

W. Zhou, D. Melamed, G. Banyai, C. Meyer, T. Tuschl, M. Wickens, J. Cao, and S. Fields. Expanding the binding specificity for rna recognition by a puf domain. *Nature Communications*, 12(1):5107, 2021.

Appendix D

Summary of the thesis (in French)

Titre de la thèse : Caractérisation structurale de la liaison de l'ARN aux domaines à Motif de Reconnaissance de l'ARN (RRM), à l'aide de l'intégration de données, de la modélisation 3D et de la simulation dynamique

Résumé étendu de la thèse en Français

Cette thèse a été réalisée dans le cadre d'un projet Européen plus vaste (ITN_RNAct) dans lequel des approches informatiques et biologiques étaient combinées pour progresser vers la conception et la synthèse de nouveaux domaines protéiques capables de se fixer sur des séquences spécifiques d'ARN (Acide RiboNucléique). Les domaines RRMs représentent 50% de toutes les protéines fixant l'ARN et sont trouvées dans environ 2% de toutes les protéines du génome humain. Cependant, du fait de la grande diversité des domaines à RRMs, il n'y a eu jusqu'à présent que très peu de succès rapportés dans la conception de nouveaux domaines à RRMs. La thèse vise donc à décrire la diversité des domaines RRM et de leurs modes de liaison à l'ARN, en se fondant sur des observations expérimentales publiées, et à exploiter ces données pour améliorer la modélisation des complexes RRM-ARN. Elle comprend une introduction (Chapitre 1), un chapitre de présentation des notions de bases nécessaires pour comprendre les méthodes et les ressources utilisées dans la thèse (Chapitre 2), trois chapitres de résultats originaux (Chapitres 3 à 5) et une conclusion ouverte sur les perspectives de ce travail (Chapitre 6).

I. Bioinformatique structurale des domaines à RRM.

Les protéines sont des macromolécules biologiques constituées par l'enchaînement linéaire d'acides aminés (aa) pris parmi un vocabulaire de 20 aa différents. La séquence en aa (aussi appelée chaîne polypeptidique) constitue la structure primaire de la protéine. Elle est déterminée par la séquence en nucléotides du gène qui encode cette protéine dans le génome. La base de connaissance universelle UniProt (UniProtKB) est la ressource en ligne centrale pour les séquences des protéines et leurs annotations. Elle contenait en mars 2023 plus de 200 millions de séquences dont environ ¼ correspond à des entrées revues et annotées manuellement (section SwissProt) et le reste à la traduction en protéines de séquences nucléotidiques, enrichie d'annotations automatiques non vérifiées (section TrEMBL). Au cours de la synthèse d'une protéine, des liaisons physico-chimiques non covalentes s'établissent entre les aa pour mettre en place d'abord une structure secondaire composée principalement d'hélices α, de feuillets β ou de boucles. Puis ces différents éléments s'organisent entre eux pour former la structure tertiaire de la protéine. Dans le cas de protéines multimériques, une structure quaternaire peut aussi être observée résultant de l'agencement entre eux des différents monomères. Diverses méthodes expérimentales ont été développées au cours des années pour étudier la structure tri-dimensionnelle (3D) des protéines. On peut citer ici la cristallographie aux rayons X, la spectroscopie par résonance magnétique nucléaire (NMR), la microscopie électronique avec sa variante récente de cryo-microscopie électronique (Cryo-EM). Les données de structure 3D expérimentales sont stockées dans la banque de données des protéines (PDB) qui est gérée depuis 2003 par un consortium international (wwPDB). En mars 2023, la PDB contenait environ 200 000 structures 3D.

Selon les cas, une protéine peut être organisée en un ou plusieurs domaines protéiques. Ces domaines sont des régions de la protéine qui peuvent se replier de façon indépendante du reste de la protéine et être retrouvées à l'identique ou sous des formes très voisines dans d'autres protéines selon des combinaisons différentes. Un domaine protéique est souvent associé à une fonction particulière, définissant ainsi des types de domaines comme par exemple les domaines qui lient l'ARN. Des alignements de séquences et de structures permettent de comparer entre elles les diverses occurrences (ou instances) d'un type de domaine et de créer des familles de domaines. Ainsi, selon le type d'alignement et la façon de caractériser les traits communs des membres d'une famille, les domaines peuvent devenir des classes abstraites de domaines protéiques réels, caractérisées par une topologie commune des structures secondaires (repliement structural

ou « 3D fold »), ou par une signature de séquence dans laquelle des aa conservés à des positions particulières sont représentés par une matrice de scores de position (PSSM) ou un motif de modèle de Markov caché (HMM). Des bases de données de domaine se sont construites sur l'un ou l'autre de ces principes. Les plus générales et les mieux maintenues sont les bases Pfam et CATH, fondées respectivement sur les alignements de séquence et les plis structuraux. La base de domaines intégrée InterPro s'efforce d'unifier la classification des domaines en tirant avantage de toutes les bases particulières.

Plusieurs familles de domaines présentent des propriétés de liaison à l'ARN. Parmi elles on trouve les domaines à RRM. Ce domaine a une longueur moyenne de 90 aa et présente une topologie conservée des éléments de structure secondaire : β1–α1–β2–β3-α2-β4 conduisant à une structure tertiaire à deux couches, l'une constituée des feuillets β et l'autre des hélices α (Figure 1). De plus la séquence des domaines à RRM présente deux motifs conservés : RNP1 (8aa : (R/K)-G-(F/Y)-(G/A)-(F/Y)-V-X-(F/Y)) et RNP2 (6aa : (L/I)-(F/Y)-(V/I)-X-(N/G)-L), localisés sur les brins β3 et β1 respectivement.

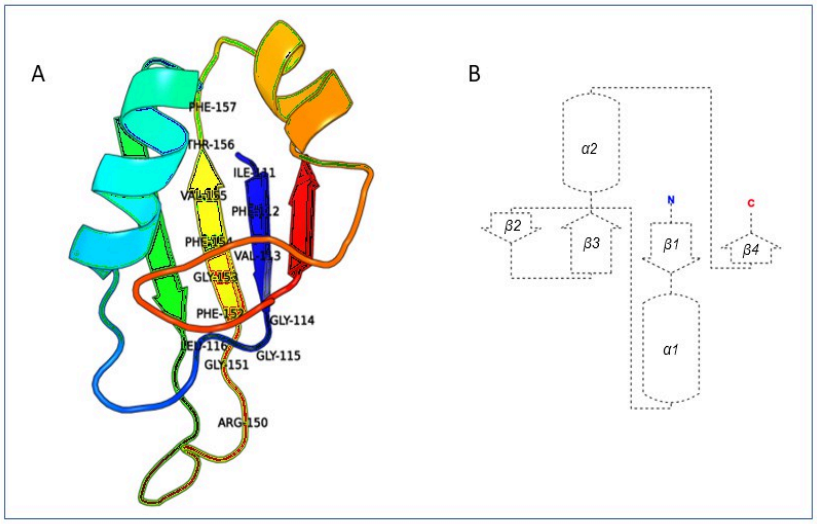


Figure 1 : Caractéristiques typiques du domaine à RRM illustrées avec l'entrée PDB 2mss (Musashi1 RBD2, NMR). A. Structure 3D en représentation ruban, colorée selon les éléments de structure secondaire, du bleu foncé côté N-terminal au rouge côté C-terminal, avec la position des motifs conservés RNP1 et RNP2 sur les brins β3 et β1 respectivement. B. Topologie conservée des éléments de structure secondaire des domaines à RRM.

Les domaines à RRM sont retrouvés dans un grand nombre de protéines régulatrices impliquées dans la régulation post-transcriptionnelle de l'expression des gènes, dans la répression de la traduction de certains gènes, dans les mécanismes de prolifération cellulaire anormale, la maintenance des cellules souches et l'activité de la télomérase. Il existe des bases de données spécifiques des protéines de liaison à l'ARN comme RBPDB, RNAct, POSTAR et RRMdb. Seule la dernière est centrée sur les domaines à RRM, avec notamment une analyse systématique des similarités de séquences entre domaines à RRM dans une perspective évolutive. Cependant elle n'est plus en ligne depuis plusieurs mois.

L'étude structurale des complexes protéine-ARN repose sur l'identification de différents types d'interaction entre les acides aminés d'une part et les nucléotides d'autre part. On distingue les interactions de Van-der-Waals, faibles et non spécifiques, les liaisons ioniques entre des groupements chargés positivement et négativement, les liaisons hydrogène, les interactions cation-π qui impliquent le nuage électronique π formé par les noyaux aromatiques de certains aa ou nucléotides, les interactions π-π entre deux nuages π aromatiques, encore appelées π-stacking quand les deux noyaux aromatiques sont dans des plans parallèles.

Du point de vue expérimental, l'énergie de liaison et sa constante d'affinité peuvent être mesurées par divers dispositifs techniques, tels que la calorimétrie à titration isothermique (ITC), la résonance de plasmons de surface (SPR), la polarisation de fluorescence (FP), le décalage de mobilité électrophorétique (EMSA).

Pour prédire la structure 3D d'un complexe protéine-ARN, il faut idéalement disposer des structures des deux partenaires isolés. Côté protéine, les structures 3D expérimentales disponibles dans la PDB ne concernent

que 0.08% des séquences enregistrées dans UniProt mais les systèmes de prédiction entraînés par apprentissage profond tels que AlphaFold ou RosettaFold sont devenus très puissants pour prédire les structures des protéines à partir des séquences. Côté ARN, il faut considérer l'ARN simple-brin comme une structure éminemment flexible qui se repliera ou se dépliera en même temps qu'elle se liera à la protéine.

Les méthodes de docking comportent deux grandes étapes : l'échantillonnage (sampling) et la notation (scoring). Il s'agit de trouver, dans l'immense espace conformationnel des poses, la ou les poses de moindre énergie représentant les conformations et orientations relatives des deux partenaires les plus plausibles pour le complexe protéine-ARN. Des méthodes à base de fragments peuvent être utilisées pour tenir compte de la flexibilité de l'ARN. Des représentations gros-grain des macromolécules sont aussi souvent adoptées pour diminuer la complexité des calculs mais elles nécessitent de définir des champs de force adaptés pour l'étape de notation des poses de docking. Egalement, des contraintes issues de données expérimentales peuvent être introduites pour limiter les recherches dans l'espace conformationnel. Les principaux programmes existants pour réaliser le docking protéine-ARN sont ATTRACT et HADDOCK. Dans le cadre de cette thèse, nous avons utilisé ATTRACT et son module ssRNA 'TTRACT qui utilise des représentations gros-grain des molécules et des librairies de conformations 3D de tri-nucléotides comme fragments pour le docking. Les poses de tri-nucléotides présentant les meilleurs scores sont conservées et deux poses sont liées entre elles quand elles se chevauchent selon un seuil donné. Le score final des chaînes d'ARN assemblées sur la protéine permet de distinguer les structures les plus plausibles. Le ou les modèles retenus sont ensuite raffinés par minimisation et/ou simulation en dynamique moléculaire.

La dynamique moléculaire est une méthode pour simuler le comportement de biomolécules isolées ou en interaction. Toutes les forces s'exerçant sur chaque atome sont exprimées grâce aux champs de force correspondant aux fonctions d'énergie des liaisons covalentes et non covalentes entre les atomes. Elles sont utilisées pour résoudre les équations de Newton déterminant le mouvement des atomes à chaque pas de temps de la simulation. Les conditions de simulation sont définies par des paramètres thermodynamiques précis qui dépendent des systèmes de simulation (NAMD, GROMACS, AMBER sont les principaux logiciels de dynamique moléculaire). La structure 3D des biomolécules étudiées est calculée à chaque pas de temps et constitue ce qu'on appelle une « frame » de la dynamique (par exemple un pas de 1 ps pour une simulation de 10 ns conduit à une trajectoire de 10 000 frames). La cinétique ainsi obtenue peut ensuite être analysée de diverses façons, en particulier pour déterminer une énergie libre d'interaction par les méthodes dites MM-PBSA (Mécanique moléculaire avec aire de la surface de Poisson-Boltzmann) ou MM-BGSA (Mécanique moléculaire avec solvatation généralisée de Born et de surface), ou par la méthode de l'énergie libre de perturbation (FEP). Cette dernière méthode a été utilisée pour étudier l'effet de mutations dans des complexes protéine-ADN mais pas encore pour les complexes protéine-ARN. De façon générale, les analyses de simulation en dynamique moléculaire viennent compléter des études expérimentales pour mieux comprendre le comportement des complexes et le rôle particulier de certains aa dans les interactions entre biomolécules.

II. InteR3M : la base de données des interactions ARN-RRM

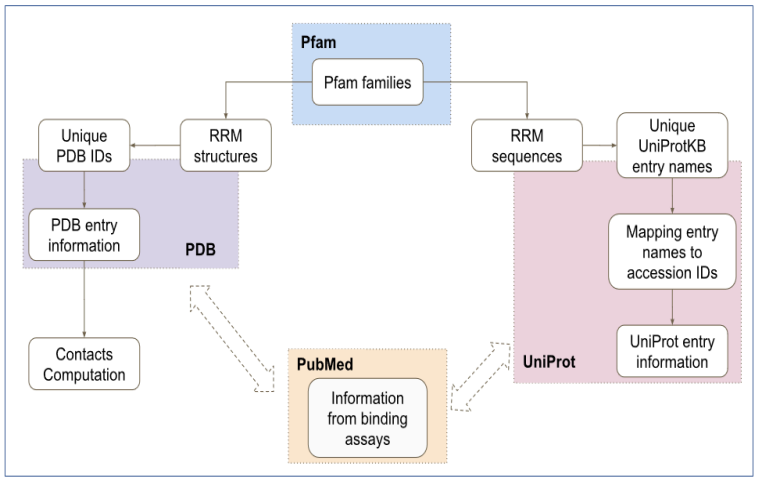
Ce chapitre décrit la conception, l'implémentation et le peuplement d'une base de données relationnelle : InteR3M, décrivant les interactions entre ARN et domaines à RRM. Par rapport aux bases de données existantes sur les protéines de liaison à l'ARN, InteR3M met l'accent sur l'analyse des structures 3D expérimentales des domaines à RRM et des complexes RRM-ARN, dans le but de contribuer à déchiffrer le code de liaison à l'ARN des domaines à RRM et d'améliorer le docking RRM-ARN. L'objectif global de la base de données InteR3Mdb est donc de collecter, d'organiser, de rendre accessible et de maintenir les informations disponibles sur les domaines à RRM.

Le travail a commencé par préciser l'inventaire des domaines à RRM parmi toutes les familles de domaines Pfam (version utilisée v.33.0). A partir d'une liste de 42 familles Pfam, issues de requêtes simples dans Pfam, du clan RRM de Pfam (CL0221) et de RRMdb, il est apparu que seules 26 de ces familles disposaient d'au

moins une occurrence structurale (notée StI pour « Structural Instance ») dans la PDB. La famille Pfam RRM_1 (PF00076) est la plus abondamment peuplée en séquences et en StIs (246 502 séquences et 1069 structures 3D). Elle contient les domaines à RRM les plus étudiés expérimentalement et deux de ces domaines (1A9N_B et 2A3J_A) ont servi de prototypes pour inspecter manuellement les StIs des autres familles selon 3 critères. (i) Le domaine correspondant dans les bases de domaines orientées structure SCOP et CATH est un domaine de liaison à l'ARN ; (ii) la topologie des éléments de structure secondaire est alignable avec la topologie classique des RRM (β1–α1–β2–β3-α2-β4) ; et (iii) l'alignement structural par le programme Kpax donne un M-score supérieur à 0.60 (seuil déterminé empiriquement). Au final, 19 familles Pfam avec StI ont été conservées pour constituer la base de données InteR3M.

La conception du modèle de données d'InteR3M s'est appuyée sur une liste de cas d'étude recueillies auprès des futurs utilisateurs de la base de données au sein du consortium RNAct. Un modèle conceptuel de données de type entité-association a été élaboré puis transformé en modèle relationnel logique puis enfin implémenté physiquement en utilisant le gestionnaire de base de données libre PostGreSQL. Les données stockées concerneront les instances de domaines à RRM, leurs ligands (principalement des ARN) et les linkers qui peuvent séparer deux domaines à RRM dans une protéine.

Le peuplement de la base de données à partir des 19 familles Pfam sélectionnées est décrit schématiquement dans la Figure 2 selon quatre blocs : (i) les séquences des domaines à RRM (des 19 familles Pfam) avec les informations associées extraites d'UniProt, (ii) les structures 3D des domaines à RRM correspondant à certaines de ces séquences, avec les informations associées extraites de la PDB, (iii) les informations de contacts atomiques entre RRM et ARN, extraites des entrées de la PDB contenant des complexes RRM-ARN, (iv) les informations sur les affinités de liaison, extraites des publications sélectionnées par le consortium RNAct. La sélection des familles Pfam de départ et l'extraction des informations sur les affinités de liaison dans les articles PubMed ont été réalisées manuellement. Toutes les autres données sont collectées automatiquement à l'aide de scripts Python réutilisables qui serviront à la mise à jour de la base de données.



Pour les contacts entre RRM et ARN (ou ADN), un aa et un nucléotide (nt) sont considérés comme interagissant si au moins un atome de chaque sont trouvés à une distance de moins de 5Å l'un de l'autre. Pour toutes ces paires (aa, nt), 4 types de contacts ou d'interactions ont recherchés : liaisons de Van-der-Waals, liaisons Hydrogène, liaisons ioniques et π-stacking.

Figure 2 : Schema de la collecte des données pour le peuplement de la base de données InteR3M.

Une interface d'interrogation et d'exploration du contenu de InteR3Mdb a également été développée en langage PHP (moteur de développement Twig) et est disponible à l'adresse suivante (https://inter3mdb.loria.fr).

La base de données InteR3M (v0.0.1) contient un total de 400 892 instances de domaines à RRM correspondant à 256 266 protéines uniques et répertoriées dans 19 familles Pfam. De toutes ces instances, seulement 303 ont été étudiées d'un point de vue structural et ont donné lieu à 727 entrées PDB, pour un total de 1456 StI. La plupart de ces structures 3D appartiennent à la famille PF00076 (RRM_1 ; 1334 StI), la 2 famille la plus peuplée est PF13893 (RRM_5) avec seulement 42 StI. Des informations de contact (459 859

interactions dont 1926 π-stacking entre aa et nt, le reste en contacts atomiques) ont pu être extraites de 550 StI en complexe avec un ARN ou un ADN. Au total, 311 acides nucléiques différents (ARN ou ADN) sont présents dans les expériences structurales ou fonctionnelles rapportées dans InteR3M.

En conclusion, ce chapitre a décrit la création d'une base de données « domaine-centrique » permettant d'accéder à toutes les données relatives aux domaines à RRM et notamment les données structurales et de contact entre RRM et ARN. Par opposition à des bases de données généralistes s'intéressant à tous les types de domaines, InteR3M présente l'avantage de contenir des données fiables, revues manuellement et pouvant être utilisées pour des études évolutives ou de design de nouvelles protéines. L'automatisation de la collecte des données permettra de maintenir InteR3M à jour dès que de nouveaux domaines à RRM ou de nouvelles structures expérimentales de complexes RRM-ARN seront décrites. Il est aussi prévu d'intégrer les données des structures prédites par Alpha-Fold dans InteR3M mais ces données ne concernent pas les complexes RRM-ARN et ne pourront donc pas enrichir les connaissances sur les contacts et les modes de liaison de l'ARN aux domaines à RRM.

III. CroMaSt : un workflow pour vérifier la classification des domaines protéiques par assignation croisée entre les bases de données et alignement structural

Ce chapitre fait l'objet d'un article accepté dans Bioinformatics Advances.

Dhondge Hrishikesh, Chauvot de Beauchêne Isaure and Devignes Marie-Dominique. CroMaSt : A workflow for assessing protein domain classification by cross-mapping of structural instances between domain databases and structural alignment. Bioinformatics Advances. 3 :vbad081, 2023.

Le développement de bases de données particulières, centrées sur un type de domaine particulier, suppose de s'assurer que les familles de domaines extraites des bases de domaines à partir de requêtes simples comprenant le nom du type de domaine considéré correspondent bien à ce type de domaine. L'expérience acquise avec les domaines à RRM nous a montré qu'il n'en était pas toujours ainsi et a conduit au développement du workflow CroMaSt (Cross-Mapping of Structural instances).

L'objectif de CroMaSt est de pouvoir attribuer un type de domaine d'intérêt à un domaine particulier dont on a une structure 3D et à travers lui à toute la famille de domaines à laquelle il appartient. En effet, un type de domaine d'intérêt peut être représenté par plusieurs familles de domaines. Mais les bases de domaines généralistes peuvent avoir affecté à tort une famille de domaines à un type de domaine d'intérêt.

L'approche repose sur une connaissance a priori du type de domaine étudié comme par exemple les domaines à RRM. On supposera donc qu'il existe une définition consensus de ce type de domaine et quelques structures 3D validées. Dans notre exemple c'est le cas de la famille RRM_1 (PF00076) dans la base de domaine Pfam, qui comprend les exemples types des domaines à RRM. La première hypothèse de CroMaSt est donc qu'il est possible de comparer des structures 3D (instances structurales ou StI) de domaines avec une structure 3D de référence pour un type de domaine d'intérêt, et d'en déduire automatiquement, avec des seuils de score appropriés, si ces StIs correspondent ou non à ce type de domaine d'intérêt. Cependant si ce travail doit être effectué pour chaque StI individuellement, il peut devenir très fastidieux. C'est pourquoi CroMaSt utilise également les bases de domaines généralistes et bien documentées que sont Pfam et CATH. La deuxième hypothèse de CroMaSt est que si une StI appartient à la fois à la famille Pfam et à la famille CATH de référence pour le type de domaine d'intérêt, il est vraisemblable qu'il s'agit d'une vraie instance de ce type de domaine. Si par contre une StI appartient uniquement à la famille Pfam de référence ou uniquement à la famille CATH de référence, alors il convient de rechercher par assignation croisée si cette StI est tout simplement absente de l'autre classification ou présente dans une autre famille de domaines. Dans les deux cas, il conviendra d'analyser cette StI et la nouvelle famille par alignement structural avec la structure 3D de référence pour déterminer si la StI isolée et/ou sa famille peuvent être considérées comme appartenant au type de domaine d'intérêt.

Ainsi, le workflow CroMaSt va classifier les StI de domaines en 4 catégories selon le tableau ci-dessous.

Tableau 1 : Les catégories de CroMaSt selon les caractéristiques des instances structurales de domaine

Catégorie	Communes entre les familles de référence initiales	Assignées de façon croisée dans l'autre base de domaines	Présentant une similarité significative par alignement structural avec la structure 3D de référence
Noyau (« Core »)	Oui	-	-
Vraie (« True »)	Non	Oui	Oui
Pseudo-vraie (« Domain-like »)	Non	Non	Oui
En échec (« Failed »)	Non	Oui ou Non	Non

Les différentes étapes du workflow sont schématisées dans la Figure 3. Des itérations sont possibles tant que de nouvelles familles ont été découvertes par assignation croisée.

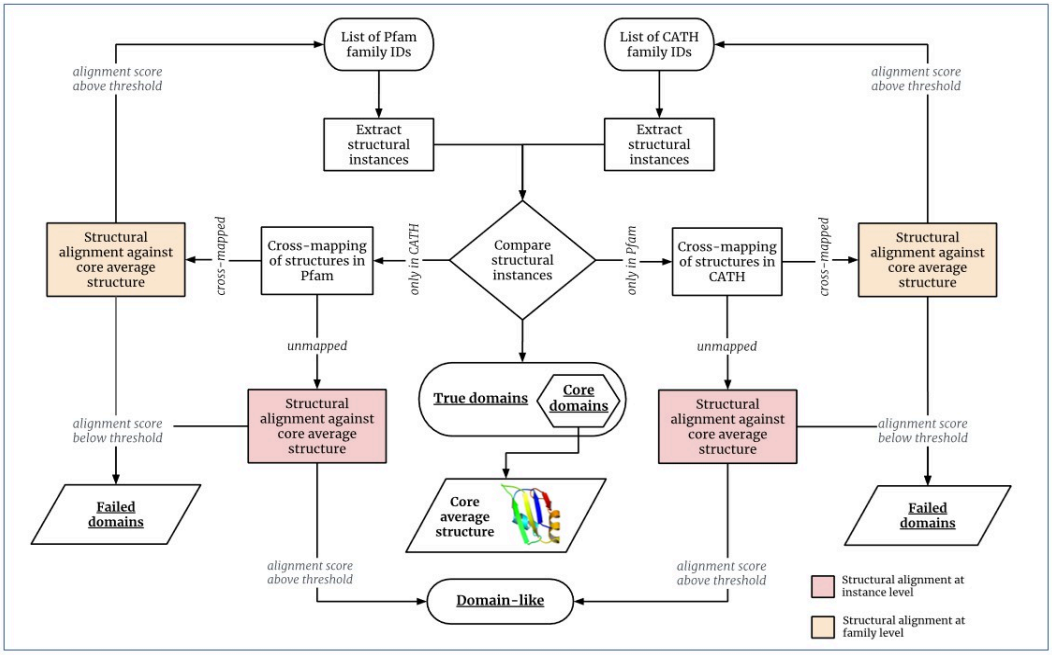


Figure 3. Schema du workflow CroMaSt.

Parmi les difficultés à résoudre pour implémenter ce workflow, il faut citer la nécessité d'introduire un format unifié pour identifier les StI dans les deux bases de domaines et le calcul des structures 3D moyennes au niveau des instances elle-mêmes et au niveau d'une famille de domaines.

Le format unifié de description des StI comprend l'identifiant dans la PDB, y compris l'identifiant de chaîne, le nom du domaine ou son numéro d'ordre lorsqu'il y en a plusieurs, l'identifiant de la famille de domaines, les positions de début et de fin dans la numérotation du fichier PDB, l'identifiant UniProt et les numéros de début et de fin dans la séquence du fichier UniProt. Les deux numérotations (PDB et UniProt) sont utilisées respectivement par CATH et Pfam. La correspondance entre ces deux numérotations est calculée automatiquement d'après les fichiers fournis par la ressource SIFTS. Une différence de 30 aa est acceptée pour les positions de début et de fin du domaine entre les deux bases de domaines. Cette valeur peut être modifiée en particulier si l'on travaille avec des domaines de petite taille.

L'alignement structural est réalisé par Kpax qui est aussi capable de calculer des structures moyennes à partir d'un alignement structural multiple. La structure 3D de référence pour le type de domaine d'intérêt est

calculée par la moyenne des structures de catégorie noyau (« Core »), c'est-à-dire partagées entre les deux familles de référence initiales. Pour éviter les biais liés aux instances de domaines très étudiées pour lesquelles il existe de nombreuses StI pour la même instance en séquence, une première structure moyenne dite « au niveau de l'instance » est calculée à partir de toutes les StI correspondant à la même séquence. Puis une structure moyenne, dite « au niveau de la famille », est calculée comme la moyenne des moyennes, à partir de l'ensemble des structures moyennes des instances d'une famille donnée.

L'ensemble du workflow a été implémenté dans le formalisme standard CWL (Common Workflow Language) qui respecte les principes FAIR de la science ouverte. Le workflow CroMaSt est disponible dans l'entrepôt WorflowHub sous l'identifiant suivant https://workflowhub.eu/workflows/390?version=1.

Les résultats obtenus pour le type de domaine à RRM sont les suivants. Lors de la première itération, les familles PF00076 (RRM_1) de Pfam et 3.30.70.330 (RRM domain) de CATH ont été sélectionnées comme familles de départ. Un total de 1333 et 1204 StI utilisables ont été extraites respectivement de chaque famille, parmi lesquelles 883 sont communes aux deux familles. Ces 883 StI ont permis de construire la structure 3D moyenne de référence pour les domaines à RRM. Les 450 StIs propres à Pfam ont été recherchées par assignation croisée dans CATH sans succès. Une explication pour cela serait que CATH utilise une version antérieure d'UniProt, incomplète donc par rapport à Pfam. Ces 450 StIs propres à Pfam ont donc été testées par alignement structural avec la structure de référence et 443 ont été trouvées avec un M-score au-dessus du seuil. Ces StIs sont donc classées dans la catégorie des StI pseudo-vraies (« Domain-like »). Les 7 StIs qui ne s'alignent pas correctement sont dites en échec (« Failed »). Les 321 StIs propres à CATH ont été recherchées par assignation croisée dans Pfam et 243 ont pu être assignées à 17 familles de Pfam. Le test d'alignement structural a montré que 79 StIs seulement, correspondant à 14 des 17 familles Pfam, sont correctement alignées avec la structure de référence et peuvent donc être classées dans la catégorie des vraies StI (« True »). Les 164 restantes (243 – 79) sont donc en échec ainsi que les 3 familles Pfam correspondantes. Quant aux 78 StIs (321-243) qui n'ont pas été assignées à des familles Pfam, toutes sont alignées correctement avec la structure moyenne de référence et viennent donc dans la catégorie pseudo-vraies.

Les 79 vraies StIs trouvées dans CATH et assignées à 14 familles Pfam seront utilisées dans la seconde itération pour tester si des membres de ces familles Pfam peuvent être trouvées dans CATH. Lors de cette seconde itération, toutes les StIs des 14 familles Pfam sont extraites, soit 100 StIs, parmi lesquelles 79 sont communes entre CATH et Pfam (catégorie « True »). Les 21 StIs propres à Pfam ne trouvent pas de nouvelle famille CATH en assignation croisée. Elles sont donc testées individuellement et 20 d'entre elles présentent un alignement correct (pseudo-vraies) contre une qui reste en échec.

Ainsi l'exécution du workflow CroMaSt avec les deux familles principales de domaines à RRM de Pfam et CATH a conduit à explorer 14 familles Pfam et une seule famille CATH et à identifier 962 vraies StI (incluant les 883 StI noyaux), 541 StI pseudo-vraies (443 + 20 de Pfam et 78 de CATH), et 172 StIs en échec (8 de Pfam et 164 de CATH). Tous les fichiers de calculs intermédiaires et en particuliers les résultats des alignements structuraux sont fournis avec les résultats de CroMaSt. Ceci permet de visualiser les structures et de comprendre le cas échéant pourquoi certaines StI, classées comme domaines à RRM sont trouvées en échec. Ces alignements permettent aussi de mieux se rendre compte de la diversité des domaines à RRMs vrais et pseudo-vrais, ce qui peut se révéler intéressant lors du design ou de l'ingénierie de nouvelles protéines contenant ce type de domaine.

En conclusion, CroMaSt peut être appliqué à l'étude de n'importe quel type de domaine pour lequel on dispose d'au moins une familles représentatives dans Pfam et CATH. Les résultats de CroMaSt peuvent être utilisés en retour vers les bases de domaines pour leur signaler de possibles erreurs d'annotation de certaines familles de domaine, ou au contraire pour proposer l'attribution d'un type de domaine à une famille encore peu annotée.

IV. Modélisation des complexes ARN-RRM à partir des données existantes

Ce chapitre décrit plusieurs exploitations des données collectées et intégrées dans InteR3Mdb.

La première utilisation a été de contribuer à déchiffrer le code de correspondance entre RRM et ARN. Il s'agit ici d'un travail collaboratif réalisé lors d'un stage de 5 mois réalisé dans le goupe Bio2Byte du Professeur Wim Vranken à Bruxelles (Vrije Universiteit Brussel), aux côtés d'un autre doctorant du projet RNAct. La contribution de cette thèse à ce travail a été de pouvoir disposer d'un jeu exhaustif de structures 3D de domaines à RRM, nettoyées et bien délimitées, ainsi que de la liste exhaustive des contacts entre aa et nt pour les complexes RRM-ARN connus.

Les contacts aa-nt ont été positionnés sur un alignement structural complet de toutes les structures 3D des RRM. Ceci a permis par la suite à l'équipe Bio2Bytes de former des clusters en fonction de la localisation des contacts et de caractériser ainsi le mode de liaison majoritaire ou canonique des ARN aux domaines à RRM. Pour ce mode de liaison, les contacts ont été recensés par paire (aa, nt) et des probabilités ont été calculées. De la sorte, pour toute séquence de domaine à RRM, il est possible de calculer un score prédictif pour les ARN pouvant se lier sur ce domaine, et de sélectionner le meilleur ARN ciblé par ce domaine. La fiabilité de la prédiction dépend bien sûr du nombre de contacts disponibles pour le mode de liaison majoritaire. RRMscorer pourra être actualisé chaque fois qu'InteR3Mdb sera mis à jour avec de nouvelles structures 3D de complexes RRM-ARN, donnant lieu à l'extraction de nouveaux contacts.

Une deuxième utilisation d'InteR3Mdb a été la modélisation par homologie de nouveaux domaines à RRM. Cette application a été débutée avant la mise à disposition du logiciel AlphaFold2. Il fallait alors rechercher dans InteR3Mdb le domaine à RRM le plus proche en séquence et pour lequel une structure 3D était disponible. Puis le nouveau domaine était aligné sur la structure 3D modèle (« template ») et modélisé en 3D à l'aide du logiciel Modeller. Les chaînes latérales étaient réparées par le logiciel SCWRL4 et le nouveau modèle 3D évalué par un score DOPE. Les résultats obtenus étaient encourageants et révélaient la nécessité de modéliser les boucles qui pouvaient être très différentes en taille entre le domaine à modéliser et le domaine template. Le workflow nommé RRMpip n'a plus beaucoup d'intérêt pour les domaines à RRM libres puisque le logiciel AlphaFold2 permet maintenant de prédire avec une grande fiabilité la structure 3D des protéines à partir de leur séquence. Toutefois, AlphaFold2 ne permet pas de prédire la structure des complexes protéine-ARN. L'approche par homologie pourrait conserver un intérêt pour modéliser les complexes RRM-ARN.

En troisième lieu, les contacts collectés dans InteR3Mdb ont été exploités dans une approche de docking de l'ARN sur les domaines à RRM avec « points d'ancrage ». Les approches par fragments pour le docking protéine-ARN donnent lieu à une combinatoire gigantesque de solutions possibles mais le problème peut être rendu moins complexe en utilisant des contraintes pour guider le docking et l'assemblage des fragments. Ces contraintes peuvent consister en des points d'ancrage, à savoir des configurations déjà observées de paires aa-nt, notamment lorsque l'interaction est de type π-stacking. Les contacts d'InteR3Mdb ont donc été filtrés pour identifier ceux qui concernent les aa aromatiques des deux signatures conservées RNP1 (position 5) et RNP2 (position 2) des domaines à RRM. Un total de 496 structures de complexes RRM-ARN présents dans InteR3Mdb a été sélectionné, parmi lesquels 257 structures présentent une interaction de π-stacking à l'une ou l'autre position. Ces 257 structures proviennent de 52 protéines uniques et 72 domaines à RRM uniques. Les structures élémentaires du nt en interaction de π-stacking avec l'aa ont été extraites après alignement sur une structure de référence (1B7F_A) selon deux groupe beta1 (pour RNP2) et beta3 (pour NRP1). Puis les nucléotides ont été transformés en représentation gros-grain selon ATTRACT, et répartis en cluster selon un algorithme agglomératif ascendant (méthode Radius) développé dans l'équipe CAPSID. Les prototypes de chaque cluster peuvent alors être utilisés comme des contraintes pour positionner les fragments d'ARN au cours du docking et de l'assemblage des fragments. Cette dernière partie du travail a été effectuée en collaboration avec une autre doctorante du projet RNAct.

A la fin de ce chapitre sont présentés les travaux réalisés à partir de simulation en dynamique moléculaire dans le but (i) de pouvoir différencier une interaction stable d'une interaction labile et (ii) de calculer l'énergie libre de liaison entre un domaine à RRM et un ARN particulier.

La trajectoire de dynamique moléculaire d'un complexe RRM-ARN natif a été comparée à celle du même complexe dans lequel l'ARN a été modifié pour ne plus pouvoir s'associer au domaine à RRM. Les frames ont été clusterisées pour identifier les différents états visités au cours de la dynamique moléculaire et une analyse de contacts a été effectuée. Malheureusement, dans les conditions de simulation utilisées, aucune différence n'apparaît dans les résultats d'analyse, comme si la dynamique moléculaire n'arrivait pas, dans ces conditions classiques, à distinguer les ligands forts des ligands faibles. Les mêmes résultats décevants ont été obtenus lorsque la dynamique moléculaire a été réalisée dans des conditions de recuit simulé (« simulated annealing »).

Les calculs d'énergie libre ont été réalisés à partir des trajectoires de dynamique moléculaire en appliquant la technique FEP (« Free Energy Perturbation »). Ces calculs devraient nous aider à vérifier l'effet de mutations ponctuelles du domaine à RRM sur son activité de liaison de l'ARN. Grâce aux données expérimentales collectées dans InteR3Mdb, nous avons pu trouver rapidement un complexe RRM-ARN pour lequel des données de mutagenèse associées à des mesures d'affinité de liaison sont disponibles. Il s'agit du domaine RRM du facteur SRSF2 humain (« Serine/Arginine-rich splicing factor » ; UniProt :Q01130). Nous avons ainsi comparé la diminution de l'affinité (augmentation de la constante de dissociation K_D) rapportée dans la publication avec la différence d'énergie libre calculée à partir de nos trajectoires, quand on passe de la structure native aux formes mutées. Les résultats obtenus montrent que la diminution de l'affinité chez les mutants est toujours associée à un calcul d'énergie libre défavorable pour le domaine à RRM mutant par rapport au domaine à RRM natif, sauf pour une mutation (D48A) située dans une boucle entre les brins β2 et β3. Plus de temps serait nécessaire pour optimiser le calcul d'énergie libre par FEP et mieux comprendre les corrélations ou les divergences par rapport aux mesures d'affinité. Cependant ces résultats sont encourageants et pourront servir de base à des études ultérieures.

Conclusion et perspectives

Cette thèse a permis d'explorer la diversité des domaines à RRM en vue de mieux comprendre leurs caractéristiques de liaison à l'ARN. Une base de données intégrées et exhaustive a été développée et rendue accessible à la communauté scientifique. Un workflow permettant de vérifier l'appartenance d'un domaine ou d'une famille de domaines à un type de domaine d'intérêt a été implémenté et testé. Il est également disponible publiquement. La collecte d'informations de contact entre RRM et ARN a permis de contribuer à un travail de déchiffrage du code de reconnaissance entre RRM et ARN. Par ailleurs, les informations disponibles dans la base de données InteR3M ont aussi contribué au développement d'un workflow de docking RRM-ARN avec points d'ancrage. De plus, cette thèse rapporte la description de plusieurs protocoles de dynamique moléculaire réalisés dans le but d'évaluer la qualité des modèles RRM-ARN.

Ce travail ouvre de nombreuses perspectives, que ce soit par rapport à l'étude des domaines RRMs ou de façon plus générique. Pour les domaines RRMs, la base de données InteR3M constitue une ressource de qualité qu'il faudra maintenir et enrichir, en particulier avec les structures de RRM prédites par AlphaFold2. Cependant, il faut souligner que cet enrichissement ne concernera pas les contacts RRM-ARN car Alphafold2 ne prédit pas les complexes entre protéine et ARN. Pour cela, une veille devra être réalisée sur les structures expérimentales de la PDB. Au fur et à mesure de l'enrichissement d'InteR3Mdb, les outils de prédiction des interactions RRM-ARN comme RRMscore et RRM-RNA-dock devront être mis à jour. Il faudra aussi améliorer les protocoles de calcul d'énergie libre pour trouver un moyen de valider les modèles 3D. L'utilisation de la structure 3D de référence pour les RRM permettra de découvrir de nouvelles protéines et peut-être de nouvelles fonctions pour les RRM en interrogeant avec cette structure la base de données des structure AlphaFold (AlphaFoldDB). Une analyse évolutive des domaines RRMs pourra être envisagée, en collaboration avec les créateurs de la base RRMdb. Enfin, il sera important d'analyser l'impact de cette base

de données et du travail de validation des domaines à RRM sur les expériences de design de nouvelles protéines de liaison à l'ARN.

D'un point de vue plus générique, les contributions bioinformatiques de cette thèse peuvent être appliquées à d'autres domaines que les domaines à RRM et à d'autres bases de domaines que CATH et Pfam. En particulier, la base de domaines ECOD pourra être utilisée à la place de CATH ou de Pfam et permettre d'accéder aux relations évolutives entre les domaines, tout en restant dans un type de domaine d'intérêt. Par ailleurs il sera intéressant d'adapter CroMaSt à l'environnement d'une base de données intégrée comme InterPro, car cela garantira d'avoir la même version de la base des séquences protéiques UniProt, quelle que soit la base de domaines considérée. Enfin, l'outil CroMaSt et la méthodologie utilisée pour créer InteR3Mdb pourront être réutilisés pour n'importe quel autre type de domaine d'intérêt et accélérer la création d'une ressource particulière domaine-centrique, en vue de conduire des expériences d'apprentissage automatique ou de design de nouvelles protéines.

Structural characterization of RNA binding to RNA recognition motif (RRM) domains using data integration, 3D modeling and molecular dynamic simulation

Abstract

This thesis was carried out in the frame of a larger European project (ITN RNAct) in which computer science and biology approaches were combined to make progress towards the synthesis of new protein domains able to bind to specific RNA sequences. The specific goal of this thesis was to design and develop computational tools to better exploit existing knowledge on RNA Recognition Motif (RRM) domains using 3D modeling of RRM-RNA complexes. RRMs account for 50% of all RNA binding proteins and are present in about 2% of the protein-coding regions of the human genome. However, due to the large diversity of RRMs, there have been very few successful examples of new RRM design so far.

A central achievement of this thesis is the construction of a relational database called 'InteR3M' that integrates sequence, structural and functional information about RRM domains. InteR3M database (https://inter3mdb.loria.fr/) contains 400,892 RRM domain instances (derived from UniProt entries) and 1,456 experimentally solved 3D structure (derived from PDB entries) corresponding to only 303 distinct RRM instances. In addition, InteR3M stores 459,859 atom-atom interactions between RRM and nucleic acids, retrieved from 656 3D structures in which the RRM domain is complexed with RNA or DNA.

During the data collection procedure, inconsistencies were detected in the classification of several RRM instances in the popular domain databases CATH and Pfam. This led me to propose an original approach (CroMaSt) to solve this issue, based on cross-mapping of structural instances of RRMs between these two domain databases and on the structural alignment of unmapped instances with an RRM structural prototype. The CroMaSt CWL workflow is available on the European Workflow hub at https://workflowhub.eu/workflows/390.

Sequence and structural information stored in InteR3M database was then used to align RRM domains and map all RRM-RNA interactions onto this alignment to identify the different binding modes of RNA to RRM domains. This led to the development, with RNAct partners at VUB (Vrije Universiteit Brussel), of the 'RRMScorer' tool. This tool contributes to decipher the RRM-RNA code by computing binding probabilities between RNA nucleotides and RRM amino acids at certain positions of the alignment. Atomic contacts between RRMs and RNA were also used to identify anchoring patterns, i.e. prototypes of 3D atomic positions (relative to the protein backbone) of a nucleotide stacked on a conserved aromatic amino acid. These anchors can be used as constraints in anchored docking protocols. The 'RRM-RNA dock' docking pipeline is presented here and integrates both anchoring patterns extracted from InteR3M and binding scores from RRMScorer.

Finally, molecular dynamic (MD) simulation is another computational tool tested in this thesis to contribute to the 3D modeling of RRM-RNA complexes. Promising preliminary MD protocols are described as attempts to distinguish between strongly and weakly binding RRM-RNA complexes.

Keywords: structural bioinformatics, protein domain, RNA Recognition Motif, protein design, 3D modeling, protein-RNA docking, database, data integration, bioinformatic workflow

Caractérisation structurale de la liaison de l'ARN aux domaines à Motif de Reconnaissance de l'ARN (RRM) à l'aide de l'intégration de données, la modélisation 3D et la simulation dynamique moléculaire

Résumé

Cette thèse a été réalisée dans le cadre d'un projet Européen plus vaste (ITN RNAct) dans lequel des approches informatiques et biologiques étaient combinées pour progresser vers la synthèse de nouveaux domaines protéiques capables de se fixer sur des séquences spécifiques d'ARN. L'objectif spécifique de cette thèse était de concevoir et développer des outils informatiques pour mieux exploiter les connaissances existantes sur les domaines à Motif de Reconnaissance de l'ARN (RRM) lors de la modélisation 3D des complexes RRM-ARN. Les domaines RRMs représentent 50% de toutes les protéines fixant l'ARN et sont trouvées dans environ 2% de toutes les régions codantes du génome humain. Cependant, du fait de la grande diversité des domaines RRMs, il n'y a eu jusqu'à présent que très peu de succès rapportés dans la conception de nouveaux domaines RRMs.

La contribution centrale de cette thèse est la construction d'une base de données relationnelle appelée (InteR3M) qui intègre des informations de séquence, de structure et de fonction sur les domaines RRMs. La base de données InteR3M (https://inter3mdb.loria.fr/) contient 400,892 instances de domaines RRM (dérivées d'entrées UniProt) et 1,456 structures 3D déterminées expérimentalement (dérivées d'entrées PDB), qui correspondent à seulement 303 instances distinctes de domaines RRM. De plus, InteR3M contient 459,859 interactions atomiques entre RRM et acides nucléiques, dérivées de 656 structures 3D dans lesquelles le domaine RRM forme un complexe avec un ARN ou un ADN. Au cours du processus de collecte de données, des incohérences ont été détectées dans la classification de plusieurs instances de domaines RRMs dans les bases de données de domaines protéiques populaires CATH et Pfam. Ceci m'a conduit à proposer une approche originale (CroMaSt) pour résoudre ce problème, à partir de la mise en correspondance des instances structurales de domaines RRMs entre ces deux bases de données et de l'alignement structural des domaines sans correspondance avec une structure prototype du domaine RRM. Le workflow CroMast est disponible sur le Workflow Hub Européen (https://workflowhub.eu/workflows/390).

Les informations de séquence et de structure intégrées dans la base de données InteR3M ont ensuite été utilisées pour aligner entre eux tous les domaines RRM et cartographier toutes les interactions RRM-ARN sur cet alignement en vue d'identifier les différents modes de liaison de l'ARN aux domaines RRM. Ceci a conduit au développement, avec nos partenaires RNAct de VUB (Vrije Universiteit Brussel), de l'outil 'RRMScorer'. Cet outil contribue au déchiffrage du code de reconnaissance RRM-ARN en calculant les probabilités de liaison entre les nucléotides de l'ARN et les acides aminés des domaines RRM à certaines positions de l'alignement. Les contacts atomiques entre RRMs et ARN ont aussi été utilisés pour identifier des motifs d'ancrage, c'est-à-dire des prototypes des positions 3D atomiques (relatives au squelette protéique) d'un nucléotide interagissant par empilement ('stacking') avec un acide aminé aromatique conservé. Ces ancres peuvent être utilisées comme des contraintes dans un protocole d'amarrage ancré ('anchored docking'). Le pipeline 'RRM-RNA dock' est présenté ici et il intègre à la fois les motifs d'ancrage extraits de la base de données InteR3M et les scores de liaison de RRMScorer.

Finalement, la simulation en dynamique moléculaire (MD) est un autre outil informatique testé dans cette thèse pour contribuer à la modélisation 3D des complexes RRM-ARN. Des protocoles MD préliminaires mais prometteurs sont décrits au titre d'essais visant à distinguer entre les complexes RRM-ARN à liaison forte ou faible.

Mots-clés: Bioinformatique structurale, domaine protéique, motif de reconnaissance de l'ARN (RRM), conception de protéines, modélisation 3D, amarrage protéine-ARN, base de données, integration de données, workflow bioinformatique