Multi-dimensional probing for RNA secondary structure(s) prediction

Afaf Saaidi

université PARIS-SACLAY

NNT : 2018SACLX067

Thèse de doctorat

Multi-dimensional probing to predict the RNA secondary structure

Thèse de doctorat de l'Université Paris-Saclay
préparée à École Polytechnique

Ecole doctorale n°573 Interfaces (Approches interdisciplinaires / fondements, applications et innovation)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Palaiseau, le 01 Octobre 2018, par

AFAF SAAIDI

Composition du Jury :

M. Bruno Sargueil Directeur de Recherche, CNRS-Université Paris Descartes	Président
M. Mathieu Giraud Chargé de Recherche, CNRS-Université de Lille	Rapporteur
M. Alain Laederach Professeur, Université de Caroline du Nord à Chapel Hill	Rapporteur
M. Fabrice Leclerc Chargé de Recherche, I2BC paris Saclay	Examinateur
M. Pierre Peterlongo Chargé de Recherche, Inria-Irisa Rennes	Examinateur
M. Ronny Lorenz Chercheur, Institut de Chimie Théorique de l'Université de Vienne	Examinateur
Mme. Mireille Régnier Directrice de recherche, LIX-CNRS	Directeur de thèse
M. Yann Ponty Chargé de Recherche, CNRS-LIX	Co-directeur de thèse

Multi-dimensional probing for RNA secondary structure(s) prediction

Afaf Saaidi

LIX – Laboratoire d'Informatique
Ecole Polytechnique
Palaiseau, France

Supervisors:

Yann Ponty, Chargé de Recherche

Mireille Régnier, Directrice de Recherche

LIX, CNRS

To my darling Mohamed Amine

Preface

In structural bioinformatics, predicting the secondary structure(s) of ribonucleic acids (RNAs) represents a major direction of research to understand cellular mechanisms. A classic approach for structure postulates that, at the thermodynamic equilibrium, RNA adopts its various conformations according to a Boltzmann distribution based on its free energy. Modern approaches, therefore, favor the consideration of the dominant conformations. Such approaches are limited in accuracy due to the imprecision of the energy model and the structure topology restrictions.

Experimental data can be used to circumvent the shortcomings of predictive computational methods. RNA probing encompasses a wide array of experimental protocols dedicated to revealing partial structural information through exposure to a chemical or enzymatic reagent, whose effect depends on, and thus reveals, features of its adopted structure(s). Accordingly, single-reagent probing data is used to supplement free-energy models within computational methods, leading to significant gains in prediction accuracy. In practice, however, structural biologists integrate probing data produced in various experimental conditions, using different reagents or over a collection of mutated sequences, to model RNA structure(s). This integrative approach remains manual, time-consuming and arguably subjective in its modeling principles. In this Ph.D., we contributed in silico methods for an automated modeling of RNA structure(s) from multiple sources of probing data.

We have first established automated pipelines for the acquisition of reactivity profiles from primary data produced through a variety of protocols (`SHAPE`, DMS using Capillary Electrophoresis, `SHAPEMap`/Ion Torrent). We have designed and implemented a new, versatile, method that simultaneously integrates multiple probing profiles. Based on a combination of Boltzmann sampling and structural clustering, it produces alternative stable conformations jointly supported by a set of probing experiments. As it favors recurrent structures, our method allows exploiting the complementarity of several probing assays. The quality of predictions produced using our method compared favorably against state-of-the-art computational predictive methods on single-probing assays.

Our method was used to identify models for structured regions in RNA viruses. In collaboration with experimental partners, we suggested a refined structure of the HIV-1 Gag IRES, showing a good compatibility with chemical and enzymatic probing data. The predicted structure allowed us to build hypotheses on binding sites that are functionally relevant to the translation. We also proposed conserved structures in Ebola Untranslated regions, showing a high consistency with both SHAPE probing and evolutionary data. Our modeling allows us to detect conserved and stable stem-loop at the 5'end of each UTR, a typical structure found in viral genomes to protect the RNA from being degraded by nucleases.

Our method was extended to the analysis of sequence variants. We analyzed a collection of DMS probed mutants, produced by the Mutate-and-Map protocol, leading to better structural models for the GIR1 Lariat-capping ribozyme than from the sole wild-type sequence. To avoid systematic production of point-wise mutants, and exploit the recent SHAPEMap protocol, we designed an experimental protocol based on undirected mutagenesis and sequencing, where several mutated RNAs are produced and simultaneously probed. Produced reads must then be re-assigned to mutants to establish their reactivity profiles used later for structure modeling. The assignment problem was modeled as a likelihood maximization joint inference of mutational profiles and assignments, and solved using an instance of the "Expectation-Maximization" algorithm. Preliminary results on a reduced/simulated sample of reads showed a remarkable decrease of the reads assignment errors compared to a classic algorithm.

Perspectives of this work include the optimization of the read assignment algorithm, and the development of structure prediction algorithm dedicated to multiple SHAPEMap probed mutants.

Résumé Substantiel

L'ADN est le réservoir génétique de tout espèce vivant sur la terre. Une meilleure compréhension du mécanisme de transcription et de la traduction entraînerait une visibilité sur les éléments dont la présence est nécessaire pour assurer les fonctions requises, ainsi une comparaison entre les espèces et une meilleure compréhension des différentes fonctions biologiques.

Les ARNs sont des biopolymères connus par leurs capacités à coder pour des protéines. Réputés pour leurs rôle d'intermédiaires pour la transmission du code génétique, ces molécules se sont montrés beaucoup plus actifs sur le plan transcription/traduction assurant des fonctions catalytiques et régulatrices chez certaines espèces notamment les viruses. L'abondance des ARNs séquencées et la révélation de certaines des fonctions ont permi une classification d'ARN en famille fonctionnelle, favorisant ainsi la construction de plusieurs base de données tel `RFAM`.

La fonction d'un ARN dépend de sa structure représentée par des interactions d'hydrogène entre les nucléotides le constituant et fréquemment par des sites d'interaction avec d'autres molécules. L'identification de la structure demeure une étape primordial pour comprendre le mode de fonctionnement des molécules d'ARNs. l'analyse aux `rayons X` et l'analyse par résonance magnétique nucléaire `RMN` sont les deux techniques expérimentales les plus répandus et qui ont permi jusqu'à nos jours une détermination d'un large nombre de structures d'ARNs. Toutefois, la complexité de la structure et le temps requis pour réaliser ces expériences rendent ces méthodes faillible devant certain ARNs. De plus, la capacité d'une séquence d'ARN à assurer plusieurs fonctions à la fois, en alternant de conformation, a ouvert la porte pour des nouvelles techniques prometteuses qui permettent une capture de cette diversité structurale à moindre coût.

SHAPE est l'une des techniques expérimentales à la pointe des approches expérimentales qui tirent bénéfice des avancées du séquençage haut-débit qui à son rôle, a connu un essor depuis le début du 21e siècle. SHAPE permet de caractériser la structure à travers un profil de réactivité. Des réactivités résiduelles élevées traduisent une accessibilité du nucléotide, permettant ainsi de renseigner sur le contexte structural. Le profile de réactivités est un signal moyen qui reflète

la diversité structurale pour une séquence d'ARN, la déconvolution de ce signal pour pouvoir en extraire de l'information est donc nécessaire.

En parallèle de la complexité de la déconvolution du signal, d'autres problématiques reliées aux séquençages hauts débits et à l'optimisation des protocoles expérimentales ont vu le jour. L'un des grands challenges en Bioinformatique, qui fait notamment l'objectif de cette thèse, est le traitement des données issues d'un protocole expérimentale couplé à un séquençage de l'ARN, en assurant à la fois un traitement automatique et optimisé de ces données, dites de sondage et une meilleure interprétation de ces données à travers des approches prédictives.

Les données de sondage servent d'éléments de support à intégrer dans l'algorithme de prédiction pour pouvoir améliorer la précision des structures prédites. La variété d'approches expérimentales cache une complémentarité structurale entre divers sources de sondage. De ce fait, l'intégration de différent données de probing dans le processus de prédiction constitue l'une des directions les plus sollicités pour améliorer les prédictions.

Sous l'hypothèse de la complémentarité entre différent données de sondage, nous avons développé une nouvelle méthode intégrative qui permet en effet de d'utiliser plusieurs sources de données de sondage, produites dans diverses conditions expérimentales, avec différent réactifs ou encore associées à un ensemble de variants d'ARN. Notre approche intégrative est basée sur un échantillonnage de l'espace des structures compatibles avec les données de sondage et sur un clustering des structures permettant ainsi de récupérer la(les) structure(s) dominante(s), stable(s) et à l'intersection des différent types de données de sondage considérées.

Notre méthode nous a permis de prédire la structure pour un ensemble d'ARNs avec une précision comparable sinon meilleure que celle obtenue à travers de méthodes de pointe. Nous avons ainsi suggéré des structures pour des régions fonctionnelles chez le VIH-1 Gag, qui a fait l'objet d'un article publié [Deforges et al., 2017], compatible avec des données de sondage chimiques/enzymatiques et confirmé par des données d'évolution.

Contents

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisor Dr. Yann Ponty for his continuous support during the Ph.D period, for his patience, passion, enthusiasm and sharing his immense knowledge. His perspicacity allowed me to be more efficient and get a clearer vision about the research subject at the earlier stage of this thesis. I could not have imagined having a better supervisor for my Ph.D study, he was the person with whom I interacted the most and without doubt a leader idol from whom I learned how to stay humble, to listen and to be passionate. Beyond being a supervisor, Yann is a close friend always listening and encouraging me to go through the challenges. Thanks to him I am now able, more than anytime before, to define my personal and professional objectives.

Besides my advisor, I would like to thank Dr. Mireille Regnier, my thesis director, for her support, her engagement towards this thesis' manuscrit and I really appreciate the time she spent to bring valuable corrections to this memory. Mireille is also a person with a great character, from whom I learned how to deal with complicated situations particularly when dealing with administrative issues.

I would like also to thank Dr.Bruno Sargueil, Nathalie Chamond for their insightful discussions and comments which incited me to widen my research from various perspectives.

I thank my fellow labmates for the stimulating discussions, Juraj for being such quiet and present in need, Christelle for her nice advices on how to succeed in daily life, Alice and Amelie for the different nice activities we have shared together. Also, a great thanks to Delphine Allouche for being a patient and lovely co-worker. A special thanks goes to Victor and Dimitri who made the working days more pleasant.

Last but not least, I would like to thank my family: my parents, especially my mother, thanks to whom I came to such a feat, my brother Oussama who let me feel an important element when it comes to take hard decisions in his personal and professional life. Without forgetting my little sister for her lovely cookies that she made every time I went home :) A particular thanks goes to my aunts Souad and

Jamila with whom I spent a pleasant time during the 3 years of my PhD, with all the nice trips we have made and adventures we went through. I could say that her presence have lightened the hard life a foreign student could have!

List of Figures

List of Tables

List of acronyms

Symbol	Definition
cDNA	complementary DNA
CMCT	1-Cyclohexyl-3-(2-(4-Morpholinyl)ethyl) Carbodiimide Tosylate
DMS	DiMethyl Sulfate
DNA	Deoxy-riboNucleic Acid
DP	Dynamic Programming
HIV	Human Immunodeficiency Virus
IPANEMAP	Integrative Probing Analysis of Nucleic Acids Empowered by Multiple Accessibility Profiles
MCC	Matthews Correlation Coefficient
MFE	Minimum Free Energy
MEA	Maximum Expected Accuracy
M&M	Mutate and Map
mRNA	messenger RNA
PCR	Polymerase Chain Reaction
ncRNA	non-coding RNA
RNA	RiboNucleic Acid
RT	Reverse Transcription
rRNA	ribosomal RNA
SHAPE	Selective 2'-Hydroxyl Acylation analyzed by Primer Extension
sRNA	small RNA
tRNA	transfer RNA
UTR	UnTranslated Region

Part I

The purpose of this first part is to introduce the necessary notions and concepts evoked in this thesis. and to highlight the interest of the developed approaches that are the subject of the Part II. This first thesis part is divided into three chapters:

In Chapter 1, we start by giving an overview of the computational approaches to predict the structure of the RNA in the absence/presence of probing data. Then, we discuss some limitations and issues related to the use of probing data for the purpose of the prediction of the structure, while remaining focused on the levels on which we were able to intervene during this Ph.D. At the end of this chapter, we announce the outline of this manuscript.

In Chapter 2, we provide a concise description of the widely used computational and experimental approaches to inform about the RNA structure.

In Chapter 3, we discuss the interpretation of profiling data. Then, we present some of the probing data integrative modeling approaches. By the end of this chapter, we present some of the computational approaches that use probing data alongside evolutionary data to improve the accuracy of predicted RNA structures.

Chapter 1

Thesis Introduction

"Science is not only a disciple of reason but, also, one of romance and passion."

Stephen Hawking

Historically, the focus of RNA research pertains to its role as a messenger, a medium of genetic information. However, RNA has been shown to play multiple other roles, including the regulation of gene expression [Moore and Steitz, 2011]. The discovery of these different functionalities allowed for a wider categorization of functional roles played by RNA: messenger RNA (mRNAs), transfer RNA (tRNAs) and ribosomal RNA (rRNAs). Riboswitches also deserve a particular mention due to their capacity to undergo conformational change upon binding small metabolites, leading to the regulation of a set of mechanisms such as transcription, translation, and splicing [Serganov and Nudler, 2013]. A major property of an RNA is to fold into a highly complex structure that tend to be preserved throughout evolution in order to conserve its function [Nowakowski and Tinoco, 1997]. This stable structure is made of pairs of nucleotides where the interaction between two paired nucleotides is mediated by hydrogen bonds.

Recent years have witnessed an explosion of structure probing data, produced using a variety of competing technologies and protocols. Consequently, a large amount of computational approaches have been developed in order to use this data as a support point for predicting the structure of the RNA. Yet, processing probing data still requires a preliminary formalization allowing to address questions such as: given an RNA sequence, to which extent are probing data able to explain its structural properties? How should those data be processed to extract the informative part of the signal? most importantly, how to effectively integrate the

resulting processed data within structure prediction frameworks? The need for such a formalization propelled the availability of a set of models and methods. Contrasting with the comparative approaches based on structural or sequential similarity, those models could be qualified as generic *i.e.* those methods could be applied to any type of RNA regardless of its sequence and its global structural context. Thus, they allow for a more direct revelation of the information carried by the probing data at the local structural level. At first, certain structure prediction models were developed to account for probing data as *hard* constraint assuming a correspondence between probing data and the structure. Then, many suggested approaches were derived to use this data as *soft* constraints to bias the structural ensemble towards a subset deemed compatible with probing data.

Probing data constitute a stochastic signal. The reactivity of a given nucleotide is a quantification of its accessibility to the chemical reagent. This accessibility is constrained by the 3D structure of an RNA, and its interaction(s) with other RNA(s) or protein(s), potentially leading to structural changes occurring upon binding by small ligand or proteins. In addition, the ability of an RNA to adopt multiple conformations induces convoluted reactivity profiles, contributing to intrinsic difficulties in their exploitation.

The aim of the present thesis is to explore multi-dimensional approaches to alleviate this stochasticity and correctly interpret it in the context of the RNA structure. Mirroring the practices of experimentalists, we used multiple experimental data, probing data produced in different experimental conditions and with different reagents, to enhance structure predictions and to diminish profile noises. We also developed a new protocol based on the use of probing data from a set of RNA variants under `SHAPEMap` protocol, that induces mutations primarily on single stranded regions, named differential-`SHAPE` in this work.

1.1 Generalities on RNA structure and probing

Classic methods to observe the structure of RNA at high resolution include X-ray crystallographic analysis [Golden, 2007], and Nuclear Magnetic Resonance (NMR), which have shown to be useful to reveal the tertiary structure of viral RNAs and riboswitches [Houck-Loomis et al., 2011]. Despite the effectiveness of these experimental approaches, many RNA structures are still not resolved yet, due to the prohibitive cost of experimental methods, along with their limited lifespan and stability. Consequently, wet-lab methods are complemented by the development of computational approaches and recently by the design of dedicated biochemical protocols.

Computational approaches In silico, the secondary structure can be computationally predicted at the thermodynamic equilibrium, using an energy model called Turner model [SantaLucia and Turner, 1997] that allows to assign any given structure a numerical value called its free-energy. The global free energy value for a given secondary structure is typically calculated as the sum of the partial free energies of its small recognizable structural domains that include hairpin, loops, bulges, and internal loops. When the RNA reaches the thermodynamic equilibrium, the thermodynamic potential induces a Boltzmann distribution based on the free-energy, where the most probable conformation is the one of lowest free-energy. Thus, RNA in silico structure prediction aims to report the Minimum Free Energy (**MFE**) structure. The prediction of the MFE structure can be performed using a variety of available dynamic programming algorithms [Zuker and Stiegler, 1981]. The most prominent advantage of this approach lies in its ability to accurately predict structures for RNA sequences of length below $700nts$ with a sensitivity of about 73% [Mathews, 2004] in a matter of seconds on a personal computer.

In vivo, an RNA may adopt alternative functional conformations. However, a major drawback of MFE-based modeling, that predicts the most stable structure at the Boltzmann equilibrium, resides is its inability to capture the structural diversity that may be required for the function of some RNAs. Conserved alternative structures are featured within RNAs associated with switching behaviors, and are increasingly considered by kinetics studies, as transient structures adopted by nascent transcripts can be crucial to channel the folding towards the correct energy basin.

On its way to the thermodynamic stability, RNA undergoes a dynamic process through which it may alternate conformations with different frequencies. The frequency of a structure in the ensemble is quantified as a Boltzmann probability. Consequently, structures with high probability are likely to be reflective of stable alternative conformations. This sampling is ensured through a dynamic programming algorithm [McCaskill, 1990] where a partition function is calculated to assign a Boltzmann probability to each possible base pairs and consequently to each suboptimal formed structure. RNA structure sampling could be ordered by decreasing free energy from the MFE [Wuchty et al., 1999], or stochastically generated [Ding and Lawrence, 2003]. These two methods allow to cover a large set of stable structures. The sampling algorithm allowed to reveal alternative biological structures. In addition, it allowed to reveal probability profiling of single-stranded regions in RNA secondary structure which could be used as a basis to predict RNA-RNA interaction.

RNA secondary structure is vital to ensure many biological process such as gene regulation, protein synthesis and RNA-RNA interactions. The conservation

of such vital functions requires preserving the ability of the RNA to adopt a predefined secondary structure across evolution. To enhance computational predictions, one of the imminent explored improvement direction is the development of comparative approaches. Indeed, for many RNA families, a common secondary structure is highly conserved throughout evolution [Hofacker et al., 2002] and can be used as basis to model the 3D structure. Comparative approaches aim to compute the consensus structure from a set of aligned RNA sequences while considering both thermodynamic stability and sequence covariation. This approach allowed to resolve the structure of `5s rRNAs` where the achieved accuracy was over 80%. In addition, this comparative analysis was used to assess non-canonical base pair conformations [Gautheret and Gutell, 1997]. However, compared to the MFE based predictions, this approach mainly benefits from the observation of compensatory mutations, interpreted as a selective pressure towards the conservation of base pairs. It is therefore very sensitive to the alignment quality, size and the distance (in sequences and potentially in structures) separating the different RNA in the alignment.

Probing data Despite the progress of computational approaches and their capacities to ensure accurate predictions, a refinement and enhancement of the thermodynamic model is still required. Additional data may include structural constraints derived from experimental data, at nucleotide or sub-structural level. The need for such auxiliary data to enrich the prediction model opened the door for the development of a wide set of experimental techniques to inform about the local structural context. The process of using these techniques with a set of RNA replicates to generate characteristic structural patterns is known as **probing**. In vitro, the most popular experimental probing are `SHAPE` chemical probing through sequencing [Merino et al., 2005], and enzymatic probing through parallel analysis of RNA structure PARS [Kertesz et al., 2010]. In all experimental protocols, RNA molecules are treated with a structure-specific reagent, either chemical or enzymatic, targeting specific nucleotides, where the accessibility is a function of both the pairing status of the nucleotide, and the global geometry of the RNA backbone. The molecule-reagent interaction results into either the formation of an adduct, in the case of chemical protocol, or the cleavage of the RNA in the case of enzymatic experimental protocol. The resulting signatures, associated with each residue, allow to generate a global reactivity profile throughout the RNA. This led to questioning to what extent are signatures from different protocols compatible with each other?

To guide predictions with probing data, the most popular method consists into converting the local reactivities into thermodynamic local potentials that reflect the local accessibility. Different methods to incorporate probing data as pseudo-

potential were proposed [Deigan et al., 2009], [Washietl et al., 2012] and [Zarringhalam et al., 2012], explained in detail in Chapter 2. Specifically, the integration of `SHAPE` data as pseudo-potentials has proven its efficiency to propose unerring predicted secondary structure, and is routinely used as a basis to infer the tertiary structure [Cruz et al., 2012].

1.2 Towards increasing the accuracy of predicted RNA structures with the use of probing data

Probing data (chemical or enzymatic) present a non negligible source of structural information. However, the inference of the structure from such data is rather delicate and sensitive to the experimental noise and the computational calibration. In the case of enzymatic probing, data-guided predictions only consider structures verifying reactivity constraints. This can lead to wrong predictions in the case of missed experiment. Moreover, reactivity profiles represent an averaged signal, and may sometimes be impacted by the existence of more than one single conformation. Finally, the exponential increase of probing data due to the use of High-throughput sequencing, prompted the development of new approaches. While allowing for a better interpretation of the reactivity profiles and a boost in the predicted structures accuracies, those experimental protocols reveal different structural features, some targeting unpaired positions while some other inform about nucleotides involved in a double strand. For those reasons, there is a need to develop an integrative multi-probing data modeling.

A first step towards such an integrative method requires the automation of probing data processing. Such analyses represent a recurrent area of interest where the most pressing question is: how to design probing data analysis pipelines that provide a faithful picture of the observed phenomenon? Multiple processing steps are unavoidably required to obtain reactivities further used to guide the structure predictions. Generally, the processing of probing data starts with the collection of raw signals, the response to a chemical/enzymatic reaction. These raw signals might be subject to accumulated noise due, among other reasons, to the experimental setup, the sequencing errors and the profile recovery method. A small change in the reactivity profile would have a direct consequence on the predicted structural ensemble. This sensitivity makes the processing of the probing data one of the interesting point addressed in this thesis. The processing of probing data produced through `HTS` can be decomposed into three steps: Firstly, the mapping of sequenced reads, or transcription stops, onto the RNA of reference; Secondly, the calculation of reactivities that quantify the response at nucleotide level to a

specific experimental reaction in function of the structural context (Paired/Unpaired). The last step, that remains the most difficult to establish, concerns the conversion of the reactivity values into pseudo-energy contributions to drive the structure prediction.

The points (including both faced issues and contributions) addressed in this thesis can be summarized as:

1. An automation of `NGS` probing data processing: from mapping of reads to the construction of reactivity profiles;
2. A new mapping algorithm based on the use of mutational profiles in the case of a simultaneous sequencing of RNA mutants `SHAPEMap` modified;
3. A new integrative approach that both exploits the coherence aspect between different probing data sources, and also takes into account the multiple conformations adopted by RNA(s);
4. An extension of the developed integrative approach to study the agreement between reactivity profiles from RNA mutants under the assumption of the conservation of the functional structure.

After this short introduction, we introduce in Chapter 2 by introducing some of the characteristics of the RNA with a focus on the structure and the pre-existing computational approaches to model this structure. We also provide a brief summary of the experimental protocols that inform about the RNA structure. We also present the corresponding tools to predict the RNA structure while integrating this informative data into the prediction model. In Chapter 3, we conclude the first part by presenting some state-of-the art approaches that use probing data to infer the RNA structure.

The second part of this thesis is dedicated to the description of different workflows and methods developed over the course of this PhD. We start in Chapter 4 by presenting different automated pipelines for the acquisition of reactivity profiles from primary data produced using a variety of protocols (`SHAPE`, DMS using Capillary Electrophoresis, `SHAPEMap`/Ion Torrent).

In Chapter 5, we describe a versatile method for structure prediction, that simultaneously integrates multiple probing profiles. Based on a combination of Boltzmann sampling and structural clustering, it produces alternative stable conformations jointly supported by a set of probing experiments. As it favors recurrent structures, our method exploits the agreement between several probing assays.

In Chapter 6, we present a novel `SHAPEMap`-based protocol based on performing undirected mutagenesis, where several mutated RNAs are produced using PCR

error-prone and simultaneously probed. The simultaneous inference of probing profiles requires an accurate mapping, while reads produced using this approach induce specific issues, and reveal considerably to be challenging to classic mapping algorithms. We modelled the assignment problem as a likelihood maximization joint inference of mutational profiles and assignments, and solved it using an instance of the "Expectation-Maximization" algorithm.

The third part of this thesis describes an extensive validation of the predictive capacities of our approaches to produce novel biological insight, and concludes with future extensions. Chapter 7 demonstrates the ability of `IPANEMAP` to predict near-native structures on simulated and real datasets, both using multiple probing sources or in a mono-probing setting.

In Chapter 8, we present an ensemble of applications of our integrative approach on real-world data, focusing on three applications:

- First, we describe an application of `IPANEMAP`, which was published in Deforges et al. [2017], to refine a model of the internal ribosome entry site (IRES) of the Gag region in HIV-1, using a combination of `SHAPE` and enzymatic probing;
- Then, we take advantage of a comprehensive set of probing data, available bot for the wild type and mutants, to model the GIR1 Lariat-capping ribozyme using our integrative method;
- Finally, we model the untranslated regions of the Ebola genome, using both `SHAPE` data and evolutionary information to infer candidate secondary structures for these unresolved yet RNAs models.

Chapter 9 concludes this thesis with a discussion of the current shortcomings of `SHAPE` computational protocols, and offers some directions for future research.

Chapter 2

Bioinformatics concepts and tools

2.1 RNA 2D Bioinformatics

2.1.1 RNA secondary structure

An RNA can be abstracted as a succession of building blocks called **nucleotides**. In vivo or in vitro, an RNA folds in a complex way leading to the adoption of a specific **tertiary structure** responsible for a specific activity of the RNA molecule. This tertiary structure is mainly mediated by hydrogen bonds, denoted by **base-pairs**, between **compatible** nucleotides: pairs involving Adenine (A) and Uracil (U), or Cytosine (C) and Guanine (G), are known as **Watson-Crick** pairing, while a pair of G and U is know as **Wobble** pairing. Within most computational approaches, an RNA molecule is characterized by a linear structure (the sequence), and one or several **secondary structure**(s). In the absence of a conventional definition, a secondary structure for a given RNA molecule is considered as a planar projection of the tertiary structure, subject to further restrictions described below.

Let S be a sequence of bases of length n with $S = b_1, b_2, ..., b_n$ where the i^{th} base is noted b_i with $b_i = A, U, C$ or G sequence of nucleotides. A secondary structure is defined as a list of base pairs (i, j), denoting the pairing of positions i and j, formed by complementary bases and verifying $i < j$. Positions that are not involved in any base-pairs are considered as being **unpaired**. For computational reasons [Lyngsø and Pedersen, 2000; Sheikh et al., 2012], existing computational approaches further restrict the secondary structure by enforcing the following constraints:

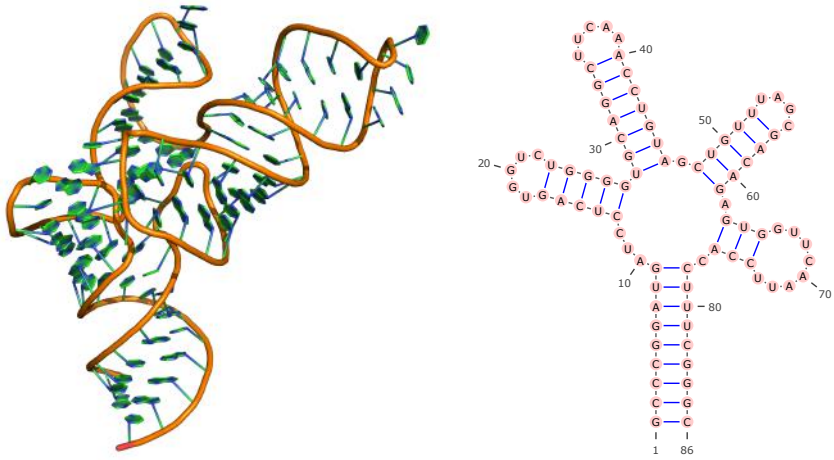
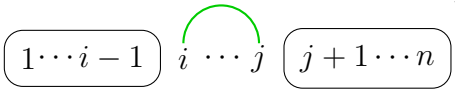


Figure 2.1: 3D model of an atypical selenocysteine-specific tRNA in the Mouse (PDB ID: 3RG5:A, left) and associated secondary structure (right).

Figure 2.2: Example of interactions that are not covered with the classic structure prediction approches.

- **Exclusivity condition**: A nucleotide can form base pairs with at most one base. Thus if $(i_1, j_1), (i_2, j_2)$ are two pairs, one has $i_1 \neq i_2$ and $j_1 \neq j_2$;
- **Non-crossing condition**: Structures that contains two base pairs (i_1, j_1) and (i_2, j_2) whith $i_1 < i_2 < j_1 < j_2$ as illustrated in Figure 2.2, are called **pseudo-knotted**, and are not considered by the classic prediction approaches.

RNA secondary structure is predicted without accounting for tertiary base pairs or for pseudo-knots; those restrictions are usually considered later on to model the tertiary structure. In the absence of crossing base pairs, each pair (i, j) subdivides the structure into two separate parts: In between the pair $(i+1, j-1)$ and the exterior region including $(1, i-1)$ and $(j+1, n)$. Therefore, the two substructures could be treated separately as follows:



This consideration was behind the development of a recursive decomposition scheme that formed the basis of all **Dynamic Programming** (DP) approaches to resolve the RNA secondary structure [Waterman and Smith, 1978]. DP is a powerful technique that allows to find the optimal solution for a given problem by combining sub-solutions for sub-problems. It can be expressed as a recursion or more expressively as overlapping sub-problems. The recursive decomposition scheme was first used to count the number of compatible structures for a certain sequence. Indeed, let $N_{i,j}$ is the number of possible structures in the sequence range $[i, j]$, one has

$$N_{i,j} = N_{i+1,j} + \sum_{\substack{k=i+1 \text{ s.t.} \\ b_i \text{ comp. with } b_k}}^{j} N_{i+1,k-1} N_{k+1,j}, \forall 1 \leq i < j \leq n$$

and $N_{i+1,i} = 1$. Starting from sequences of unit length, the algorithm proceeds iteratively over subsequences of increasing length until reaching $N_{1,n}$ the number of compatible structures with the whole RNA sequence. The complexities of the algorithm is in $\Theta(n^3)$ for time, and $\Theta(n^2)$ for space.

2.1.2 Computational methods for 2D structure prediction

The questions around which computational approaches were developed concern the prediction of one or several conformations from an RNA sequence, potentially supplemented with additional experimental data. For the sake of simplicity, we will illustrate the principles underlying the main prediction paradigms on a simple base pair based model akin to the one used in the work of Nussinov et al. [1978], using the unambiguous decomposition scheme of Waterman and Smith [1978] to allow for a computation of the partition function (and derived quantities).

Energy minimization. A first category of approaches considers the **minimal free-energy (MFE)** structure, the most stable conformation a given RNA sequence may adopt with respect to a given energy model. Indeed, Nussinov et al. [1978] developed the first algorithm dedicated to predict the MFE structure: a DP algorithm that returns an optimal structure as the one with the maximal number of base pairs, by a backtracking procedure. The DP scheme corresponds to:

$$M_{i,j} = \min \begin{cases} M_{i+1,j} \\ \min_k (E_{i,k} + M_{i+1,k-1} + M_{k+1,j}). \end{cases}$$

where

$$E_{i,j} = \begin{cases} -1, & if \quad \{i,j\} \in \{A,U\},\{C,G\},\{G,U\}. \\ \infty, & otherwise. \end{cases}$$

Later on, Zuker and Stiegler [1981] revisited the problem and suggested a new version of the `Nussinov` algorithm to retrace the structure with minimal free-energy in the **Turner energy model**. The Turner group experimentally evaluated the energy of characterized RNA substructures and brought to the RNA community an energy data base where each particular substructure has an experimentally determined energy value. A substructure could be either a stacked base pairs or a loop with a set of categories.

We remind here that an admissible structure is defined as a set of base pairs (i,j) excluding tertiary interactions and pseudo-knots. Each base pair (i,j) contributes to define its own looped region $[i+1, j-1]$. A k**-loop** is formed by U unpaired bases and $k-1$ pairs excluding the closing pairs. Loops are classified as:

Hairpin	for	$k = 1$
Stacked pairs	for	$k = 2$
Bulges/Interior loops	for	$k = 2$ and $U \neq 0$
Multi-loops	for	$k > 2$

The free energy for a given structure s is the cumulative energy associated with its loops:

$$E(s) = \sum_{L \in s} E(L)$$

where $E(L)$ is the free energy of k-loops, for $k \leq 2$, obtained from the Turner experimental model.

Assuming that the `MFE` structure remains a suboptimal structure from the ensemble and is slightly similar to the native structure. In [Zuker, 1989], the authors pioneered a new method not to solely report one structure but rather a subset of **suboptimal structures** offering a limited level of redundancy. This work was later extended by Wuchty et al. [1999] to produce the complete ordered set of suboptimals within a given energy interval, a version of this algorithm is implemented in the `Vienna package` [Hofacker, 2009].

Boltzmann distribution. Recent *in silico* methods for RNA 2D structure modeling rely on the assumption of a thermodynamic equilibrium. This paradigm accounts for stochastic fluctuations, causing a transcript not to necessarily adopt a

single stable conformation. At the thermodynamic equilibrium, each admissible structure can be theoretically observed, albeit with very low probability in the case of unstable structures. This leads to the formal definition of the Boltzmann ensemble, where the probability of observing a given structure s with energy E(s) is given by its **Boltzmann probability**

$$P(s) = \frac{e^{\frac{-E(s)}{RT}}}{\mathcal{Z}}. \tag{2.1}$$

with T the temperature, typically expressed in kcal.mol^{-1} and R the Boltzmann constant, expressed in kcal.(mol.K)$^{-1}$. The normalization term $\mathcal{Z}$ is called the **partition function**, and is defined as

$$\mathcal{Z} = \sum_{s' \in \mathcal{S}} e^{\frac{-E_{s'}}{RT}}$$

where $\mathcal{S}$ is the ensemble of secondary structures for an RNA S.

The probability of a given base pair (i,j) can be deduced as:

$$P(i,j) = \sum_{\substack{s' \in \mathcal{S} \\ (i,j) \in s'}} P(s')$$

Despite summing over a number of conformations which scales exponentially with the sequence length, the partition function can be efficiently computed, owing to the recursive algorithm suggested by McCaskill [1990]. Indeed, this algorithm enables a **computation of the partition function** $\mathcal{Z}$ of an RNA in polynomial time. The additivity of free energy induces a multiplicativity in the contributing terms to the partition function $\mathcal{Z}$. It follows that a `DP` scheme for the partition function can be adapted directly from the energy minimization scheme, by simply replacing the $(\min, +)$ operators with $(+, \times)$, and exponentiate the constant terms contributing to the energy.

Namely, $\mathcal{Z}_{i,j}$, the partition function restricted to a region $[i, j]$, can be inductively computed as:

$$\mathcal{Z}_{i,j} = \mathcal{Z}_{i+1,j} + \sum_{i \leq k < j} e^{\frac{-E(i,k)}{RT}} \mathcal{Z}_{i+1,k-1} \mathcal{Z}_{k+1,j} \tag{2.2}$$

with $\mathcal{Z}_{i,i-1} := e^{-0/RT} = 1$ for the base case. $\mathcal{Z}_{i,j}$ is calculated iteratively starting from the shortest segment until reaching $\mathcal{Z}_{1,N}$ with N the length of the RNA sequence and $\mathcal{Z} := \mathcal{Z}_{1,N}$.

Stochastic sampling. Given an ensemble of predicted structures what would be the set of optimal ones? Ding and Lawrence [2003] have suggested a **stochastic approach** to generate statistically representative subsets of the conformation space, the **Boltzmann ensemble of low energy**. The partition function allows to sample structures from the ensemble with respect to their Boltzmann probabilities through backtracking over the partition function matrix $\mathcal{Z}$. Therefore, the sampling is a **stochastic backtrack**, based on the selection of base pairs by random choice of a decomposition step with respect to the Boltzmann probability.

For instance, the decomposition of $\mathcal{Z}_{i,j}$ from Equation 2.2 allows either to consider j as being unpaired or paired with a nucleotide k. From the first contribution, the probability of choosing i to be unpaired during the backtrack over $[i,j]$ is set to

$$\mathbb{P}(i \text{ unpaired} \mid i,j) = \frac{\mathcal{Z}_{i+1,j}}{\mathcal{Z}_{i,j}} = \sum_{s \in \mathcal{S}_{[i+1,j]}} \frac{e^{-E(s)/RT}}{\mathcal{Z}_{i,j}} = \sum_{\substack{s' \in \mathcal{S}_{[i,j]} \text{ s.t.} \\ b_i \text{ unpaired}}} \frac{e^{-E(s')/RT}}{\mathcal{Z}_{i,j}},$$

and indeed coincides with the cumulated probability of structures featuring i unpaired for the interval $[i,j]$. If this case is chosen, then the backtrack should proceed recursively over the $[i+1,j]$ interval and return the structure in the interval.

In the case of being paired, one still needs to decide which partner k to choose to form the base pair with i. Again, the key idea is to choose a given case in the decomposition, a partner k with a probability proportional to its contribution to the partition function, namely

$$\mathbb{P}(i \text{ paired to } k \mid i,j) = \frac{e^{\frac{-E(i,k)}{RT}} . \mathcal{Z}_{i+1,k-1} . \mathcal{Z}_{k+1,j}}{\mathcal{Z}_{i,j}}. \tag{2.3}$$

In practice, in order to limit the required number of random bits, a random value V is generated in $[0, \mathcal{Z}_{i,j}[$ and terms associated to the different values for k will be subtracted from V until it becomes strictly negative. It is then easy to show that the probability of choosing a given k is indeed the one stated in Equation (2.3). Once a given k is chosen, the backtracking procedure proceeds by performing two independent backtracks on $[i+1,k-1]$ and $[k+1,j]$, merging the two substructures, adding the chosen base pair (i,k) and returning the complete structure.

The use of the recursive algorithm over all the randomly chosen decompositions allows to generate a sample of m secondary structures in $\mathcal{O}(n^3 + m.n^2)$ worst-case time and $\mathcal{O}(n^2 + m.n)$ space, with n the sequence length. Remarkably, the average case complexity of this algorithm is in $\mathcal{O}(n^3 + m.n\sqrt{n})$ time, while the worst case complexity can be decreased to $\Theta(n^3 + m.n \log n)$ [Ponty, 2008].

Moreover, when performing a stochastic sampling the resulting structures correspond to the most probable ones. This stochastic sampling can lead to a redundancy issue where the frequency of a given structure is proportional to its Boltzmann probability. An efficient in-house non-redundant stochastic sampling, to explore in depth the structural ensemble while avoiding biases, was recently proposed by [Michálik et al., 2017].

Maximum Expected Accuracy (MEA). Given a set of structures, it is a natural question to ask for a representative structure, which can be used to formulate functional hypotheses. This question can be formalized as: Given an ensemble of structures, either abstractly described or generated from an underlying distribution, what is the structure with the maximal expectation to be drawn from the set? To answer this question, the `DP` algorithm, previously used to calculate the partition function, was extended to generate the MEA structure by maximizing the expected base-pair accuracy as suggested by Lu et al. [2009] and for which Hamada et al. [2009] have proposed novel centroid estimators.

The **expected accuracy** of a structure s is, essentially, its expected overlap with another structure, generated from a background distribution $\mathcal{B}$. When considering only base-pairs, the positive predictive value of a structure s with respect to a reference structure $s^\star$ admits a simple formulation:

$$PPV(s \mid s^\star) = \frac{TP}{P+N} = \frac{|s \cap s^\star|}{|s^\star| + |\bar{s}^\star|}$$

where $s \cap s'$ denotes the base pairs in the intersection of s and s', and $|s|$ denotes the number of base-pairs in s. Now, remind that the number of possible base-pairs is given by $\binom{n}{2} = \mathcal{N}(\mathcal{N}-1)/2$. Thus, $|s^\star| + |\bar{s}^\star| = \binom{\mathcal{N}}{2}$. The **expected accuracy** (in reality, expected PPV) is then expressed as:

$$\begin{aligned}
\mathbb{E}(PPV(s) \mid \mathcal{B}) &= \sum_{s' \in \mathcal{S}} P(s' \mid \mathcal{B}) \cdot PPV(s \mid s') \\
&= \frac{\sum_{s' \in \mathcal{S}} P(s' \mid \mathcal{B}) \cdot |s \cap s'|}{\binom{\mathcal{N}}{2}} \\
&= \frac{\sum_{(i,j) \in s} \sum_{\substack{s' \in \mathcal{S} \text{ s.t.} \\ (i,j) \in s'}} P(s' \mid \mathcal{B})}{\binom{\mathcal{N}}{2}} \\
&= \frac{\sum_{(i,j) \in s} P_{bp}(i,j)}{\binom{\mathcal{N}}{2}}
\end{aligned}$$

where $P_{bp}(i,j)$ is the probability of positions (i,j) to form a base pair. The probability of a base i to be unpaired is expressed as

$$P_{un}(i) = 1 - \sum_{j} P_{bp}(i,j).$$

The **maximum expected accuracy** structure for a sequence A is then expressed as:

$$MEA(A) = \underset{s \in \mathcal{S}}{\operatorname{argmax}}\, \mathbb{E}(PPV(s) \mid \mathcal{B})$$

Extending the relevant features to include unpaired bases leads to the MEA concept introduced by Lu et al. [2009].

Let $W(i,j)$ be the maximum expected accuracy for a sequence from nucleotide i to nucleotide j including i and j. The term $W(i,j)$ can then be computed as

$$W(i,j) = \max \begin{cases} P_{un}(i), & if \quad i = j, \\ P_{un}(i) + W(i+1,j), & \\ P_{un}(j) + W(i,j-1), & \\ V(i,j), & \\ W(i,k) + W(k+1,j)^{1}, & for \quad i \leq k < j. \end{cases} \tag{2.4}$$

where $V(i,j)$: the maximum expected accuracy for a sequence from nucleotide i to nucleotide j including i and j and is calculated as:

$$V(i,j) = \max \begin{cases} 0, & j - i + 1 < \text{ minimum hairpin loop}, \\ {}^{2}\gamma \times 2P_{bp}(i,j) + W(i+1,j-1), & \text{i and j are susceptible to pair} \\ -\infty, & \text{i and j can not pair.} \end{cases}$$

$W'(i,j)$ is additionally computed and it corresponds to the exterior region i.e. from 1 to i and from j to $\mathcal{N}$. The recursive algorithm allows to get `V` and `V'` values for each canonically authorized pair. Structures are then determined through a trace-back procedure where the MEA for a given substructure a delimited by a pair (i,j) is calculated as:

$$MEA(a_{(i,j)}) = V(i,j) + V'(i,j) - P_{bp}(i,j)$$

As a consequence, base pairs that belong to the substructure a and show a high expected accuracy are identified. Due to the back-tracing procedure, the MEA structure for the whole sequence that corresponds to $MEA(a_{(1,\mathcal{N})})$ is deduced.

[1] This term allows to identify the multi-branch loop

[2] This definition corresponds to the generalized $\gamma-$ centroid estimator [Hamada et al., 2009] that is equivalent to the centroid estimator as defined by Lu et al. [2009] when $\gamma = 1$

Software for RNA structure prediction The ensemble of algorithms to predict the RNA secondary structure have an implemented version in a set of software. Among the prominent suites, we find `Mfold/UnaFold` [Zuker, 2003], `RNAstructure` [Reuter and Mathews, 2010] and `Vienna package` [Hofacker, 2009]. Over the course of this PhD, we chose to use the `Vienna package`, due to its user-friendliness, its extremely rich combination of options, and its free access to the source code. In addition, the Turner model energy parameters are up-to date and prediction parameters are easily parametrized. All those factors prompted us to choose this package with a focus on the following programs:

- `RNAfold`, a program dedicated to compute the MFE structure (an implementation of `Zuker` algorithm) that returns the partition function (an implementation of `McCaskill` algorithm).
- `RNAsubopt`, a program to sample structures from the ensemble either with a decreasing energy order or stochastically (`Ding` and `Lawrence` algorithm).
- `RNAeval`to calculate the energy for a given RNA sequence-structure using the Turner model.

2.2 Wet-lab experiment for structure modeling

RNA function depends on its tertiary structure and on the information encoded in its Watson-Crick base-pairing potential. Powerful methods to determine structural properties of small and large RNAs have emerged for decades. Experimental methods to determine the RNA structure could be classified into 3 categories: spectroscopic, physical and chemical/enzymatic probing.

Spectroscopic experiments aim to define the whole biomolecular structure by making use of electromagnetic radiation. X-ray crystallograpgy [Golden, 2007], Cryo-Electron Microscopy `Cryo-EM` [chen Bai et al., 2015] and `NMR` [Scott and Hennig, 2008] remain the most popular used methods.

Physical methods, such as sedimentation velocity [Su et al., 2003] and single-molecule pulling experiments with laser tweezers [Manosas and Ritort, 2005], are used to get the information about the size and the shape of a molecule or a complex by measuring its movement in solution.

These methods are generally accurate but remain time-consuming and often induce experimental biases. In addition, they allow to capture one particular RNA state (crystal) from the landscape of possible conformations. Probing methods

came to overcome those limitations. Indeed, probing methods allow to produce an image of the RNA structural diversity.

2.2.1 Experimental probing

The most popular **chemical probing** method is **hydroxyl radical footprinting**. Hydroxyl radical aims to probe the solvent accessibility of the RNA backbone by abstracting hydrogen from the C5' position of the backbone, eventually leading to strand scission, with 3'-phosphate and 5'-aldehyde products [Ingle et al., 2014]. Hydroxyl radical technique is capable of detecting changes in the tertiary structure [Tullius and Greenbaum, 2005]. Despite the speed of its reaction, Hydroxyl radical is not appropriate to detect changes in secondary structure because of its inability to probe the bases.

In footprinting protocol, the modification of the RNA backbone leads to the cleavage of the RNA molecule i.e. the strand scission at specific sites. A popular method for RNA cleavage uses **RNase enzymes**. RNase enzymes cut at their binding sites, resulting in the formation of a $2',3-$cyclic phosphate and a $5'-$hydroxyl. RNase `V1` cuts double-stranded regions [Wan et al., 2013] where RNase `T1` recognizes single-stranded RNA sequences and cuts at a guanosine residues [Peng et al., 2012]. RNase probing is limited by the footprints of the enzymes. Instead, researchers have turned to small molecules for higher-resolution probing experiments by extending the footprinting method to detect positions sensitive to chemical attack [Ziehler and Engelke, 2000]. As a consequence, several alternative chemical probes were used to target the accessibility of the bases. Alkylating reagents were the first probing reagents used to probe single-stranded RNA. Dimethyl sulfate (DMS), is one of the pioneer and commonly used reagent [Lempereur et al., 1985]. It alkylates the N1 position of adenosine and the N3 position of cytidine. Other probes of bases are 1-cyclohexyl-3-(2-morpholinoethyl) carbodimide metho-p-toluene sulfonate CMCT, which reacts primarily with N3 of U and N1 of G modifying two sites responsible for hydrogen bonding on the bases [Burgstaller et al., 1995] and diethyl pyrocarbonate DEPC, which reacts primarily with N7 of A, and kethoxal which reacts with N1 and N2 of G.

An analogous chemical probing method was proposed with complementary oligonucleotides, which allowed the measurement of the accessibility for small part of the RNA sequence (stretches of 10 nucleotides) within a folded RNA [Zarrinkar and Williamson, 1994].

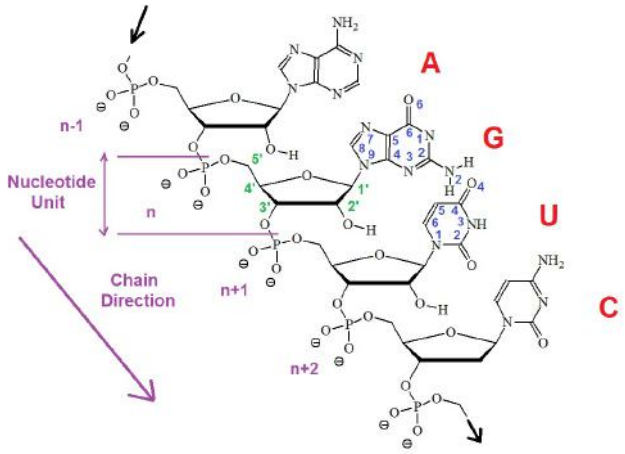


Figure 2.3: **The numbering convention of RNA nucleotides:** an example of RNA fragment with nucleotides A, G, C and U linked by 3',5'-phosphodiester bonds. The chain direction from 5' to 3' is indicated by an arrow. The atom numbering scheme is indicated in the nucleotide units.

2.2.2 SHAPE probing

New techniques have been developed to monitor the flexibility of each $2'-OH$ group within the ribose backbone by measuring the ability to cleave at the adjacent phosphodiester linkage (in-line probing [Soukup and Breaker, 1999]) or to attack an added electrophile (SHAPE [Merino et al., 2005]). The `SHAPE` chemistry is less selective than base-specific chemical probing protocols. Therefore, this allows to interrogate the global sequence and to provide direct measurements of the RNA backbone flexibility. `SHAPE` technology is exceptionally useful to inform about the secondary structure, which explain the adoption of this experimental protocol for High throughput use [Mortimer and Weeks, 2007].

SHAPE Diverse methodologies and technologies have been developed to assess the structural profile of the RNA molecule as previously described. In these biochemical techniques, RNA molecules are treated with a reagent. The interaction with the reagent results either into the formation of an adduct as for the chemical probing or into the cleavage of the RNA as for the enzymatic probing. This interaction happens at the nucleotide level. Therefore, the resulting reactivity profile informs about the flexibility of each single nucleotide.

Compared to other techniques, `SHAPE` makes it possible to extract more information on the RNA conformation [Merino et al., 2005]. `SHAPE` is characterized

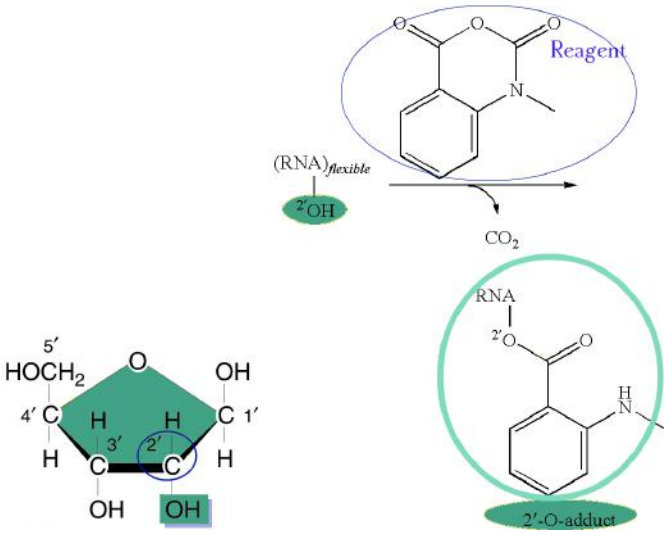


Figure 2.4: **SHAPE reagent interaction with the nucleotide ribose:** In the left, a nucleotide ribose where the OH component at the 2′ position is highlighted. In the right, the reaction of the SHAPE reagent with one molecule of RNA induce the formation of an adduct.

by its independence to the solvent accessibility [McGinnis et al., 2012], by its insensitivity to the nucleotide nature [Wilkinson et al., 2009] and, most notably, by its ability to inform about the local nucleotide dynamics with high accuracy [Gherghe et al., 2008]. Beyond its ability to infer the structure, SHAPE chemistry has been used to investigate the folding kinetics [Mortimer and Weeks, 2009] and to support sequence design [Lee et al., 2014].

SHAPE Mechanism Unlike other chemical probes that target the nucleobases, SHAPE interacts with the backbone. The most used reagents to perform SHAPE are 1M7 [3] [Turner et al., 2013] and NMIA [4]. Considered as anhydrid molecules, they selectively acylate the OH component at the 2′ level.

SHAPE experimental protocols are organized in two steps. First, RNA molecules are incubated with a chemical reagent that presumably reacts selectively with the OH component at the 2′ level of flexible RNA nucleotides. The acylation, both in vitro [Merino et al., 2005] and in cells [Spitale et al., 2013], results into the formation of an adduct. As a second step, a reverse transcription is performed: DNA primers are provided and a primer extension process is ensured by the presence

[3] The most useful and robust reagent for routine SHAPE experiments

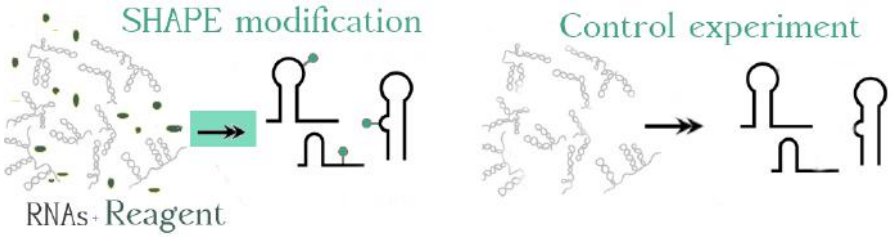
[4] A SHAPE reagent with a long half-life in solution that interacts with nucleotides showing a slow dynamics

of a polymerase, a necessary molecule to proceed for the transcription. During the reverse transcription, the polymerase considers each encountered adduct as a stop. This results in the generation of fragments of various lengths starting from 3' extremity. Fragments are filtered through the capillary gel electrophoresis. Given that the reverse transcription can stop spontaneously due to the processivity decay of the polymerase or to some structural constraints that prevent the polymerase from advancing in the sequence reading, an additional experiment is considered to count for the natural termination of the primer extension process: the control experiment. In this experiment, the reverse transcription is performed by relying only on the RNA molecules in the absence of a chemical reagent. The two experiments (`SHAPE` and control) are then read out by highly parallel sequencing. As a last step, a residual reactivity score is calculated. This score reflects the degree of interaction with the reagent for each single nucleotide. The reactivity values are obtained by considering the exposure intensity to the reagent from which the intensity of the background is removed, here the intensity value for a given position `i` corresponds to the frequency of the fragments of length `i`.

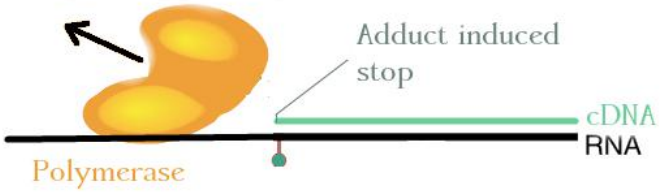
Resulting intensity signals are normalized in order to compare reactivities at the molecule level, i.e. between different residues in the RNA sequence. This normalization is also necessary to allow for a comparison between different experiment outputs. The normalization is ensured through a method introduced by Deigan et al. [2009] to pre-process the probing data. First, outliers are eliminated: an outlier corresponds to a value greater than 1.5 times the interquartile range. Then, each value is divided by the mean of the top 10% of the data.

`SHAPEMap` mechanism `SHAPEMap` technology exploits experimental conditions that force the polymerase to integrate a non-complementary nucleotide to the original sequence at adduct level in the resulting cDNA. The adduct locations are thus revealed as mutations. In `SHAPEMap` experiment, RNA is treated with a `SHAPE` reagent and a control experiment is performed in parallel to account for mutations that might occur spontaneously, presumably due to a sequence-specific bias.

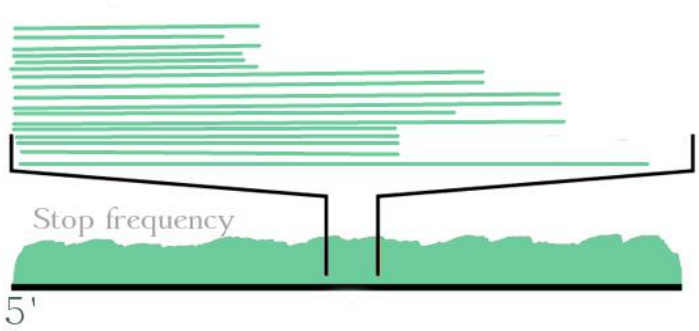
Additionally, in `SHAPEMap` experiment, a denatured experiment is performed. It consists into adding a `SHAPE` reagent to the RNA molecules under denaturing conditions aiming to count for sequence-specific biases. RNA molecules from each of the three experimental conditions undergo a reverse transcription; then the resulting cDNAs are introduced in the library to perform a massively parallel sequencing. In the context of `SHAPEMap`, reactivities are computed by comparing three mutation rates observed in different experimental conditions: presence (`SHAPE`)/absence (Control) of SHAPE reagent, and in the absence of structure (Denatured). The reactivity for a given position i is calculated as a normalized `SHAPE`



(a) SHAPE protocol with two chemical experiments: the SHAPE experiment in the presence of a reagent and a control experiment to count for the spontaneous stops.



(b) The polymerase proceed with a reverse transcription resulting in a cDNA strand truncated at the level of the adduct.



(c) Calculation of the end extremity (the adduct position) frequencies from the aligned cDNAs fragments.

Figure 2.5: SHAPE **chemical protocol and processing steps to construct the intensity signal .**

mutation residual rate formulated as:

$$\mathsf{Reactivity}(i) = \frac{\mathrm{m}_{\mathsf{SH}}(i) - \mathrm{m}_{\mathsf{Control}}(i)}{\mathrm{m}_{\mathsf{Denatured}}(i)}.$$

m_{SH} (resp. $\mathrm{m}_{\mathsf{Control}}$ / $\mathrm{m}_{\mathsf{Denatured}}$) is the mutational rate under the SHAPE (resp. Control/ Denatured) condition. The corresponding standard error is calculated as:

$$StdErr(i) = \frac{1}{\mathrm{m}_{\mathsf{Denatured}}} \sqrt{\frac{\mathrm{m}_{\mathsf{SH}}}{\mathrm{r}_{\mathsf{SH}}} + \frac{\mathrm{m}_{\mathsf{Control}}}{\mathrm{r}_{\mathsf{Control}}} + \frac{\mathrm{m}_{\mathsf{Denatured}}}{\mathrm{r}_{\mathsf{Denatured}}} (\mathrm{m}_{\mathsf{SH}} - \mathrm{m}_{\mathsf{Control}})^2}$$

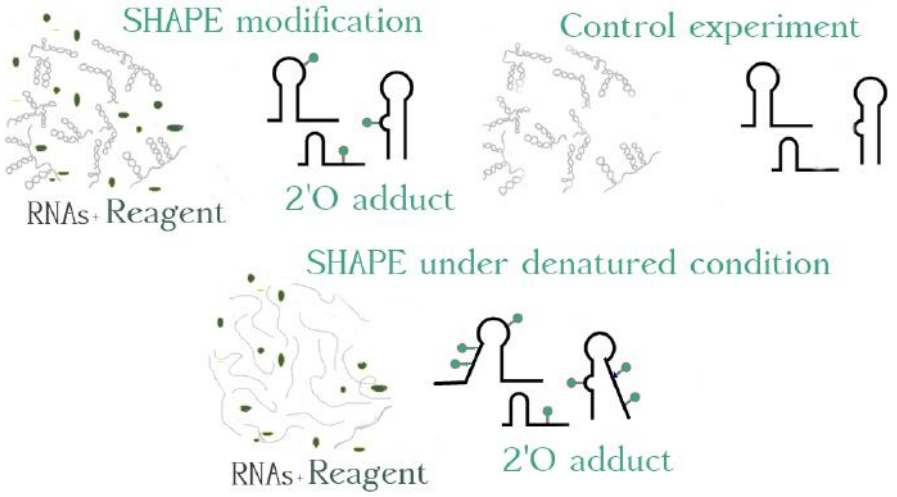
where r_X is the read depth for condition X at the considered position i.

Mutate and Map The integration of `SHAPE` data allowed to get accurate predictions [Gherghe et al., 2008]. However, the reactivity profile could be sensitive to the local modification either due to the interaction with the environment or to the adoption of alternative conformation(s). The loss of a hydrogen link between two nucleotides would increase the exposition of the two nucelotides to the reagent. A mutation of a paired nucleotide is likely to release the nucleotide-partner and consequently raises its accessibility to the chemical reagent. Mutate and map $M\&M$ [Kladwang et al., 2011] is a strategy to infer base pairs that is based on point-wise, potentially structurally disruptive, single mutation; then a change in the reactivity profile is examined and quantified. Ideally, a high reactivity is detected at the level of the nucleotide-partner. $M\&M$ consists into probing the WT sequence, systematically mutating a single nucleotide into its complement within the sequence and performing a probing experiment. The choice of mutating a paired nucleotide to its complement guaranties the elimination of the base pair [Duarte et al., 2003]. The $M\&M$ remains a robust approach to infer base pairs. It has been shown to significantly contribute to the enhancement of the final predicted structure when applied to a small model hairpin loop [Kladwang et al., 2011]. An example of predicted structure for Adenine riboswitch, V. vulnificus, through the use of $M\&M$, is displayed in Figure 2.7.

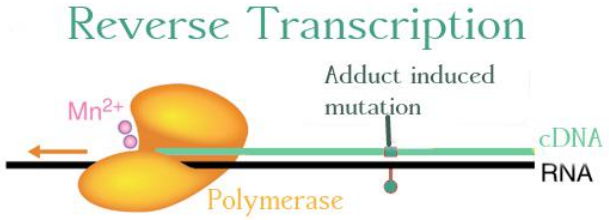
2.2.3 Integrating probing data in computational predictions

The incorporation of probing data in the prediction model has led to remarkable improvement in the accuracy of predicted structures, as for the complete HIV-1 genome (10kb) [Wilkinson et al., 2008].

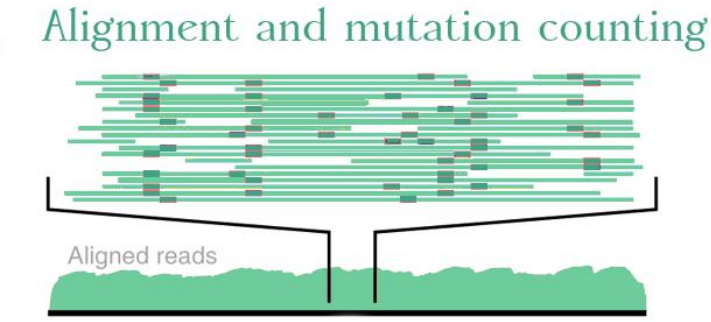
Hard constraints. The most direct approach to consider probing data as auxiliary information is to use it as a hard constraint [Zuker and Stiegler, 1981], where each nucleotide with a reactivity above certain threshold is considered as being absolutely single. Thus, it is prevented from forming possible pairing. Subsequently, all structures showing non-tolerated base pairs are eliminated from the Boltzmann ensemble.



(a) `SHAPEMap` chemical experiments under a `SHAPE`, a control and a denatured conditions.



(b) The polymerase proceed in a reverse transcription, the adduct induces a mutation.



(c) Calculation of the mutation rate after the alignment of the cDNAs strands.

Figure 2.6: `SHAPEMap` **protocol and processing to construct the mutational profile.**

The `DP` scheme used to constrain the predicted ensemble under the assumption of hard constraints is the same as for the Boltzmann probability calculation 2.2 while skipping some cases in the recursion: bases constrained to be unpaired are directly specified by the first term in Equation 2.2. On the other hand, bases

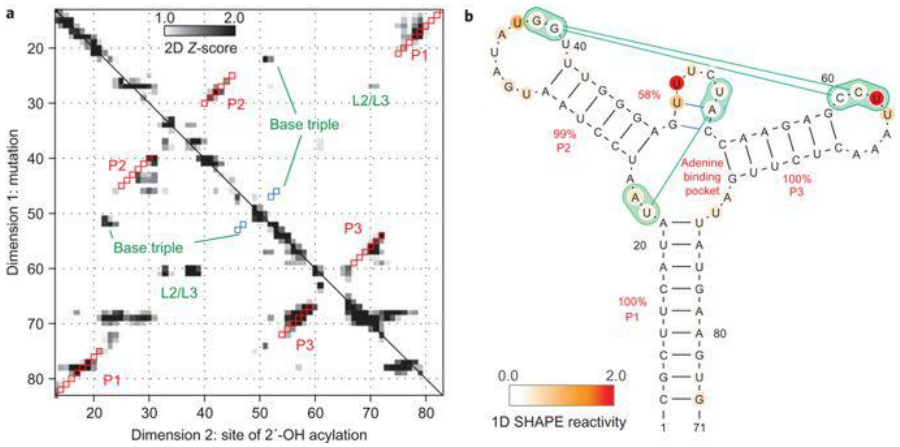


Figure 2.7: *M&M* **2D plot and the resulting predicted structure for add. riboswitch:** the entire *M&M* data set across 71 single mutations, plotted in grey scale and the Secondary structure derived from incorporating `Z-scores` (the number of standard deviations from mean at each residue) into the `RNAstructure` modeling. Squares show secondary structure model guided by *M&M* data(red, match to crystallographic Watson-Crick stems; blue, match to non-Watson-Crick stem). Additional tertiary contacts inferred from a separate clustering analysis are given in green. The figure was imported from [Kladwang et al., 2011].

constrained to be paired contribute to avoid the checking step for the partner nucleotide k.

However, the simplicity of this method hides a high level of complexity in the interpretation. Indeed, the resulting binary classification (paired vs. unpaired) is highly sensitive to the choice of the threshold. In addition, this assumption makes it hard to decipher the stochasticity of the reactivity profile.

Pseudo-potentials. Deigan et al. [2009] suggested a *softer* approach to deal with probing data where the reactivity scores are converted into a **pseudo free-energies** and integrated in the thermodynamic model.

The pseudo-energy contribution δG of a stacked nucleotide `i` is calculated according to Equation 2.5.

$$\delta G(i) = m \times ln(reactivity(i) + 1) + b \tag{2.5}$$

m and b, the slope and the intercept were parametrized on a collection of 23S rRNA by using predicted structure obtained from a comparative analysis [Deigan

et al., 2009].

It is worth reminding that the energy bonus affects only stacked nucleotides. A high reactivity value for a nucleotide predicted as being paired, would result into bringing a positive energy preventing the formation of the base pairs and vice versa. The intercept b is negative, thus supports a pairing state for a nucleotide with a low SHAPE reactivity. Meanwhile, the slope m has a positive value, which contributes to a positive energy when the `SHAPE` reactivity is high. The optimal parameters for folding large RNAs have found to correspond to $m = 2.6$ and $b = -0.8$ [Deigan et al., 2009].

The integration of probing data as pseudo-potentials could have other forms. A formal framework was suggested to reconcile information from both prediction algorithms and probing experiments where pseudo-energies were introduced to minimize the discrepancy between the turner model based prediction and the probing experiment [Washietl et al., 2012]. An other approach suggested to consider a pseudo-energy term for all nucleotide positions rather than solely stacked bases [Zarringhalam et al., 2012]. However, when compared to the approach of Deigan et al. [2009], these approaches do not typically induce a substantial improvement for the accuracy of the predicted structures. Thus, in this thesis, when pseudo-energies are mentioned without further details, we refer to the above-mentioned computed term developed by Deigan et al. [2009].

2.3 Accuracy assessment tools

2.3.1 NGS output mapping quality assessment

The availability of different HTS technologies leads to a diversity in produced reads: paired end/single end, complete/short... To reconstruct the reactivity profile, one must analyze those reads, and either assemble the genomic sequence *de novo*, or map the reads to the RNA of reference. The mapping algorithms are based on indexing technique that rely on small data structures used for large texts to solve many tasks within an optimal time. The widely used pattern matching algorithm is known as the Burrows-Wheeler Alignment (`BWA`) [Li et al., 2008]. Many programs were developed based on this mapping algorithm such as `Bowtie` [Langmead et al., 2013] and Torrent Mapping Alignment Program `TMAP`.

Definitions:

- **Sequencing** Sequencing is about generating bits of sequences called reads. In function of the sequencing technology, two types of reads could be characterized:

  Single-end reads that result from sequencing one end of a cDNA fragment.

  Paired-end reads consists into sequencing the same fragment twice i.e. each end of the fragment is sequenced. This results in the formation of pairs of reads.

- **Mapping** The mapping is a crucial procedure to extract information from the read-out reads. It consists into projecting reads from sequencing to the RNA of reference.

The quality of mapping is a function of nucleotide distance to the reference, the length of the read and the uniqueness of mapping position. This quality is quantified as a MapQ score.

$$MAPQ = -10 \log_{10} Probability(\text{mapping position is wrong})$$

This information will be useful to get the most accurate reads after the mapping step.

NGS data processing is a crucial step towards accurate reactivity profiles, especially for protocols based on the quantification of mutations. Indeed, within a mutational context, it is likely to reject strongly mutated reads because of their low $MAPQ$ values which would, for example, affect the recovery of mutation rates in the case of `SHAPEMap`. The choice of the cut-off value for the $MAPQ$ is therefore very important.

2.3.2 Evaluation of the accuracy of predicted RNA structure

To evaluate the accuracy of the predicted RNA structures, we calculated the `sensitivity` that measures the percentage of correct predicted base pairs from the all predicted base pairs, and the Positive Predictive Value `PPV` that measures the percentage of correctly predicted base pairs from known base pairs. We also reported the Matthews Correlation Coefficient `MCC` and sometimes, for comparison

reason with other competing tools, the geometric mean of the `sensitivity` and the `PPV`.

Definitions:
Let X (resp. Y) be the set of base pairs from the native (resp. predicted) structure. Let $\bar{X}$(resp. $\bar{Y}$) be the complementary set that corresponds to the possible base pairs between different nucleotides in the sequence out of X (resp.Y).

- True Positive `TP`: is defined as the set of predicted base pairs figuring in the real set:
$$TP = Y \cap X$$
- False Positive `FP`: the set of predicted base pairs that are not present in the real set:
$$FP = Y \cap \bar{X}$$
- False Negative `FN`: the set of base pairs present in the real set but are not predicted:
$$FN = \bar{Y} \cap X$$
- True Negative `TN`, base pairs that are neither predicted nor found in the real set:
$$TN = \bar{Y} \cap \bar{X}$$
The possible number of base pairs for a given sequence of length n is $\frac{n(n-1)}{2}$. Thus,
$$\mid TN \mid = \frac{n(n-1)}{2} - \mid X \mid - \mid Y \mid + \mid TP \mid$$

The choice of `MCC` to be the accuracy metric is justified by its ability to both measure and maximize the overall accuracy while remaining informative and easily interpretable. The `MCC` is a correlation coefficient between the predicted and the native base pairs expressed as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

The geometric mean of `sensitivity` and `PPV` corresponds to :

$$\Pi = \sqrt{Sensitivity \times PPV}$$

$$\text{with} \quad Sensitivity = \frac{TP}{TP+FN} \quad \text{and} \quad PPV = \frac{TP}{TP+FP}$$

Chapter 3

Probing data integrative modeling

3.1 Modeling challenges

3.1.1 Probing data

Modeling of the RNA structure from the sequence alone does not allow to extract features such as binding sites, non-canonical and long-range tertiary interactions that influence the RNA folding. Guided prediction with enzymatic and chemical probing have shown to resolve those features. Due to its ability to regroup information on where the RNA is single or double-stranded in a single reactivity profile and due to its particularity to target the ribose in the sugar phosphate backbone, allowing a more confident information about the flexibility of the base, `SHAPE` chemical probing has surpassed its competitors, leading to exciting development of methods and computational approaches. In parallel, RNA community has been working on the development of a set of computational approaches dedicated to other probing techniques such as the DMS-mapping that remains one of the preferable techniques for some research laboratory. Regardless the type of the chemical mapping, the stochasticity aspect of the reactivity profiles persists. This stochasticity is a direct consequence of the superposition of as many profiles as conformations. The deconvolution of such profile allows a better revelation of the RNA structure(s). Therefore, the important question to address is about finding feasible methods to make this embedded reactivity profile more understandable and subsequently interpretable.

3.1.2 What is measured by probing?

The reactivity reflects the structural diversity in solution at the time of probing. For a given residue, the reactivity indicates the fraction of molecules with one specific state at equilibrium. Therefore, understanding the meaning of a reactivity value constitutes a milestone to the development of adequate methods to correctly integrate this score in the prediction model. Reactivity is a normalized numerical value lying in the range of [0,1] where values below 0.3 are considered as being nonreactive and those above 0.7 as being highly reactive. Nucleotides engaged in a `Watson-Crick` pairs are usually showing low reactivity except for closing pairs of helices that are likely to show moderate reactivity. The moderate reactivity can also translate the existence of multi-conformations in the experimental pool. Thus, the structural diversity could be confirmed, for example, in the case where one side of a helix is showing a low reactivity and the opposite side is presenting a moderate to a high reactivity value. Unpaired nucleotides are likely to show high reactivity. However, they can often be non-reactive to the reagent because of nearest neighbour interactions that can limit the motion of single strands. The understanding of such specifications is of a major importance to infer accurate RNA structure.

3.2 The evolution of probing data integrative methods

The integration of `SHAPE` data as pseudo-energies has shown to remarkably improve the accuracy of predicted RNA structures [Deigan et al., 2009]. A leverage for the automated modeling through the Pseudo-energy framework for prior chemical mapping such as `DMS` motivated the work from [Cordero et al., 2012] where it has been shown that the incorporation of `DMS` probing data as pseudo-energies, either as favorable energies or as penalties, allowed to get similar or better information content compared to `SHAPE` data. The existence of distinct `SHAPE` and `DMS` signatures at nucleotides engaged in a non-`WC` interactions explained the best performance of the DMS mapping. Therefore, `DMS`-guided prediction has shown to be more accurate than `SHAPE`-guided prediction for regions where `SHAPE` data can not differentiate between `WC` and non-`WC` base pairs. In addition, [Cordero et al., 2012] have used CMCT probing data as pseudo-energies and they recorded a poorer accuracy compared to `SHAPE` or to `DMS` guided predictions. From the benchmark result, they have shown the ability of the `DMS` probing to achieve comparable accuracies to `SHAPE` using the pseudo-energy framework, and suggested that the structure modeling with both `DMS` and `SHAPE` data, when carried-out in

parallel, will permit a rapid assessment of the prediction errors and guaranty a more accurate inference.

The predictive capacity of the different probing mapping on the gold-standard secondary structures from [Cordero et al., 2012] is displayed on Table 3.1.

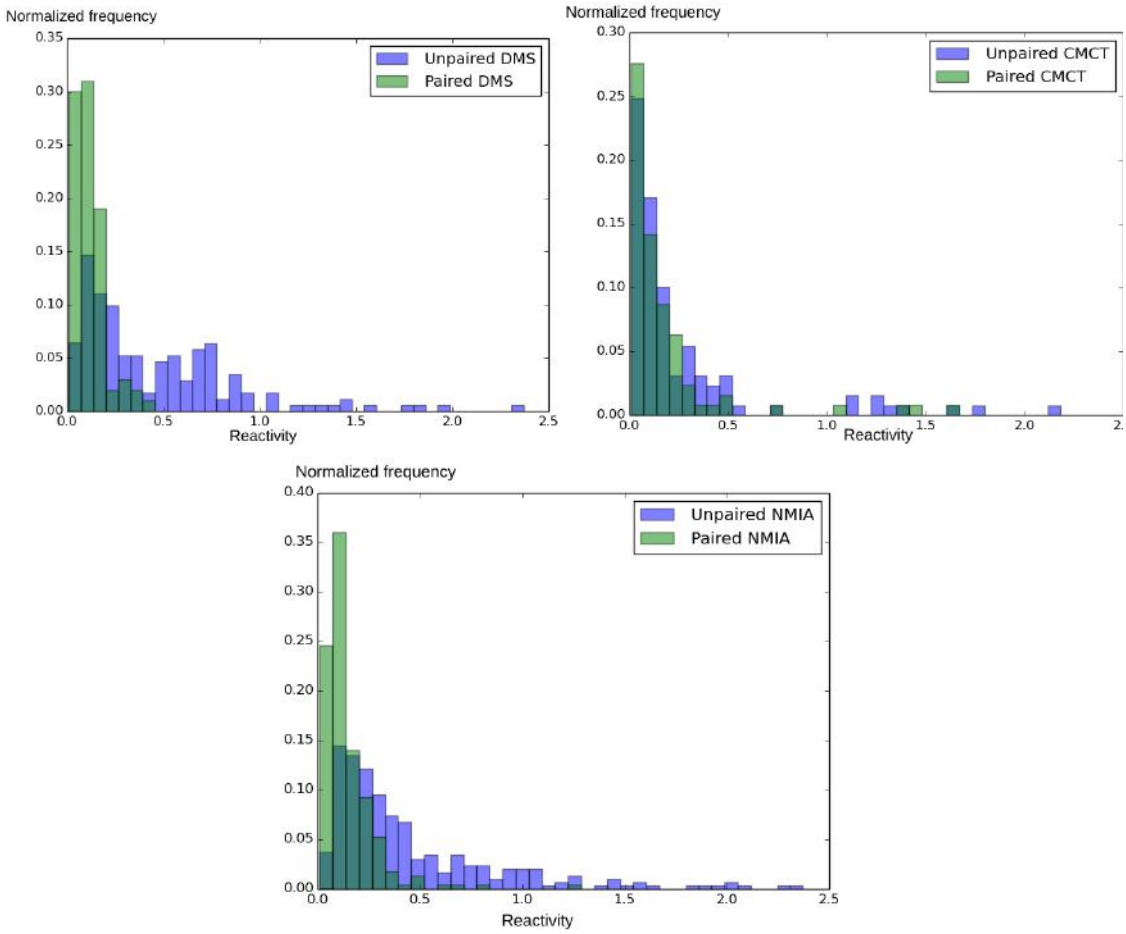


Figure 3.1: Distribution of reactivities for paired vs. unpaired nucleotides. Three probing reagent are considered: `DMS`, `CMCT` and `NMIA`. Data was extracted from [Cordero et al., 2012]. `SHAPE` NMIA probing data better reflects the correlation with the local structure but less well than the `DMS` data, whereas `CMCT` data do not differentiate between paired and unpaired nucleotides.

[Rice et al., 2014] suggested a multi-dimensional `SHAPE` guided predictions. They found that the use of three `SHAPE` reagents probing data allowed to obtain

RNA ID	MFE	MFE-DMS	MFE-NMIA	MFE-CMCT
5SRNA,E.Coli	0.241	0.686	0.686	0.254
glycineriboswitch,F.nucleatum	0.306	0.313	0.313	0.395
cidGMPriboswitch,V.Cholerae	0.77	0.77	0.77	0.667
P4P6domain	0.837	0.808	0.714	0.773
adenineriboswitch,add	1	0.333	1	0.41
tRNAphenylalanineyeast	0.976	0.976	0.976	0.976

Table 3.1: Accuracies of predicted RNA structures with the thermodynamic modeling, with DMS/NMIA and CMCT guided predictions using `RNAfold` with the default parameters. The `DMS` and `SHAPE`-guided predictions are showing better accuracies compared to the CMCT-guided predictions. An expected result due to the inability of CMCT data to characterize the local structural context. The accuracy is computed as the square root of the sensitivity and the PPV respectively corresponding to the fraction of the base pairs predicted correctly, and the fraction of correct base pairs that occur in the structure.

structure models that exceeded in average accuracy the models obtained with the chemical `DMS` and `CMCT` probing from [Cordero et al., 2012]. In addition, the incorporation of differential reactivities from `NMIA` and `1M6` reagents along with `1M7` in the prediction model, has shown to allow for highly accurate secondary structures compared to the prediction modeling guided by a single probing data. The performance of the mutli-probing guided predictions could be explained by the ability of `1M7-SHAPE` to measure local nucleotide flexibility, of `NMIA-SHAPE` to interact with nucleotides that experience slow dynamics and of `1M6-SHAPE` to stack with RNA nucleobases that are not interacting with other nucleobases. This differential approach allowed to identify, besides the canonical base pairs, the non-canonical and the tertiary interactions. [Rice et al., 2014] developed a new pseudo-free energy expression that include contributions from slowly dynamic and stacked bases. The differential energy with the `1M7` pseudo-free energy yielded a substantial improvement in the sensitivity and the `PPV` for a set of RNAs considered as being predictively challenging. In this work, they assumed that the consistent observed accuracy improvement suggest that both `1M7-SHAPE` guided modeling and differential-`SHAPE` contribute orthogonally to the construction of the RNA structure. The high information content of the differential-`SHAPE` allowed to model the most challenging RNA in agreement with accepted structures. As a result, predicted models have shown comparable if not better accuracies next to the approaches developed around the use of probing data from a set of mutants as $M\&M$ [Kladwang et al., 2011].

These prior works, focusing on probing guided predictions, assumed the dominance of one single structure. Thus, they do not consider the ability of the RNA sequence to arrange into many and often energetically close structures. In a recent work, [Spasic et al., 2018] suggested a new probing data based method to predict RNA conformers for sequences that populate multiple structures at equilibrium. Their method is based on sampling structures from the ensemble for which the estimated reactivities match the experimental ones. Their developed method (`'Rsample'`) has allowed to suggest multiple clusters that correspond to the experimentally reported known conformations for a set of RNAs. Predicted structures with `'Rsample'` achieved an accuracy of about 80% and allowed to provide more accurate structures when compared to the stochastic sampling method in the absence of supporting experimental data as suggested by Ding and Lawrence [2003].

3.3 Probing data and evolutionary covariation

In a multi-probing study, the first focus concerns the alignment of the reactivity profiles. Profiles alignment has shown to be useful to adjust a sequence alignment or to directly inform about structured versus unstructured regions, and more interestingly, to bring a completing structural information to strengthen the information encoded in a Multi-Sequence Alignment (MSA).

Lavender et al. [2015] have developed a model based on high-throughput chemical probing comparison. The use of `SHAPE` profiles alignment has shown to lead to comparable prediction accuracies next to the classic comparative prediction with an MSA. In addition, the combination of both a reactivity profile alignment and an MSA resulted on more accurate predicted structures. Indeed, the aligned `SHAPE` data profiles allowed to adjust the sequence-based alignment, and subsequently the 'corrected' MSA and the `SHAPE` data contribute both to the pseudo-free energy term. The `SHAPE` profile comparative model was validated against the $16S$ and the $23S$ rRNAs where some critical differences at the level of functional regions were noticed. An observation that may lead to build hypothesis stipulating the existence of unknown functions. In this work, they claimed that the use of the `SHAPE` profiles alignment is a new step towards the discovery of new functions and motifs.

Kutchko and Laederach [2017] studied some of the variants from `5'UTR` of RB1 RNA, and found out that the corresponding `SHAPE` reactivity profiles shared a considerable amount of highly similar regions. This observation came to strengthen the hypothesis about the ability of `SHAPE` profiles alignment to indicate the struc-

tural similarity between the sequences without knowing the structure. Besides the perceived quality of the SHAPE profiles to get easily aligned, the ability of the arising alignment to inform about specific structural properties and to make assumptions about the existence of unknown functions drove the development of new probing based alignment methods.

An in house developed approach [Reinharz et al., 2016], based on the combination of an MSA and a multiple chemical probing, allowed to detect binding sites for structured RNAs. The proposed method takes benefit from the $M\&M$ method. We remind here that one reactivity profile is considered as an uni-dimensional projection of the structure where $M\&M$ protocol probe the sequence with an ensemble of mutants leading to a multi-dimensional projection. The late projection was used to detect hotspots that involve a conformational change susceptible to disturb the WT-structure. In this work, it was assumed that an hyper variable SHAPE profile is a sign of a structural gap to the WT structure. This puts into question the conservation of the function. In parallel to the aligned SHAPE profiles and in order to preserve the structure and the interactions, an MSA was used to impose the selective pressure on the sequence. This alignment combination allowed to get more accurate predictions for a set of highly diverse RNAs, including a subset of riboswitches, compared to the consideration of only one of the two contributing alignments.

$M\&M$ protocol has shown to be once again a good approximation to detect the variability of reactivity profiles through a set of mutants. In a recent version M2-seq [Cheng et al., 2017], $M\&M$ protocol was enhanced through the use of an error-prone PCR that promotes the formation of accidental mutations. DMS-based probing using M2-seq protocol has shown to increase both the visibility and the detectability of base-pairs. In this analysis, M2-seq allowed to get more accurate helices for GIR1 Lariat-capping ribozyme compared to the classic $M\&M$ method.

The detection of conserved and variable regions by exploiting reactivity profiles from intentionally installed mutations was the first motivation that guided the present Ph.D work. We developed a new approach based on the use of SHAPE profiles alignment from a set of RNA variants obtained through an error-prone PCR. The SHAPE experimental protocol and the sequencing procedure were both performed in one pool. The implementation of this approach brought us to face a reads assignment issue due to the omnipresence of mutations. Surprisingly with the M2-seq method, intentional mutations did not prevent an accurate assignment of the sequencing output or at least Cheng et al. [2017] did not report any reads mapping difficulties in their analysis.

Part II

Outline

The acquisition of the reactivity profile is the initial step to proceed for the structure modeling guided by probing data. In Chapter 4, we present some automated analysis pipelines to recover reactivity profiles from raw data for a variety of protocols and technologies.

A considerable set of RNA functions depends on the appropriation of stable conformations. Current research interests, beyond being focused on revealing the RNA conformational diversity, are driven to introduce auxiliary probing data in the prediction model. However, this probing data relies substantially on the utilized experimental technique. Such characteristic leads, in certain cases, to predict diverse structures in function of the type of the considered probing data.

We present in Chapter 5 an automated integrative structural modeling approach, whose principle is to use concurrently a set of experimental data from diverse probing sources to guide RNA secondary structure predictions. Our developed approach `IPANEMAP` allows both to explore the structural landscape corresponding to miscellaneous probing conditions and to take advantage of the complementarity that exists between those conditions. `IPANEMAP` offers the possibility to model structure(s) at the intersection between deemed reliable probing data. Thus, `IPANEMAP` grants to identify stable conformations through a multi-probing ansatz while minimizing the effect of non-structurally supporting or noisy experimental data.

Starting from the intuition that the combination of unsupervised mutagenesis with `SHAPEMap` will form a complete construct of RNA structures by reporting both conserved base pairs and unpaired bases, we developed a new protocol spreading over two steps: first, RNA mutants were generated through a biased error-prone `PCR`, likely favouring mutations at the level of non conserved regions. Then, the resulting variants have undergone a `SHAPEMap` protocol favouring mutations likely to happen, this time, at the level of unpaired nucleotides.

We refer to our suggested protocol as **"Differential-`SHAPE`"**. Besides its use of non-directed mutagenesis that allowed to avoid the systematic production of pointwise mutants as in $M\&M$, its particularity is indeed related to the probing technique: mutated RNAs are probed simultaneously via high-throughput sequencing. Produced reads must then be re-assigned to mutants in order to establish their reactivity profiles. Ultimately, they are used to reconstruct compromise reactivity profile for which we hope to provide accurate inference to the structure. However, the hybrid mutation nature of the produced reads, both originated from the PCR and from the `SHAPEMap` protocol, made it challenging to correctly re-assign the reads to their RNA of reference and subsequently to verify such hypothesis.

We present in Chapter 6, a modeling of the read-assignment problem through a likelihood maximization joint inference of mutational profiles and assignments.

Chapter 4

Probing data analysis

Experimental probing aims to interrogate the structure of the RNA through the formation of adducts in nucleotides depending on their structural context (paired/unpaired). Adduct positions are then detected by Reverse Transcription (RT) and converted into stops or into mutations. In both cases, the RT results in the generation of a set of cDNAs.

Statistical analysis of stops or mutations frequency is used to estimate a measure of a reactivity at the nucleotide level. Mapping the cDNAs on the sequence of reference is the first step towards calculating residual reactivities. This chapter is dedicated to the presentation of profiling data preprocessing workflows, from the mapping of cDNA fragments to the final production of reactivities.

4.1 Capillary electrophoresis data

Wetlab experiments on RNA molecules allows to prepare a library of cDNAs. In the case of High-throughput `SHAPE`, each of the cDNA strands is labelled with a fluorophore and analysed with `capillary electrophoresis` [Wilkinson et al., 2008]. The separation by capillary has shown to be technically more practical than the manual manipulation of radioactive sequencing gels. This property allowed to apply the technique to the conventional biochemical and enzymatic probing [Mitra et al., 2008].

The extraction of data probing from the `capillary electrophoresis` intensity signal output is a multi-step process. `Qushape` [Karabiber et al., 2013] is a dedicated and pioneering tool that offers an automated and accurate `capillary`

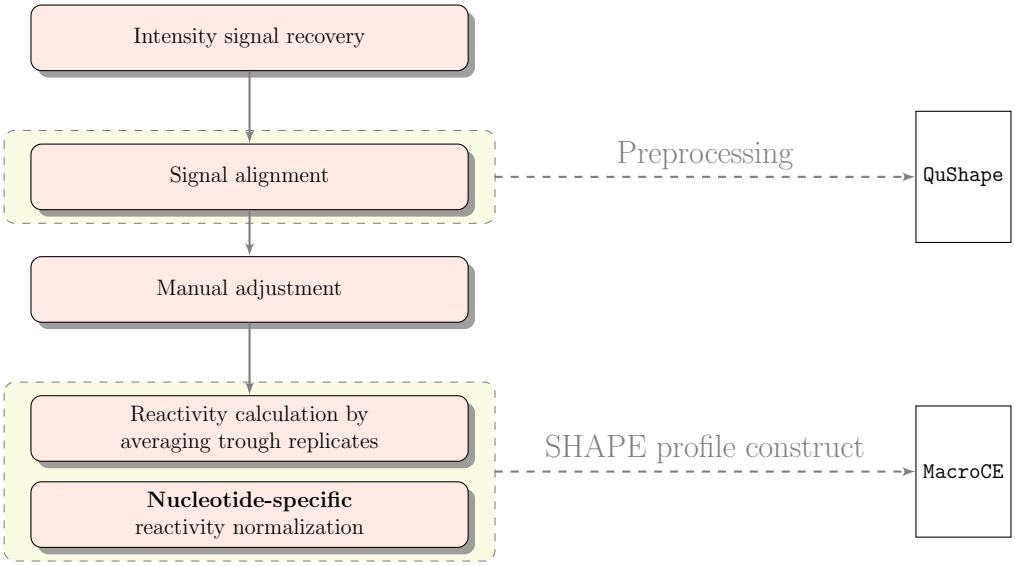


Figure 4.1: **Pipeline for computing `SHAPE` reactivities using `capillary electrophoresis` technique**.

`electrophoresis` data analysis. It includes a set of algorithms for signal decay correction, for signal alignment across capillaries and for signal scaling. Despite the robustness of the `Qushape` tool to extract accurate intensity signal information, the user supervision is required for further adjustments. After decoding the intensity signals, residual reactivities are calculated as an average of intensity values over all RNA replicates.

One contribution of this Ph.D. is the development of the `MacroCE` tool to automate the reactivity computation. The overall pipeline to treat the `capillary electrophoresis` data is illustrated in Figure 4.1. We made available an open-source implementation of `MacroCE` program at:

https://github.com/afafbioinfo/Macro_CE

A normalization of the raw reactivities along the RNA sequence, as proposed by McGinnis et al. [2012], is required for a better structure inference. The backbone flexibility is generally indicative of single-stranded regions. These regions tend to show a reactivity in the range of $[0.7, 1]$. However, it is possible to observe highly reactive nucleotide being associated with values above 2. Nucleotides in dynamic regions are likely to belong to this family. McGinnis et al. [2012] observed that 2% of nucleotides are hyper-reactive, and attribute this property to their flexibility, or to the constrained backbone within a specific conformation that favors the accessibility of the 2' hydroxyl group. Those nucleotides are treated as outliers, and subsequently discarded from the normalization contribution. Thus,

the normalization term includes the average intensity of 8% from the most highly reactive nucleotides after the elimination of the outliers.

What is the particularity of the `MacroCE` tool ? When dealing with different probing sources, we found out that the use of probing nucleotide-specific reagent, such as DMS that targets nucleotides C and A, should be carefully treated to build the corresponding reactivity profile. We assumed that the reactivity calculation, and subsequently the normalization, should only affect nucleotides explicitly targeted by the chemical reagent.

As one way to assess the restriction of the normalized reactivity calculation to the reagent targeted nucleotides, we computed the MFE structure from a DMS-guided prediction. First we considered normalized reactivities obtained through a normalization along the RNA sequence, and at a second time, we considered only normalized reactivity values for specific nucleotides, namely `A` and `C`, where the normalization term contains only contributions from those specific nucleotides. Table 4.1 shows the predictive performances of these two normalizations, by reporting the base pair distance between the resulting MFE structures and the native structure over a set of 7 RNAs. The obtained predicted structures confirmed the

RNAs	MFE-NormDMS	MFE-NormSpecDMS
GIR1 Lariat-capping ribozyme	32	33
5SRNA,E.Coli	57	57
adenineriboswitch,add	2	2
tRNAphenylalanineyeast	1	1
glycineriboswitch,F.nucleatum	52	**3**
cidGMPriboswitch,V.Cholerae	20	**12**
P4P6domain	53	**15**

Table 4.1: **The accuracy of predicted models with nucleotide selective normalization:** Base pairs distance of predicted structures to the native structure through a DMS-guided prediction. NormDMS refers to the case where normalized data covers the whole sequence where NormSpec is about considering only normalized values for A and C to bias the predicted ensemble. The three RNAs in the bottom show the interest of restricting reactivities to the specific targeted nucleotides by the probing reagent. Indeed, it allows to reduce drastically the considered base-pair distance, thus leading to more accurate predictions.

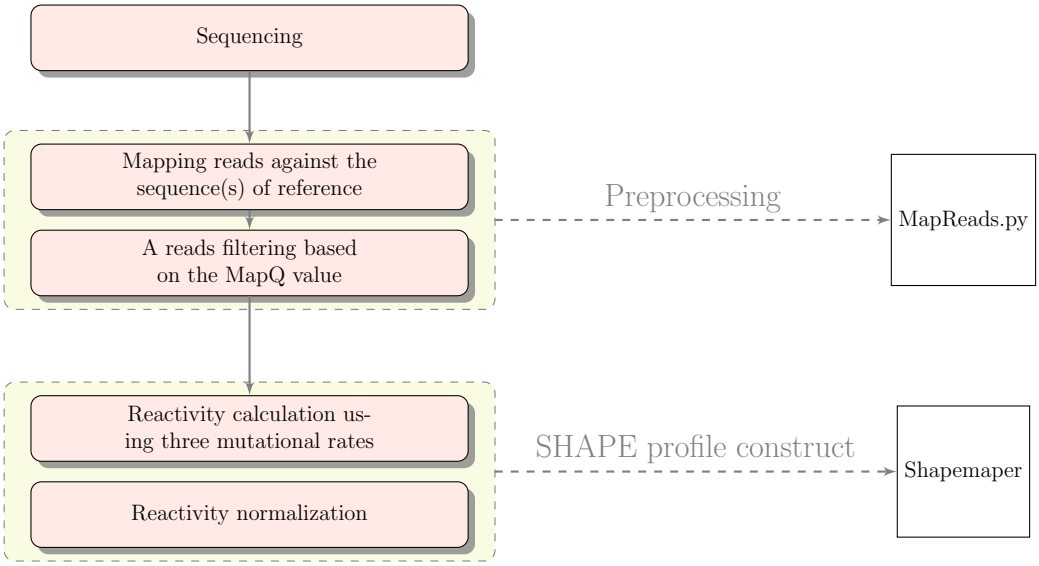


Figure 4.2: `SHAPEMap` **reactivities from HTS output**.

necessity to take into account the nucleotide nature for a probing data guided prediction with nucleotide-specific reagent. One possible explanation of the observed improvement, when considering the nucleotide nature, is that when considering only responding nucleotides to the reagent for the normalization term, the resulting normalized reactivities better reflect the structural context by allowing a larger scope of structural diversity.

4.2 High-throughput data

Parallel massive sequencing with the `SHAPEMap` method is the first step towards the mutational profile recovery. The processing of `SHAPEMap` data was performed through `ShapeMapper` [Smola et al., 2015], a software that proceeds at multiple level:

1. The mapping of reads.
2. The computation of mutation rates.
3. The calculation of `SHAPE` reactivity.

`ShapeMapper` is a powerful tool allowing to calculate the reactivity from the sequencing output. The vital step is about mapping the reads. `Bowtie2` [Langmead

et al., 2013] is a reads aligner integrated in `ShapeMapper` and is able to achieve a combination of high speed, sensitivity and accuracy. As `ShapeMapper` was designed to process **Illumina** paired-end reads, mapping setting parameters were trained to allow for a trade-off between the coverage of residues and the mapping quality for this specific read category. The command line to run `Bowtie2` with the optimized setting is:

```
bowtie2 --local -D 20 -R 3 -N 1 -L 15 -i S,1,0.50 --score-min
G,20,8 --ma 2 --mp 2,2 --rdg 5,1 --rfg 5,1 --dpad 100 --maxins
500 -p 4 -x fastafile -U fastqfile
```

When dealing with single-end reads from **Ion-Torrent** technology[1], the mapping principle should be revised in order to avoid uncovered artefacts. This prompted us to look for a more suitable and ideally **Ion Torrent** dedicated mapping tool. **T**orrent **M**apping **A**lignment **P**rogram (**TMAP**) is a program offering a set of performing and optimized mapping algorithms for Ion Torrent sequenced data. We choose the map4 module that is based on the Burrows-Wheeler Aligner BWA [Li and Durbin, 2009] mapping algorithm and allows a fast detection for the maximum exact matches reads-reference.

```
tmap mapall -n 24 -f fastafile -r fastqfile -v -Y -u -a 3 -s
samfile -o 0 stage1 map4
```

We made an open-source implementation of `MapReads` script available at:

https://github.com/afafbioinfo/ReadsMap

We evaluated the impact of the utilized mapper in the case of single-end reads. Using both `Bowtie2` and `TMAP`, we mapped a set of sequenced reads from 3 experimental conditions: `SHAPE`-1M7, Denatured-1M7 and the control, which we designated by the untreated condition. Then we evaluated the mapping percentage for each tool. We operated on the GIR1 Lariat-capping ribozyme RNA.

Impact of the mapping quality: A case-study. The GIR1 Lariat-capping ribozyme RNA spans over 188 nts with additional 20 polyA. In some of our previous experiment, the recovered mutational profile did not allow to express the residual reactivity for some nucleotides. This led us to analyze the length of sequenced reads and consequently to verify the percentage of complete reads. We visualized the distribution of the reads based on their length, the result is depicted in Appendix

[1] known also as Ion semiconductor sequencing, it is a category of DNA sequencing that relies on the detection of released hydrogen ions during the polymerization of DNA.

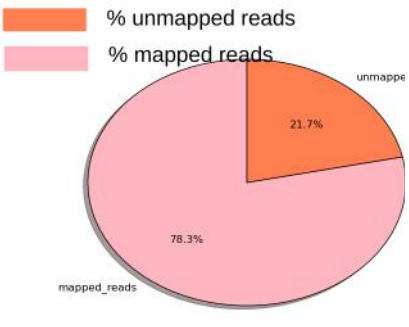
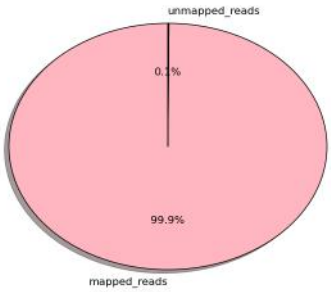
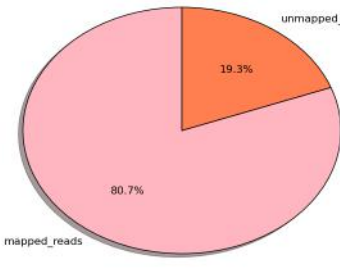
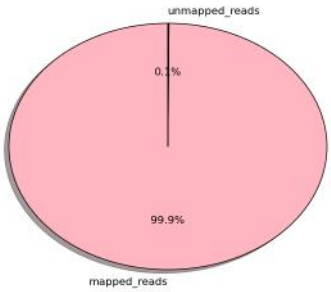
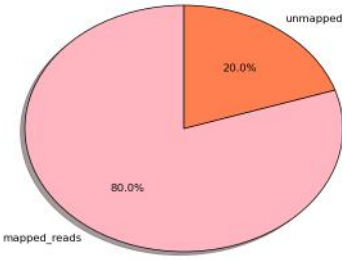
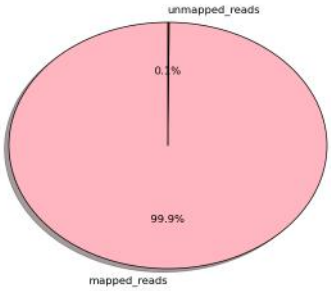
Experiment	Bowtie2	TMAP
SHAPE-1M7		
Denatured-1M7		
Untreated		

Table 4.2: **Comparison of the reads mapping percentage TMAP vs. Bowtie2.** Pink areas correspond to the percentage of mapped reads. A remarkable reads loss is observed with Bowtie2 compared to an almost complete assignment of the reads with TMAP.

[10*A*]. From the resulting reads distribution we concluded that more than one half of the reads are complete and exceeds largely the 5000 instances: the minimal required number of reads to build an accurate `SHAPE` profile [McGinnis et al., 2012]. Therefore, a coverage problem should not be faced. To assess the impact of the mapping tool, we computed the reactivity through an **Ion-Torrent** adapted version of `SHAPEMapper` and subsequently fed the resulting reactivities to RNAfold to calculate the partition function. The obtained results are displayed in Figure 4.3. The first observation is that the MFE structure is closer to the native when opting for a mapping step with the use of `Bowtie2` tool. However, from the partition function visualization, we observed the capacity of TMAP-Data guided predictions to detect more patterns that are compatible with the native signatures.

4.3 Conclusion

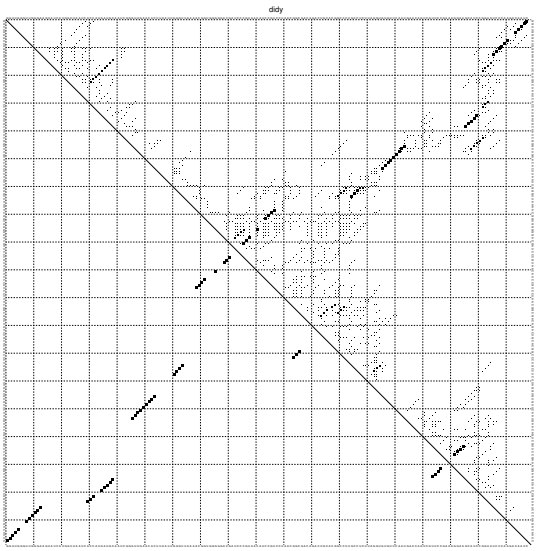
We presented two developed analysis pipelines to generate profiling data from stop-adduct and mutation-adduct experiments.

The normalization of probing data is a critical step towards "accurate" reactivities. We have found that the final normalized reactivity is nucleotide sensitive as long a the reagent is nucleotide sensitive. Hence, there is an interest to consider the nucleotide nature for both the normalization (to eliminate possible artefacts that may affect non-reactive nucleotides to the utilized chemical reagent) and the integration of probing data in the prediction model.

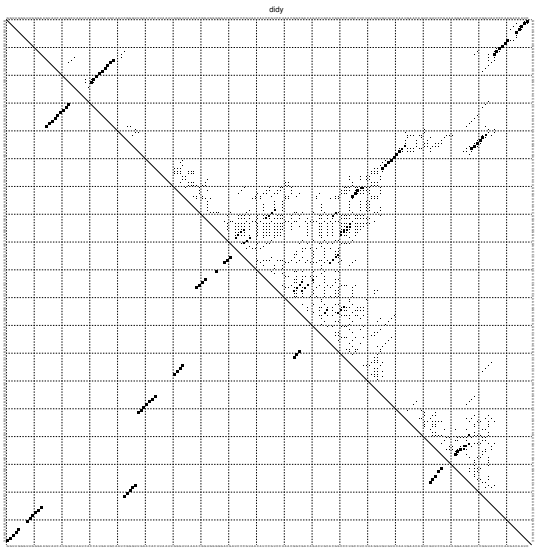
We also shed light on the importance to use an appropriate mapper tool when dealing with single-end reads from NGS data. Indeed, the optimal mapping parameters suggested by McGinnis et al. [2012] with the use of `Bowtie2` have shown to lead to a colossal loss in the reads recovery of up to 25% for the all 3 experimental conditions as shown in Table 4.2. A result that is not expected for the case of untreated experiment that contains only a few mutations.

Moreover, when we compared the resulting ensembles from a guided prediction with the profiling data issued from a `TMAP` mapping (TMAP-Data) and Bowtie2-Data, we noticed that the use of TMAP-Data to guide predictions has allowed to recover more patterns in accordance with the native structures compared to the use of Bowtie2-Data.

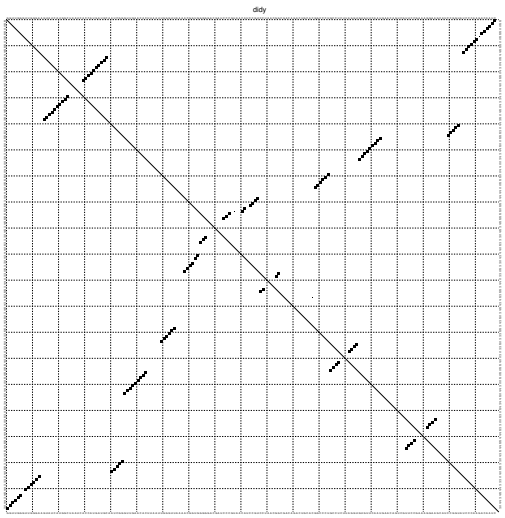
We concluded that the sensitivity to the mapping tool must be taken into account to get "precise" reactivities. In the absence of evidence regarding the optimal mapping tool to use, we decided to opt for `TMAP` mapper as long as it is dedicated `Ion-torrent output` processing.



TMAP-data guided prediction



Bowtie2-data guided prediction



Native structure

Figure 4.3: **Assessment of the predicted ensemble in function of the mapping tool:** Base pairs probability dot-plots, for the native structure (at the bottom) and for 2 `SHAPE`-guided predictions (on the top). Bowtie2-Data (resp.TMAP-Data) corresponds to guided prediction with reactivities obtained from `ShapeMapper` with `Bowtie2` (resp. `TMAP`). For each dot-plot, the lowest triangular matrix corresponds to the computed MFE structure, where the upper matrix displays the base pairing probability in the ensemble. This pairing probability is reflected by the dot intensity.

Chapter 5

IPANEMAP: Integrative Probing Analysis of Nucleic Acids Empowered by Multiple Accessibility Profiles

RNA probing encompasses a wide variety of experimental protocols providing partial structural information through the exposure to either a chemical or an enzymatic reagent, whose effect on the RNA reveals its structural features. The agreement between different probing assays has always driven the RNA structure(s) modeling by structural biologists, either through considering probing data produced in various experimental conditions, using different reagents or even over a collection of mutated sequences. However, such an integrative approach remains largely manual, time-consuming and frequently leads to models that are influenced by, arguably subjective, modeling choices. During this Ph.D., these different facets of the integrative modeling have been treated with a focus on developing in silico methods for an automated modeling of the RNA structure.

There is no doubt on the capacity of experimental probing data to provide information about the structural context at the nucleotide granular level. However, when dealing with such data two main issues could be faced. First, the complexity of the RNA 3D structure prevents encapsulated nucleotides from interacting with the chemical reagent, leading to a partial data probing information. An issue that could be aggravated by the use of a chemical/enzymatic reagent targeting specific nucleotides (such as `DMS` for nucleotides `A` and `C`)/ specific local structure (such as

V1-Rnase for paired nucleotides). Second, the ability of an RNA to switch from one conformer to another contributes to the stochasticity of the emitted probing data. Therefore, inferring the RNA structure from profiling data is still a big challenge.

The comparison of probing data from different sources contributes to ensure a maximum sequence coverage and allows to assess the agreement of data derived from multiple sources of probing. One of the widespread methods among structural biologists that use miscellaneous probing data boils down to conducting a guided prediction with one of the probing data, then projecting the remaining probing data on the structure to manually refine the predicted structure. However, this method remains sensitive to the chosen probing data to guide the structure prediction.

In this chapter, we present a probing-data integrative approach: `IPANEMAP` that aims to remedy the shortcomings of manual methods.

5.1 Towards a multi-probing integrative approach

5.1.1 Problem statement

Given an RNA sequence with a set of experimental probing data $\mathcal{D}$, one may define K structures representing K partitioning of the ensemble $\mathcal{S}_\mathcal{D}$: the ensemble of structures obtained from a $\mathcal{D}$-guided prediction. This problem is formalized as:

Input: RNA sequence, probing data $\mathcal{D}$, integer K

Output:

$$\{s_1^*, .., s_K^*\} = \text{argmin} \sum_{s \in \mathcal{S}_\mathcal{D}} P(s) \min_{i \in [1,K]} \mathcal{BP}(s, s_i^*)$$

with $\{s_1^*, .., s_K^*\}$:the set of K structures characterizing the partitioning of the Boltzmann ensemble. $P(s)$: the Boltzmann probability of the structure s compatible with the experimental condition $\mathcal{D}$, $\mathcal{BP}(s, s^*)$: the base pairs distance separating the two structures s and s^*.

A trivial solution for the above problem consists into listing all the structures, assessing their inter-distance then proceed for a ranking of the structures based on the minimization of this distance. However, this solution has an exponential cost. As an alternative, we proposed a stochastic sampling of the Boltzmann ensemble

followed by a grouping of the structures. The advantage of using a stochastic sampling lies in its ability to cover the whole structural landscape while returning stable structures with high frequency.

5.1.2 Integrative probing principle

In this section we describe our integrative approach driven by the hypothesis of the agreement between various probing data.

Assuming that the RNA functional structure(s) should be both energetically stable and supported by several experimental conditions, we coupled a stochastic sampling from the probing-guided ensembles, with a clustering across the diverse considered probing data, to produce reliable structural models. The first step of our method consists into sampling structures that are compatible with the constraints imposed by the various probing data, from the Boltzmann ensemble [Ding and Lawrence, 2003]. In order to detect recurrent RNA architectures, the ensemble of sampled models were clustered using the classic Base-Pair distance ($\mathcal{BP}$) as a measure of structural dissimilarity. To identify recurrent structures, we used the clustering algorithm Mini-Batch K-means [Sculley, 2010], implemented in the `scikit-learn` Python package [Pedregosa et al., 2012]. In order to define the number of clusters to generate, we developed a new iterative clustering approach. We then sought to identify clusters that are homogeneous, stable and well supported by experimental evidences, leading to the identification of the two following objective criteria:

- Represented conditions: We aimed to favor clusters encompassing structures generated from many experimental conditions. However, the larger sampled sets required for reproducibility tend to populate each cluster with structures from all conditions. We thus associated with each cluster the number of represented conditions, defined as the number of conditions for which the accumulated Boltzmann probability in the cluster exceeds a predefined threshold.
- Boltzmann weight: Clusters might be populated with unstable structures. To favor clusters with likely stable structures, we computed the cumulated normalized Boltzmann probabilities for each given cluster.

The next step of our method consists into restraining the analysis to a subset of clusters that are not strictly dominated by another cluster with respect to the two above-mentioned criteria: such clusters contribute to build the `2D` Pareto Frontier [Mattson and Messac, 2005] that maximizes both the number of represented

experimental conditions and the Boltzmann weight. After the determination of the optimal Pareto cluster(s), representative structure(s) for each elected cluster was computed as the MEA structure [Lu et al., 2009], more precisely as the combination of base pairs with the highest accumulated Boltzmann probability inside the elected cluster. A schematic representation of the `IPANEMAP` method is depicted in Figure 5.1.

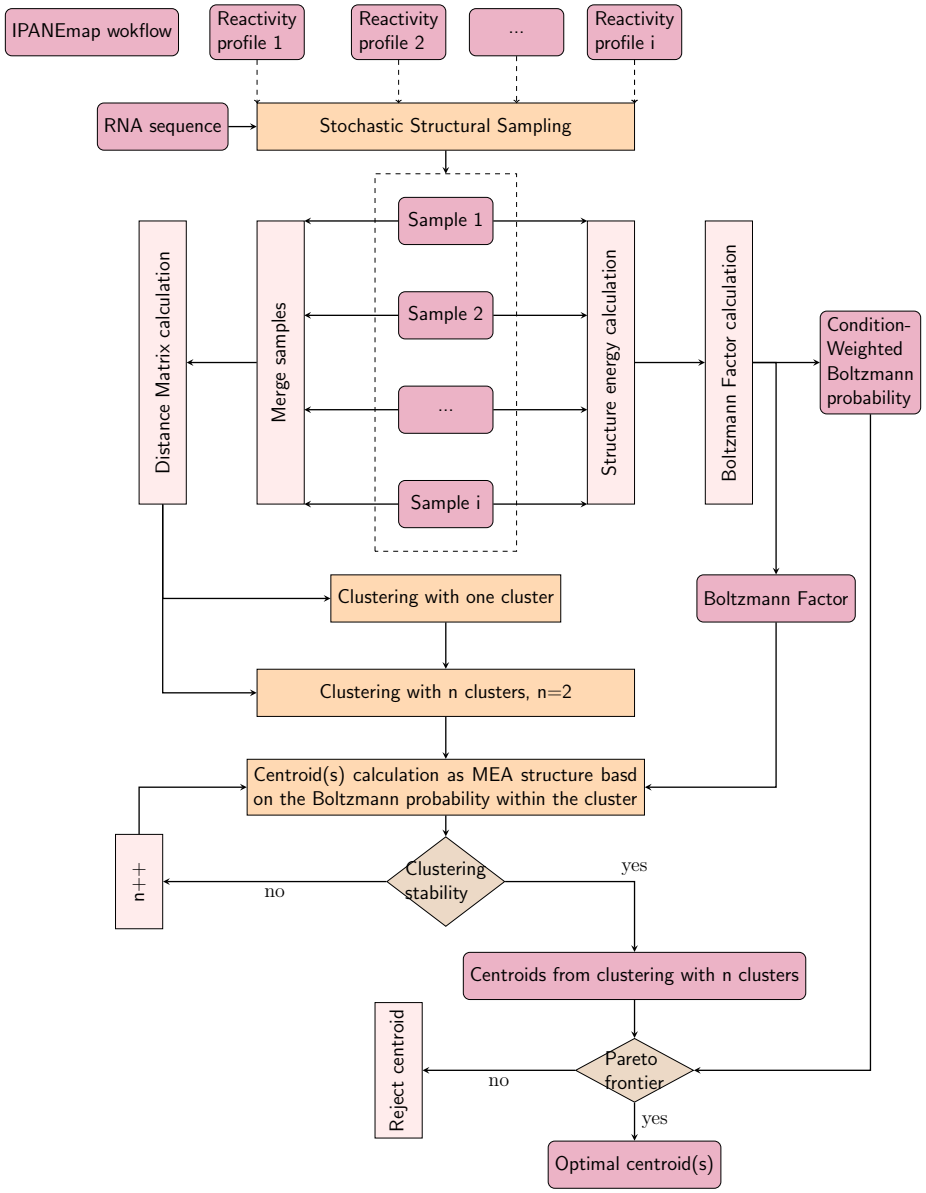


Figure 5.1: **Schematic diagram of** `IPANEMAP`. For a given RNA sequence and i relative structural profiling data (Reactivity profile 1, Reactivity profile 2...) as input data, `IPANEMAP` proceeds with a stochastic sampling for each of the predicted data-driven ensemble apart. Leading to a set of i samples (sample 1, sample 2...) representing the i experimental conditions. Then, the Boltzmann probability is calculated for each structure from each sample apart. We refer to this probability as the Conditional Weighted Boltzmann probability. In Parallel, samples are mixed in one set while retaining the label for the condition of origin of each of the structures. Then, the distance separating different structures from the set are assessed to allow `IPANEMAP` to proceed for the iterative clustering process that involves a clusters stability checking step and returns the number of clusters with their corresponding structure contents where each cluster is characterized by its centroid structure `MEA`. Centroids from clusters with their respective cumulated Weighted Boltzmann probability are used for the optimality check (Pareto frontier). Resulting optimal centroid(s) form the predicted structures through our integrative approach.

5.1.3 Distance evaluation between ensembles

As a first evaluation of the agreement between miscellaneous probing data, we started by analysing the compatibility of the structural models induced by probing data through the assessment of the similarity between the sets of generated ensembles. This evaluation was done by comparing their respective Boltzmann probability matrices where at the thermodynamic equilibrium each secondary structure from the ensemble is characterized by a Boltzmann probability [McCaskill, 1990]. To evaluate the impact of the nature of probing data on the predicted ensemble, we performed a four steps procedure:

1. **Data Incorporation:** Probing data were incorporated in the prediction model either as hard or as soft constraints where the late was converted into a pseudo-energy penalty [Deigan et al., 2009].

2. **Conformational ensemble generation:** Secondary structures were computed using `RNAfold`, from `ViennaPackage2.2.5`, with default parameters and structural constraints option to count for the experimental constraint. In addition, the option$-$`p` from `RNAfold` was used to generate base pairing probabilities.

3. **Distance calculation:** to evaluate the distance between experimentally constrained conformational spaces, we opted for an euclidean distance $\mathcal{ED}ist$. This metric is defined for two given probing conditions $(\mathcal{D}_i, \mathcal{D}_j) \in \mathcal{D}^2$ as:

$$\mathcal{ED}ist(\mathcal{D}_i, \mathcal{D}_j) = \sum_{x=1}^{\mathcal{N}} \sum_{y=\mathcal{N}-x}^{\mathcal{N}} (P_{bp}^{\mathcal{D}_i}(x,y) - P_{bp}^{\mathcal{D}_j}(x,y))^2.$$

   with $\mathcal{N}$ the length of the RNA sequence, $P_{bp}^{\mathcal{D}_i}(x,y)$(resp. $P_{bp}^{\mathcal{D}_j}(x,y)$) the Boltzmann probability to form the base pair(x,y) under the experimental condition $\mathcal{D}_i$ (resp.$\mathcal{D}_j$).

4. **Similarity evaluation:** to assess the compatibility of the different structural models two by two, a spectral bi-clustering was applied. The clustering of the different constrained structural ensembles based on the euclidean distance matrix was performed using the Python package [`scikit` $-$ `learn`].

5.2 Sampling and Clustering

5.2.1 Sampling structural models

We used the stochastic sampling mode of `RNAsubopt`, from `ViennaPackage2.2.5` with default parameters to stochastically generate conformers satisfying reasonable trade-off between the thermodynamic stability and the probing data compatibility.

After sampling structures from the Boltzmann ensembles, structures were gathered in one set while labelling each structure with the respective probing original condition. Ding and Lawrence [2003] have found that for an ensemble of $\sim 10^{303}$ structures, a sample of 1000 structures yielded statistical reproducibility. Therefore, we fixed the number of sampled structures for each single experimental condition to 10^3.
In parallel, `RNAeval` from `ViennaPackage2.2.5`, was used to evaluate the folding energy of each secondary structure from the set.

5.2.2 Clustering

After sampling structures constrained by various types of probing data, the next step was the identification of an occurring similarity between RNA structures in the sample.

Distance matrix. As a first step in the clustering method, a dissimilarity matrix illustrating the base pairs distance between each pair of structures was computed. For two given structures s_1 and s_2 from the ensemble $\mathcal{S}$, the base pair distance is calculated as:

$$\mathcal{BP}(s_1, s_2) = \sum_{1 \neq x < y \neq \mathcal{N}} |P_{x,y,1} - P_{x,y,2}| = s_1 \oplus s_2$$

where structures are represented by the set of their base pairs: $s_1 := \{P_{x,y,1}\}$, $s_2 := \{P_{x,y,2}\}$ with (x, y) a permissible base pair. This metric calculates the number of base pairs to remove from and to add to s_1 to obtain the structure s_2.

The second step consists into using `Mini-Batch k-means` algorithm designated by $\mathtt{M-Kmeans}$. The implementation of $\mathtt{M-Kmeans}$ is available in the $\mathtt{scikit-learn}$ package. One of the advantages of this k-means algorithm variant resides in its low computational cost: on the contrary of k-means algorithm that uses all the dataset at each iteration, $\mathtt{M-Kmeans}$ uses only a subsample

from the dataset while keeping a good performance for the partitioning. In addition, we tested our integrative method on a benchmark of 24 RNAs from [Hajdin et al., 2013] with a variety of hierarchical clustering algorithms along side with k-means algorithm where we observed a reproducibility of the partitioning results. Therefore, we choose to ultimately apply the unsupervised learning $\texttt{M}-\texttt{Kmeans}$ algorithm.

So arises the question about the number of clusters to set. We developed a new adaptive iterative clustering method that detects the number of clusters to choose. Before presenting this iterative clustering method, we define some of the designed features to characterize a given cluster of RNA structures, namely the *cumulated Boltzmann probability* and the *centroid*.

Weighted Boltzmann probability. Let $\mathcal{S}_{\mathcal{D}j}$ be a representative sample for a given probing experimental condition $\mathcal{D}_j$.
The Boltzmann factor $\mathbb{BF}$ for a given structure $s \in \mathcal{S}_{\mathcal{D}j}$ is calculated as:

$$\mathbb{BF}(s \mid s \in \mathcal{S}_{\mathcal{D}j}) = e^{\frac{-E(s)}{RT}}$$

where R is the Boltzmann constant, T the temperature and $E(s)$ the free energy of the structure s evaluated according to the Turner model.

We remind here that when probing data are used as soft constraints in the prediction model, the considered energy contains also a term accounting for the probing data contribution. For each experimental condition $\mathcal{D}_j$, a normalization factor $\mathbb{Z}_{\mathcal{D}_j}$ is computed as the sum of Boltzmann factors overall structures arising from the sample $\mathcal{S}_{\mathcal{D}j}$:

$$\mathbb{Z}_{\mathcal{D}_j} = \sum_{s_k \in \mathcal{S}_{\mathcal{D}j}} \mathbb{BF}(s_k \mid \mathcal{S}_{\mathcal{D}j}).$$

Factor $\mathbb{Z}_{\mathcal{D}_j}$ is used to define the weighted Boltzmann probability $\mathbb{P}$.

For a structure s arising when considering the experimental condition $\mathcal{D}_j$ as a structural constraint, the corresponding $\mathbb{P}$ is calculated as:

$$\mathbb{P}(s \mid \mathcal{D}_j) = \frac{\mathbb{BF}(s \mid s \in \mathcal{S}_{\mathcal{D}j})}{\mathbb{Z}_{\mathcal{D}_j}}$$

Cluster features

1. *Pseudo-weighting:* Each cluster $c \subset \mathcal{C}$ may contain structures s from different conditions that could be characterized by their weighted Boltzmann

probability $\mathbb{P}(s \mid \mathcal{D})$. In order to compare different structures that satisfy different conditions $\mathcal{D}$, a re-normalization is required. The normalization factor is calculated as the sum of weighted Boltzmann probabilities over all non redundant structures in each condition present in the cluster:

$$\mathbb{Z}_c = \sum_{s \in c} \mathbb{P}(s \mid c) \tag{5.1}$$

s being a non redundant structure in the cluster c.

2. *Centroid structure:* Each cluster $c \subset \mathcal{C}$ can be also characterized by a gravity center called the centroid around which structures from the cluster are distributed. We choose a centroid structure to represent a cluster because it is a good approximation for the over all structural diversity in the cluster. For each resulting cluster c from the clustering step, a centroid structure c_o is assessed as the `MEA` structure. A version of the `DP` algorithm proposed by Lu et al. [2009] was implemented. Instead of counting for the frequency of a given base pair to deduce its pairing probability, we rather reasoned on its weighted Boltzmann probability within the cluster. This new base pair probability calculation aimed to attenuate, if not eliminate, the effect of frequent base pairs from thermodynamically unstable structures in the cluster. We used the same recursive algorithm described in Equation 2.4. At variance with Lu et al. [2009] proposal, we defined the base pairing probability as:

$$P_{bp}(x, y) = \frac{1}{\mathbb{Z}_c} \cdot \sum_{s/(x,y) \in s} \mathbb{P}(s \mid c)$$

where $\mathbb{P}(s \mid c)$ is the weighted Boltzmann probability of a structure s from the cluster c and $\mathbb{Z}_c$ is calculated via Equation 5.1. We remind here that the computation time complexity for the `MEA` calculation is $\mathcal{O}(n^3)$.

Choosing a suitable number of clusters. In this paragraph, we address the following question: As the number of clusters to chose is a particular aspect of a clustering method, what kind of criteria should be verified in order to obtain both a meaningful and representative number of clusters, particularly in the RNA structures specific context? In order to tackle this question, we define a set of stopping criteria, mirroring usual modeling practices, for a procedure which iteratively refines the clustering.

As described above, each cluster from the structural clustering is characterized by a centroid that is a median structure forming the core of the cluster, and by a cumulated Boltzmann probability that forms a thermodynamic stability indicator.

Our clustering methodology consists into performing an initial clustering with one cluster then iteratively increase the number of clusters while assessing at each iteration the stability of the resulting clusters. The iterative clustering process stops once the clustering stability is reached.

To define the stopping criteria for the iterative clustering, we assume that the number of clusters should reflects a stability in the clustering, in other terms, the number of clusters that allows to keep the same assignment of the structures even if we go beyond this value.

Criterion 1: centroids, at a given iteration n, are all mutually exclusive. Let $\mathcal{BP}(c_o, c'_o)$ denote the base pair distance that separates two centroids c_o and c'_o at iteration n . Let ϵ be a tolerance parameter that we fixed to 1 to tolerate one base-pair distance to decide for centroids similarity. Therefore, n is the adequate number of clusters if:

$$\exists c_o \neq c'_o \in ({\mathcal{C}_{\mathcal{O}}}^{n+1} \times {\mathcal{C}_{\mathcal{O}}}^{n+1}) \quad \text{such that} \quad \mathcal{BP}(c_o, c'_o) \leq \epsilon$$

with ${\mathcal{C}_{\mathcal{O}}}^{n+1}$ the set of centroids defined at iteration $n+1$.

Criterion 2: At which iteration the clustering can no more get refined?
We assume that a low cumulated Boltzmann is associated with unstable structures. We considered this assumption as a filter to monitor the formation of new clusters. The assessment of the distance is performed solely between clusters represented by thermodynamically stable structures. This allows for a formulation of the second stopping criterion as :

$$\forall c_o \in {\mathcal{C}_{\mathcal{O}}}^{n} \quad \text{such that} \quad \mathbb{CP}_c \geq \epsilon' \quad , \exists \quad c'_o \in {\mathcal{C}_{\mathcal{O}}}^{n+1} : \mathcal{BP}(c_o, c'_o) \leq \epsilon$$

ϵ corresponds to unit base pair distance. $\mathbb{CP}_c$: the cumulated Boltzmann probability of the cluster c, computed according to Equation 5.2. ϵ' is a parameter that was trained over `IPANEMAP`. A trade-off between a reasonable number of iterations and accurate resulting centroids was found for a value equal to the third of the cumulated Boltzmann probability from the initial cluster for $n = 1$.

5.2.3 Election of optimal clusters

We designated by **optimal** cluster, the cluster with the most stable structures and covering the maximum of experimental conditions. The determination of optimal cluster(s) was performed through the evaluation of two factors, we judged to be enough to assess the coherence and the diversity of the clusters.

Let n be the number of resulting clusters. For each subset c from the set of clusters $\mathcal{C}$, we compute:

Algorithm 1 Iterative clustering algorithm to elect a number of clusters

procedure OptimalNumClusters($\mathcal{S}, \epsilon, \epsilon'$)
  $n \leftarrow 2$ ▷ Current number of clusters
  $SC^\star \leftarrow \varnothing$ ▷ Previous subset of stable clusters
  $Stop \leftarrow False$ ▷ Stopping condition
  while not $Stop = False$ **do**
    $\mathcal{C} \leftarrow$ Clustering($\mathcal{S}, n$) ▷ Clusters $c_1, \ldots, c_\mathcal{N}$
    $\mathcal{C}_\mathcal{O} \leftarrow$ Centroids($\mathcal{C}$) ▷ Centroids $c_{o1}, \ldots, c_{o\mathcal{N}}$
    $B \leftarrow$ BoltzProb($\mathcal{C}, \mathcal{S}$) ▷ Cumulated Boltz. prob. $\mathbb{CP}_c1, \ldots, \mathbb{CP}_c\mathcal{N}$
    $SC \leftarrow \{c_{oi} \mid i \in [1, n] \ \wedge \ \mathbb{CP}_c i \geq \epsilon'\}$ ▷ Subset of stable clusters
    if $\exists c_o, c'_o \in {\mathcal{C}_\mathcal{O}}^2 \cup (SC^\star \times SC), c_o \neq c'_o$ **such that** $\mathcal{BP}(c_o, c'_o) \leq \epsilon$ **then**
      $Stop \leftarrow True$
    else
      $SC^\star \leftarrow SC$ ▷ Save previous subset of stable clusters
      $n \leftarrow n + 1$ ▷ Onwards to the next iteration
    end if
  end while
  return n
end procedure

1. **The number of substantially represented condition** $\mathcal{I}_c$: the number of experimental conditions $\mathcal{D}_j$ represented by an acceptable number of structures thus closing the door to the outlier structures. The number of conditions in the cluster is computed as:

$$\mathcal{I}_c = |\{\mathcal{D}_j \mid \sum_{\substack{s \in c \\ s \in \mathcal{D}_j}} \mathbb{P}(s \mid \mathcal{D}_j) \geq \epsilon_\mathcal{D}\}|,$$

   whit $\epsilon_\mathcal{D} = \frac{1}{n+1}$, n being the number of clusters.

2. **The cumulated weighted Boltzmann probability** $\mathbb{CP}_c$ to count for conditions not sufficiently presented in term of number of structures but might encompass stable structures. This factor reflects, indeed, the thermodynamic stability of the cluster.

$$\mathbb{CP}_c = \sum_{s \in c} \sum_{s \in \mathcal{D}_j} \mathbb{P}(s \mid \mathcal{D}_j) \tag{5.2}$$

A cluster is considered as being optimal if it is not dominated by other clusters where the definition of **dominance** is as follow: a cluster c, abstracted as its two

components $(\mathcal{I}_c, \mathbb{CP}_c)$, dominates another cluster $c' =: (\mathcal{I}_{c'}, \mathbb{CP}_{c'})$ iff:

$$(\mathcal{I}_c > \mathcal{I}_{c'} \text{ and } \mathbb{CP}_c \geq \mathbb{CP}_{c'}) \quad \text{or} \quad (\mathcal{I}_c \geq \mathcal{I}_{c'} \text{ and } \mathbb{CP}_c > \mathbb{CP}_{c'})$$

To define dominant clusters, we considered each cluster candidate one by one then we updated the list of the optimal clusters For an inspected cluster candidate two situations may occur: the candidate is dominated by at least one of the optimal clusters, in this case the cluster candidate is eliminated. If the candidate is not dominated by any element from the list of optimal clusters, then any element from the list of optimal clusters dominated by this element is discarded and the new candidate is embedded to the list of optimal clusters.

Algorithm 2 Block-nested-loop algorithm to elect clusters on the Pareto front

```
1: procedure ParetoFront(C)
2:     C_max ← ∅
3:     while C ≠ ∅ do
4:         c ← pop(C)                          ▷ For each time pick up one cluster
5:         dominated ← False
6:         for all d ∈ C_max do
7:             if c dominates d then
8:                 C_max ← C_max \ d;
9:             else if d dominates c then
10:                dominated ← True
11:            end if
12:        end for
13:        if not dominated then
14:            C_max ← C_max ∪ {c}
15:        end if
16:    end while
17:    return C_max
18: end procedure
```

The algorithm remains simple but it requires to be able to perform the test of dominance between any two clusters. The computation time complexity for the dominance test is $\mathcal{O}(n^2)$ as it requires to compare each element from the set $\mathcal{C}$ to each single element from the progressively built set C_{max}.

5.2.4 Implementation details and availability

`IPANEMAP` is written in `Python2.7` and heavily depends on the `[scikit-learn]` library and the `Vienna Package 2.2.5+`. `IPANEMAP` is freely available at:

https://github.com/afafbioinfo/IPANEMAP

Conclusion

In this chapter, we presented a new integrative method to resolve the RNA secondary structure in the presence of supporting auxiliary data. The novelty of our method lies in both its ability to simultaneously integrate numerous probing data and its capacity to detect a set of reasonable number of accurate conformers. As described in this chapter, our method is based on performing a sampling of the structural landscape resulting from guided predictions with probing data and allows to predict a set of accurate conformers. As it favors recurrent structures that are jointly supported by several experimental conditions, our method allows, for the first time, to exploit the complementary nature of several probing assays which allows to smooth the effect of possible artefacts caused by some probing data. The effect of probing reactivities on the Boltzmann ensemble can indeed be substantial, and largely diverse across conditions, as revealed by new metrics based on base-pair probability distance matrices.

One other strength of our method lies in its ability to treat mono-probing assays. Indeed, the accuracy of predicted structures obtained through our method compares favourably against state-of-the-art computational predictive methods as we will as presented below in Chapter 7.

Chapter 6

Differential-SHAPE assignment

In the presence of HTS data, an accurate mapping of reads against the sequence of reference is mandatory to grant an accurate structure inference. However, in the case of simultaneous sequencing, especially when dealing with RNA variants, the task becomes hard to handle. Many algorithms have been developed to overcome the issue of simultaneous mapping reads against a set of homologous sequences but the problem is not fully resolved, particularly in the case of short reads. The issue addressed in our study is much more challenging: in addition to the parallel assignment issue in the presence of short reads, RNA variants molecules used for the library sequencing preparation undergo a `SHAPEMap` treatment thus, causing the formation of mutations at the level of accessible nucleotides. Those mutations are likely to induce a miss-mapping issue where a given read derived from a given RNA variant becomes closer in term of base pair distance to another RNA variant.

In this chapter, the focus is on describing an **E**xpectation **M**aximization (**EM**) algorithm that we developed to tackle this unprecedented known assignment problem. Our suggested assignment approach is based on the characterization of each RNA variant by its corresponding `SHAPE` mutational profile instead of being merely characterized by the sequence of nucleotides. The EM algorithm aims to maximize a joint likelihood: the probability of a read to belong to a specific RNA variant and its contribution to build the RNA variant associated mutational profile.

We addressed the issue of read mapping uncertainty in the presence of many mutants. It is mandatory to thoroughly understand the mapping errors intrinsic to the mutants presence and to the `SHAPE` modification. Indeed, the high sequence similarity between variants and the undergone mutation during `SHAPE` process both contribute to generate mapping errors. One classic method to map reads with those

specific conditions consists simply on the elimination of ambiguous reads likely to get mapped to more than one RNA variant. However, this rescue approach does not solve the problem, only reads with multiple assignment choice will be subject to elimination while their presence is significant. We are suggesting a statistical approach to treat the assignment of sequencing outputs in the presence of many sources of ambiguity that is more rigorous.

6.1 The assignment problem statement

Let us consider an RNA sequence modeled as a string $\mathcal{W} = w_1 w_2 w_3 w_4 w_5 .. w_{\mathcal{N}}$ on the alphabet Σ of 4 letters $\{C, G, A, U\}$.

First, a non-directed mutagenesis PCR process was performed, leading to a proliferation of mutations: this corresponds to bring arbitrary letters modifications to the string $\mathcal{W}$.

The PCR process could be modeled as a branching process: at each cycle we have at maximum two new copies of the sequence; consequently a total of 2^k variants are generated by the end of this process.

The resulting variants are used as a basis for an additional experimental protocol `SHAPEMap` that favor letter modification for some specific positions with a residual mutation rate around 2%.

By the end of the two modification processes, a sequencing step is performed. The resulting reads are subject to degradation leading to a significant number of short reads. The problem now is to correctly assign those reads to their RNA variant of origin.

We assume that data come as a set of `r` reads of length `l` generated from `d` variants. For ease of notation, we give the expression of our EM model parameters as:

$\mathcal{N} \rightarrow$ Sequence length
V $\rightarrow$ Variants; $d :=\mid V \mid$
X $\rightarrow$ Reads; $r :=\mid X \mid$, each $x_i \in X$ is delimited by positions s_i and e_i from the sequence of reference
Z $\rightarrow$ Missing values (mapping reads to Variants)
$\theta \rightarrow$ Parameters (emission probabilities)

Let $X = (x_1, x_2 ..., x_r)$ be a set of r independent reads, and let $Z = (z_1, z_2, ..., z_r)$ be the latent variable that determines the variant of origin for each read.

Given an integer p, a read x_i and a variant v_j, one denotes $x_{i,p}$ and $v_{j,p}$ the

nucleotides at position p in x_i and v_j, respectively.
We define a function ψ to express the mutational state of the nucleotides as:

$$\psi(x_{i,p}, v_{j,p}) = \begin{cases} \bar{m} & \text{if } x_{i,p} = v_{j,p} \\ m & \text{otherwise.} \end{cases}$$

We choose to characterize each variant v_j by a mutational profile, denoted by M_j. M_j is modeled as $[1, \mathcal{N}] \times \{m, \bar{m}\}$ matrix of probability values $P(m, w_k)$ and $P(\bar{m}, w_k)$.

$$P(\bar{m}, w_k) = 1 - P(m, w_k)$$

We consider the problem of estimating the statistical model parameter set:

$$\theta = \{M_j\}_{j=1}^{d}$$

We consider the following assumptions:

- Reads are independent.
- Reads contribute equally to build the mutational profile regardless their length and the covered part of the sequence.

We define the probability density function:

$$f(x_i; M_j) = \prod_{k \in [s_i, e_i]} M_j(k, \psi(x_{i,k-s_i}, v_{j,k}))$$

. We define $T_{j,i}^{(t)}$ the probability of the read x_i to be generated from the variant v_j given the current estimate of the parameter $\theta^{(t)}$:

$$T_{j,i}^{(t)} := P(Z_i = j \mid X_i = x_i; \theta^{(t)}) = \frac{f\left(x_i; M_j^{(t)}\right)}{\sum_{j'=1}^{d} f\left(x_i; M_{j'}^{(t)}\right)}$$

Note that $\sum_{j=1}^{d} T_{j,i}^{(t)} = 1$.

We define the complete-data likelihood function as:

$$\begin{aligned} L(\theta; X, Z) &= \prod_{i=1}^{r} \prod_{j=1}^{d} \left[\frac{f(x_i; M_j)}{d}\right]^{\mathbb{1}_{z_i=j}} \\ &= exp(-r.\log(d) + \sum_i \sum_j \mathbb{1}_{z_i=j} \times \log f(x_i; M_j)) \end{aligned}$$

with $\mathbb{1}$ an indicator function.

The assignment problem could be formalized by:

Input: V: set of d variants, X: set of r reads where a given read x is characterized by `SHAPE` mutational profile.

Output: $\underset{\theta}{argmax}\ L(\theta; X, Z)$
with Z: the assignment of the r reads to the d variants , $L(\theta; X, Z)$: the complete-data likelihood

6.2 The EM parameters estimation

The first step in the EM algorithm is to use the current parameter value of θ to find the posterior distribution of the latent variable Z. The posterior probability of the random variable given the observation while keeping $\theta^{(t)}$ fixed is given by:

$$
\begin{aligned}
Q(\theta \mid \theta^{(t)}) &= E_{Z|X,\theta^t}[\log L(\theta; X, Z)] \\
&= E_{Z|X,\theta^{(t)}}[\sum_i \log L(\theta; x_i, Z_i)] \\
&= \sum_i E_{Z|X,\theta^{(t)}}[\log L(\theta; x_i, Z_i)] \\
&= \sum_i \sum_j T_{j,i}^{(t)} \times [\log L(M_j; x_i, Z_i)] \\
&= \sum_i \sum_j T_{j,i}^{(t)} \times \left(\log \left[\prod_{k \in [s_i, e_i]} M_j(k, \psi(x_{i,k-s_i}, v_{j,k})) \right] - \log(d) \right)
\end{aligned}
$$

As $\sum_i \sum_j T_{j,i}^{(t)} = 1$,

$$
\begin{aligned}
Q(\theta \mid \theta^{(t)}) &= \sum_i \sum_j T_{j,i}^{(t)} \left[\sum_{k \in [s_i, e_i]} \log M_j(k, \psi(x_{i,k-s_i}, v_{j,k})) \right] - \log(d) \sum_i \sum_j T_{j,i}^{(t)} \\
&= \sum_i \sum_j T_{j,i}^{(t)} \left[\sum_{k \in [s_i, e_i]} \log M_j(k, \psi(x_{i,k-s_i}, v_{j,k})) \right] - \log(d) \times r \\
&= \sum_i \sum_j T_{j,i}^{(t)} \left[\sum_{k \in [s_i, e_i]} \log M_j(k, \psi(x_{i,k-s_i}, v_{j,k})) \right] + \mathit{constant}
\end{aligned}
$$

Now we need to compute our next estimate $\theta^{(t+1)}$, defined as

$$
\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta \mid \theta^{(t)})
$$

First, we observe that the constant term does not depend on θ, and can therefore be safely neglected for the optimization of $Q(\theta \mid \theta^{(t)})$.

We are left to optimize the contributions of the $M_j(k, S)$, *i.e.* optimize the mutational profile associated with each position for each variant.

Remark that:

$$\sum_i \sum_j T_{j,i}^{(t)} \left[\sum_{k \in [s_i, e_i]} \log M_j(k, \psi(x_{i,k-s_i}, v_{j,k})) \right]$$

$$= \sum_j \sum_i T_{j,i}^{(t)} \left[\sum_{k \in [s_i, e_i]} \log M_j(k, \psi(x_{i,k-s_i}, v_{j,k})) \right]$$

$$= \sum_j \sum_i T_{j,i}^{(t)} \left[\sum_{k \in [1, n]} 1_{k \in [s_i, e_i]} \times \log M_j(k, \psi(x_{i,k-s_i}, v_{j,k})) \right]$$

$$= \sum_j \sum_{k \in [1, \mathcal{N}]} \sum_i \left[1_{k \in [s_i, e_i]} \times T_{j,i}^{(t)} \times \log M_j(k, \psi(x_{i,k-s_i}, v_{j,k})) \right]$$

$$= \sum_j \sum_{k \in [1, \mathcal{N}]} \sum_{S \in \{\mathsf{m}, \bar{\mathsf{m}}\}} \sum_i \left[1_{k \in [s_i, e_i]} \times 1_{\{\psi(x_{i,k-s_i}, v_{j,k}) = S\}} \times T_{j,i}^{(t)} \right] \times \log M_j(k, S)$$

Additionally, note that the constraints in the optimization are local, and can be reduced to:

$$\forall j \in [1, d], \forall k \in [1, \mathcal{N}] : M_j(k, S) \in [0, 1] \text{ and } \sum_{S \in \{\mathsf{m}, \bar{\mathsf{m}}\}} M_j(k, S) = 1.$$

It follows that the contributions of the terms having equal values for j and k can be regrouped, and consequently the $M_j(k, .)$ be independently optimized, since

$$\sum_j \sum_k \sum_S \sum_i \left[1_{k \in [s_i, e_i]} \times 1_{\{\psi(x_{i,k-s_i}, v_{j,k}) = S\}} \times T_{j,i}^{(t)} \right] \times \log M_j(k, S)$$

$$= \sum_j \sum_k \sum_S \alpha_{j,k,S} \times \log M_j(k, S)$$

where

$$\alpha_{j,k,S} := \sum_i \left[1_{k \in [s_i, e_i]} \times 1_{\{\psi(x_{i,k-s_i}, v_{j,k}) = S\}} \times T_{j,i}^{(t)} \right]$$

so that there exist constant terms $\alpha_{j,k,S}$, $j \in [1, d], k \in [1, \mathcal{N}], S \in \{\mathsf{m}, \bar{\mathsf{m}}\}$ and we are left to optimize $2 \times d$ independently contributing functions of the form:

$$\alpha_{j,k,\mathsf{m}} \cdot \log M_j(k, \mathsf{m}) + \alpha_{j,k,\bar{\mathsf{m}}} \cdot \log M_j(k, \bar{\mathsf{m}})$$

The maximization step for θ is to solve:

$$maximize \quad \alpha^{(t)}_{j,k,\mathsf{m}} \cdot \log M^{(t)}_j(k,\mathsf{m}) + \alpha^{(t)}_{j,k,\bar{\mathsf{m}}} \cdot \log M^{(t)}_j(k,\bar{\mathsf{m}})$$

subject to :

$$\sum_{S\in\{\mathsf{m},\bar{\mathsf{m}}\}} M^{(t)}_j(k,S) = 1,\ M^{(t)}_j(k,S) \geq 0,\ S \in \{\mathsf{m},\bar{\mathsf{m}}\}$$

The solution of our constrained optimization problem is obtained using the Lagrangian method. To solve $M_j(k)$: the mutational profile for a given variant v_j at a given position k, we form the Lagrangian:

$$\begin{aligned} J(M^{(t)}_j(k),\lambda) &= J(M^{(t)}_j(k,\mathsf{m}), M^{(t)}_j(k,\bar{\mathsf{m}}),\lambda) \\ &= \sum_{S\in\{\mathsf{m},\bar{\mathsf{m}}\}} \alpha^{(t)}_{j,k,S} \cdot \log M^{(t)}_j(k,S) + \lambda \times (1 - \sum_{S\in\{\mathsf{m},\bar{\mathsf{m}}\}} M^{(t)}_j(k,S)) \end{aligned}$$

$M^{(t)}_j(k,S)$ values, $S \in \{\mathsf{m},\bar{\mathsf{m}}\}$ that maximize $J(M^{(t)}_j(k),\lambda)$ depend on the value of λ.

Let us denote this optimizing value of $M^{(t)}_j(k,S)$ by $M^{(t)}_j(k,S,\lambda)$.

Since $J(M^{(t)}_j(k),\lambda)$ is a concave [1] function of $M^{(t)}_j(k,S)$ for $S \in \{\mathsf{m},\bar{\mathsf{m}}\}$, it has a unique maximum at a point where :

$$\frac{\partial J}{\partial M^{(t)}_j(k,S)} = \frac{\alpha^{(t)}_{j,k,S}}{M^{(t)}_j(k,S)} - \lambda = 0 \ for\ all\ S \in \{\mathsf{m},\bar{\mathsf{m}}\}$$

Thus

$$M^{(t)}_j(k,S,\lambda) = \frac{\alpha^{(t)}_{j,k,S}}{\lambda}$$

Observe that $M^{(t)}_j(k,S,\lambda) > 0$ for $\lambda > 0$ and that

$$\sum_{S\in\{\mathsf{m},\bar{\mathsf{m}}\}} M^{(t)}_j(k,S) = \sum_{S\in\{\mathsf{m},\bar{\mathsf{m}}\}} \frac{\alpha^{(t)}_{j,k,S}}{\lambda}$$

[1] $f(x) = \log(x)$, $\ddot{f} = -\frac{1}{x^2}$, so the log function is concave for $x \in (0,\infty]$.

By selecting $\lambda^\star = \sum_{S \in \{\mathsf{m}, \bar{\mathsf{m}}\}} \alpha^{(t)}_{j,k,S}$ such that the constraint $\sum_{S \in \{\mathsf{m}, \bar{\mathsf{m}}\}} M^{(t)}_j(k, S, \lambda^\star) = 1$ is satisfied and

by applying the Lagrangian sufficiency theorem, we assume that $M^{(t)}_j(k, S, \lambda^\star)$ is the optimal solution for our optimization problem.

which yields:

$$M^{(t+1)}_j(k, S) = M^{(t+1)}_j(k, S, \lambda^\star) = \frac{\alpha^{(t)}_{j,k,S}}{\sum_{S' \in \{m, \bar{m}\}} \alpha^{(t)}_{j,k,S'}}$$

6.3 The EM assignment algorithm

6.3.1 Algorithm and complexity

Let us consider:

- $\{\mathsf{m}, \bar{\mathsf{m}}\}$ the set of states, with m for a detected mutation and $\bar{\mathsf{m}}$ for a matching nucleotide.
- C the ensemble of complete reads x_i mapped to the variant v_j with high quality. Reads from C would be considered as static reads and will not participate in the reassignment.
- A the ensemble of short and ambiguous reads.

The EM reads assignment algorithm is performed through 5 steps:

1. **Preprocessing**:
   - Output reads from sequencing are mapped against the Wild-Type (WT) sequence to define the best mapping locations indexed by MapQ scores.
   - Complete reads with good MapQ score are then assigned to the unique variant with minimal distance. The assignment (complete reads, variant) allows to build the set C henceforth considered as the training data set.
   - Short reads, complete reads with low MapQ scores and ambiguous reads were used to generate the set A.

2. **Initialization**: for each read x_i choose the initial estimates $M_j(k,c)^{(0)}$, $c \in N$, and compute the initial log-likelihood

$$L^{(0)} = \sum_i \sum_j \mathbb{1}_{z_i=j} \times \log f(x_i; M_j^{(0)}) \quad = \sum_i \sum_j T_{j,i}^{(0)} \times \sum_{k \in [s_i, e_i]} \log M_j^{(0)}(k, \psi(x_{i,k-s_i}, v_{j,k})))$$

with

$$T_{j,i}^{(0)} = \begin{cases} \frac{f\left(x_i; m_j^{(0)}\right)}{\sum_{j'} f\left(x_i; m_{j'}^{(0)}\right)} & \text{if } i \in \mathsf{A} \\ 1 & \text{if } (i,j) \in \mathsf{C} \\ 0 & \text{otherwise.} \end{cases}$$

3. **E_step**: for each read x_i and variant v_j compute

$$f(x_i; M_j^{(t)}) = \prod_{k \in [s_i, e_i]} M_j^{(t)}(k, \psi(x_{i,k-s_i}, v_{j,k})),$$

4. **M-step**: Update $T_{j,i}$ and compute $\alpha_{j,k,S}$

$$\alpha_{j,k,S}^{(t)} = \sum_i \left[\mathbb{1}_{k \in [s_i, e_i]} \times \mathbb{1}_{\{\psi(x_{i,k-s_i}, v_{j,k}) = S\}} \times T_{j,i}^{(t)} \right]$$

Compute the new estimates $M_j^{(t+1)}(k,S)$, $S \in \{\mathsf{m}, \bar{\mathsf{m}}\}$ as:

$$M_j^{(t+1)}(k,S) = \frac{\alpha_{j,k,S}^{(t)}}{\sum_{S' \in \{\mathsf{m}, \bar{\mathsf{m}}\}} \alpha_{j,k,S'}^{(t)}}$$

5. **Convergence test**: Compute the log-likelihood at $t+1$:

$$L^{(t+1)} = \sum_i \sum_j \mathbb{1}_{z_i=j} \times \sum_{k \in [s_i, e_i]} \log M_j^{(t+1)}(k, \psi(x_{i,k-s_i}, v_{j,k}))$$

If $\mid L^{(t+1)} - L^{(t)} \mid > \epsilon$ for a given ϵ, return to step 2; otherwise end the algorithm.

6.3.2 The EM-assignment algorithmic complexity

For a given read x_i of length $l_i = e_i - s_i$. and a given variant v_j the estimation of one mutational profile M_j costs the cumulated length of all reads that we denote by L. Thus, the complexity for the EM algorithm during one iteration is $\mathcal{O}(dL)$ where d is the number of variants. Our suggested iterative EM-mapping algorithm is computationally greedy with a linear algorithmic complexity of $\mathcal{O}(dLt)$ where t is the order of number of iterations.

6.3.3 The EM-assignment validation

The EM-assignment was implemented in `Python 2.7` and is freely available at:

https://github.com/afafbioinfo/EM-assignment.

We choose the GIR1 Lariat-capping ribozyme, an RNA of length 188, to be our system model.

Differential `SHAPE` . We performed a differential `SHAPE` protocol, where different mutants were synthesized through a non-directed mutagenesis PCR of 30 steps. At the end of the PCR process, 92 mutants were arbitrarily chosen. Figure 6.1 shows the distribution of the resulting residual mutations along the set of chosen variants where

The mutation rate does not exceed 0.1 per nucleotide. Mutations are mainly located at nucleotide positions in the range $[25, 168]$. When a position is mutated, the nature of resulting mutation, except in few cases, remains the same through the set of variants which is in accordance with the PCR branching process: if a mutation occurs during the cycle C_i, the two derivative branches at cycle C_{i+1} might keep the mutation.

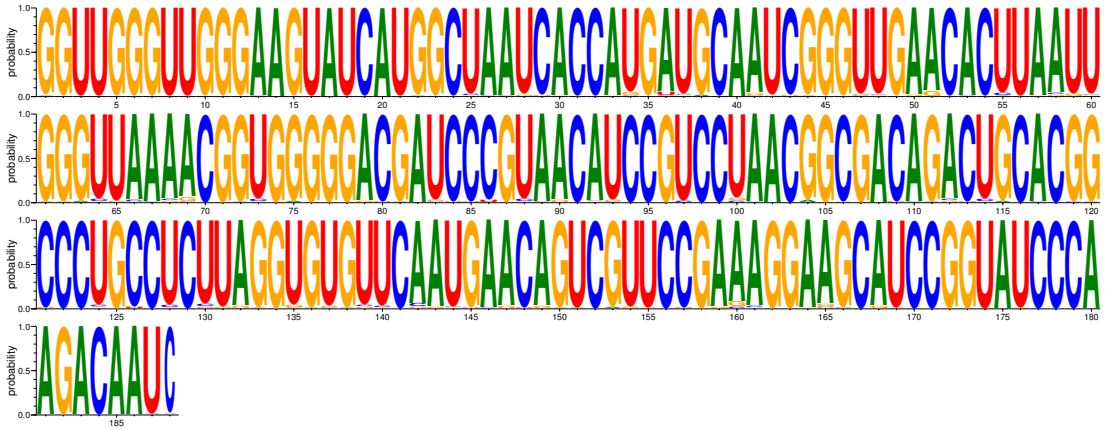


Figure 6.1: **Mutation distribution after a non-directed mutagenesis PCR of 30 steps over 92 chosen mutants.**

As a second experimental step, variants molecules were simultaneously treated by `SHAPEMap`, then a cDNAs library preparation was performed to start sequencing. For size limitation reasons, solely a sample of molecules from the library, for which we assumed the presence of the all variants, was sequenced.

For the analysis, a simultaneous reads mapping was performed with `TMAP` against the set of mutants; it was based on the minimal nucleotide distance. We first noticed a large amount of data loss due to the inability of classic mapping algorithms to deal with the ubiquitous mutations. And, surprisingly reads supposed to be mapped with a MAP-Q value above 20 are indeed miss-mapped.

We tested our suggested EM algorithm to overcome this mapping issue. Due to the complexity of the algorithm we could not perform the reads assignment. Then, we concluded that our greedy EM-assignment algorithm needs to be optimized. In meanwhile, to evaluate the mapping ability of our suggested EM-assignment, in the presence of a multi-source of mutations, we built a set of simulated reads derived from a limited number of variants.

Validation of the EM assignment with simulated mutated reads. To simulate `SHAPE`-mutated reads, we started by arbitrary picking up 5 RNA variants from the pool including the WT. The corresponding alignment for the set of chosen RNAs is illustrated in Figure 6.2.

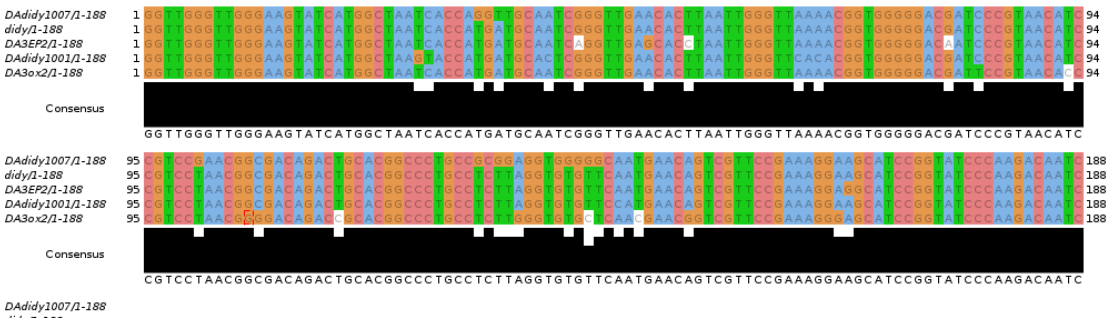


Figure 6.2: **Alignment of 5** GIR1 Lariat-capping ribozyme **mutants sequences used to build the test dataset. The figure was obtained using** `Jalview` **software.**

To generate `SHAPE` mutated reads we adopted the following framework:

1. Random generation of a secondary structure for each single RNA in the set, where the rate of a position to be unpaired has been fixed to 0.005.

2. Random generation of `SHAPE` mutations for a set of 10000 replicates where the mutation rate for an unpaired position was set to 0.002. At the end of this step, only non-redundant replicates were selected as input for the sequencing simulator.

3. HTS using the sequencing read simulator `Curisim` [Caboche et al., 2014], with the command line:
`java -jar CuReSim.jar -f variants.fa -m 188 -n 1000000`
We considered complete reads with the full length of 188 and set the number of generated reads to 10^7.

After the generation of the sets of simulated reads, we first mapped back the reads to the WT sequence using `TMAP` and then we retrieved all reads by setting a `MapQ` value to 0. As the first iteration of our algorithm suggests, a mapping based on the minimal distance to all variants was performed.
For our benchmark, we simulated three sets of `SHAPE`-mutated reads.

To assess the reads assignment performance, we run our EM-assignment program with 1000 iterations. We reported, in Table 6.1, the number of correctly mapped reads after the first iteration (that corresponds to a simple assignment based on the minimal distance to all variants), after the second and the 1000^{th} iterations, that correspond respectively to the first and last instance of the EM algorithm, and at the i^{th} iteration where the difference in the log-likelihood to the previous iteration "shift" is almost null.

At first glance, the most surprising results concern the colossal loss of reads with a fairly low MapQ value in the case of reads mapping with `TMAP`. On the other side, our method has shown promising results: for the three sets,a noticeable improvement was detected as early as the second iteration, i.e. after the first EM instance.

For the two sets: 1 and 2, the maximal value for correctly mapped reads was achieved after the 1000^{th} iteration. In addition, at iterations with almost a null likelihood shift, good performances were achieved. For the pool number 3, a fluctuation in the number of correctly mapped reads through the iterations was observed.

				EM-assignment		TMAP (MapQ ≤ 1)	
Set	Iteration	shift	ET (s)	correct	incorrect	correct	incorrect
1	1	390587		292	2144	239	148
	2	2126		295	2141		
	117	0.02		1444	992		
	1000	0.4	5713.5	1538	898		
2	1	584765		983	1243	215	2
	2	2089		1096	1130		
	295	0		1122	1104		
	1000	-1.75	5537.8	1138	1088		
3	1	135363.		1176	785	44	4
	2	41.9		1293	668		
	42	0.75		604	1357		
	133	-0.05		569	1392		
	382	0.34		637	1324		
	1000	0.95	4905.67	716	1245		

Table 6.1: **Number of correctly/incorrectly mapped reads using `TMAP` and the EM-assignment algorithm.** The Execution Time (ET) is reported for the last iteration.

Conclusion

The promising preliminary results prompt to test our algorithm on a real-life examples. However, an optimization seems to be necessary for the algorithm that has shown to be time consuming; It requires an execution time of about $5s$ per iteration for 2000 reads with an average read length of 188. The convergence of the EM-assignment was not guaranteed in all cases: it is possible to converge to degenerate solutions (the case of the aberrant mutation profiles that get most of the reads assigned to it thus causing a divergence from the optimal solution(s)) or to a local maxima. There is also a need to consider further parameters in the EM model such as a Bayesian prior to penalize non credible `SHAPE` mutational rates. Finally, defining the adequate number of iterations to perform remains one of the point to be addressed.

Symbol	Definition
d	Number of RNA variants
$L(.)$	Log-likelihood
m	Mutated state
$M_j(.)$	Mutational profile of the variant j
$\mathcal{N}$	Length of RNA sequence
$P(m, w_k)$	probability of the position w_k to have the state m
r	Number of reads
W	RNA sequence
θ	Emission probabilities
V	Set of variants
X	Set of reads
x_i	A read delimited by $[s_i, e_i]$
Z	Missing values

Table 6.2: **Glossary of symbols for EM-assignment algorithm**

Part III

In this part, we first describe a validation of the predictive capacity of `IPANEMAP` in the presence of both mono-probing and multi-probing data (Chapter 7), then we present concrete applications on RNA models including some with unresolved structures and little-known biological functions (Chapter 8). We conclude with a discussion of areas for improvement (Chapter 9).

Chapter 7

Validating the predictive capacity of IPANEMAP

Our objective here was to test with IPANEMAP the complementarity of several sources of probing data and to gain in the precision of predictions.

We run IPANEMAP with a set of RNAs for which probing data were available and structures were known.

As the set of RNAs with publicly available sets of probing data is small, we benchmarked IPANEMAP on a set of RNAs with resolved structures (experimentally or by comparative approaches), for which we simulated the probing data. We compared the performance of IPANEMAP with Rsample. This competitive tool has shown to outperform other tools while integrating structural constraints ensuing from one source of probing data, that we referred to by 'the mono-probing' case.

7.1 Validating IPANEMAP

7.1.1 Dataset

To evaluate the performance of our program IPANEMAP with mono-probing data or bi/multi-probing data sources, we considered 6 RNAs from Cordero et al. [2012] listed in Table 7.1, distinguished by their known structures and three experimental probing: NMIA-SHAPE, DMS and CMCT. We also considered 24 sequences from [Hajdin et al., 2013] with known structures and one available chemical probing data 1M7-SHAPE. This dataset involves a variety of organisms spanning on different

PDB ID	Description	Length
1EVV	yeast phenylalanine Transfer RNA	76
1Y26	A-riboswitch Vibrio vulnificus	70
2GDI	thiamine pyrophosphate-specific riboswitch	78
3IRW	c-di-GMP riboswitch from V. cholerae	88
3P49	Glycine Riboswitch from Fusobacterium nucleatum	159
2R8S	P4-P6 RNA ribozyme domain	161

Table 7.1: **PDB IDs for RNAs in the test dataset from Cordero et al. [2012].**

length ranges with a considerable amount of riboswitches.

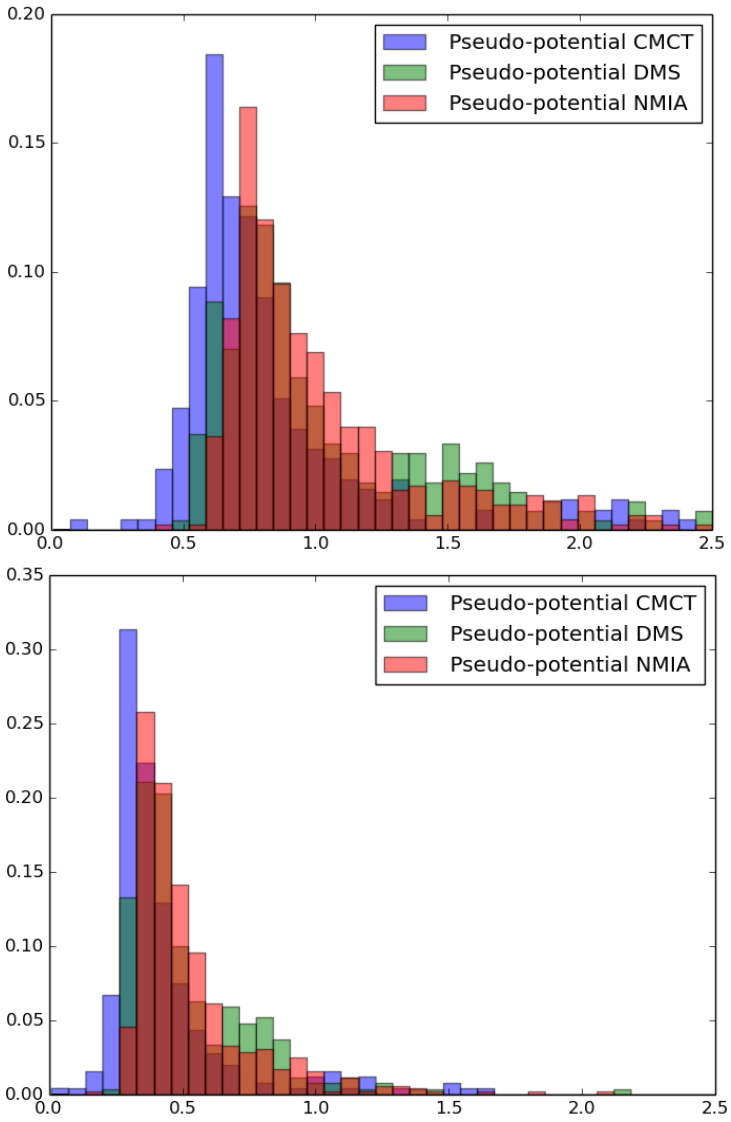
7.1.2 Probing data contribution

Probing data were converted into a pseudo-free energy δG as suggested by Deigan et al. [2009]. The authors suggest Equation 7.1 to consider the soft-constraints in the prediction model:

$$\delta G = m \times ln(reactivity + 1) + b \tag{7.1}$$

The optimal values for m and b were found to be respectively equal to 2.6 and -0.8 when benchmarking against a dataset of `SHAPE` data.

Pseudo-potentials, calculated from the reactivity values according to Equation 7.1 have shown to follow a normal distribution as depicted in Figure 7.1. However, when comparing pseudo-potentials from different probing sources, we noticed that their distributions differ in means and in kurtosis. This led us to assume the sensitivity of the pseudo-energy contribution to the probing data nature. Thus, there is a need for a correction term to ensure a common interpretability of the pseudo-energy for all considered probing data sources.
One may think about considering the `NMIA-SHAPE` distribution as a distribution of reference then estimate the quotient distribution. This quotient is a ratio that allows a given distribution to converge to a given distribution.
However, this trick is subject to fitting errors accumulation. To ensure the convergence of all distributions to the one of reference, we rather choose to act on m and b values. We attributed different values for the two parameters within a fixed range



X-axis corresponds to pseudo-potentials and Y-axis to their respective normalized frequencies

Figure 7.1: **Pseudo-energy distribution for three probing data from 6 RNAs** [Cordero et al., 2012] with the parameters of the Equation 7.1 set to `m=2.6, b=-0.8` [left] and `m=1.3, b=-0.4` [right].

and derived pseudo-potentials were compared. We found that `(m=1.3,b=-0.4)` allowed to obtain comparable distributions between the tested probing sources as illustrated in Figure 7.1. This new parametrization, that indeed divides the

pseudo-energy contribution by 2, allowed for a more accurate predicted structures using `IPANEMAP` when tested with 12 RNAs from [Hajdin et al., 2013], RNAs for which the use of probing data has shown to lead to noticeable accuracy improvement, with 1M7-SHAPE probing as shown in Table 7.2. Moreover, `(m=1.3,b=-0.4)` belongs to the accepted range of optimal parameters defined by Deigan et al. [2009].

			MCC					
RNA	Length	Unpaired	δG	$\frac{\delta G}{2}$	$\frac{\delta G}{3}$	$\frac{\delta G}{4}$	$\frac{\delta G}{6}$	$\frac{\delta G}{8}$
cyclicdiGMPriboswitchVcholerae	97	42	**0.9474**	**0.9474**	0.8421	0.8421	0.8421	0.8421
GroupIIntronTthermophila	425	169	**0.8941**	0.8231	0.8137	0.8137	0.8106	0.8106
TPPriboswitchEcoli	79	36	**0.913**	**0.913**	0.8095	0.7143	0.8095	0.8095
P546domainbI3groupIintron	155	41	0.9636	**0.9818**	0.955	0.955	0.9464	0.9464
5domainof16SrRNAEcoli	530	234	**0.8489**	0.8397	0.8285	0.8526	0.7832	0.7727
RNasePBsubtilis	401	177	0.6789	**0.7**	0.6878	0.6244	0.614	0.614
SARScoronaviruspseudoknot	82	37	**0.8**	**0.8**	**0.8**	**0.8**	**0.8**	0.766
HIV15primepseudoknotdomain	500	204	0.4842	0.5086	0.4983	**0.5172**	0.3793	0.3724
HepatitisCvirusIRESdomain	336	134	**0.8615**	0.8557	0.8557	0.8557	0.6599	0.7513
5SrRNAEcoli	120	50	0.9296	**0.9444**	0.8	0.8	0.8	0.274
SignalrecognitonparticleRNAhuman	301	101	0.58	0.58	**0.5918**	0.3385	0.3402	0.3385
Telomerasepseudoknothuman	42	23	0.7368	**0.8**	**0.8**	**0.8**	0	0

Table 7.2: **Accuracy of the predicted structures through `IPANEMAP` with different pseudo-energy contributions**: the structure of 12 RNAs was predicted with mono-probing `SHAPE`-1M7 data. Different pseudo-potential δG contributions were tested with δG corresponding to the pseudo-energy obtained by sitting parameters to $m = 2.6$ and $b = -0.8$. The contribution of a pseudo-energy of $\frac{\delta G}{2}$ is outperforming the other contribution values. Maximal accuracy values are highlighted in bold.

7.1.3 Method and results

We used 6 RNAs from [Cordero et al., 2012]. First, we predicted RNA structures with the classic prediction model using RNAfold, the resulting MCC are presented in Table 7.3. The use of DMS probing data has shown to remarkably improve the accuracy of predicted structures compared to CMCT or to SHAPE-NMIA data. In a second time, we used IPANEMAP under 3 cases: without probing data (to evaluate the effect of thermodynamic ensemble, we denoted this case by "sampling"), with a mono-probing data and with the all possible combinations of probing data.

We aimed at comparing our integrative method, for mono/bi/multi-probing data, to the classic approach and at showing its ability to lead to more accurate predictions in the presence of more than one probing data.

Compared to the "sampling" case, the use of probing data with IPANEMAP allowed to achieve better MCC for all RNAs apart for the adenine riboswitch: an RNA for which the use of probing data weakened the predictive power of our modeling and even of the classic prediction model.

Structure prediction for the 6 RNAs through IPANEMAP allowed to obtain comparable if not better accuracies in the presence of more than one probing data.

Additionally, in the presence of multi-probing data only 2 clusters were formed during the clustering process (except for one of the bi-probing combination involving NMIA and CMCT data, that returned 3 clusters) as shown in Table 7.5. Moreover, over all the possible combinations, solely one structure has found to be the optimal. This rapid assessment of the clustering process and the uniqueness of the predicted structure strengthen the hypothesis of complementarity between various probing data, especially when the considered probing data inform about a unique conformation.

In the next paragraph, we present in details the resulting structures from our modeling for 3 RNAs (for which a noticeable difference in accuracy was detected). In particular, the glycine riboswitch seems to be the good candidate to approve our integrative modeling. Indeed, the MEA structure predicted with a classic probing guided modeling reached in maximum an accuracy of 0.7 (surprisingly with the less performing probing reagent: the CMCT) while the use of multi-probing data with IPANEMAP allowed for an accuracy of 0.87 and more interestingly a value of 0.95 with DMS mono-probing.

RNA	MFE	MEA	MFE-DMS	MEA-DMS	MFE-CMCT	MEA-CMCT	MFE-NMIA	MEA-NMIA
5S rRNA	0.241	0.241	0.686	0.686	0.254	0.269	0.686	0.686
glycine rbs.	0.306	0.593	0.313	0.593	0.395	0.693	0.313	0.6
cdGMP rbs.	0.77	0.77	0.77	0.72	0.667	0.667	0.77	0.77
P4-P6 RNA	0.837	0.845	0.808	0.808	0.773	0.781	0.714	0.722
adenine rbs.	1	1	0.333	0.333	0.41	0.356	1	0.306
tRNA^{phi}	0.976	0.334	0.976	0.976	0.976	0.976	0.976	0.976

Table 7.3: **Accuracy of the predicted `MFE` and `MEA` with classic modeling:** the accuracy is reported as the geometric mean of the `sensitivity` and the `PPV`. Two conditions were distinguished: with and without one source of probing data.

RNA	Sampling	DMS	CMCT	NMIA	DMS+CMCT	DMS+NMIA	NMIA+CMCT	DMS+NMIA+CMCT
5S rRNA	0.25	0.244	0.238	0.247	0.247	0.247	0.254	0.241
glycine rbs.	0.568	0.952	0.627	0.868	0.658	0.868	0.868	0.868
cdGMP rbs.	0.654	0.77	0.77	0.654	0.77	0.654	0.654	0.77
P4-P6 RNA	0.864	0.864	0.856	0.881	0.856	0.864	0.864	0.856
adenine rbs.	1	0.977	0.956	1	0.956	1	1	1
tRNA^{phi}	0.286	0.746	0.746	0.976	0.746	0.976	0.976	0.976

Table 7.4: **Accuracy of the predicted structures through `IPANEMAP`:** the accuracy is reported as the geometric mean of the `sensitivity` and the `PPV`. Several conditions were analysed: in the absence of probing data and under the case of mono-, bi- and multi-probing data.

RNA	sampling	DMS	CMCT	NMIA	DMS+CMCT	DMS+NMIA	NMIA+CMCT	DMS+NMIA+CMCT
5S rRNA	**2**	4	**2**	**2**	**2**	**2**	**2**	**2**
glycine rbs.	**2**	**2**	**2**	**2**	**2**	**2**	3	**2**
cdGMP rbs.	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
P4-P6 RNA	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
adenine rbs.	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
tRNA^{phi}	5	**2**	**2**	**2**	**2**	**2**	**2**	**2**

Table 7.5: **Optimal cluster numbers reported by the clustering module integrated in `IPANEMAP`.**

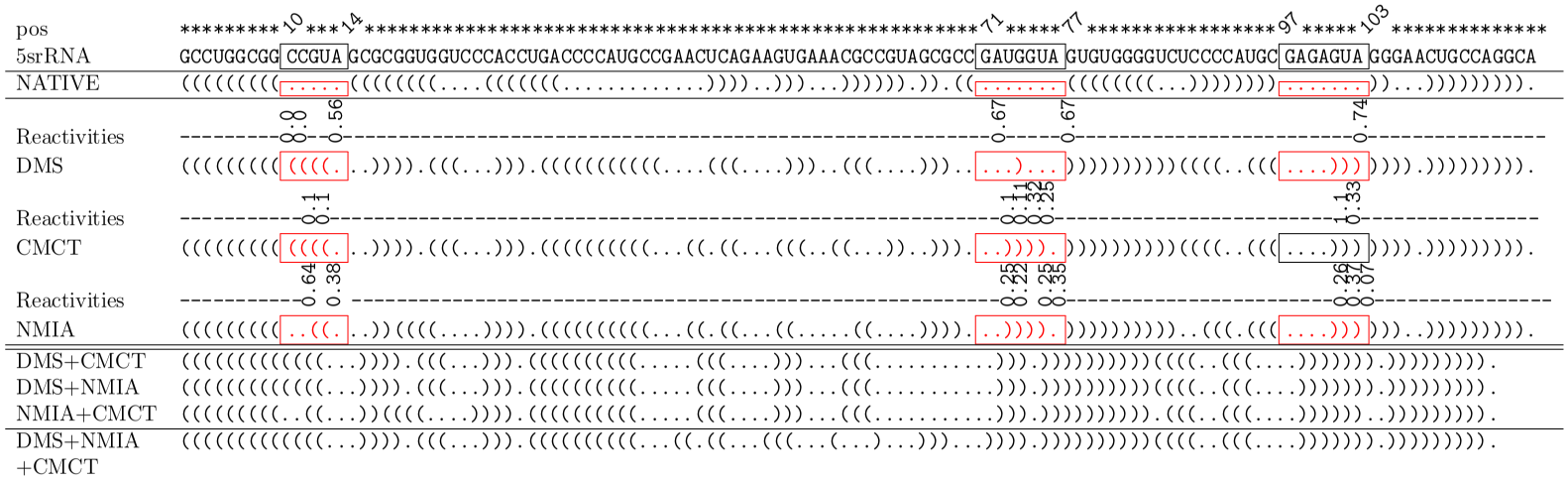


Figure 7.2: **`5s rRNA` predicted structures using `IPANEMAP`:** we identified 3 single-strand regions in the native structures (in red) where the dominant nucleotides are `G` and `U`, and for which `NMIA` and `CMCT` conditions are showing a moderate reactivity. The lack of `DMS` covering and the moderate values reported by the other probing data favored the prediction of stacked nucleotides and subsequently the prediction of an alternative conformation. Although our predictions do not allow to get acceptable `MCC` as observed with the classic probing guided predictions, the combination of 3 probing data allowed for an accuracy surpassing the minimal value obtained under the mono-probing case.

cdGMP rbs.	CACGCACAGGGCAAACCAUUCGAAAGAGUGGGACGCAAAGCCUCCGGCCUAAACCAGAAGACAUGGUAGGUAGCGGGGUUACCGAUGGCAAAAUG
NATIVE	((((......((...((((((....))))))...))...(((.(((((((..((.........)))))))..))))))...)).)).........
sampling	...((...((((...((((((....))))))...))..((((((.((((..((((.......))))))))..).))))))..))....)).....
DMS	...((...((((...((((((....))))))...))..((((.((((((..((((.......))))))))..))))))..))....)).....
CMCT	...((...((((...((((((....))))))...))..((((.((((((..((((.......))))))))..))))))..))....)).....
NMIA	...((...((((...((((((....))))))...))..((((((.((((..((((.......))))))))..).))))))..))....)).....
DMS+CMCT	...((...((((...((((((....))))))...))..((((.((((((..((((.......))))))))..))))))..))....)).....
DMS+NMIA	...((...((((...((((((....))))))...))..((((((.((((..((((.......))))))))..).))))))..))....)).....
NMIA+CMCT	...((...((((...((((((....))))))...))..((((((.((((..((((.......))))))))..).))))))..))....)).....
DMS+NMIA +CMCT	...((...((((...((((((....))))))...))..((((.((((((..((((.......))))))))..))))))..))....)).....

Figure 7.3: **cdGMP rbs. predicted structures**: through all the tested cases, IPANEMAP returned two dominant conformations (pointed by red/blue colors) with a difference of two base pairs. In mono-probing case with NMIA, the predicted centroid is the same as from "sampling". This structure is also predicted when NMIA intervenes in a bi-probing combination. As the NMIA probing data covers the whole sequence, one may think that DMS or CMCT probing data is less contributing then that of NMIA, which could be true since the combination of the three probing data allowed to find again the structure supported by DMS and CMCT that remained the closest to the native structure.

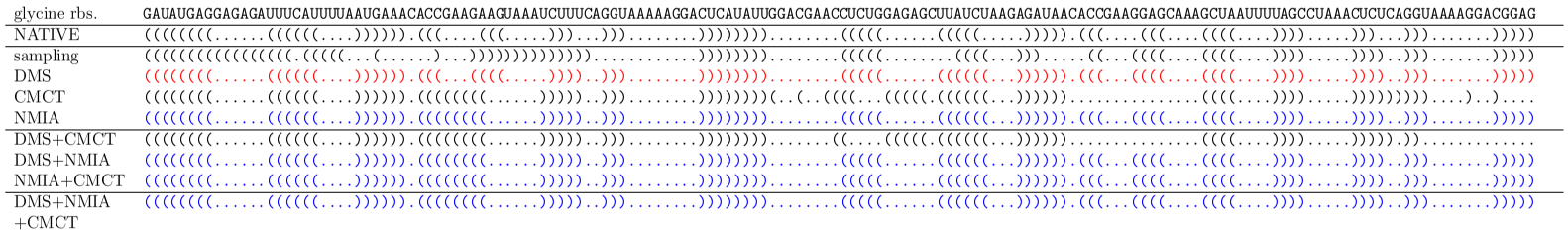
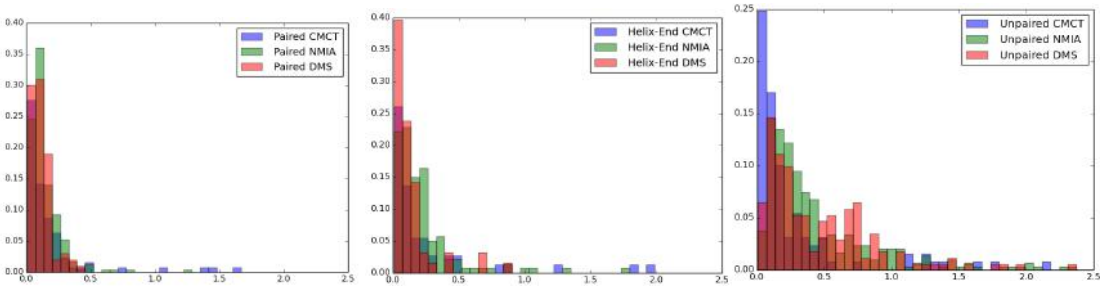


Figure 7.4: **Predicted structures for glycine riboswitch:** this RNA is characterized by the existence of two hairpin-loop substructures. In the absence of probing data: "sampling", the first hairpin loop is not correctly predicted. `DMS` mono-probing allowed to predict an accurate structure (in red) with 3 more base pairs (located at the helix-end level) compared to the native, thus achieving an accuracy of 95%. In the case of `CMCT` mono-probing, the second hairpin loop is not correctly predicted. However, the use of an additional probing data allowed to gain in accuracy especially with `NMIA` probing. The use of `NMIA` probing in all possible combinations leaded to the same structure (in blue) thus, reflecting a dominance of this probing data.

7.2 Benchmark on simulated probing data

7.2.1 Simulation of probing data

The aim of this section is twofold: to visualize the impact of the probing data nature on the structure at the nucleotide level, and to construct a probabilistic model to simulate probing data. To this purpose, probing data from experimentally resolved structures [Cordero et al., 2012] were used to build the normalized distribution of reactivities for three structural context namely paired, helix-end and unpaired. Resulting distributions are displayed in Figure 7.5.



X-axis corresponds to Reactivities and Y-axis to their respective normalized frequencies

Figure 7.5: **Reactivity distributions for three structural contexts: paired, helix-end and unpaired**: for `DMS`, `CMCT` and `NMIA-SHAPE` probing data.

We used these distributions to build our probabilistic model. First, we discretized the space with 500 bins then we computed the `mid-value` for each bin, a parameter considered later on with its corresponding `probability` to build the generative model. To simulate the probing data we used the `random` function from `numpy` library for each single structural category and each specific probing data:

```
import numpy as np
def Random(mid-value,probability):
    return  np.random.choice(mid-value,1,p=probability)
```

7.2.2 Method and results

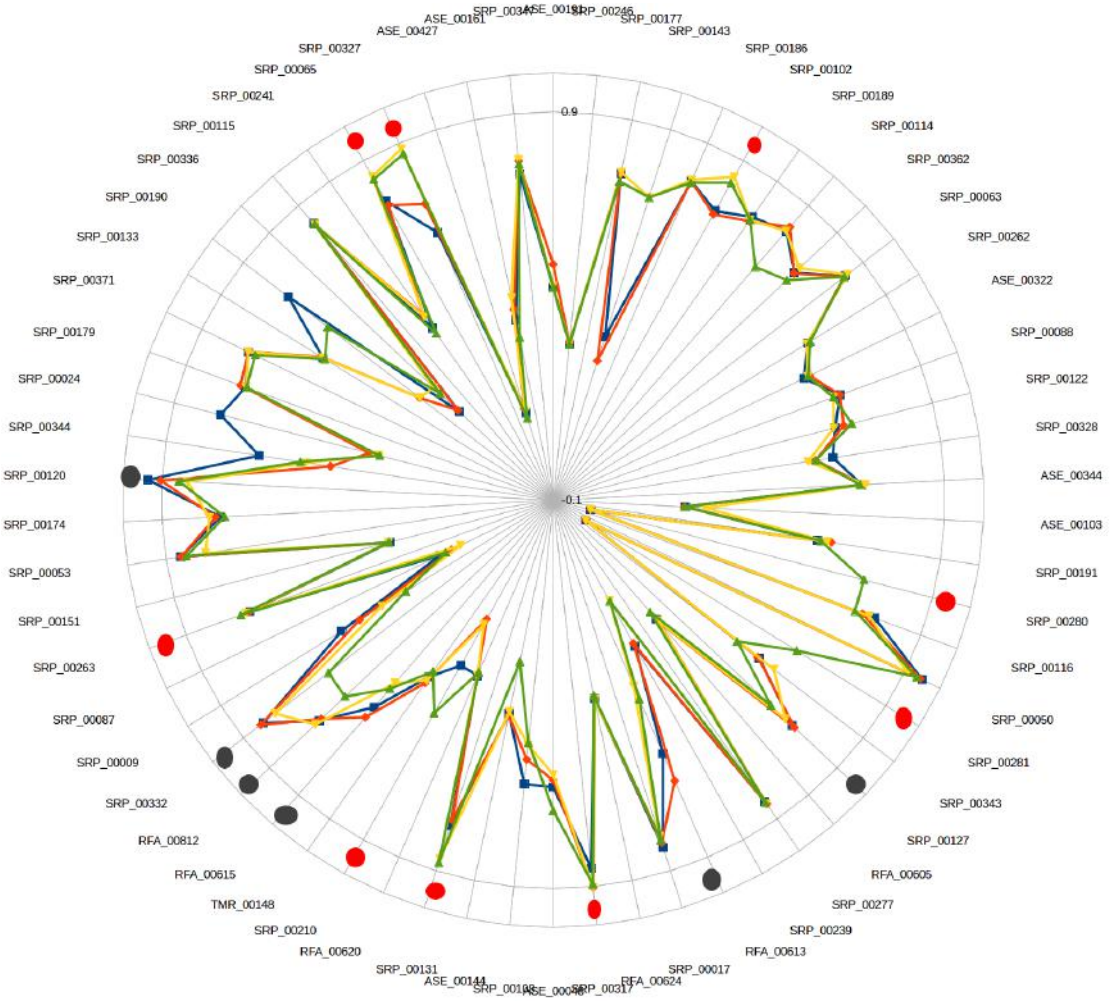
We imported 62 RNAs from `RNAstrand` database [Andronescu et al., 2008], for which the structure is resolved either experimentally or through comparative analysis. The chosen parameters with their correponding values are displayed in Table 7.6.

Parameter	Value/Range
Length	Between 150 and 500
Duplicates	Non redundant sequence only
Number of multi-loops per molecule	Greater than or equal to 1
Number of pseudoknots per molecule	Less or equal to 0

Table 7.6: `RNAstrand` **parameters setting to get RNAs with resolved structures.**

We simulated probing data for this set of RNAs via our generative model. Then, we used the resulting probing data with the RNA sequences to model structures via `IPANEMAP`. The performance of our predictive tool was assessed as the `MCC` then compared to the classic modeling. From the results of the benchmark, we observed the ability of `DMS` probing to lead to more accurate structures compared to the classic predictions (as shown in Figure 7.6) and to other simulated probing data (as displayed in Figure 7.7). Furthermore, we observed that the use of `CMCT` simulated data does not improve the accuracy but on the contrary deteriorates the quality of the prediction 7.7.

The use of `SHAPE-NMIA` along with `DMS` and `CMCT` probing data allowed a slight gain in accuracy for some RNAs as shown in Figure 7.8. In addition, `IPANEMAP` with `DMS` mono-probing has shown to be as efficient as with the multi-probing case.



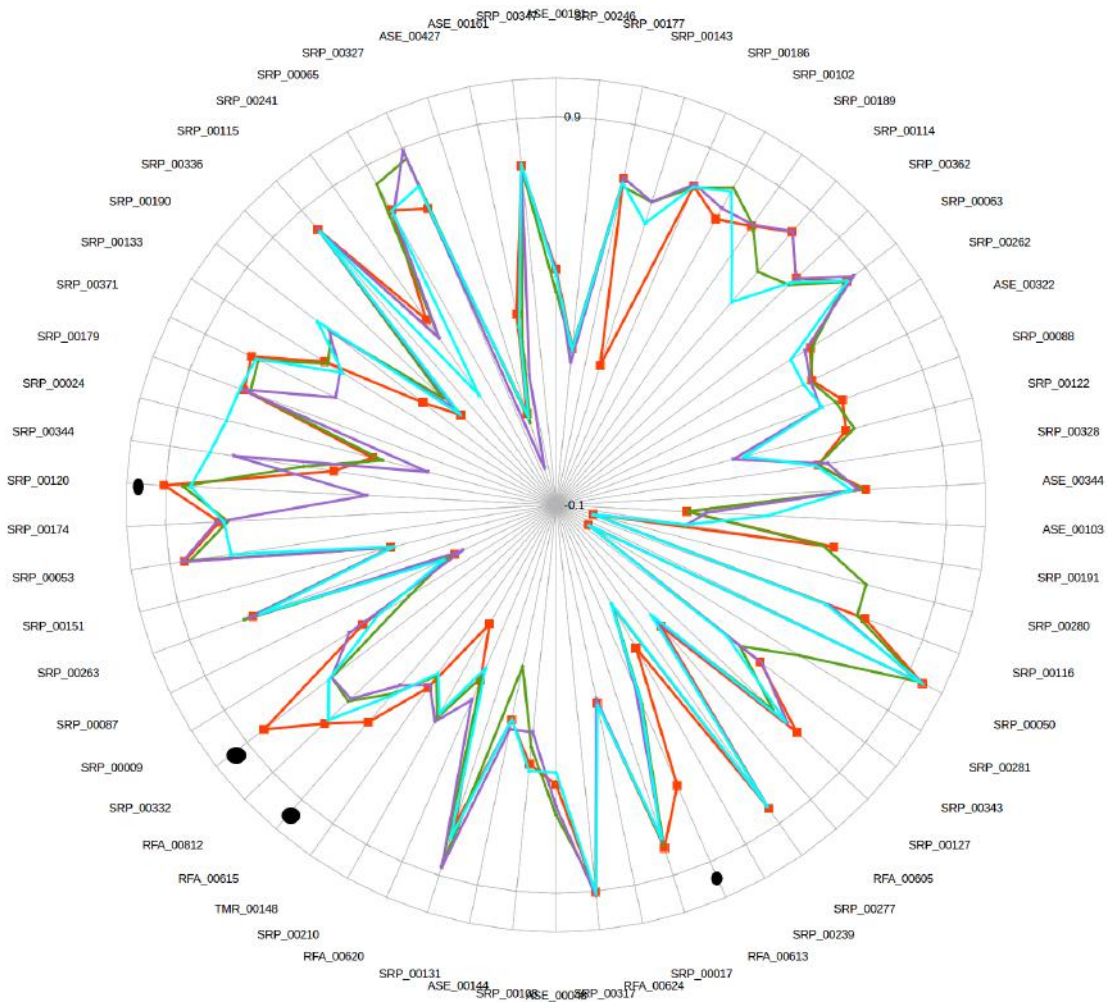
↪: MFE without probing data

↪: IPANEMAP sampling without probing data

↪: MEA without probing data

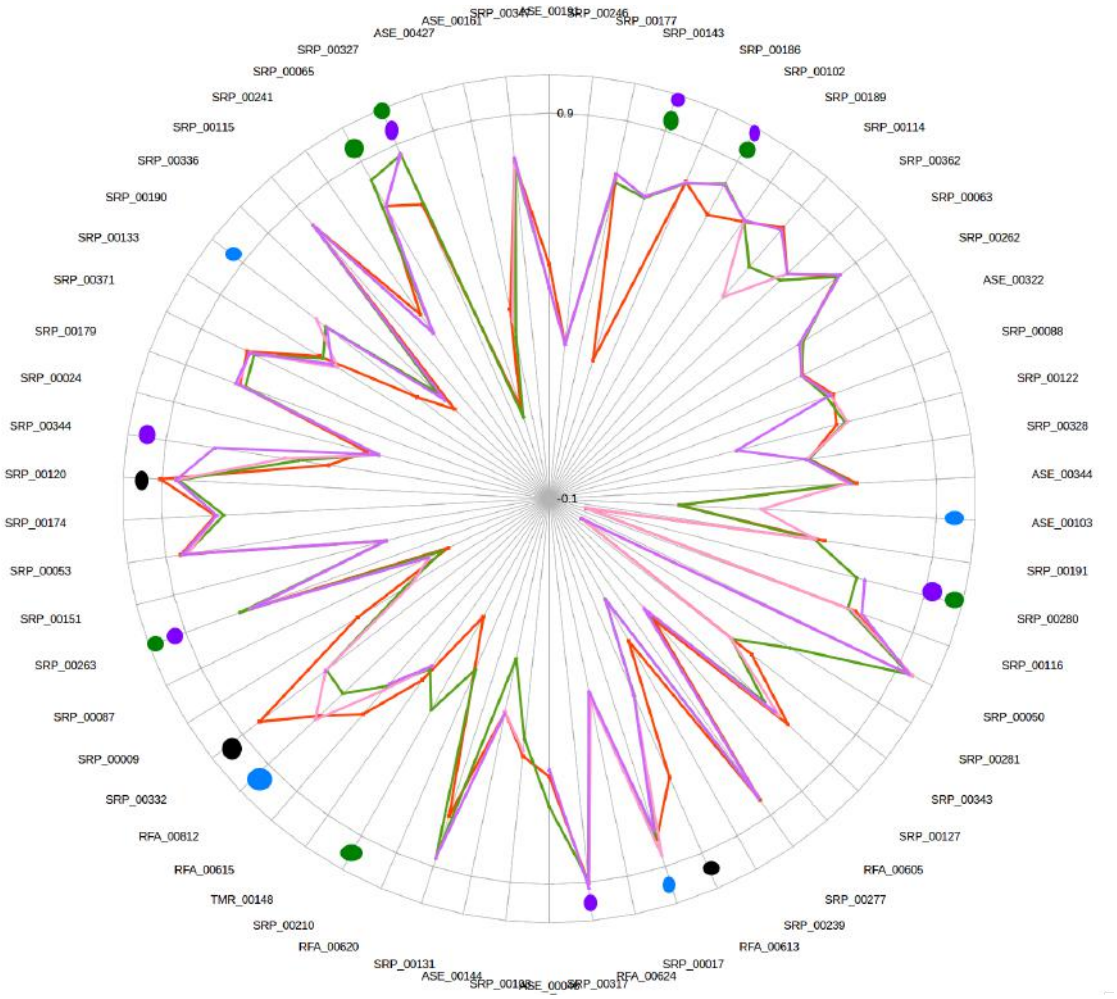
↪: IPANEMAP simulated DMS mono-probing case

Figure 7.6: **MCC for predicted structures through IPANEMAP with simulated DMS probing data compared to the MFE and the MEA structures:** a comparable performance of IPANEMAP with the DMS mono-probing is observed in comparison with the MFE and the MEA structures obtained through a classic prediction except for few RNAs pointed by black circle. A noticeable improvement with the use of our tool was detected for certain RNAs (pointed by red circle).



↪: MEA without probing data ↪: IPANEMAP simulated DMS mono-probing
↪: simulated NMIA mono-probing ↪: IPANEMAP simulated CMCT mono-probing

Figure 7.7: **Accuracy of predicted structures in mono-probing case using IPANEMAP with simulated data:** with the exception of a few RNAs, a comparable if not a better performance is observed when using IPANEMAP, with one of the simulated probing data, compared to the predicted MEA with RNAfold. Circles in black shows the cases where the MEA is performing better than IPANEMAP with one probing data.



↪: MEA without probing data ↪: IPANEMAP simulated DMS mono-probing
↪: simulated NMIA-DMS ↪: IPANEMAP simulated NMIA-DMS-CMCT

Figure 7.8: **Accuracy of predicted structures in the multi-probing case with simulated data using IPANEMAP.**: blue circles show RNA where the NMIA-DMS combination allowed to obtain better representative structures. green (resp. purple) circles highlight the out-performance of DMS mono-probing (resp. NMIA+DMS+CMCT multi-probing).

7.3 Comparison of IPANEMAP with other tools

To position our developed method IPANEMAP in relation to similar mono-guided sampling/clustering based algorithm such as Rsample [Spasic et al., 2018], we performed a test with a dataset of 24 non coding RNAs with resolved structures and one source of probing data: 1M7-SHAPE.
Although Rsample is dedicated to detect RNA multi-conformers, it has shown to outperform up to date methods for predicting single conformation within a mono-probing case. The comparison of the predicted structures represented by their accuracies is illustrated in Figure 7.9. Over all predictions, a remarkable improvement is detected with the use of IPANEMAP notably for the 5S rRNA Ecoli structure which could not be previously resolved with other source of probing data (Figure 7.2). The PPV, the Sensitivity and the accuracy values are reported in Appendix [10B].

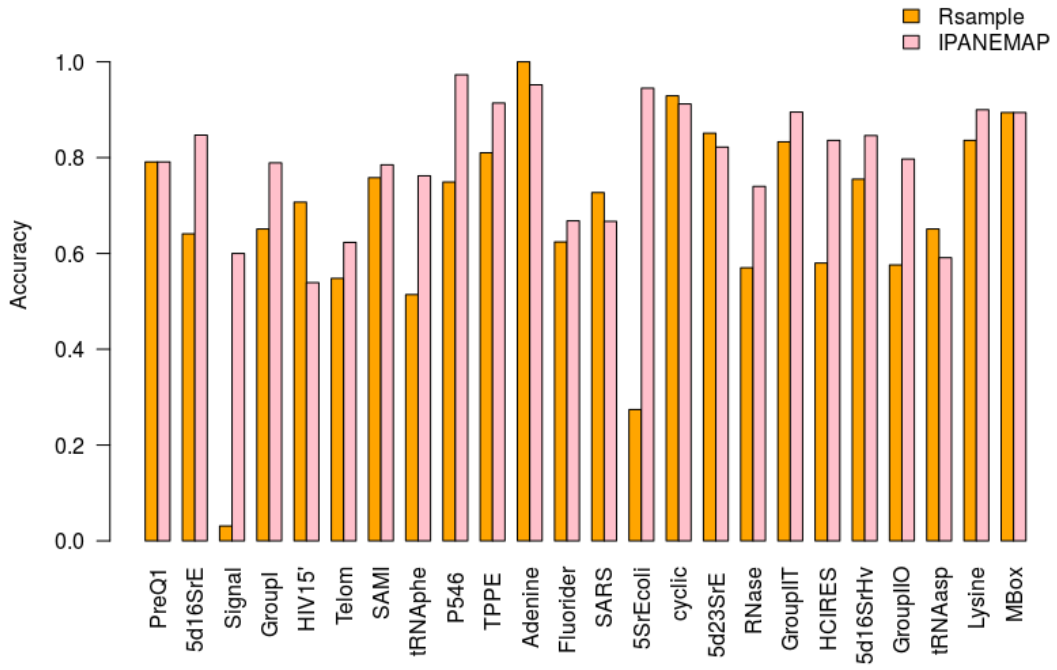


Figure 7.9: **Comparison of the predicting power between `Rsample` and `IPANEMAP` in the mono-probing case:** reported accuracy is assessed as the geometric mean of the `Sensitivity` and the `PPV`.

Chapter 8

Applications

In order to validate the experimental protocols and to produce structural insight matching the expertise of our collaborators, we have specifically studied the structure for three RNA models: GIR1 Lariat-capping ribozyme, `HIV-1 gag` and `Ebola UTRs`.

GIR1 Lariat-capping ribozyme is a ribosomal RNA characterized by an enzymatic function where the structure seems to be indispensable to ensure its activity. This RNA has a set of pseudo-knots that make it challenging to predict its structure.

From the other side, viruses are creatures subject to undergo rapid mutations by excellence. However, to maintain their functional activity they tend to conserve the structure. One of the particular characteristics of viruses concerns the initiation of the translation mechanism that might occur at different sites from the 5' end. This property has been shown to be the case for the `HIV-1 gag` where the identification of the structure through our multi-probing integrative method helped in this finding.

Furthermore, the molecular events leading to `Ebola` virus mRNA translation are unknown. As a first step towards the discovery of this aspect of Ebola virus life cycle, we were committed to characterize the structure of the 5' and 3' UTRs of the seven mRNAs from the virus with the aim to recognize structural motifs, common features or even structures known to be functional.

8.1 HIV-1 Gag-IRES

8.1.1 Dataset

The `HIV-1 Gag-IRES` RNA considered sequence is:

AUGGGUGCGAGAGCGUCGGUAUUAAGCGGGGGAGAAUUAGAUAAAUGGGAAAAAAUUCG
GUUAAGGCCAGGGGGAAAGAAACAAUAUAAACUAAAACAUAUAGUAUGGGCAAGCAGGG
AGCUAGAACGAUUCGCAGUUAAUCCUGGCCUUUUAGAGACAUCAGAAGGCUGUAGACAA
AUACUGGGACAGCUACAACCAUCCCUUCAGACAGGAUCAGAAGAACUUAGAUCAUUAUA
UAAUACAAUAGCAGUCCUCUAUUGUGUGCAUCAAAGGAUAGAUGUAAAAGACACCAAGG
AAGCCUUAGAUAAGAUAGAGGAAGAGCAAAACAAAAGUAAGAAAAAGGCACAGCAAGCA
AGCAGCUGACACAGGAAACAACAGCCAGGUCAGCCAAAAUUACCCUAUAGUGCAGAACC
UCCAGGGGCAAAUGGUACAUCAGGCCAUAUCA

The structure of the 444 nucleotides-long RNA was probed using several types of experimental techniques:

- SHAPE probing experiments with `NMIA` and `1M7` reagents;
- Enzymatic probing with `V1` RNAses and `T1` RNAses targeting respectively paired and unpaired nucleotides.

Probing data was used in our integrative model as pseudo-energy penalties for `SHAPE` data, and as hard constraints for the enzymatic data. In the late case, reactivities were compared to an arbitration value ϵ to decide for the residual structure i.e. a given nucleotide is considered to be paired/unpaired in function of its reactivity value compared to ϵ.

The integration of the enzymatic probing data as hard constraint was inspired by the best practice from experimentalists. A choice that could be also justified by the lack of evidence regarding the extension of the pseudo-energy contribution for probing data from non chemical protocols. To decide for the ϵ optimal value, a range of training values was fixed then the ensuing predicted ensembles were compared to the `SHAPE`-guided predicted ensemble and to the RNA structure from the literature. This comparison was performed via the calculation of the euclidean distance between the ensembles. The resulting distances are displayed in Figure 8.1. The ϵ value allowing the minimal distance to both the `SHAPE`-guided predicted ensemble and to the resolved structure while ensuring a reasonable number of constraints was selected.

The two chosen ϵ values were 1.0 and 1.5. Structural constraints for these values were then produced as described below with the example of `T1++` condition.

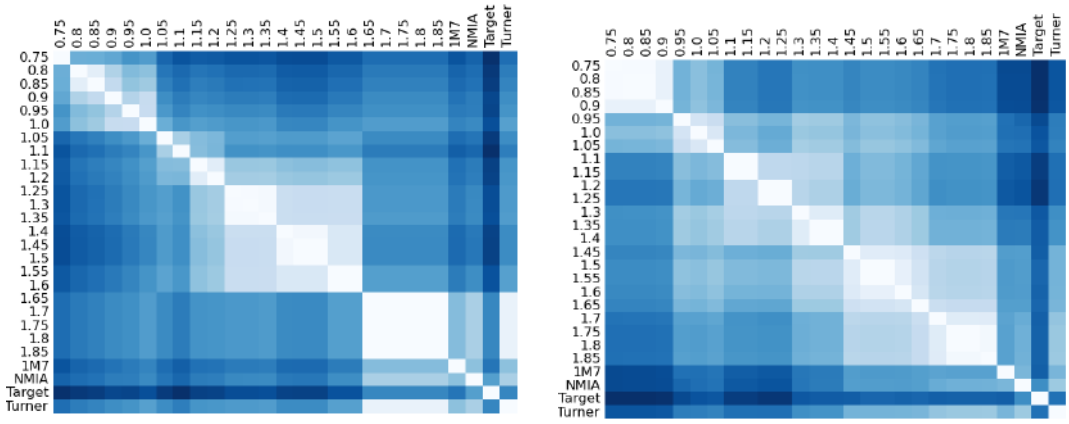


Figure 8.1: **Euclidean distance between predicted ensembles resulting from hard constraints guided predictions with the enzymatic probing `V1` (left) and `T1` (right) for different ϵ values**. In both heat-maps the darker the cell, the more distant the corresponding coordinates. Training ϵ values were ranging from 0.75 to 1.85. The distance to the `1M7`, `NMIA` , `Turner` and the proposed structure in the literature (denoted by `target`) predicted ensemble were also assessed.

ϵ	symbol	Constraints T1	Constraints V1
1.0	`+-`	31	26
1.5	`++`	14	8

Table 8.1: The number of residual structural constraints verifying a reactivity value above ϵ.

List of nucleotides highly reactive (`++`), case of enzymatic probing with `T1`:

[40, 103, 124, 133, 156, 173, 183, 207, 211, 212, 217, 262, 264, 272, 273, 277, 286, 294, 295, 298, 304, 309, 332, 336, 348, 352, 355, 403, 435, 436]

Structural constraint generated in Fasta format, with signs: 'x' to force the corresponding base to be unpaired, and '.' to not impose any constraint:

```
> T1++
AUGGGUGCGAGAGCGUCGGUAUUAAGCGGGGGAGAAUUAGAUAAAUGGGAAAAAAUUCGGUUAA
GGCCAGGGGGAAAGAAACAAUAUAAACUAAAACAUAUAGUAUGGGCAAGCAGGGAGCUAGAACG
AUUCGCAGUUAAUCCUGGCCUUUUAGAGACAUCAGAAGGCUGUAGACAAAUACUGGGACAGCUA
CAACCAUCCCUUCAGACAGGAUCAGAAGAACUUAGAUCAUUAUAUAAUACAAUAGCAGUCCUCU
AUUGUGUGCAUCAAAGGAUAGAUGUAAAAGACACCAAGGAAGCCUUAGAUAAGAUAGAGGAAGA
GCAAAACAAAAGUAAGAAAAAGGCACAGCAAGCAGCAGCUGACACAGGAAACAACAGCCAGGUC
AGCCAAAAUUACCCUAUAGUGCAGAACCUCCAGGGGCAAAUGGUACAUCAGGCCAUAUCA
.........................................x.......................
.....................................x.................x....
....x.....................x.............x.........x.........
.............x...xx....x.....................................
.....x.x.......xx...x........x........xx..x.....x....x.........
..........x...x..........x...x..x.....................x.......
.................x.............................xx........
```

When dealing with `V1` constraints, the symbol '|' is used to force a specific base to be paired.

The two experimental approaches led to 6 parameterizations for the structure prediction methods: `1M7`, `NMIA`, `V1++`, `V1+-`, `T1++`, and `T1+-` as described in Table 8.2.

Conditions	Description	Structural constraints	Incorporated as
V1++ V1+-	Enzymatic cleavage V	Most likely paired nucleotides Likely paired nucleotides	Hard constraints
T1++ T1+-	Enzymatic cleavage T	Most likely unpaired nucleotides Likely unpaired nucleotides	Hard constraints
1M7 NMIA	SHAPE with 1M7 SHAPE with NMIA	Reactivity score	Soft constraints (Pseudo-energy)

Table 8.2: `HIV1 gag` considered probing data and their respective integration mode in the prediction modeling with `RNAfold`.

8.1.2 Methods

We used the Multi-Dimensional Scaling (MDS) algorithm [Wickelmaier, 2003] , implemented in `scikit-learn` package [Pedregosa et al., 2012], to visualize the euclidean distance as a chosen measurement of the compatibility of structural models across the different probing conditions. The 2D distance matrix visualization is displayed in Figure 8.2.

We noticed that hard constraints in the mid-range `(+-)` case reduce the number of structures in the corresponding ensemble. A result that was expected given the number of imposed constraints (Table 8.1). We can not conclude for the structures verifying this condition neither to be accurate nor to be outliers. The presence of such "noisy" data and the absence of validation method could lead to erroneous results. This motivated us to explore in depth the ensemble of structures for each of the six studied conditions. Hence, the interest of our integrative method `IPANEMAP` that allows a close and sophisticated analysis of the predicted ensembles is omnipresent.

Sampling parameters

For each condition, an ensemble of 2000 structures was stochastically generated. This results into a set of 12000 structures covering the six conditions.
We used `RNAsubopt` from `Vienna` Package version 2.3, with options listed in Table

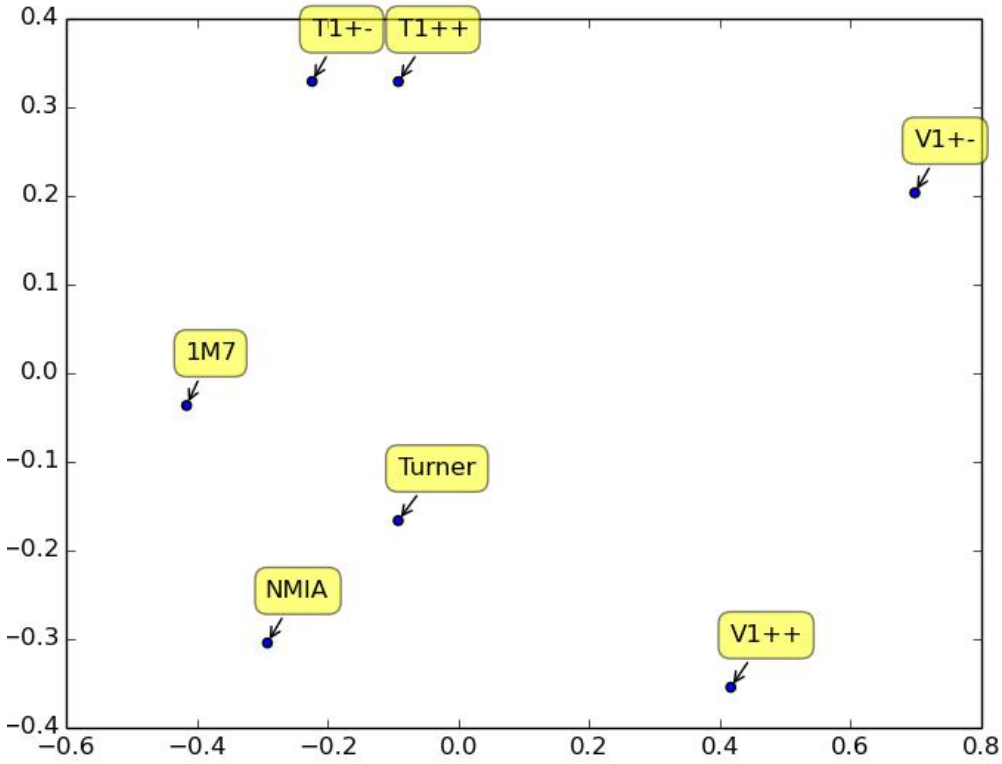


Figure 8.2: **Multi-Dimensional Scaling (MDS) to display the euclidean distance between predicted ensembles** where the horizontal axes present high-dimensional objects in typically two dimensions. In addition to the six conditions, the case without constraints pointed by `Turner` is considered. We observed the formation of a cluster, involving conditions from `[NMIA,1M7,V1++,T1++,T1+-]`, around the `Turner` model where the condition `[V1+-]` is distant from the cluster.

8.3. In furtherance of evaluating the folding energy of each generated structure, `RNAeval` from `Vienna` Package 2.3 was used.

Clustering and optimal structure definition

As the `HIV1-Gag IRES` predicted model was generated with a first version of `IPANEMAP` , the iterative clustering algorithm was not conceptualized yet. Therefore, the number of clusters was set to 6. Moreover, the minimization of the mean distance between structures in a given cluster was used as an additional criterion to decide for the optimal centroid(s). This criterion informs about the internal

Options	Description	Value
`-p`	Produce a random sample of suboptimal structures	2000
`-s`	Sort the suboptimal structures by energy	
`-T`	Temperature of experiments	37
`-C` `--enforceConstraint`	Calculate structures subject to constraints Enforce base pairs constraints	structural constraint file
`--shape`	Consider SHAPE reactivity data to guide structure predictions	SHAPE file

Table 8.3: RNAsubopt parametrization

coherence of the cluster and was calculated according to Equation 8.1.

$$\delta \mathsf{D}[\mathsf{c}] = \frac{\sum_{\mathsf{s_m},\mathsf{s_n} \in \mathsf{c}^2} \mathsf{P}(\mathsf{s_m}) \cdot \mathsf{P}(\mathsf{s_n}) \cdot \mathcal{BP}(\mathsf{s_m},\mathsf{s_n})}{\frac{\mathcal{I}_c^2}{2}}. \tag{8.1}$$

with $\mathcal{I}_c$ the probing condition cardinal of the cluster c, $P(s)$ the Boltzmann probability of the structure s in the sample, and $\mathcal{BP}(s_m, s_n)$ the base pairs distance separating the two structures s_m and s_n.

8.1.3 Results

One centroid figuring on the $3D$ Pareto frontier was found to be the optimal. To refine our model, we further analysed its base pairs in a comparative setting, using `R-Chie` [Lai et al., 2012] on sequences from the Los Alamos compendium [Yusim et al., 2016] to produce covariations. A visualization of the covariation is displayed in Figure 8.4. Most of our predicted base pairs were confirmed by the result of the covariation analysis. Figure 8.3 shows the structural model suggested through `IPANEMAP` with further refinement added by our experimentalists collaborators where a few base pairs that appeared to be somewhat reactive to one or another probe were eliminated.

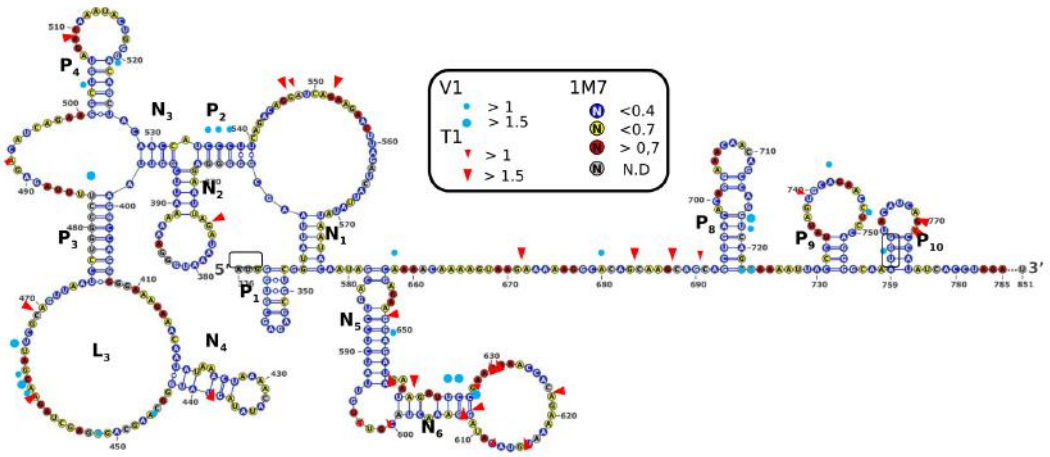


Figure 8.3: `HIV1 gag IRES` **predicted structure using** `IPANEMAP` **in the presence of** `SHAPE` **and enzymatic probing data as presented in the article** [Deforges et al., 2017].

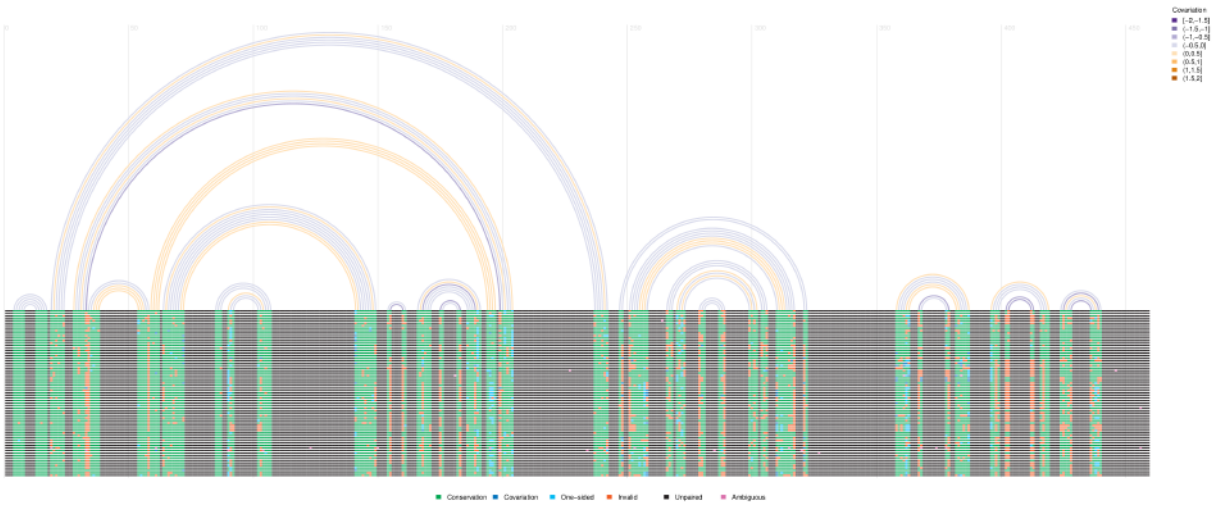


Figure 8.4: **Covariations over the** `MSA` **for** `HIV-1 gag` **considering** `IPANEMAP` **predicted structure.**

8.2 GIR1 Lariat-capping ribozyme

GIR1 Lariat-capping ribozyme, D. iridis is an RNA sequence of length 188. In addition to its known $3D$ structure, the availability of probing data generated in the laboratory of our collaborators made it a good candidate to produce a proof of concept for our integrative approach `IPANEMAP`.

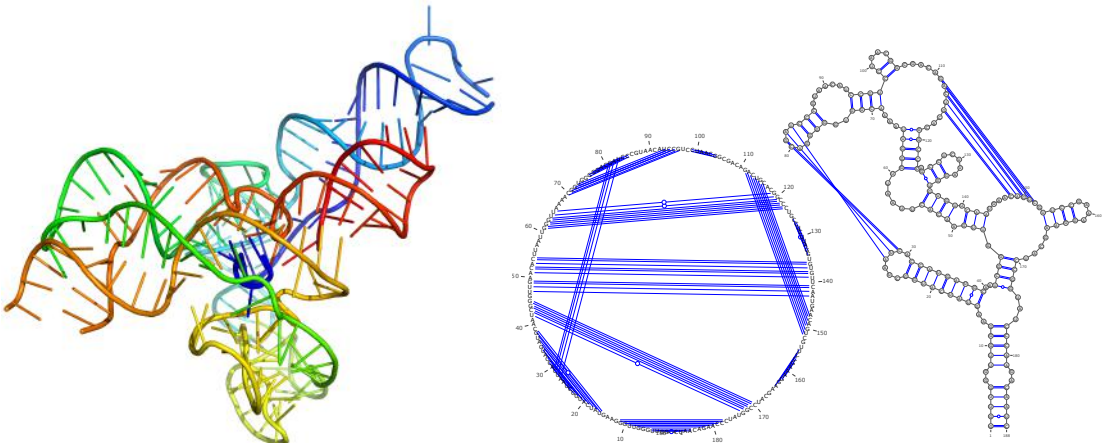


Figure 8.5: Visualization of GIR1 Lariat-capping ribozyme 3D structure with `PyMOL` [Stockwell, 2003] on the left, and its 2D projection with `VARNA` [Darty et al., 2009] on the right.

8.2.1 Dataset

We considered a set of 14 various experimental constraints while varying both the technology and the reagent as described in Table 8.8. Data analysis pipelines presented in Chapter 4 were used in order to generate reactivities for the different considered conditions.

8.2.2 Method and Results

First, we studied the influence of the probing data type when considered with the classic prediction by assessing the **Shannon entropy**: a measure to quantify the

Probing condition	Reagent	Technology
1M7ILUMg	1M7 + Mg^{2+}	SHAPEMap Illumina
1M7ILUMg3	Pool 3x amplified, 1M7 + Mg^{2+}	SHAPEMap Illumina
1M7ILU	1M7	SHAPEMap Illumina
1M7ILU3	Pool 3x amplified, 1M7	SHAPEMap Illumina
1M7	1M7	SHAPEMap IonTorrent
1M7Mg	1M7 + Mg^{2+}	SHAPEMap IonTorrent
NMIA	NMIA	SHAPEMap IonTorrent
NMIAMg	NMIA + Mg^{2+}	SHAPEMap IonTorrent
NMIAMgCE	NMIA + Mg	Chemical probing
BzCNMg	BzCN + Mg^{2+}	Chemical probing
CMCTMg	CMCT + Mg^{2+}	Chemical probing
DMSMg	DMS + Mg^{2+}	Chemical probing
NaiMg	Nai + Mg^{2+}	Chemical probing
Nai	Nai	Chemical probing

Table 8.4: GIR1 Lariat-capping ribozyme **probing conditions**

local structural variability of a given nucleotide. This entropy is computed as:

$$H = -\sum_{s} P(s) \ln P(s)$$

where $P(s)$ is the Boltzmann probability of the structure s in the predicted ensemble.

From the resulting entropy shown in Figure 8.6, we observed that the contribution of probing data is slightly weakened when considering $\frac{\delta G}{2}$ which allowed to tolerate for higher diversity in the predicted ensemble.
As there is an interest to consider the structural diversity within our integrative approach based on sampling and clustering of the structures, a moderate contribution of probing data seemed to be beneficial. Thus, we considered a pseudo-energy of $\frac{\delta G}{2}$ to analyse the GIR1 Lariat-capping ribozyme structure with `IPANEMAP`.

Mono-probing case. Predicted models with `IPANEMAP` were compared to the native structures where the corresponding MCC values were reported in Table 8.5. We noticed that the use of probing data with `IPANEMAP` as auxiliary information allowed to obtain more accurate structures compared to the "sampling" case where the treated structure sample has no constraints. At the top of accurate predictions we find `SHAPE` and `DMS` probing conditions.

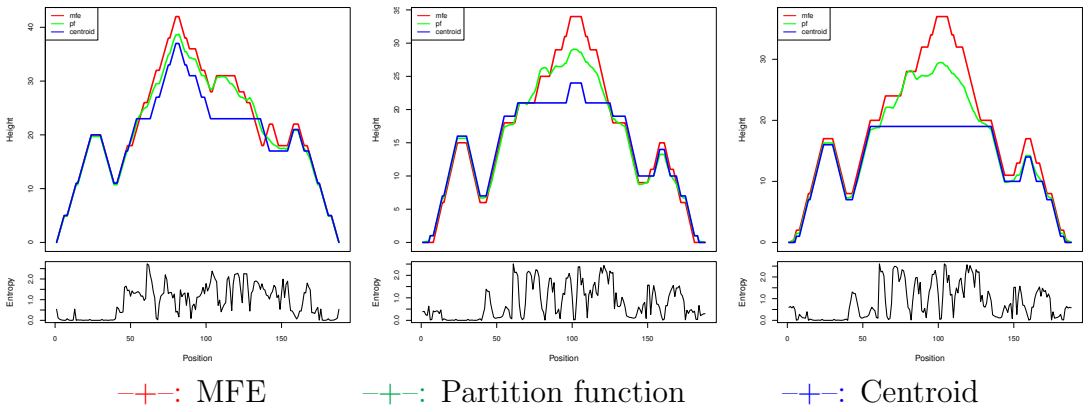


Figure 8.6: **Shannon entropy from predictions with** `RNAfold` in the absence of probing data (on the left), with `1M7ILU3` data for a pseudo-energy of δG (on the middle) and $\frac{\delta G}{2}$ (on the right). The consideration of probing data allowed for less fluctuations of the local structure. Changing the pseudo-free energy contribution by half does not affect the `MFE` structure. However, a slight difference has been detected at both extremities of the sequence. Figures were generated with `RNAfold.`

A better characterization of the structure was detected over all probing conditions in the presence of `Magnesium` Mg^{2+} with an estimated gain of 0.04 in average. For example, adding Mg^{2+} to `1M7IlU` condition allowed to reduce the base pairs distance of the predicted structure by 1 (a base pairs distance of 23 instead of 24 with respect to the native structure). Using Mg^{2+} with condition `1M7` yields a base pairs distance to the native of 34 rather than 37 when `1M7` is probed without Mg^{2+}. An expected result since the use of Mg^{2+} allows to bring out tertiary interactions.

To check for the performance of our predictions in comparison with the classic modeling, we evaluated the accuracy of predicted structures (`MFE/MEA`) with `RNAfold` while considering probing data as soft constraints. The resulting accuracies are depicted in Table 8.6. With our integrative method a better performance was achieved for the set of the considered experimental probing conditions with an average accuracy of 0.702 that exceeds that of the `MFE` (0.635).

Multi-probing case. We integrated pairs of probing data from GIR1 Lariat-capping ribozyme two by two in `IPANEMAP`, the resulting `MCC` and its comparison with the average of mono-probing `MCC` for each condition in the pair is presented in Appendix [10C]. In general a better performance is recorded when combining two source of probing data.

Condition	PPV	Sens	Π	MCC
sampling	0.5161	0.5161	0.5161	0.515
1M7ILUMg	0.8421	0.7742	0.8074	0.807
1M7ILU	0.8519	0.7419	0.795	0.7946
DMS Mg	0.7966	0.7581	0.7771	0.7766
NMIAMg	0.7797	0.7419	0.7606	0.76
NMIA	0.7667	0.7419	0.7542	0.7536
NMIAMgCE	0.7333	0.7097	0.7214	0.7208
BzCNMg	0.7097	0.7097	0.7097	0.709
1M7Mg	0.7414	0.6935	0.7171	0.7164
1M7	0.7358	0.629	0.6803	0.6797
NaiMg	0.6774	0.6774	0.6774	0.6767
1M7ILU3Mg	0.7222	0.629	0.674	0.6733
1M7ILU3	0.7561	0.5	0.6149	0.6142
Nai	0.5873	0.5968	0.592	0.5911
CMCTMg	0.5667	0.5484	0.5575	0.5564

Table 8.5: **Comparison of** GIR1 Lariat-capping ribozyme **predicted structures with** `IPANEMAP`**:** in the absence of probing data noted by "sampling" and with considering one source of probing. Π is the geometric mean of `PPV` and `sensitivity`. Experimental probing conditions are arranged by decreasing accuracy.

We then proceeded with clustering probing conditions based on the euclidean distance that separates their respective predicted ensembles. The clustering was achieved using `M-Kmeans` with a fixed number of 8. The matrix distance is depicted in Figure 8.7 where the content of each cluster is displayed in Table 8.7.

Three macro states resulting from the clustering were found to be populated with more than one condition. Conditions figuring in clusters number 3 and 5 are in accordance with the calculated `MCC` for mono-probing deriving cases. The cluster number 4 combines two conditions sharing the same technology and the same reagent, and where the difference lies in the size of the replicates.

We exploited this grouping of conditions to create several combinations of probing data conditions internally to each cluster and between clusters. Tested combinations with their respective prediction accuracies are illustrated in Figure 8.8.

The eye-catching result from the accuracy of predicted structures concerns the combination of conditions from single cluster. Indeed, this combination allowed a gain of 1.3 in `MCC` compared to the average of `MCC` values over the contributing

Condition	MCC IPANEMAP	MCC MFE guided	MCC MEA guided
No probing data	0.515	0.514	**0.64**
1M7ILUMg	**0.807**	0.777	0.777
1M7ILU	**0.7946**	0.78	0.79
DMSMg	**0.777**	0.369	0.749
NMIAMg	**0.76**	**0.76**	**0.76**
NMIA	**0.754**	0.656	0.656
NMIAMgCE	**0.721**	0.66	0.71
BzCNMg	**0.709**	0.703	0.692
1M7Mg	**0.716**	0.68	0.56
1M7	**0.68**	0.5	0.51
NaiMg	**0.677**	0.623	0.671
1M7ILU3Mg	0.673	**0.72**	0.714
1M7ILU3	**0.614**	0.543	0.592
Nai	0.591	0.581	0.59
CMCTMg	0.556	0.544	0.544
Average	**0.702**	0.635	0.666

Table 8.6: **Comparison of predicted structures for GIR1 Lariat-capping ribozyme with IPANEMAP and with classic modeling.** The best performance is highlighted in bold. The 'average' column contains averaged MCC over all conditions for a given type of probing data.

Cluster ID	Condition(s)
1	1M7
2	Nai
3	DMSMg, NMIA, NMIAMg, 1M7ILU
4	1M7ILU3Mg, 1M7ILUMg
5	BzCNMg, NMIAMgCE, NaiMg
6	CMCTMg
7	1M7ILU3
8	1M7Mg

Table 8.7: Probing conditionsclustered by the proximity of their pseudo-Boltzmann ensemble.

conditions.

When considering this combination with lowly performing conditions from clus-

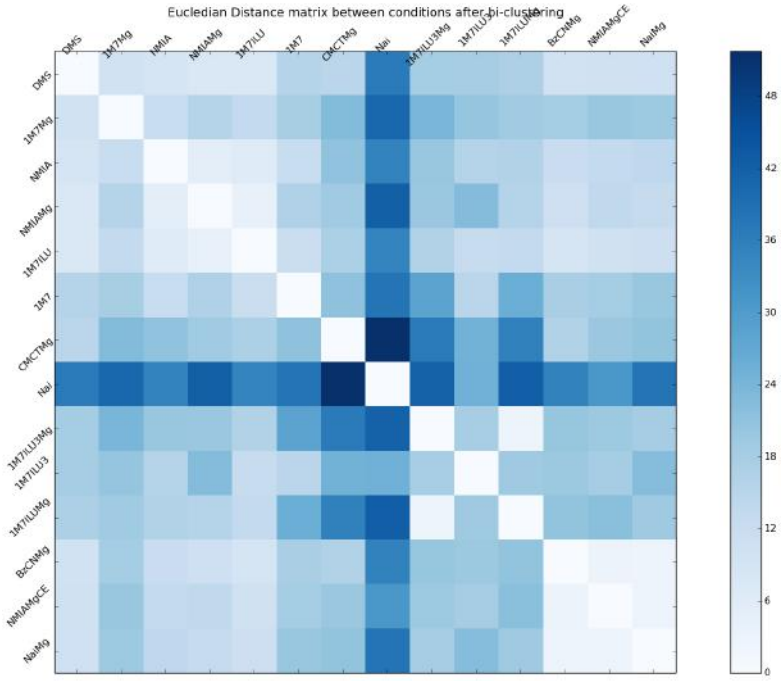


Figure 8.7: **Bi-clustering of probing conditions according to the Euclidian distance:** the distance matrix is rearranged to regroup similar structural ensembles.

ters 3 and 5 two structures were found to be optimal.

In the other side, when highly performing conditions from clusters 3 and 5 were considered along the "unique condition" clusters, a gain in `MCC` is more noticeable.

A net improvement was roughly observed across the different combinations. Although, it is not possible to deduce rules from this test about the experimental conditions to consider in order to get accurate predictions, the last combination that is dominated by `SHAPE` conditions `1M7` /`NMIA` with Mg^2+ and characterized by the absence of `DMS` probing condition, might give the hope for the ability of combined `SHAPE` probing data to tune the structure predictions with `IPANEMAP`.

Combinations	Description	PPV	Sens	MCC	Average mono-probing
1	`Conditions from cluster 3`				
	DMSMg, NMIA, NMIAMg, 1M7ILU	0.807	0.742	**0.77**	**0.77**
	NMIA, NMIAMg, 1M7ILU	0.793	0.742	0.766	**0.77**
	DMSMg, NMIAMg, 1M7ILU	0.807	0.742	**0.77**	**0.77**
	DMSMg, NMIA, NMIAMg	0.78	0.742	**0.76**	**0.76**
	DMSMg, NMIA, 1M7ILU	0.767	0.742	0.753	**0.77**
2	`Conditions from cluster 5`				
	BzCNMg, NMIAMgCE, NaiMg	0.721	0.71	**0.714**	0.7
3	`Conditions from clusters with single element`				
	1M7,Nai,CMCTMg,1M7ILU3,1M7Mg	0.815	0.71	**0.76**	0.63
4	`Conditions from all clusters with lowest MCC`				
	1M7,Nai,CMCTMg,1M7ILU3,1M7Mg,NMIA,NaiMg	0.59	0.597	0.59	
		0.767	0.742	**0.753**	0.65
5	`Conditions from all clusters with highest MCC`				
	1M7,Nai,CMCTMg,1M7ILU3,1M7Mg,1M7ILU, NMIAMgCE	0.852	0.742	**0.794**	0.67

Figure 8.8: **Comparison of predicted structures through `IPANEMAP`** : reported accuracies for some combinations chosen from the resulting macro-states with their correponding `MCC` average over the mono-probing cases. Best performances are highlighted in bold.

8.3 Ebola UTRs

Translation of mRNAs into proteins is one of the last steps in gene expression. Although it was thought not to be regulated, we have now the evidence that not all mRNAs are translated in the same way and with the same efficiency.

Beyond the sequence that specify the composition of the proteins produced, mRNAs carry a wealth of information in their structure. Such structural signals interact with cellular protein and/or the cellular translation machinery to regulate its activity. Although these structural signals may be anywhere in the mRNA, most of them are located in 5′ UnTranslated Region (5′`UTR`) and 3′ UnTranslated Region (3′ `UTR`).

Viruses as obligatory cellular parasite need to use most of the cellular machineries to express their genetic code. In particular they require the ribosome to produce their proteins. Once in the cytoplasm, viral mRNA have to compete with cellular mRNA. Evolution has selected many different mechanisms for the viruses to high-jack the translation machinery to their benefit. Some damage the cellular mRNAs, as the influenza virus, some other, as the poliovirus, damage the cellular machinery itself in such a way that it is only able to translate the viral mRNA. In all cases, the viral mRNA structure brings a key information to the process.

The molecular events leading to Ebola virus mRNA translation are unknown, and for obvious reasons they are not easily amenable to in vivo or even in cellulo experimentation.

For this reason, as a first step towards the discovery of this aspect of Ebola virus life cycle, we undertook to characterize the structure of the 5′ and 3′ UTRs. The Ebola virus contains 7 mRNAs as depicted below:

8.3.1 Dataset

To predict the structures for the 14 UTRs , we used two sources of probing data: `1M7-SHAPEMap` with and without Mg^{2+}. Moreover, we integrated evolutionary data as a third constraint in `IPANEMAP`.

Condition	Description
1M7Mg	1M7-SHAPEMap in presence of Mg^{2+}, sequencing with Illumina . Reactivity profiles generated with SHAPEMapper2 [Busan and Weeks, 2018]
1M7	1M7-SHAPEMap, sequencing with Illumina. Reactivity profiles generated with SHAPEMapper2
MSA	Multi-Sequence Alignment treated with ClustalW

Table 8.8: **Constraints integrated in IPANEMAP to resolve Ebola UTRs**

8.3.2 Method

We used IPANEMAP to predict the 14 UTR structures by considering two probing data sources: 1M7 and 1M7Mg along with evolutionary data.

The ability of the evolutionary data to inform about conserved structures motivated the extension of IPANEMAP to further consider this structural informative data alongside with probing data. The new structural informative dimension is expected to strengthen the predictive power of IPANEMAP. We integrated a new option in IPANEMAP tool to allow sampling from an MSA guided prediction with RNAalifold.

8.3.3 Results

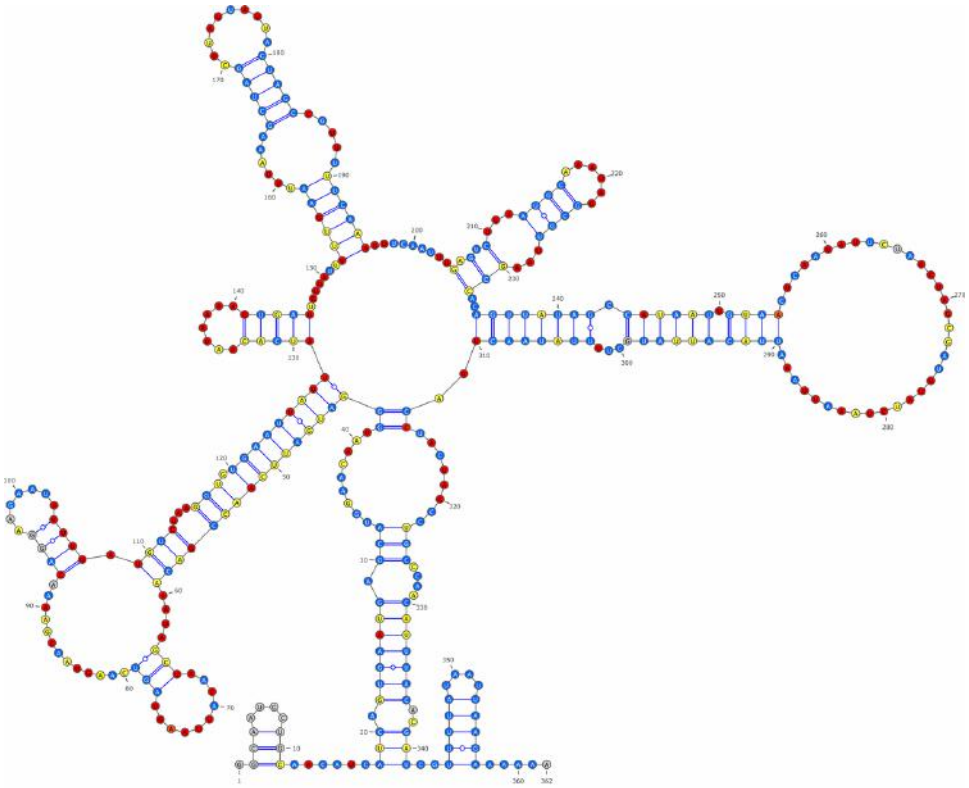


Figure 8.9: **Candidate structure(s) for the 3' UTR of the NP gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 4 dominant clusters. One cluster was found to be the optimal with two dominant conditions: `1M7Mg` and `MSA`. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq 0.4 \rightarrow$ blue, $\geq 0.7 \rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

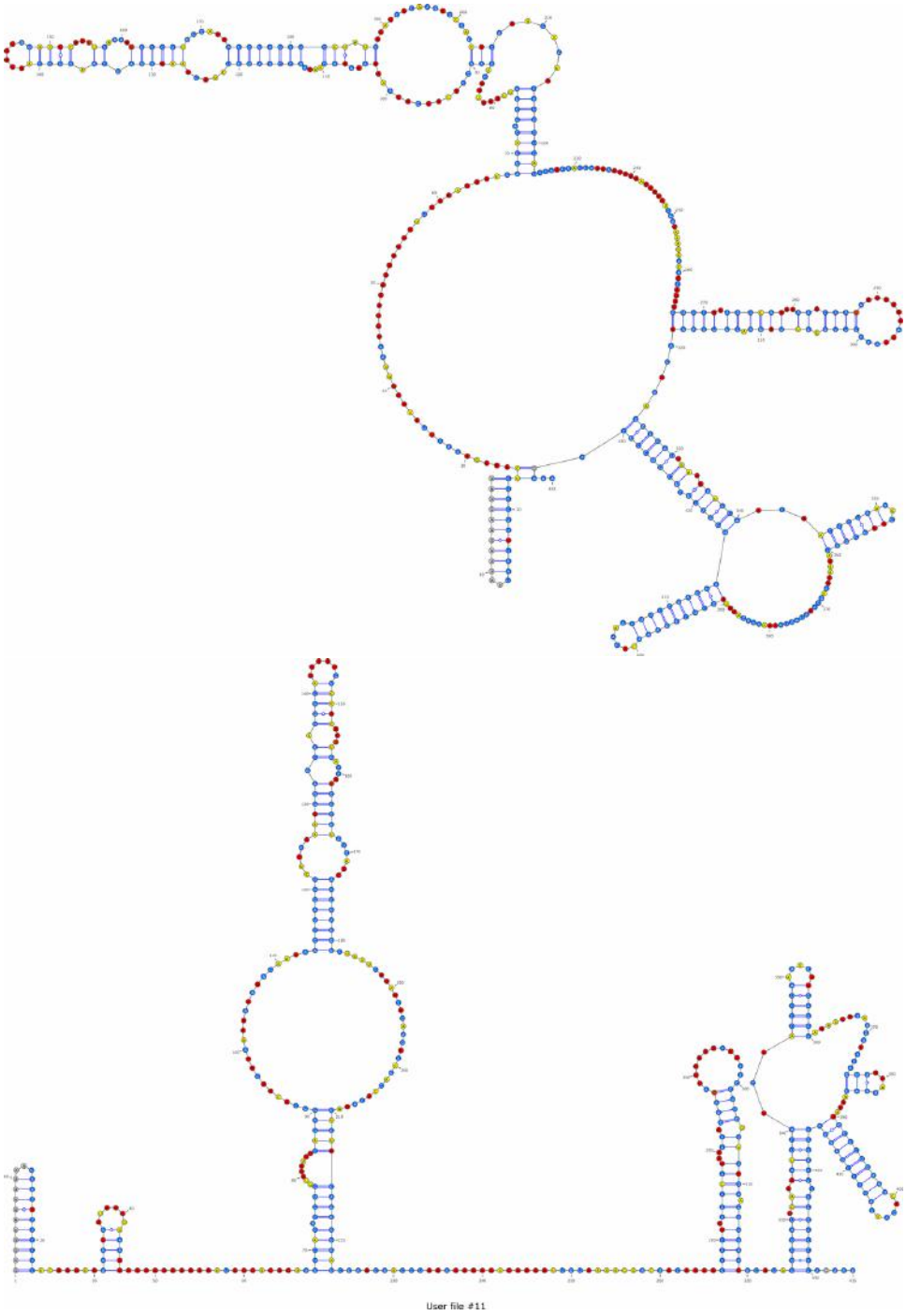


Figure 8.10: **Candidate structure(s) for the 5' UTR of the NP gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 20 dominant clusters. Two clusters were found to be optimal: One cluster has two representative conditions `1M7Mg` and `1M7`, the resulting centroid is illustrated on the left. The second cluster has `MSA` as dominant condition with the ensuing centroid presented on the right. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq$0.4 $\rightarrow$ blue, $\geq$0.7 $\rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

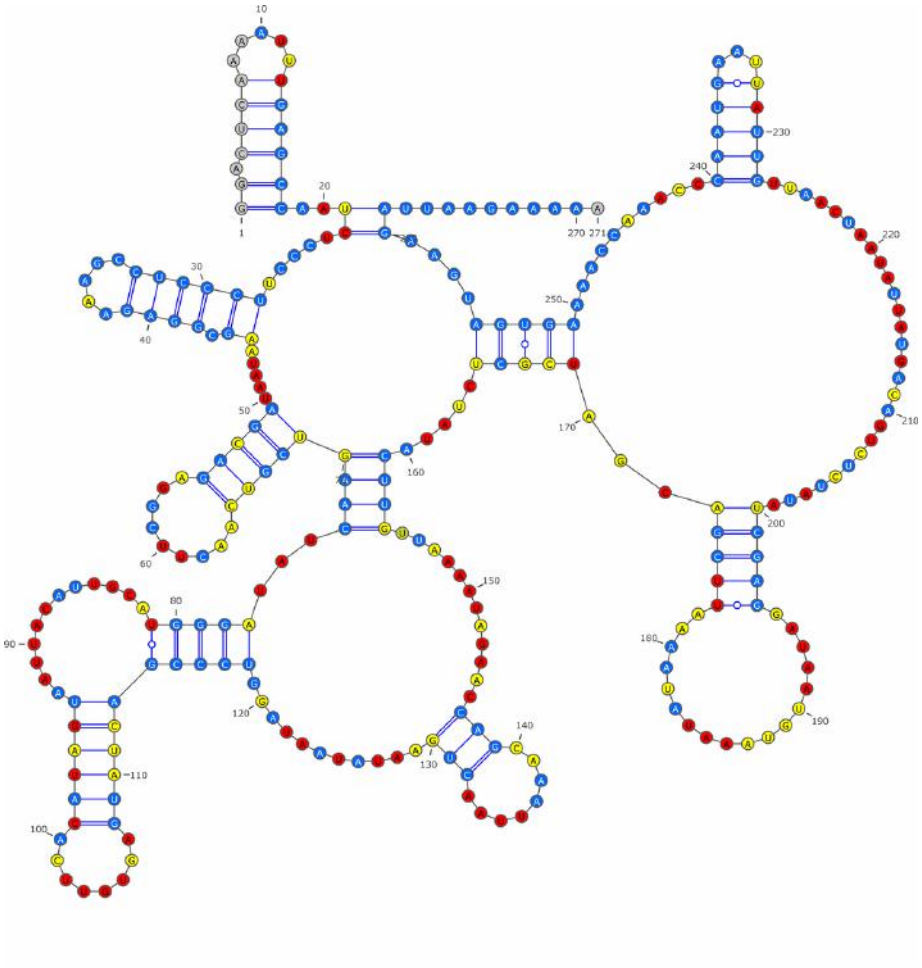


Figure 8.11: **Candidate structure(s) for the 3' UTR of the VP35 gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 3 dominant clusters. One cluster has found to be optimal with 2 dominant conditions: `1M7Mg` and `1M7`. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq$0.4 $\rightarrow$ blue, $\geq$0.7 $\rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

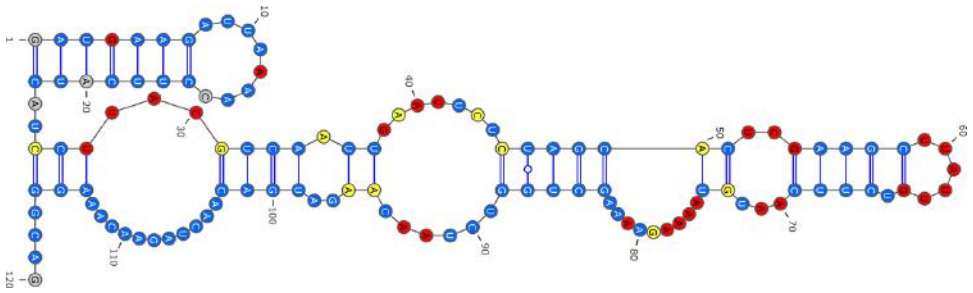


Figure 8.12: **Candidate structure(s) for the 5' UTR of the VP35 gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 3 dominant clusters. One optimal cluster with all conditions: `1M7Mg` ,`1M7` and `MSA` as representative. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq$0.4 $\rightarrow$ blue, $\geq$0.7 $\rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

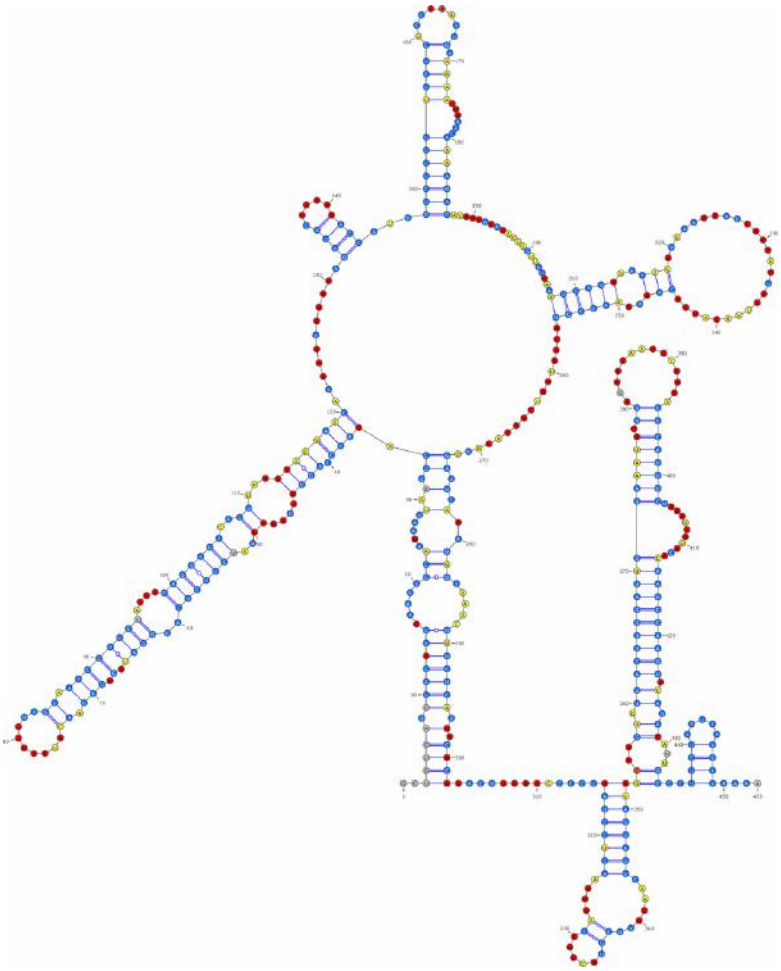


Figure 8.13: **Candidate structure(s) for the 3' UTR of the VP40 gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 20 dominant clusters. One cluster has found to be the optimal with two dominant conditions: `1M7Mg` and `1M7`. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq$0.4 $\rightarrow$ blue, $\geq$0.7 $\rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

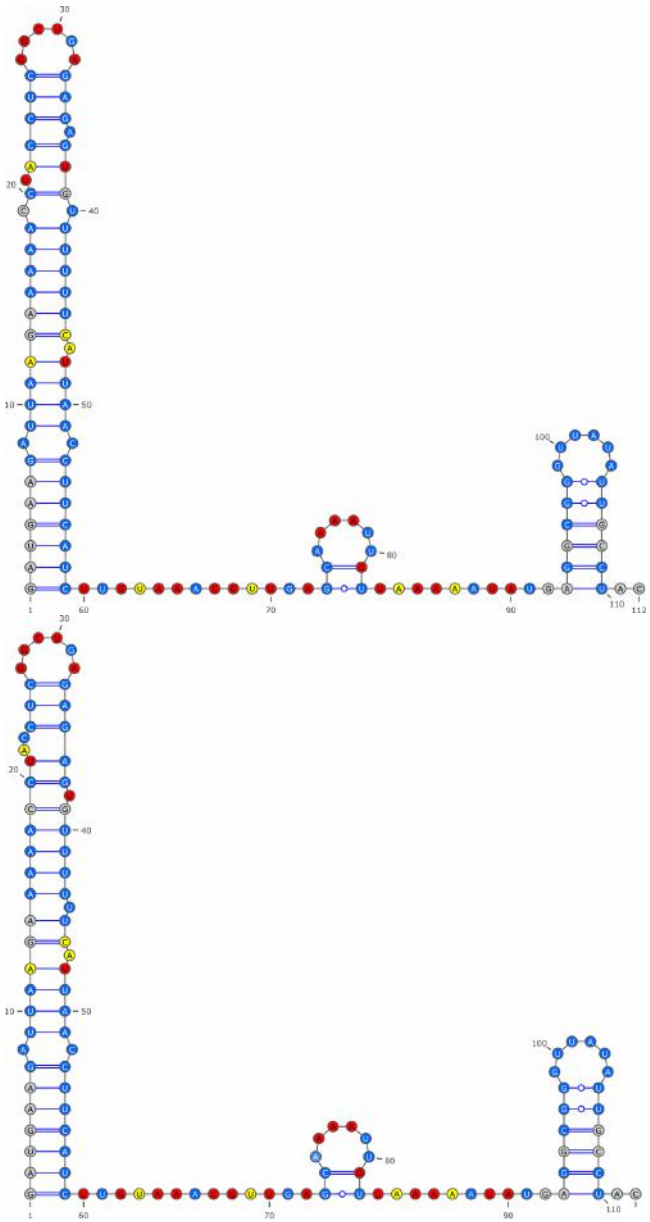


Figure 8.14: **Candidate structure(s) for the 5' UTR of the VP40 gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 3 dominant clusters. Two clusters were found to be optimal: The cluster associated with the top centroid has two representative conditions `1M7Mg` and `1M7`; The cluster associated with the bottom centroid represents three dominant conditions: `MSA`,`1M7Mg` and `1M7`. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq$0.4 $\rightarrow$ blue, $\geq$0.7 $\rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

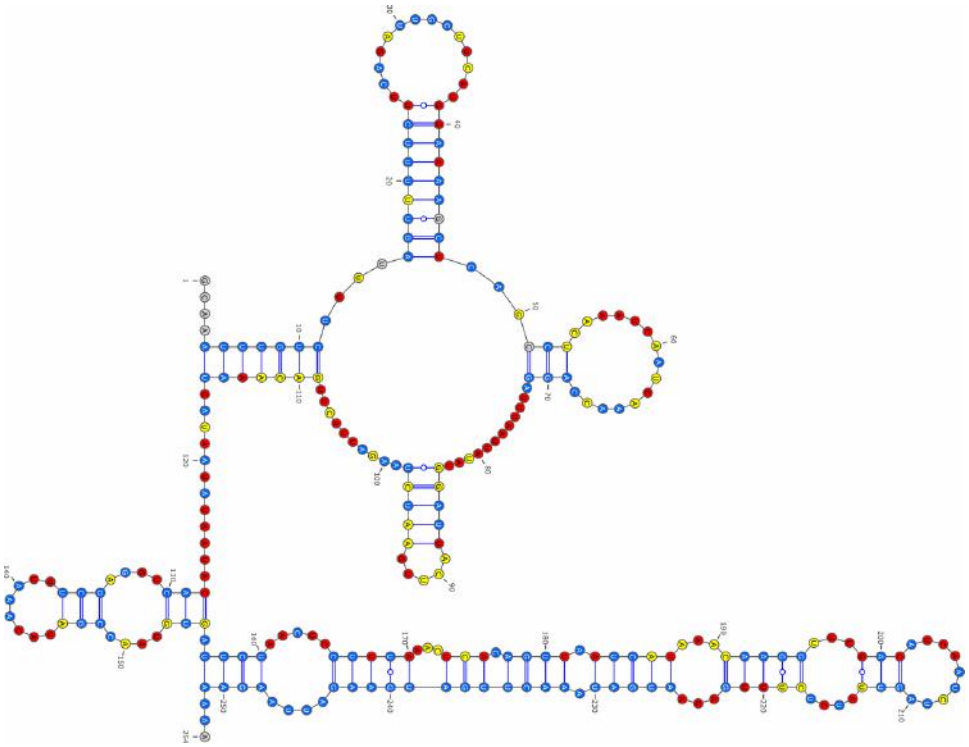


Figure 8.15: **Candidate structure(s) for the 3' UTR of the GP gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 4 dominant clusters. One cluster was found to be optimal with 2 dominant conditions: `1M7Mg` and `1M7`. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq$0.4 $\rightarrow$ blue, $\geq$0.7 $\rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

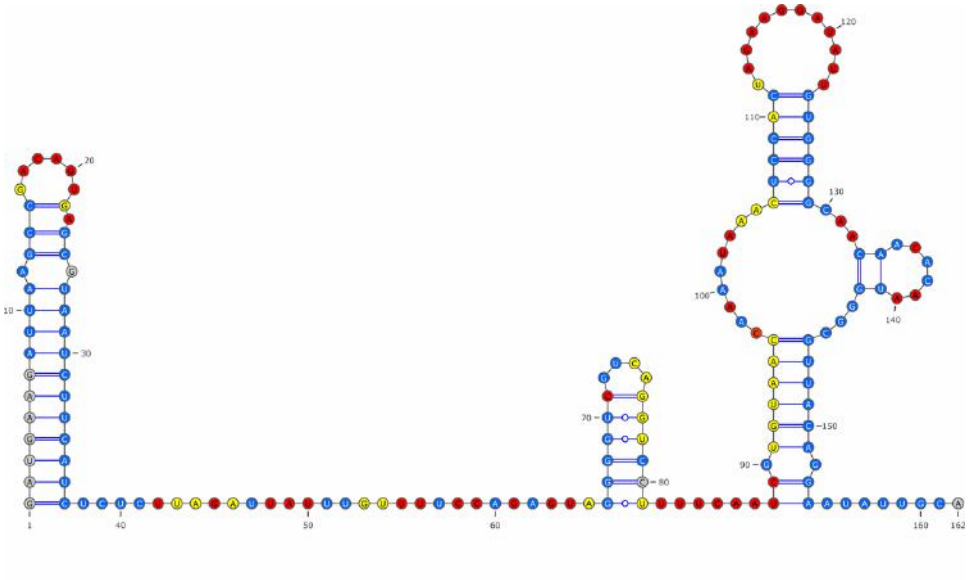


Figure 8.16: **Candidate structure(s) for the 5' UTR of the GP gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 3 dominant clusters. One optimal cluster was found with 2 dominant conditions `1M7Mg` and `1M7`. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq$0.4 $\rightarrow$ blue, $\geq$0.7 $\rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

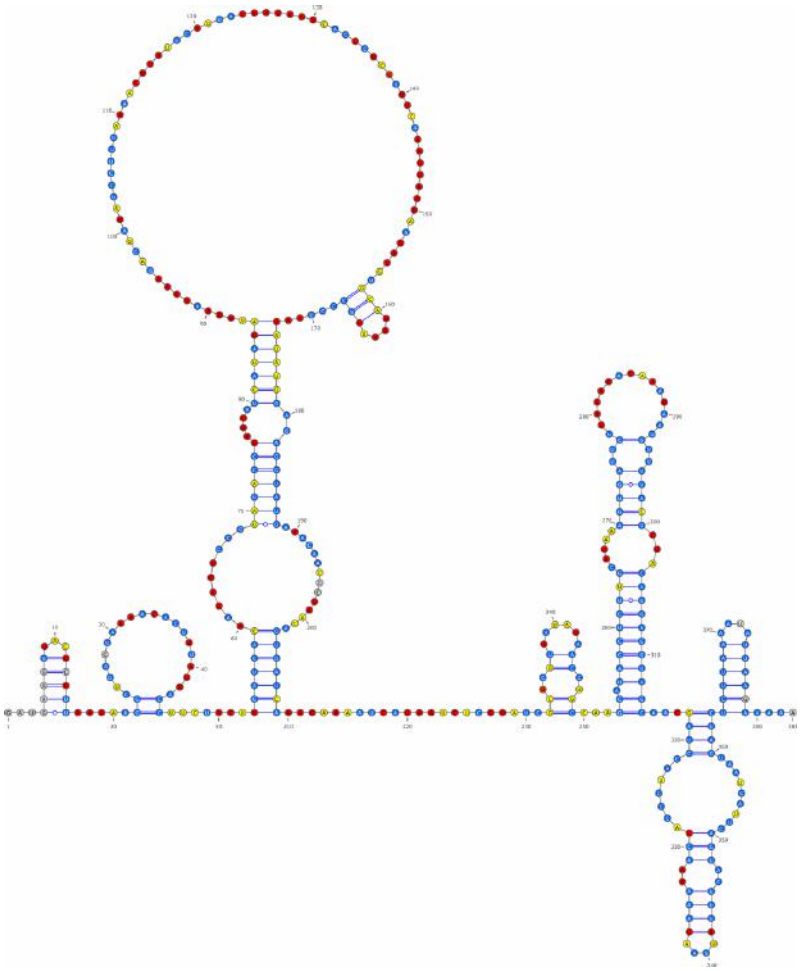


Figure 8.17: **Candidate structure(s) for the 3' UTR of the VP30 gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 4 dominant clusters. One cluster was found to be optimal with one dominant condition: `1M7Mg`. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq 0.4 \rightarrow$ blue, $\geq 0.7 \rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

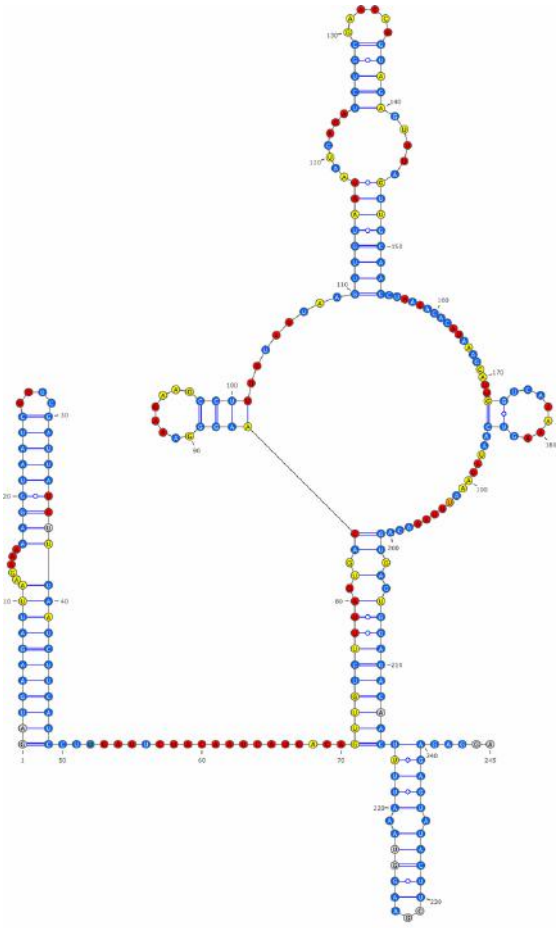


Figure 8.18: **Candidate structure(s) for the 5' UTR of the VP30 gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 2 dominant clusters. One optimal cluster dominated with 2 represented conditions `1M7Mg` and `1M7`. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq$0.4 $\rightarrow$ blue, $\geq$0.7 $\rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

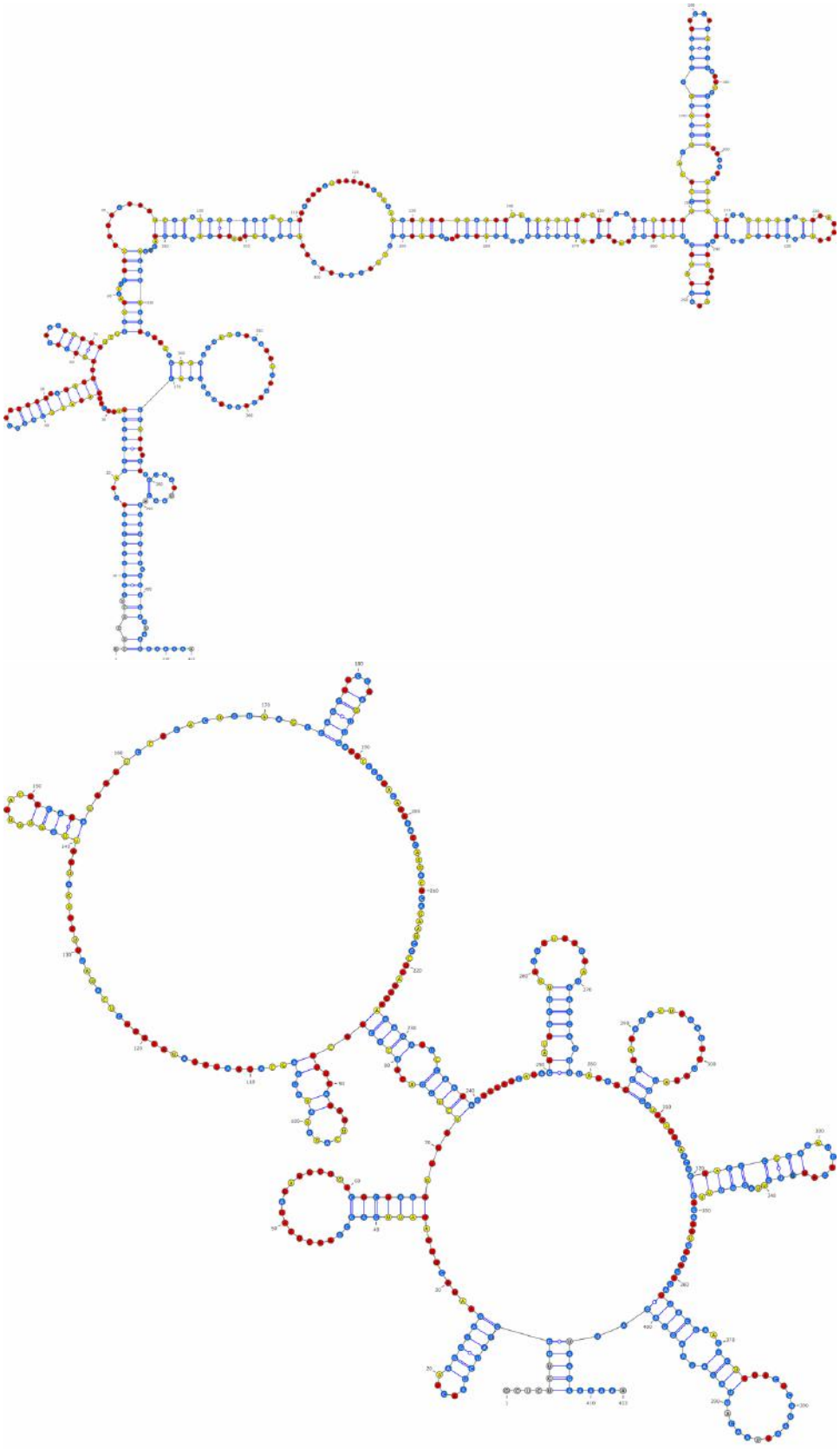


Figure 8.19: **Candidate structure(s) for the 3' UTR of the VP24 gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 20 dominant clusters. Two optimal clusters were identified. The top centroid corresponds to a cluster dominated by `MSA` and on the right the cluster dominated by `1M7Mg` and `1M7` Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq 0.4 \rightarrow$ blue, $\geq 0.7 \rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

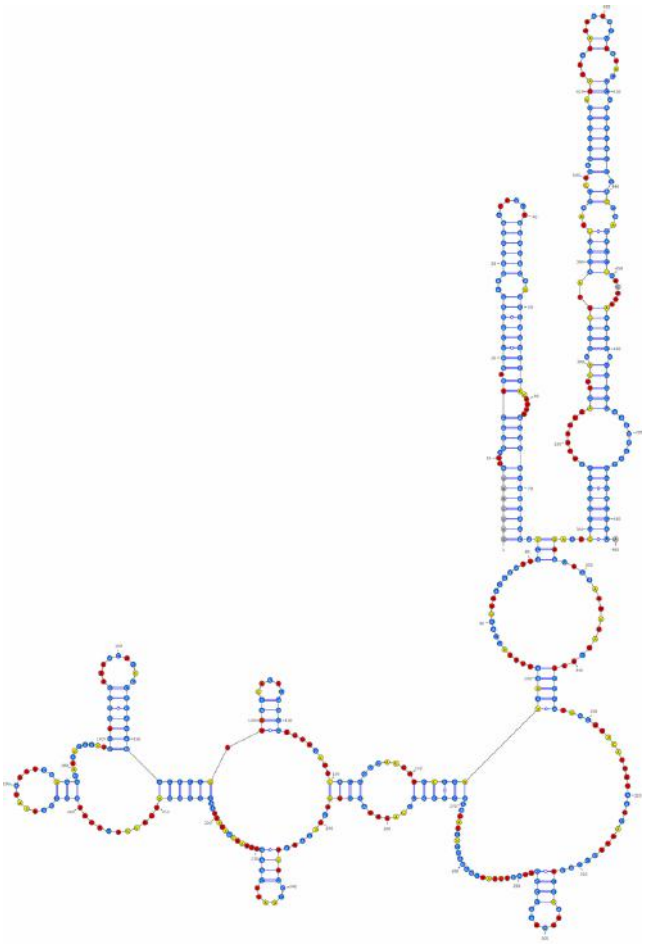


Figure 8.20: **Candidate structure(s) for the 5' UTR of the VP24 gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 3 dominant clusters. One cluster has found to be the optimal with two dominant conditions: `1M7Mg` and `1M7`. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq$0.4 $\rightarrow$ blue, $\geq$0.7 $\rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

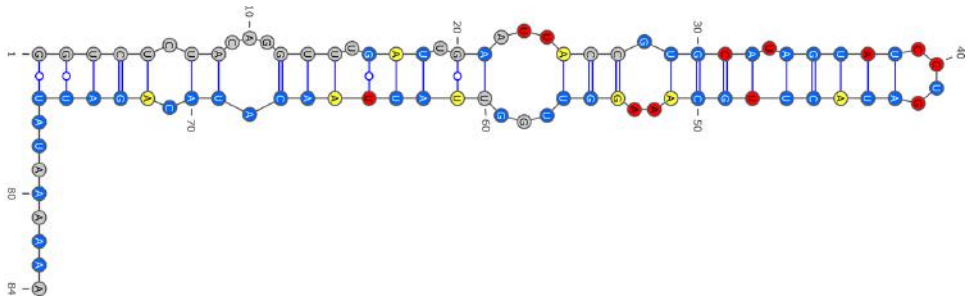


Figure 8.21: **Candidate structure(s) for the 3' UTR of the L gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 2 dominant clusters. One cluster has found to be optimal with all conditions `MSA`, `1M7Mg` and `1M7` as dominant. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq$0.4 $\rightarrow$ blue, $\geq$0.7 $\rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

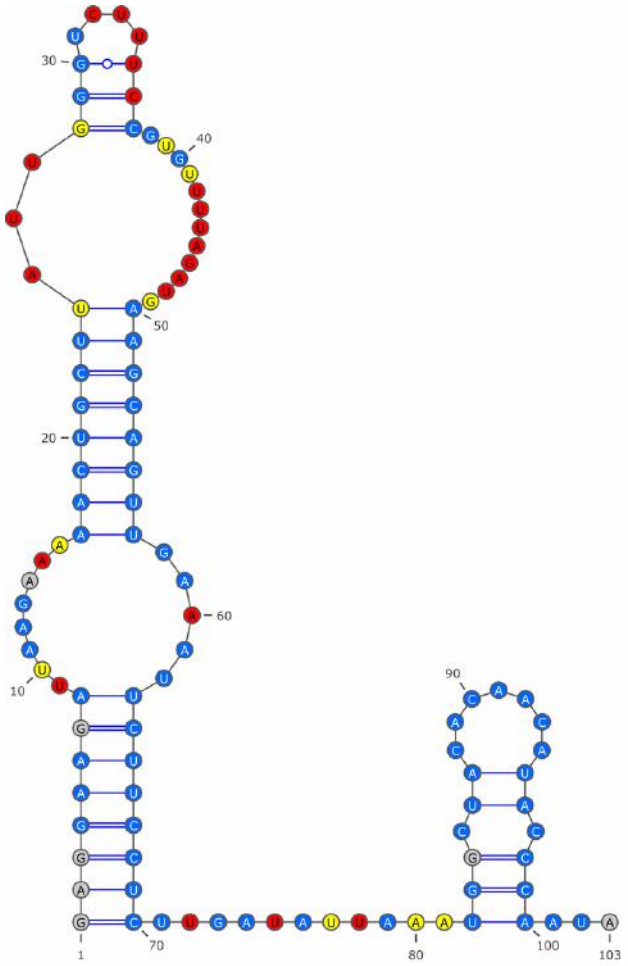


Figure 8.22: **Candidate structure(s) for the 5' UTR of the L gene, predicted using `IPANEMAP`:** `IPANEMAP` identified 2 dominant clusters. One cluster has found to be optimal with all conditions `MSA`, `1M7Mg` and `1M7` as dominant. Nucleotides are color-coded based on their normalized SHAPE reactivities ($\leq$0.4 $\rightarrow$ blue, $\geq$0.7 $\rightarrow$ red, mid range $\rightarrow$ yellow; missing values in gray)

UTR structures prediction with `IPANEMAP` led in most cases to two clusters: one dominated by the `MSA` and the other dominated by the probing data. However, both predicted conformations visually appear to be equally supported by probing data. For some UTRs (3'L, 3'NP, 5'L, 5'VP35 and 5'VP40), `IPANEMAP` reported one structure supported by both probing and evolutionary data. Overall, the good agreement between predictions informed by SHAPE data and evolutionary information is encouraging, and suggests the existence of stable structured regions within Ebola UTRs. In particular, we recover Stem-and-loop structures previously reported by Brauburger et al. [1993] at the 5' end of each mRNA. Interestingly, our predictions apparently contradict the common-sense assumption of recurrent regulatory motifs across Ebola UTRs.

However, in term of our predicted structure, a caveat of our analysis resides in the extreme sequence identity observed within most available alignments. For such UTRs, the absence of compensatory mutations leads to a overwhelming dominance of the free-energy model within the scoring scheme of `RNAalifold`. In other words, the `MSA` contribution is very limited, and the stochastic sampling converges towards the classic Boltzman sampling of Ding and Lawrence [2003]. For the same reason, we could not establish the statistical significance of predicted base pairs using classic software such as `R-scape` [Rivas et al., 2016].

In order to validate our analysis, a direction would be to go beyond the Ebola RNA model, and include in the MSA more distant homologs, such that the UTRs of the Marburg virus while being aware that the extreme rate of phylogenetic evolution witnessed in viruses may lead to a loss of function.

Chapter 9

Discussion and perspectives

9.1 Contributions

In the present thesis, we developed a new algorithm to characterize the RNA structure through the simultaneous use of various probing data that differ either in the adopted experimental protocol or in the deployed HTS technique.

In Chapter 4, we presented various frameworks dedicated to the simultaneous use of probing data to infer the RNA structure. In Chapter 5, we described our integrative approach leveraging the agreement between different profiling data issued from the transcriptome in the presence of one dominant conformation.

Our novel approach is implemented in `IPANEMAP` tool. Given an RNA sequence with a set of reactivity profiles, `IPANEMAP` explores the structural landscapes resulting from a probing data-driven prediction and allows to detect a set of stable cluster(s). `IPANEMAP` includes the implementation of an iterative clustering algorithm, a new implementation of the MEA calculation based on the Boltzmann probability and an adapted version of nested algorithm to define dominant clusters.

Performing predictions for a set of RNA models with `IPANEMAP` led us to formulate several questions at several levels that will be addressed in the discussion section bellow.

In Chapter 6, we presented a new `SHAPE`-based protocol within a hybrid mutational context where two sources of mutations were considered: mutations that characterize variants from the WT and those that result from a mutation-prone

protocol such as SHAPEMap. The implementation of this protocol led us to deal with a yet unresolved problem of assigning reads to their variants of origin. To tackle this issue, we developed a new EM-based assignment algorithm, harnessing the mutational profile instead of the minimal base distance to the RNA of reference. Preliminary results on a set of simulated sequenced data have shown to lead to accurate reads assignation compared to the classic mapping algorithm. The EM-assignment iterative algorithm has shown to be greedy. Thus, the optimization of this algorithm is one of the point to be addressed in the near future.

9.2 Discussion

9.2.1 Inferring the structure from probing data

The emergence of diverse protocols and technologies producing reactivity profiles leads to question the quality of the produced data.

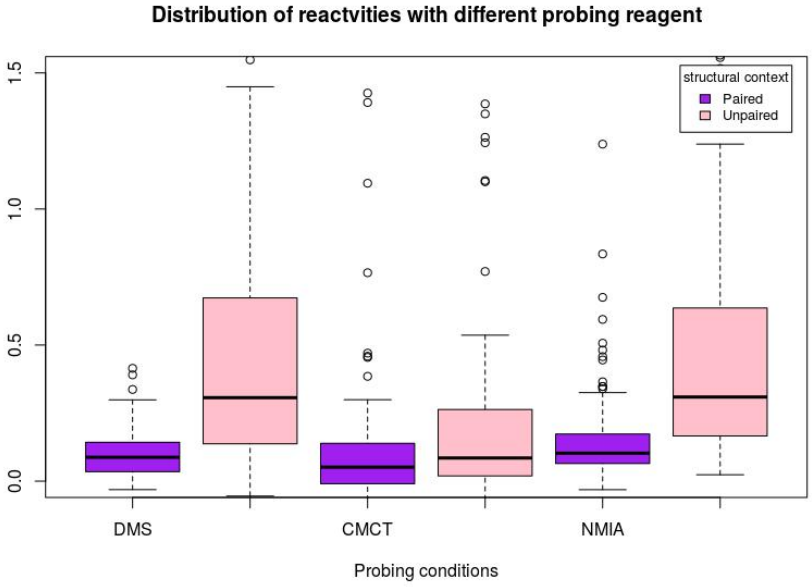


Figure 9.1: **Distribution of reactivities for different probing data:** DMS, CMCT and NMIA, in function of the stuctural context (paired/unpaired). A significant overlap between the two structural contexts is observed for the case of CMCT data.

Profiling methods allow to infer structural characteristics at the residual level. These methods differ in the nature of the experimental protocol, in the probing technique and in the data analysis. This disparity results into distinct profiling data distributions. Thus, leading to distinct interpretations while they are integrated into the thermodynamic prediction model according to the [Deigan et al., 2009] formula that was specifically trained for `SHAPE` data. Therefore, scoring the reactivity, in addition to integrating them into the structural modeling, form some of the aspects to be reviewed to ensure a better inference of the structure.

Towards direct inference of structural motifs from probing data. The inability of the current `DP` based prediction models to directly and reliably infer the structure from profiling data arouses more than ever a need for a sophisticated method that directly make use of reactivities without resorting to the thermodynamic context modeling.

Driven by the need for a common approach to interpret and mine information from experimental profiles, Ledda and Aviran [2018] have recently proposed an `HMM` based model to directly infer RNA structures motifs from experimental profiling data. The suggested pattern recognition algorithm has shown to be compatible with various profiling techniques and experimental protocols. In addition, by taking advantage of homologies between a desired target pattern and transcriptomic regions, the pattern recognition algorithm interprets data within a functional domain context and not only at the residual level. The suggested model does not make any assumption from the thermodynamic modeling to infer the structure, it entirely relies on the data from which the pairing states (paired/unpaired) probabilities for each nucleotide is estimated.

In the absence of a data-guided modeling algorithm that allows to sample structures in the presence of complex structural elements such a pseudo-knots and tertiary interactions, Ledda and Aviran [2018] have shown that the structural motifs can be directly obtained from `SHAPE` data while obviating the thermodynamic modeling. Hence, besides the direct structure inference, complex structural elements could be recovered.

Stop-induced and mutation-induced protocols provide complementary structural signatures. The use of one probing data source with `IPANEMAP` (suggested models in Chapter 8) has shown to allow for a more accurate results, especially with the use of `DMS` data profiling, compared to the multi-probing state. Combining different profiling data to eliminate those that fail to capture the complexity of the structure and to strengthen the mostly supported structure is a hypothesis that still can not be generalized. Among other reasons, this is mainly

due to the equivalent interpretation of the different probing data and their contribution weights in the prediction model. Moreover, there are no common rules to choose for the optimal data combination; should we consider only data from mutation-modification adduct, data from stop-modification adduct or a hybrid combination of both data.

Compared to the RT stops protocols, mutation-modification adduct protocols facilitate the normalization of data and allow to analyze covariation of mutational frequencies at different sites (as exploited by $M\&M$). This explains their popularity. For a long time, it has been considered that stops and mutations could be taken as commutable markers of modification. Sexton et al. [2017] questioned this finding. In the common probing data analysis, the modification of accessible nucleotides could be reported as stops or as mutations. In this work, it has been shown that a chemical modification does not always leads to either an RT stop event or a mutation event which can drastically bias the interpretation of their respective structural profiling. They assumed that the probability of a modification-induced stop or mutation depends on the sequence context and on the nature of the RT enzyme. Sexton et al. [2017] analysis of DMS-probing RT-stops and RT-mutation has shown that these two metrics are poorly correlated and completely orthogonal in some cases. The need to incorporate signatures from both RT-stops and RT-mutation motivated the statistical analysis of RT-stops and RT-mutations, led by Yu et al. [2018], to estimate a measure of probing reactivity at residual level.

Indeed, based on the previously mentioned statistical analysis by Ledda and Aviran [2018], Yu et al. [2018] suggested an extension of a maximum-likelihood derivation that results into a reactivity formula to combine two probabilities for a given nucleotide to be modified from RT-stops and RT-mutations events. The proposed formula by [Yu et al., 2018] matches the interpretation of the reactivity as a fraction of residual adduct at the end of probing reaction. In the same work, it has been assumed that a hybrid combination of adduct-stops and adduct-mutation allows a significant gain in reactivity accuracy and an invariance of reactivity to RT conditions.

This finding came to support the complementarity hypothesis: the integration of both mutations and stops in chemical probing data is one way to mitigate the protocol biases and ultimately provide greater insight into RNA structure from probing experiments. The hybrid combination of stop-induced and mutation-induced conditions was one of the point of interest previously discussed in this thesis. We did resolve the GIR1 Lariat-capping ribozyme structure in the presence of hybrid conditions where we have found that the optimal combination of conditions that allowed a significant gain in accuracy contains conditions from both stop-

induced and mutation-induced experimental protocols.

Towards a compromise between profiling data and the structural diversity. The common way to assess reactivity accuracy is by performing data-guided RNA structure modeling then evaluate the improvement in term of predicted structure. In the absence of a direct method to assess the accuracy of reactivity, this led once again, to question the identical integration of various probing data in the classic prediction model.

The use of derived pseudo-energies from probing data results into a significant improvement of the prediction performances compared to the conversion of probing data into hard constraints. When interrogating the intensity of the pseudo-energies via the analysis led in Chapter 7, we concluded that the structural landscape is sensitive to a slight variation of the pseudo-energy value. Indeed, a non moderate contribution might lead to severe restrictions on the predicted ensemble which makes it delicate to analyze the shrunken structural space. Therefore, we assume that the intensity of pseudo-energy might be too important and the need for a more robust integration method still persists.

The question arising here concerns the structural inference of probing data. By sticking to the thermodynamic model, unlike for the direct inference method [Ledda and Aviran, 2018], there is a need to foresee other factors to take into account in the pseudo-energy contribution such as the sequence context, accounting for uncertainty (standard error) in modeling. Introducing a notion of "intensity" of the reactivity during data incorporation could also be helpful: in the case of missing data one may set a 0 intensity instead of relying completely on the thermodynamic model that interprets the corresponding nucleotides as stacked.

In a recent work related to probing data guided characterization of the structural landscape, Li and Aviran [2018] have suggested a probabilistic model that reconstructs accurate conformers landscape from probing guided structural samples and estimates their unknown relative abundances.

Alternative conformations that an RNA may adopt make it difficult to computationally characterize the structural landscape. In return, exploiting the stochasticity aspect of reactivity profiles informs about the diversity of the structural landscape and ultimately improves RNA structure predictions.

Unlike classic computational models that are bounded by the prediction of nested canonical secondary structures, the probabilistic model allows to reconstruct alternative structures encompassing non-nested structures. In this case, pseudo-knots and tertiary interactions can be detected.

9.2.2 Towards a multi-variants probing analysis

In Chapter 6, we have suggested a novel structural-context mutational protocol to predict the RNA structure. This protocol is favoring mutations from an error-prone PCR and makes use of SHAPEMap protocol to populate mutations at the level of single strand regions.

In the same optic, a recent work-flow M2-seq based on DMS-$M\&M$ protocol coupled with a non directed mutagenesis error-prone PCR [Cheng et al., 2017], has shown to be able to detect more accurate RNA substructures. In particular, the introduction of intentionally installed mutations to the classic $M\&M$ protocol allowed to detect more accurate RNA helices for the GIR1 Lariat-capping ribozyme RNA.

The primary objective of this Ph.D. was to use multi-variants probing data to enhance the RNA structure prediction. As we have not yet been able to recover relative reactivities for SHAPEMap data in the context of simultaneous variants SHAPEMap analysis, we found it worthy to exploit the mutants probing data from M2-seq protocol in order to test the ability of our method IPANEMAP to be extended to conduct a multi-variant probing data guided predictions.

We used the DMS profiling data with intentional mutations for the RNA model GIR1 Lariat-capping ribozyme extracted from the RMDB database [1]. Probing data were initially renormalized to solely count for values from C and A nucleotides. First, we ran IPANEMAP in a mono-probing mode with DMS data from the WT. Then, we combined probing data from the WT with data from each single variant d_i. The set of variants contains 188 RNAs; each RNA was characterized by a point-wise mutation at the i^{th} nucleotide. The resulting MCC over the predicted structures in the case of bi-probing, is presented in Figure 9.2.

In order to assess our intuition about a possible extension of IPANEMAP approach to the case of multi-variants, under the structure conservation assumption, we built 10 arbitrary combinations of 10 variants. Then, we integrated each combination with the WT to predict the RNA structure with IPANEMAP. The performance of the predicted structures was assessed and the ensuing values were then compared to the value for the sole WT. Results are depicted in Chart 9.3.

In the context of mutational protocol, a structure fully supported by variants profiling could be inferred if the probing was performed with the dominance of the conserved structure shared by all mutants. We refer to this case as "positive selection". This is obviously the case of the multi-variants DMS probing analysis where the consideration of an additional profiling data to the WT led to a significant

[1] RMDB reference: $GIR1RZ\_DMS\_0001.rdat$

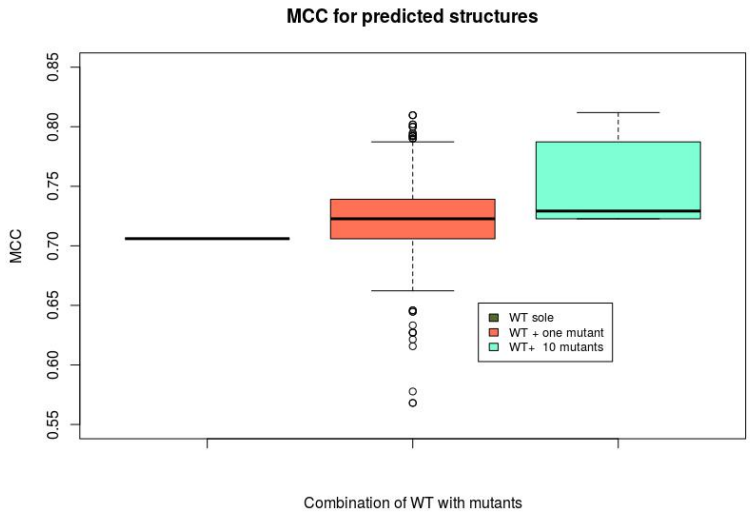


Figure 9.2: `MCC` **distribution from Differential-**`DMS with different combinations:` for mono-probing with the WT, bi-probing from one mutant with the WT and multiprobing from the combination of 10 mutants with the WT. The first quartile from the bi-probing distribution exceeds the value when the WT is used alone. Compared to the bi-probing, the multi-probing with the mutants combination allowed to get an MCC value that surpasses the median value from the bi-probing distribution.

gain in accuracy (Chart 9.3). When the dominance of one conserved structure can not be guaranteed, two cases could be faced in the presence of structural profiles from variants within a multi-variants probing analysis:

- Under the hypothesis of the conservation of the functional structure(s) across the mutants, the set of respective profiles will contribute to eliminate the structural noise that may affect one single profile, subsequently contribute to strengthen the `WT` structure.
- The embedded mutants profiles refer to structure(s) that are different from that of the `WT`. In this case a more sophisticated analysis should be set up to determine variants sequences that are responsible for such deviation from the `WT` and subsequently localize the involved mutations.

In both cases, to guaranty a more accurate inference of the structure from variants profiling data, it is necessary to consider an additional analysis dimension. Covariations that are known for their power to infer conserved pairs could be a good candidate.

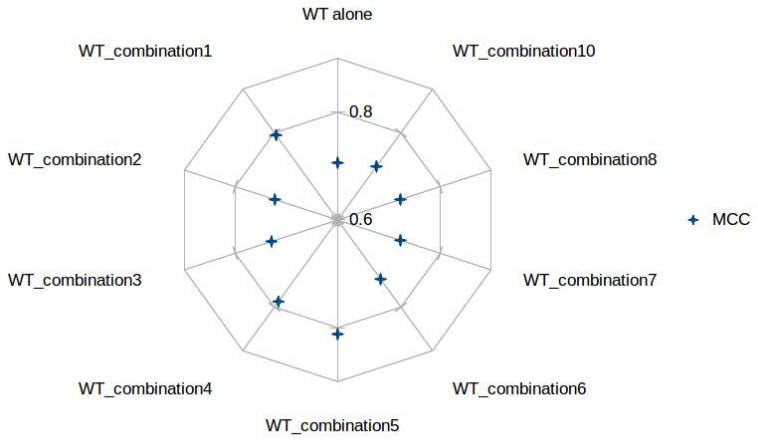


Figure 9.3: `MCC` values from predicted structures using `IPANEMAP` in the case of multi-probing with different mutants. For all tested combinations a higher accuracy is recorded compared to the mono-probing with the WT.

We did extend our predictive model to take into account such structural informative dimension in the case of multi-probing with the `WT`. An additional sample of structures generated from the `MSA` was considered in parallel to the generated samples from the probing data. This extension allowed to resolve Ebola Utrs models (Chapter 8) where predicted structures through `IPANEMAP` have shown to be in accordance with both the `MSA` (a sequence information) and the involved probing data (a structure information).

We have shown that we gain in accuracy prediction when performing a multi-variants `DMS` analysis favoring a new structure inference dimension that is due to the intentionally installed mutations.

We have also noticed a gain in the prediction precision when considering an MSA information as a further analysis dimension.

These two promising results witnessed the versatility and the robustness of our integrative modeling method `IPANEMAP` with the use of a new dimension informing about the structure. This gives a good hope to expand `IPANEMAP` method to deal with the case of multi-variants `SHAPEMap` analysis.

9.3 Conclusion

The main contribution of this thesis is the development of a new predictive method that integrates multiple probing data, issued either from different techniques or from various RNA mutants, as proxies within the thermodynamic modeling.

Through the application of `IPANEMAP` to predict various RNA models, we have been able to show the interest of using multiple sources of probing data within a sampling/clustering approach to gain in prediction accuracy. In the future, our method `IPANEMAP` could be further improved by developing and integrating an RNA structure dedicated clustering algorithm instead of relying on the use of classic clustering algorithm such as `k-means`. In addition, defining the optimal combination of probing data types to guaranty a faithful inference of the RNA structure was one of the faced challenge in this thesis. It is also one of the points that began to arouse the interest of RNA community. Being able to decide for the optimal combination for a specific RNA with a specific data probing is expected to allow for a considerable gain of accuracy with the use of `IPANEMAP`.

Our integrative method is closely dependent on the classic thermodynamic modeling. As this model still suffers from several flaws, especially when considering auxiliary information such as probing data, several improvement directions could be followed with this regards. Reviewing the interpretation of different sources of probing data with their integration as pseudo-energy is particularly interesting for considering new analysis dimensions such as the sequence context.

Because of the imperfection of probing data protocols and the imprecision of the analysis methods, there still is a long way to ensure a direct inference of the RNA structure from probing data. Developing new algorithms around the thermodynamic classic modeling while considering the agreement between different probing data to alleviate these biases is one of the way to gain in prediction accuracy and ultimately accelerate the discovery of novel functional RNAs. Still, to what extent can one gain in prediction precision while integrating auxiliary data? and are we close to reach the precision limits inherent to prediction algorithms in the presence of probing data?

Bibliography

Jules Deforges, Sylvain De Breyne, Melissa Ameur, Nathalie Ulryck, Nathalie Chamond, Afaf Saaidi, Yann Ponty, Theophile Ohlmann, and Bruno Sargueil. Two ribosome recruitment sites direct multiple translation events within HIV1 Gag open reading frame. *Nucleic Acids Research*, 2017. doi: 10.1093/nar/gkx303.

Pablo Cordero, Wipapat Kladwang, Christopher C Vanlang, and Rhiju Das. Quantitative Dimethyl Sulfate Mapping for Automated RNA Secondary Structure Inference. *Biochemistry*, 2012. doi: 10.1021/bi3008802. URL https://pubs.acs.org/doi/pdf/10.1021/bi3008802.

Peter B. Moore and Thomas A. Steitz. The roles of RNA in the synthesis of protein. *Cold Spring Harbor Perspectives in Biology*, 3(11), nov 2011.

Alexander Serganov and Evgeny Nudler. A decade of riboswitches. *Cell*, 152(1-2):17–24, jan 2013. ISSN 1097-4172. doi: 10.1016/j.cell.2012.12.024. URL http://www.ncbi.nlm.nih.gov/pubmed/23332744http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4215550.

J Nowakowski and I Tinoco. RNA structure and stability. *Seminars in Virology*, 8(3):153–165, 1997. ISSN 10445773. doi: 10.1006/smvy.1997.0118.

Barbara L. Golden. Preparation and crystallization of RNA. *Methods in Molecular Biology*, 2007. ISSN 10643745. doi: 10.1385/1-59745-209-2:239.

Brian Houck-Loomis, Michael A. Durney, Carolina Salguero, Neelaabh Shankar, Julia M. Nagle, Stephen P. Goff, and Victoria M. D'Souza. An equilibrium-dependent retroviral mRNA switch regulates translational recoding. *Nature*, 480(7378):561–564, dec 2011. ISSN 0028-0836. doi: 10.1038/nature10657. URL http://www.ncbi.nlm.nih.gov/pubmed/22121021http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3582340http://www.nature.com/articles/nature10657.

J SantaLucia and D H Turner. Measuring the thermodynamics of RNA secondary structure formation. *Biopolymers*, 44 (3):309–319, 1997. ISSN 0006-3525. doi: 10.1002/(SICI)1097-0282(1997)44:3⟨309::AID-BIP8⟩3.0.CO;2-Z.

Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 1981. ISSN 03051048. doi: 10.1093/nar/9.1.133.

David H. Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 2004. ISSN 13558382. doi: 10.1261/rna.7650904.

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

Stefan Wuchty, Walter Fontana, Ivo L. Hofacker, and Peter Schuster. Complete suboptimal folding of RNA and the stability of secondary structures, feb 1999.

Y. Ding and C.E Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301, 2003. ISSN 1362-4962. doi: 10.1093/nar/gkg938. URL http://nar.oxfordjournals.org/content/31/24/7280.long.

Ivo L. Hofacker, Martin Fekete, and Peter F. Stadler. Secondary structure prediction for aligned RNA sequences. *Journal of Molecular Biology*, 319(5):1059–1066, 2002.

Daniel Gautheret and Robin R. Gutell. Inferring the conformation of RNA base pairs and triples from patterns of sequence variation. *Nucleic Acids Research*, 1997. ISSN 03051048. doi: 10.1093/nar/25.8.1559.

Edward J. Merino, Kevin A. Wilkinson, Jennifer L. Coughlan, and Kevin M. Weeks. RNA structure analysis at single nucleotide resolution by Selective 2'-Hydroxyl Acylation and Primer Extension (SHAPE). *Journal of the American Chemical Society*, 2005. ISSN 00027863. doi: 10.1021/ja043822v.

Michael Kertesz, Yue Wan, Elad Mazor, John L. Rinn, Robert C. Nutter, Howard Y. Chang, and Eran Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 2010. ISSN 14764687. doi: 10.1038/nature09322.

K. E. Deigan, T. W. Li, D. H. Mathews, and K. M. Weeks. Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences*, 2009. ISSN 0027-8424. doi: 10.1073/pnas.0806929106.

Stefan Washietl, Ivo L. Hofacker, Peter F. Stadler, and Manolis Kellis. RNA folding with soft constraints: Reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Research*, 2012. ISSN 03051048. doi: 10.1093/nar/gks009.

Kourosh Zarringhalam, Michelle M. Meyer, Ivan Dotu, Jeffrey H. Chuang, and Peter Clote. Integrating Chemical Footprinting Data into RNA Secondary Structure Prediction. *PLoS ONE*, 2012. ISSN 19326203. doi: 10.1371/journal.pone.0045160.

José Almeida Cruz, Marc Frédérick Blanchet, Michal Boniecki, Janusz M. Bujnicki, Shi Jie Chen, Song Cao, Rhiju Das, Feng Ding, Nikolay V. Dokholyan, Samuel Coulbourn Flores, Lili Huang, Christopher A. Lavender, Véronique Lisi, François Major, Katarzyna Mikolajczak, Dinshaw J. Patel, Anna Philips, Tomasz Puton, John Santalucia, Fredrick Sijenyi, Thomas Hermann, Kristian Rother, Magdalena Rother, Alexander Serganov, Marcin Skorupski, Tomasz Soltysinski, Parin Sripakdeevong, Irina Tuszynska, Kevin M. Weeks, Christina Waldsich, Michael Wildauer, Neocles B. Leontis, and Eric Westhof. RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, 2012. ISSN 13558382. doi: 10.1261/rna.031054.111.

Rune B Lyngsø and Christian NS Pedersen. Rna pseudoknot prediction in energy-based models. *Journal of computational biology*, 7(3-4):409–427, 2000.

Saad Sheikh, Rolf Backofen, and Yann Ponty. Impact of the energy model on the complexity of rna folding with pseudoknots. In *Annual Symposium on Combinatorial Pattern Matching*, pages 321–333. Springer, 2012.

M. S. Waterman and T. F. Smith. RNA secondary structure: a complete mathematical analysis. *Mathematical Biosciences*, 1978. ISSN 00255564. doi: 10.1016/0025-5564(78)90099-8.

Ruth Nussinov, George Pieczenik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for Loop Matchings. *SIAM Journal on Applied Mathematics*, 1978. ISSN 0036-1399. doi: 10.1137/0135006.

M Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 1989. ISSN 0036-8075. doi: 10.1126/science.2468181.

Ivo L Hofacker. RNA secondary structure analysis using the Vienna RNA package. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 2009. ISSN 1934-340X. doi: 10.1002/0471250953.bi1202s26.

Yann Ponty. Efficient sampling of rna secondary structures from the boltzmann ensemble of low-energy: the boustrophedon method. *Journal of mathematical biology*, 56:107–127, January 2008. ISSN 0303-6812. doi: 10.1007/s00285-007-0137-z.

Juraj Michálik, Hélène Touzet, and Yann Ponty. Efficient approximations of RNA kinetics landscape using non-redundant sampling. In *Bioinformatics*, 2017. doi: 10.1093/bioinformatics/btx269.

Zhi John Lu, Jason W Gloor, and David H Mathews. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA (New York, N.Y.)*, 15(10):1805–1813, 2009.

Michiaki Hamada, Hisanori Kiryu, Kengo Sato, Toutai Mituyama, and Kiyoshi Asai. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics (Oxford, England)*, 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn601.

Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 2003. ISSN 03051048. doi: 10.1093/nar/gkg595.

Jessica S. Reuter and David H. Mathews. RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 2010. ISSN 14712105. doi: 10.1186/1471-2105-11-129.

Xiao chen Bai, Greg McMullan, and Sjors H.W. Scheres. How cryo-EM is revolutionizing structural biology, 2015. ISSN 13624326.

Lincoln G Scott and Mirko Hennig. RNA structure determination by NMR. *Methods in molecular biology (Clifton, N.J.)*, 2008. ISSN 1064-3745. doi: 10.1007/978-1-60327-159-2_2.

Linhui Julie Su, Michael Brenowitz, and Anna Marie Pyle. An alternative route for the folding of large RNAs: Apparent two-state folding by a group II intron ribozyme. *Journal of Molecular Biology*, 2003. ISSN 00222836. doi: 10.1016/j.jmb.2003.09.071.

M. Manosas and Felix Ritort. Thermodynamic and kinetic aspects of RNA pulling experiments. *Biophysical Journal*, 2005. ISSN 00063495. doi: 10.1529/biophysj.104.045344.

Shakti Ingle, Robert N. Azad, Swapan S. Jain, and Thomas D. Tullius. Chemical probing of RNA with the hydroxyl radical at single-atom resolution. *Nucleic acids research*, 2014. ISSN 13624962. doi: 10.1093/nar/gku934.

Thomas D. Tullius and Jason A. Greenbaum. Mapping nucleic acid structure by hydroxyl radical cleavage, 2005. ISSN 13675931.

Yue Wan, Kun Qu, Zhengqing Ouyang, and Howard Y. Chang. Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. *Nature Protocols*, 2013. ISSN 17542189. doi: 10.1038/nprot.2013.045.

Yi Peng, Toby J. Soper, and Sarah A. Woodson. Rnase footprinting of protein binding sites on an mRNA target of small RNAs. *Methods in Molecular Biology*, 2012. ISSN 10643745. doi: 10.1007/978-1-61779-949-5_13.

William A. Ziehler and David R. Engelke. Probing RNA Structure with Chemical Reagents and Enzymes. In *Current Protocols in Nucleic Acid Chemistry*. Wiley, 2000. ISBN 0471142700. doi: 10.1002/0471142700.nc0601s00.

Laurence Lempereur, Monique Nicoloso, Nadine Riehl, Chantal Ehresmann, Bernard Ehresmann, and Jean Pierre Bachellerie. Conformation of yeast 18S rRNA. Direct chemical probing of the 5' domain in ribosomal subunits and in deproteinized RNA by reverse transcriptase mapping of dimethyl sulfate-accessible sites. *Nucleic Acids Research*, 1985. ISSN 03051048. doi: 10.1093/nar/13.23.8339.

Petra Burgstaller, Michel Kochoyan, and Michael Famulok. Structural probing and damage selection of citrullineand arginine-specific RNA aptamers identify base positions required for binding. *Nucleic Acids Research*, 1995. ISSN 03051048. doi: 10.1093/nar/23.23.4769.

P. Zarrinkar and Williamson. Kinetic intermediates in RNA folding. *Science*, 1994. ISSN 0036-8075. doi: 10.1126/science.8052848.

Garrett A. Soukup and Ronald R. Breaker. Relationship between internucleotide linkage geometry and the stability of RNA. *RNA*, 1999. ISSN 13558382. doi: 10.1017/S1355838299990891.

Stefanie A. Mortimer and Kevin M. Weeks. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *Journal of the American Chemical Society*, 2007. ISSN 00027863. doi: 10.1021/ja0704028.

Jennifer L. McGinnis, Jack A. Dunkle, Jamie H.D. Cate, and Kevin M. Weeks. The mechanisms of RNA SHAPE chemistry. *Journal of the American Chemical Society*, 2012. ISSN 00027863. doi: 10.1021/ja2104075.

K. A. Wilkinson, S. M. Vasa, K. E. Deigan, S. A. Mortimer, M. C. Giddings, and K. M. Weeks. Influence of nucleotide identity on ribose 2'-hydroxyl reactivity in RNA. *RNA*, 2009. ISSN 1355-8382. doi: 10.1261/rna.1536209.

Costin M. Gherghe, Zahra Shajani, Kevin A. Wilkinson, Gabriele Varani, and Kevin M. Weeks. Strong correlation between SHAPE chemistry and the generalized NMR order parameter (S2) in RNA. *Journal of the American Chemical Society*, 2008. ISSN 00027863. doi: 10.1021/ja804541s.

Stefanie A. Mortimer and Kevin M. Weeks. Time-resolved RNA SHAPE chemistry: Quantitative RNA structure analysis in one-second snapshots and at single-nucleotide resolution. *Nature Protocols*, 2009. ISSN 17542189. doi: 10.1038/nprot.2009.126.

Jeehyung Lee, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, Snehal Gaikwad, Sungroh Yoon, Adrien Treuille, and Rhiju Das. RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences*, 2014. ISSN 0027-8424. doi: 10.1073/pnas.1313039111.

Rushia Turner, Kinneret Shefer, and Manuel Ares. Safer one-pot synthesis of the 'SHAPE' reagent 1-methyl-7-nitroisatoic anhydride (1m7). *RNA (New York, N.Y.)*, 2013. ISSN 1469-9001. doi: 10.1261/rna.042374.113.

Robert C. Spitale, Pete Crisalli, Ryan A. Flynn, Eduardo A. Torre, Eric T. Kool, and Howard Y. Chang. RNA SHAPE analysis in living cells. *Nature Chemical Biology*, 2013. ISSN 15524450. doi: 10.1038/nchembio.1131.

Wipapat Kladwang, Christopher C. VanLang, Pablo Cordero, and Rhiju Das. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nature Chemistry*, 2011. ISSN 17554330. doi: 10.1038/nchem.1176.

Carlos M. Duarte, Leven M. Wadley, and Anna Marie Pyle. RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Research*, 2003. ISSN 03051048. doi: 10.1093/nar/gkg682.

Kevin A. Wilkinson, Robert J. Gorelick, Suzy M. Vasa, Nicolas Guex, Alan Rein, David H. Mathews, Morgan C. Giddings, and Kevin M. Weeks. High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS Biology*, 2008. ISSN 15449173. doi: 10.1371/journal.pbio.0060096.

Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. SOAP: Short oligonucleotide alignment program. *Bioinformatics*, 2008. ISSN 13674803. doi: 10.1093/bioinformatics/btn025.

Ben Langmead, Steven L Salzberg, and Langmead. Bowtie2. *Nature methods*, 2013. ISSN 1548-7091. doi: 10.1038/nmeth.1923.Fast.

G. M. Rice, C. W. Leonard, and K. M. Weeks. RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA*, 2014. ISSN 1355-8382. doi: 10.1261/rna.043323.113.

Aleksandar Spasic, Sarah M Assmann, Philip C Bevilacqua, and David H Mathews. Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Research*, 46(1):314–323, jan 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1057. URL http://academic.oup.com/nar/article/46/1/314/4643371.

Christopher A Lavender, Ronny Lorenz, Ge Zhang, Rita Tamayo, Ivo L Hofacker, and Kevin M Weeks. Model-Free RNA Sequence and Structure Alignment Informed by SHAPE Probing Reveals a Conserved Alternate Secondary Structure for 16S rRNA. *Plos Computational Biology*, 2015. doi: 10.1371/journal.pcbi.1004126. URL www.chem.unc.edu/.

Katrina M Kutchko and Alain Laederach. Transcending the prediction paradigm: novel applications of SHAPE to RNA function and evolution. *Wiley Interdisciplinary Reviews RNA*, 2017. doi: 10.1002/wrna.1374.

Vladimir Reinharz, Yann Ponty, and Jérôme Waldispühl. Combining structure probing data on RNA mutants with evolutionary information reveals RNA-binding interfaces. *Nucleic Acids Research*, 2016. ISSN 13624962. doi: 10.1093/nar/gkw217.

Clarence Y Cheng, Wipapat Kladwang, Joseph D Yesselman, and Rhiju Das. RNA structure inference through chemical mapping after accidental or intentional mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 114(37):9876–9881, sep 2017. ISSN 1091-6490. doi: 10.1073/pnas.1619897114. URL http://www.ncbi.nlm.nih.gov/pubmed/28851837http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5603990.

Somdeb Mitra, Inna V. Shcherbakova, Russ B. Altman, Michael Brenowitz, and Alain Laederach. High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic Acids Research*, 2008. ISSN 03051048. doi: 10.1093/nar/gkn267.

Fethullah Karabiber, Jennifer L McGinnis, Oleg V Favorov, and Kevin M Weeks. QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA (New York, N.Y.)*, 19(1):63–73, 2013. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3527727{%}7B{&}{%}7Dtool=pmcentrez{%}7B{&}{%}7Drendertype=abstract.

Matthew J. Smola, Greggory M. Rice, Steven Busan, Nathan A. Siegfried, and Kevin M. Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nature Protocols*, 2015. ISSN 17502799. doi: 10.1038/nprot.2015.103.

Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp324.

D Sculley. Web-scale k-means clustering. *Proceedings of the 19th international conference on World wide web WWW 10*, page 1177, 2010. URL http://portal.acm.org/citation.cfm?doid=1772690.1772862.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2012. URL http://dl.acm.org/citation.cfm?id=2078195{%}5Cnhttp://arxiv.org/abs/1201.0490.

Christopher A. Mattson and Achille Messac. Pareto frontier based concept selection under uncertainty, with visualization. *Optimization and Engineering*, 6(1):85–115, mar 2005.

C. E. Hajdin, S. Bellaousov, W. Huggins, C. W. Leonard, D. H. Mathews, and K. M. Weeks. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proceedings of the National Academy of Sciences*, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1219988110.

Ségolène Caboche, Christophe Audebert, Yves Lemoine, and David Hot. Comparison of mapping algorithms used in high-throughput sequencing: Application to Ion Torrent data. *BMC Genomics*, 2014. ISSN 14712164. doi: 10.1186/1471-2164-15-264.

Mirela Andronescu, Vera Bereg, Holger H. Hoos, and Anne Condon. RNA STRAND: The RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, 2008. ISSN 14712105. doi: 10.1186/1471-2105-9-340.

Florian Wickelmaier. An introduction to MDS. *Reports from the Sound Quality Research Unit,* {{}{...}{}}, page 26, 2003. ISSN 13791176. doi: 10.1080/13546800903126016. URL http://steep.inrialpes.fr/{%}7B{%}25{%}7D7B{%}7B{~}{%}7D{%}7B{%}25{%}7D7DArnaud/indexation/mds03.pdf.

Daniel Lai, Jeff R. Proctor, Jing Yun A. Zhu, and Irmtraud M. Meyer. R-CHIE: A web server and R package for visualizing RNA secondary structures. *Nucleic Acids Research*, 2012. ISSN 03051048. doi: 10.1093/nar/gks241.

Karina Yusim, Hyejin Yoon, Brian Foley, Shihai Feng, Jennifer Macke, Mira Dimitrijevic, Werner Abfalterer, James Szinger, Will Fischer, Carla Kuiken, and Bette Korber. Integrated sequence and immunology filovirus database at Los Alamos. *Database : the journal of biological databases and curation*, 2016, 2016. ISSN 1758-0463. doi: 10.1093/database/baw047. URL http://www.ncbi.nlm.nih.gov/pubmed/27103629http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4839628.

Gareth Stockwell. PyMOL tutorial. *Biochemistry*, 2003. ISSN 09249338.

Kévin Darty, Alain Denise, and Yann Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 2009. ISSN 13674803. doi: 10.1093/bioinformatics/btp250.

Steven Busan and Kevin M. Weeks. Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA*, 2018. ISSN 14699001. doi: 10.1261/rna.061945.117.

Kristina Brauburger, Yannik Boehmann, Verena Krähling, and Elke Mühlberger. Transcriptional Regulation in Ebola Virus: Effects of Gene Border Structure and Regulatory Elements on Gene Expression and Polymerase Scanning Behavior. 1993. doi: 10.1128/JVI.02341-15.

Elena Rivas, Jody Clements, and Sean R. Eddy. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature Methods*, 2016. ISSN 15487105. doi: 10.1038/nmeth.4066.

Mirko Ledda and Sharon Aviran. PATTERNA: transcriptome-wide search for functional RNA elements via structural data signatures. *Genome Biology*, 19(1):28, dec 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1399-z. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1399-z.

Alec N. Sexton, Peter Y. Wang, Michael Rutenberg-Schoenberg, and Matthew D. Simon. Interpreting Reverse Transcriptase Termination and Mutation Events for Greater Insight into the Chemical Probing of RNA. *Biochemistry*, 2017. ISSN 15204995. doi: 10.1021/acs.biochem.7b00323.

Angela M Yu, Molly E Evans, and Julius B Lucks. Estimating RNA structure chemical probing reactivities from reverse transcriptase stops and mutations . *bioRxiv*, pages 34–37, 2018. doi: 10.1101/292532. URL http://dx.doi.org/10.1101/292532.

Hua Li and Sharon Aviran. Statistical modeling of RNA structure profiling experiments enables parsimonious reconstruction of structure landscapes. *Nature Communications*, 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-02923-8.

Chapter 10

Appendix

10A Reads distribution

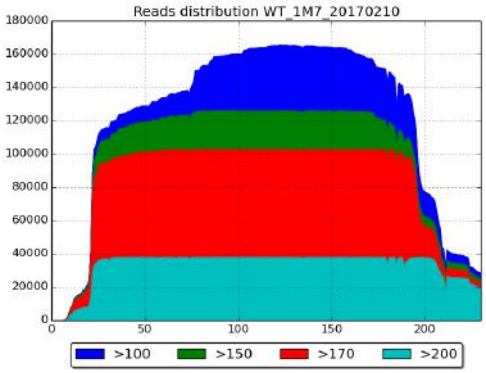
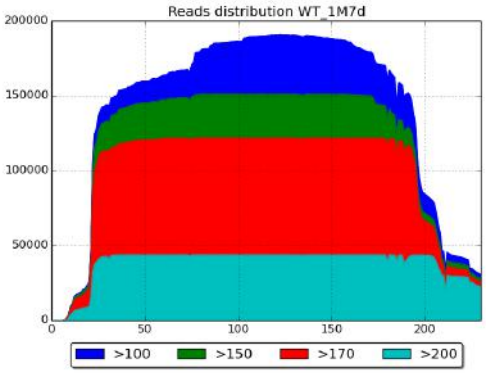
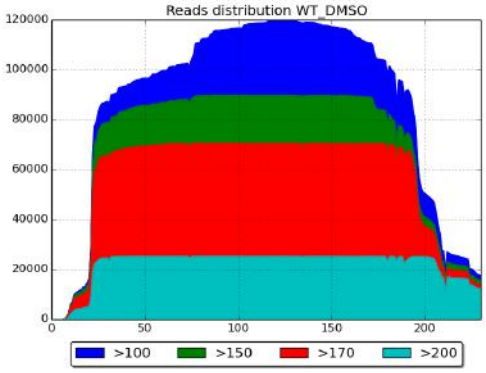
SHAPE-1M7:	
Denatured-1M7:	
Untreated:	

Table 10.1: **Reads distribution from TMAP-mapping for** GIR1 Lariat-capping ribozyme**.**

10B IPANEMAP Vs. Rsample

RNA	Rsample			IPANEMAP		
	Sensitivity	PPV	Accuracy	Sensitivity	PPV	Accuracy
PreQ1riboswitchBsubtilis	0.625	1	0.791	0.625	1	0.791
5domainof16SrRNAEcoli	0.6622	0.6203	0.641	0.8919	0.8049	0.847
SignalrecognitonparticleRNAhuman	0.03	0.0313	0.031	0.6	0.6	0.6
GroupIintronAzoarcussp	0.6349	0.6667	0.651	0.7302	0.8519	0.789
HIV15primepseudoknotdomain	0.6382	0.7823	0.707	0.5263	0.5517	0.539
Telomerasepseudoknothuman	0.4	0.75	0.548	0.5333	0.7273	0.623
SAMIriboswitchTtengcongensis	0.7179	0.8	0.758	0.7436	0.8286	0.785
tRNApheEcoli	0.4762	0.5556	0.514	0.7619	0.7619	0.762
P546domainbI3groupIintron	0.6786	0.8261	0.749	0.9643	0.9818	0.973
TPPriboswitchEcoli	0.7727	0.85	0.81	0.9545	0.875	0.914
AdenineriboswitchVvulnificus	1	1	1	0.9524	0.9524	0.952
FluorideriboswitchPsyringae	0.5625	0.6923	0.624	0.625	0.7143	0.668
SARScoronaviruspseudoknot	0.6538	0.8095	0.727	0.6538	0.68	0.667
5SrRNAEcoli	0.2857	0.2632	0.274	0.9714	0.9189	0.945
cyclicdiGMPriboswitchVcholerae	0.9286	0.9286	0.929	0.9286	0.8966	0.912
5domainof23SrRNAEcoli	0.9328	0.7762	0.851	0.8824	0.7664	0.822
RNasePBsubtilis	0.5826	0.5583	0.57	0.7304	0.75	0.74
GroupIIntronTthermophila	0.8397	0.8271	0.833	0.9084	0.8815	0.895
HepatitisCvirusIRESdomain	0.5481	0.6129	0.58	0.8077	0.866	0.836
5domainof16SrRNAHvolcanii	0.7986	0.7143	0.755	0.875	0.8182	0.846
GroupIIintronOiheyensis	0.5379	0.6174	0.576	0.75	0.8462	0.797
tRNAaspyeast	0.619	0.6842	0.651	0.619	0.5652	0.591
LysineriboswitchTmaritime	0.8095	0.8644	0.836	0.8095	1	0.9
MBoxriboswitchBsubtilis	0.875	0.913	0.894	0.875	0.913	0.894

Table 10.2: `PPV`, `sensitivity` and accuracy for predicted structures with `Rsample` and with `IPANEMAP`.

10C GIR1 Lariat-capping ribozyme

	1M7ILUMg	1M7ILU	DMS Mg	NMIAMg	NMIA	NMIAMgCE	BzCNMg	1M7Mg	1M7	NaiMg	1M7ILU3Mg	1M7ILU3	Nai	CMCTMg
1M7ILUMg	0.807	0.777	0.797	0.756	0.76	0.729	0.773	0.773	0.773	0.773	0.777	0.807	0.591	0.746
1M7ILU		0.7946	**0.804**	0.763	0.756	0.743	0.76	0.729	0.794	0.787	0.732	0.787	0.591	0.78
DMS Mg			0.7766	0.76	0.76	**0.786**	**0.786**	0.76	**0.797**	0.753	0.673	0.606	0.591	0.77
NMIAMg				0.76	0.76	0.753	0.688	0.766	0.766	0.76	0.673	0.76	0.591	0.76
NMIA					0.7536	**0.766**	0.709	0.753	0.753	**0.76**	0.753	0.667	0.591	0.572
NMIAMgCE		0.787				0.7208	0.714	**0.743**	**0.743**	0.72	0.701	0.614	0.591	0.572
BzCNMg				0.753	0.747		0.709	**0.727**	**0.749**	0.692	0.673	0.614	0.591	0.709
1M7Mg	0.7496	0.787						0.7164	0.716	**0.753**	**0.767**	0.614	0.591	0.586
1M7	0.807								0.6797	**0.76**	**0.767**	0.629	0.591	0.486
NaiMg										0.6767	**0.684**	0.614	0.591	0.572
1M7ILU3Mg											0.6733	**0.767**	0.591	0.673
1M7ILU3												0.6142	0.591	0.486
Nai													0.5911	0.591
CMCTMg														0.5564

Table 10.3: MCC values for predicted structures using `IPANEMAP` when combining probing conditions two by two. MCC values from mono-probing are highlighted in blue. Values in red correspond to the case where two clusters are found to be optimal. Values in bold correspond the case where the combination of allows to improve the accuracy of predicted structure.

	1M7ILUMg	1M7ILU	DMS Mg	NMIAMg	NMIA	NMIAMgCE	BzCNMg	1M7Mg	1M7	NaiMg	1M7ILU3Mg	1M7ILU3	Nai	CMCTMg
1M7ILUMg	0	-0.0238	**0.0052**	-0.0275	-0.0203	-0.0349	**0.015**	**0.0113**	**0.02965**	**0.03115**	**0.03685**	**0.0964**	-0.10805	**0.0643**
1M7ILU		0	**0.0184**	-0.0143	-0.0181	-0.0147	**0.0082**	-0.0265	**0.05685**	**0.05135**	-0.00195	**0.0826**	-0.10185	**0.1045**
DMS Mg			0	-0.0083	-0.0051	**0.0373**	**0.0432**	**0.0135**	**0.06885**	**0.02635**	-0.05195	-0.0894	-0.09285	**0.1035**
NMIAMg				0	**0.0032**	**0.0126**	-0.0465	**0.0278**	**0.04615**	**0.04165**	-0.04365	**0.0729**	-0.08455	**0.1018**
NMIA					0	**0.0288**	-0.0223	**0.018**	**0.03635**	**0.04485**	**0.03955**	-0.0169	-0.08135	-0.083
NMIAMgCE		**0.0293**				0	-0.0009	**0.0244**	**0.04275**	**0.02125**	**0.00395**	-0.0535	-0.06495	-0.0666
BzCNMg			**0.0102**	**0.0125**			0	**0.0143**	**0.05465**	-0.00085	-0.01815	-0.0476	-0.05905	**0.0763**
1M7Mg	-0.0121	**0.0315**						0	**0.01795**	**0.05645**	**0.07215**	-0.0513	-0.06275	-0.0504
1M7	**0.06365**								0	**.0818**	**0.0905**	-0.01795	-0.0444	-0.13205
NaiMg										0	**0.009**	-0.03145	-0.0429	-0.04455
1M7ILU3Mg											0	**0.12325**	-0.0412	**0.05815**
1M7ILU3												0	-0.01165	-0.0993
Nai													0	**0.01725**
CMCTMg														0

Table 10.4: Comparison of MCC values from the bi-probing and the average of MCC from the corresponding mono-probing predictions. Values in bold correspond to the case where the performance of the bi-probing is higher than the average on the respective MCC values in mono-probing mode.

Titre : Analyse différentielle de données de sondage pour la prédiction des structures d'acides ribonucléiques

Mots clés : Structure d'ARN, échantillonnage, données de probing, SHAPE, profil de mutation, clustering

Résumé : En bioinformatique structurale, la prédiction de la (des) structure(s) secondaire(s) des acides ribonucléiques (ARNs) constitue une direction de recherche majeure pour comprendre les mécanismes cellulaires. Une approche classique pour la prédiction de la structure postule qu'à l'équilibre thermodynamique, l'ARN adopte plusieurs conformations, caractérisées par leur énergie libre, dans l'ensemble de Boltzmann. Les approches modernes privilégient donc une considération des conformations dominantes. Ces approches voient leur précision limitées par l'imprécision des modèles d'énergie et les restrictions topologiques pesant sur les espaces de conformations. Les données expérimentales peuvent être utilisées pour pallier aux lacunes des méthodes de prédiction. Différents protocoles permettent ainsi la révélation d'informations structurales partielles via une exposition à un réactif chimique/enzymatique, dont l'effet dépend, et est donc révélateur, de la (les) structure(s) adoptée(s). Les données de sondage mono-réactif sont utilisées pour valider et complémenter les modèles d'énergie libre, permettant ainsi d'améliorer la précision des prédictions. En pratique, cependant, les praticiens basent leur modélisation sur des données de sondage produites dans diverses conditions expérimentales, utilisant différents réactifs ou associées à une collection de séquences mutées. Une telle approche intégrative est répandue mais reste manuelle, onéreuse et subjective. Au cours de cette thèse, nous avons développé des méthodes in silico pour une modélisation automatisée de la structure à partir de plusieurs sources de données de sondage. En premier lieu, nous avons établi des pipelines d'analyse automatisés pour l'acquisition de profils de réactivité à partir de données brutes produites à travers une série de protocoles. Nous avons ensuite conçu et implémenté une nouvelle méthode qui permet l'intégration simultanée de plusieurs profils de sondage. Basée sur une combinaison d'échantillonnage de l'ensemble de Boltzmann et de clustering structurel, notre méthode produit des conformations dominantes, stables et compatible avec les données de sondage. En favorisant les structures récurrentes, notre méthode permet d'exploiter la complémentarité entre plusieurs données de sondage. Ses performances dans le cas mono-sondage sont comparables ou meilleures que excèdent permettant ainsi de retrouver, sinon d'améliorer, les celles des méthodes prédictives de pointe. Cette méthode a permis de proposer des modèles pour les régions structurées des virus. En collaboration avec des expérimentalistes, nous avons suggéré une structure raffinée de l'IRES du VIH-1 Gag, compatible avec les données de sondage chimiques et enzymatiques, qui nous a permis d'identifier des sites d'interactions putatifs avec le ribosome. Nous avons également modélisé la structure des régions non traduites d'Ebola. Cohérents avec les données de sondage SHAPE et les données de covariation, nos modèles montrent l'existence d'une tige-boucle conservée et stable à l'extrémité 5', une structure typiquement présente dans les génomes viraux pour protéger l'ARN de la dégradation par les nucléases. L'extension de notre méthode pour l'analyse simultanée de variants, appliquée dans un premier temps sur des mutants produits par le protocole Mutate-and-Map et sondés par le DMS, a permis d'enregistrer une amélioration en précision de prédiction. Pour éviter la production systématique de mutants ponctuels et exploiter le protocole récent SHAPE-Map, nous avons conçu un protocole expérimental basé sur une mutagenèse non dirigé et le séquençage, où plusieurs ARN mutés sont produits et simultanément sondés. Nous avons traité l'affectation des reads aux mutants de références à l'aide d'une instance de l'algorithme "Expectation-Maximization" dont résultats préliminaires, sur un échantillon de reads réduit/simulé, ont montré un faible taux d'erreurs d'assignation.

Title : Multi-dimensional probing to predict the RNA secondary structure

Keywords : RNA structure, sampling, probing data, SHAPE, mutational profil, clustering

Abstract : In structural bioinformatics, predicting the secondary structure(s) of ribonucleic acids (RNAs) represents a major direction of research to understand cellular mechanisms. A classic approach for structure postulates that, at the thermodynamic equilibrium, RNA adopts its various conformations according to a Boltzmann distribution based on its free energy. Modern approaches, therefore, favor the consideration of the dominant conformations. Such approaches are limited in accuracy due to the imprecision of the energy model and the structure topology restrictions. Experimental data can be used to circumvent the shortcomings of predictive computational methods. RNA probing encompasses a wide array of experimental protocols dedicated to revealing partial structural information through exposure to a chemical or enzymatic reagent, whose effect depends on, and thus reveals, features of its adopted structure(s). Accordingly, single-reagent probing data is used to supplement free-energy models within computational methods, leading to significant gains in prediction accuracy. In practice, however, structural biologists integrate probing data produced in various experimental conditions, using different reagents or over a collection of mutated sequences, to model RNA structure(s). This integrative approach remains manual, time-consuming and arguably subjective in its modeling principles. In this Ph.D., we contributed in silico methods for an automated modeling of RNA structure(s) from multiple sources of probing data. We have first established automated pipelines for the acquisition of reactivity profiles from primary data produced through a variety of protocols (SHAPE, DMS using Capillary Electrophoresis, SHAPE-Map/Ion Torrent). We have designed and implemented a new, versatile, method that simultaneously integrates multiple probing profiles. Based on a combination of Boltzmann sampling and structural clustering, it produces alternative stable conformations jointly supported by a set of probing experiments. As it favors recurrent structures, our method allows exploiting the complementarity of several probing assays. The quality of predictions produced using our method compared favorably against state-of-the-art computational predictive methods on single-probing assays. Our method was used to identify models for structured regions in RNA viruses. In collaboration with experimental partners, we suggested a refined structure of the HIV-1 Gag IRES, showing a good compatibility with chemical and enzymatic probing data. The predicted structure allowed us to build hypotheses on binding sites that are functionally relevant to the translation. We also proposed conserved structures in Ebola Untranslated regions, showing a high consistency with both SHAPE probing and evolutionary data. Our modeling allows us to detect conserved and stable stem-loop at the 5'end of each UTR, a typical structure found in viral genomes to protect the RNA from being degraded by nucleases. Our method was extended to the analysis of sequence variants. We analyzed a collection of DMS probed mutants, produced by the Mutate-and-Map protocol, leading to better structural models for the GIR1 lariat-capping ribozyme than from the sole wild-type sequence. To avoid systematic production of point-wise mutants, and exploit the recent SHAPEMap protocol, we designed an experimental protocol based on undirected mutagenesis and sequencing, where several mutated RNAs are produced and simultaneously probed. Produced reads must then be re-assigned to mutants to establish their reactivity profiles used later for structure modeling. The assignment problem was modeled as a likelihood maximization joint inference of mutational profiles and assignments, and solved using an instance of the "Expectation-Maximization" algorithm. Preliminary results on a reduced/simulated sample of reads showed a remarkable decrease of the reads assignment errors compared to a classic algorithm.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France