TESI DOCTORAL UPF 2018

Approaches to characterize structural properties of RNA

Riccardo Delli Ponti

Director

Dr. Gian Gaetano Tartaglia
Gene Function and Evolution Group
Bioinformatics and Genomics Department
Centre for Genomic Regulation (CRG)

Per i miei genitori,
Pancrazio e Carmen

"Welcome to my thesis. Come freely. Go safely;
and leave something of the happiness you bring."

Modified from Bram Stoker, Dracula

Acknowledgments

If one of the committee members is reading this part, I suggest you to jump directly to the part of the real science, because in this section you would just find some feelings drived by "Oh my I am finishing the PhD!" and a bit of delirium in Italian. But if you want to go on, at least I warned you.

I was never good at talking about myself, but since I know that usually the Acknowledgments have more readings than the Discussion in the lab, I will try to put effort.

The PhD was a moment of big changes for me. I still remember when I saw the CRG terrace during the Caixa interviews thinking "Wow! It would be great to work here". At the beginning was difficult to change country and to leave my parents, my friends, and a dog that at the time was still thin (the reverse way for me). During these years of PhD I have changed and learned a lot. Sometimes when I think at the person that I was five years ago, I almost don't recognize him, luckily in a positive way.

Lot of new people from all around the world became part a of my new life: friends, colleagues and more. While the people in Italy are the backbone of my life, the new people in the sunny Barcelona enriched my life with new and positive experiences. Some moments were tough, there were difficulties to overcome, but in life nothing is just as we say in Italy "rose e fiori", and challenges are a fundamental part of it.

When things start to get tough, I always remember "Hey, once I've jumped from a plane, it cannot be worst!" (who knows me will understand). Since this part is called "Ringraziamenti", I would like to thank the people that were, are and I hope will be a positive part of my life.

I will start thanking Gian, because this path would not be possible without

him. This sounds quiet cheesy but it is true. It is thanks to Gian, who selected me during the Caixa call, that I had the possibility to start this new life in Barcelona. In Italy we are usually quiet formal and in many places the hierarchy is strong. I remember when I casually met Gian (that at the time I addressed him as Dr. Tartaglia) on my first plane coming to Barcelona. I was wearing a AC/DC t-shirt and when Gian saw me in the plane, after saying "Cool t-shirt!", started to talk about science... and the plane was close to fly. In that moment I was thinking: "Maybe this working place will be different". And it was true, and after four years of learning, jokes, training networks and AUCs, my life changed. Thanks Gian for giving me the opportunity to be here and for guiding me all these years.

I would also like to thank Michiel at RIKEN for giving me the possibility to join his group for three months. Working and living in Japan was an incredible and formative experience that I will always remember.

I was also lucky for the working environment. Seriously, I think that I could hardly have found such funny lab in any other place. At Tartaglia's lab people stay and go. Some friends are still there, while others already moved to other places. But this is life in science.

I would start thanking Davide, which was my mentor during the transition at the beginning of the PhD, the "old one" to bother for any advice, always ready for a coffee break or to try a new restaurant around Barcelona. I remember my first beer with him and Federico here in Barcelona, when I couldn't believe that in the lab there was somebody going around with a camera 24h a day... but well life is full of surprises.

Talking of other ex-Tartaglia's, I would like to thank Mimma for the dinners, the beers and the jokes. Thanks Mimma also for your ability to jump till the roof when I scared you in some random Sundays at the lab, when you were alone in the lab.

I would like to thank Nieves for the laughs, her energy, the funny talks, the coffees and for establishing the legendary "carajillo del viernes". I always remember how much I laughed when we were shooting the "French movie" for the video of Mimma and I discovered the legendary acting talent of Nieves.

I would like to thank Fernando for his craziness and all his weird, funny and surreal histories. I still think that somebody should follow him around with a camera, would be an interesting comedy show. Where is Petr when we need him

I would like to thank Alex for his help when facing catRAPID and for suggesting me the best Greek restaurant that I've ever tried in Barcelona. I would like to thank him also for being the "mojito man" at every lab party.

I would like to thank the postdocs and technicians for their advices: Teresa, Natalia, Benni, Ben and Iona. Benni for the funny pictures of Alice that you sent in the group and for helping me with the formality of the cover letters. Thanks Teresa for the calcotadas at your garden.

I would like to specifically thank Stefanie for her help with revisions of the papers and to share the duties of the RNA secondary structure predictions. A special thanks also for her help correcting the thesis, which would be a lot different, probably full of references to the cobra venom.

Thanks Elias, Maria Carla and Alessandro for the complex and funny histories made of bizarre names and angles in the mirror during lunch. I would also like to specifically thank Alessandro and Maria Carla for their comedy role as married couple always fighting, remembering me "Raimondo and Sandra" (old Italian TV show).

I would also like to thank who has been at the lab for few time, but with whom was nice to hang out for a beer or a coffee, so thanks Laura, Irene B. and Mireia.

Now is the time to thank the friends of Barcelona that are not part of the lab. Silvina for the many dinners that we had (especially for her empanadas dinners), for her modernist advices and for her acting support in many thesis videos. Alessandra for all the crazy roomescapes that we made together (luckily escaping). Fran for being always ready for a dinner out and to talk about cinema. I would also like to thank Shalu, especially for her braveness trying to teach me Bollywood.

I would also like to thank people from Gabaldon's lab for all the international dinners and parties that I joined, and especially Miguel, Marina and Veronica

for being also a good roomescape team.

Ora è tempo di scrivere un po' in italiano per ringraziare amici e familiari. Vorrei ringraziare tutti gli amici dei Castelli, ci conosciamo dalle superiori e la lista è lunga. Vorrei ringraziarvi per tutte le birre, i kebab, i capodanni, i passaggi scroccati, le nerdate, gli aggravati, gli scherzi, le risate e per essere sempre presenti, ancora oggi dopo tanti anni a prescindere da dove siamo.

In ordine alfabetico comincio ringraziando Andrea per tutti i consigli, le chiacchierate, le ospitate e per aver stoicamente resistito a tutti gli scherzi e le scarpe perse senza mai "prenderla sul personale". Brizio per tutto il cazzeggio dai tempi delle medie, per tutti i film/giochi/cartoni trovati, visti e discussi, e per tutti gli oggetti che sono ancora ostaggio in casa mia (con l'uso capione direi che ci siamo). Ringrazio Damiano per avermi fatto arrivare sano e salvo all'hotel a Praga (con il contributo di Andrea) e per aver sempre tenuto alto, ancora oggi, l'onore del Metal. Ringrazio Emanuele per le migliaia di chiacchierate, gli esami passati studiando insieme, i caffè a qualunque ora e per tutti i passaggi scroccati (ma Andrea e Brizio potrebbero dire che ne ho scroccati più a loro). Ringrazio Fulvio per tutti i Kebab di mezzanotte, per tutti i discorsi di cinema (anche se spesso in disaccordo) e per la sua "leggendaria calma". Ringrazio Joel per la sua costruttiva follia, per le sue birre di contrabbando, per tutti gli scherzi organizzati e per essere stato un degno rivale a ping-pong. Ringrazio Luigi per essere sempre presente, per tutti i "dopo questa mi devi un film", per le risikate fino a tarda notte e specialmente per il suo "leggendario" inglese. Ringrazio Simone per tutti i folli progetti iniziati (ma non conclusi), per i libri scambiati, le chiacchierate sulla scrittura, per avermi dimostrato che passate le 22 si può dormire in qualunque luogo... e aggiungo per i 3 aggravati appena inflitti (critico su Andrea).

Vorrei specialmente ringraziare Joel, Simone, Luigi e gli altri che mi sono stati vicini in uno dei momenti piú difficili.

La lista è lunga e lo spazio poco, concludo ringraziando in ordine sparso Simona, Lucia, Elena, MMM, Sara, Francesca, Francesco "Verzy", Valentina, Cecilia, Betta, Olga, Martina e tutti gli altri.

A bit of English before concluding in Italian. I would like to thank especially

Irene. I remember when we first met during the security training at CRG, and how she was wandering around the corridors with a small map to find the room. Since that day everything was different. She was my first friend here in Barcelona, and then someone even more important. These years in Barcelona would not be the same without her. I would like to thank her for all the travels, the activities, the dinners, the smiles, the laughs, the support and for any simple moment that we shared together during these years.

Prima di concludere vorrei ringraziare nonna Marietta. In questo momento avrebbe avuto 100 anni e credo sarebbe stata molto fiera di me.

Vorrei concludere ringraziando i miei genitori, Pancrazio e Carmen, per essermi stati sempre vicini e per avermi sostenuto in ogni scelta. Oltre ad essere genitori per me sono sempre stati degli amici, con cui poter parlare di qualunque cosa e a cui chiedere consiglio. Non mi hanno mai privato della libertá e mi hanno sempre permesso di scegliere con la mia testa, di sbagliare e di apprendere dalle mie decisioni. Li ringrazio per tutto il loro appoggio in ogni momento della mia vita, per esserci sempre, anche mentre sono lontano. So che avere l'unico figlio lontano, spesso viaggiando per mezzo mondo vi fa preoccupare, ma apprezzo sempre la vostra presenza mai critica o demoralizzante, ma sempre costruttiva e positiva. Non sarei mai arrivato dove sono e vissuto le esperienze fatte se non fosse stato per il vostro appoggio e per come mi avete educato. Grazie.

I don't know who will really read till here, but I will conclude thanking all the people that were a positive part of my life.

Riccardo Delli Ponti

Barcelona, September 2018

Abstract

The secondary structure of an RNA molecule is fundamental for its function. However, structural conservation and the structure of RNA *in vivo* are still poorly understood. Data from recent high-throughput experiments can provide new insights, but they have not yet been systematically exploited. The aim of my doctoral studies was to exploit these experimental data to develop computational approaches for discovering and analyzing structural properties of RNA. I developed two algorithms: CROSS predicts the secondary structure propensity profile of an RNA, and CROSSalign discovers structural similarities among different RNAs. In addition, I studied the effect of the presence of protein binding motifs on the prediction of the RNA structure *in vivo* and investigated how the propensity of RNAs to bind to proteins could be exploited to create a predictive tool. The suite of tools that I developed opens new possibilities for studying the structural properties of long RNA molecules and for investigating structural conservation in large-scale analyses.

Resumen

La estructura secundaria del ARN es fundamental para su función. Sin embargo, la conservación estructural y la estructura del ARN *in vivo* son poco conocidas. Los datos provenientes de experimentos de alto rendimiento pueden proporcionar nuevos conocimientos, pero aún no han sido usados sistemáticamente. El objetivo de mis estudios de doctorado fue emplear estos datos experimentales con el fin de desarrollar metodos computacionales para el descubrimiento y el análisis de las propiedades estructurales del ARN. Como resultado de mi tesis desarrollé dos algoritmos: CROSS, que predice el perfil de propensión de estructura secundaria de un ARN; y CROSSalign, que busca similitudes estructurales entre diferentes ARNs. Además, estudié el efecto de la presencia de dominios de unión proteínica en la predicción de la estructura del ARN *in vivo*; e investigué cómo la propensión de los ARNs a unirse a las proteínas podría usarse para crear un modelo predictivo. El conjunto de herramientas que desarrollé abren nuevas posibilidades para estudiar las propiedades estructurales de moléculas de ARN largas y para investigar la conservación estructural en análisis a gran escala.

Preface

The RNA secondary structure (RSS) is crucial for the biological activity of the RNA, from the interaction with proteins to the correct three-dimensional folding. Crystallographic techniques such as NMR and X-ray, which set the standard for the understanding of protein structures, can also be employed to probe the structure of RNA, but up to date there are very little NMR/X-ray structural data available for RNA.

To compensate for this lack of data, an increasing number of computational approaches for predicting the secondary structure based on the sequence were developed. However, the vast majority of these approaches are based on the thermodynamic properties of isolated RNA, prohibitively slow for long (>10'000 nucleotides) sequences, and built on limited low-throughput data.

In the last years, several novel experimental techniques that are based on chemical probes or on enzymes were able to profile the RNA structure genome-wide. The new results provided the scientific community with an extensive view of the functionality of the RNA secondary structure. Also the experimental profiling of several long non-coding RNAs (lncRNAs) provided information on the complex structure of these large molecules.

The recent flow of data from high-throughput experimental techniques has not yet been systematically exploited for the development of computational approaches to predict structural properties of the RNA. In **Chapter I**, I will present CROSS, an algorithm trained on high-throughput experimental data. The tool is able to profile multiple RNAs at single nucleotide resolution and without sequence length restriction. CROSS was trained on multiple experimental datasets and each model can reproduce a specific technique. CROSS is a powerful tool that can be applied on long non-coding RNAs or on complete transcriptomes. CROSS technology was also applied to study other

complex structural properties of the RNA. The conservation of the RNA secondary structure is still a debated topic, especially in lncRNAs, where also sequence conservation is not well established. In **Chapter II** I will present CROSSalign, an algorithm based on CROSS that is able to assess structural similarities between RNAs of different lengths. The tool was applied to study the structural conservation of important lncRNAs such as *HOTAIR* and *Xist*, and also to identify possible regulatory regions common to single-stranded RNA viruses.

While the processes behind the RNA folding *in vitro* are well defined, the *in vivo* folding is a complex system affected by several features. However, although some techniques are able to profile the RNA structure *in vivo*, computational approaches are still not able to use and to predict the *in vivo* data. In **Chapter III**, I present an analysis of the effect of the presence of proteins and of a crowded environment on the RNA secondary structure *in vivo*. The results have indicated for the first time that binding to proteins has an influence on secondary structure folding and that knowledge of protein binding properties can thus improve the RNA secondary structure prediction. The approach behind this analysis will be used to build CROSSalive, an algorithm able to predict *in vivo* structural data with higher performances.

The **Discussion** of my thesis will highlight the main results of **Chapter I**, **Chapter II**, and **Chapter III**, and their importance for the scientific community. In Future perspectives I will give an overview of my personal opinion regarding the possible future development of the field. I will close my thesis with the **Conclusions**, where I will summarize my findings and their importance. In the **Appendix** section I will include the complete list of publications and the supplementary materials of the main chapters.

Contents

Introduction

1 The RNA and its structure

1.1 The nucleotides and the primary structure

The RNA is a single stranded molecule composed of four base nucleotides. Each nucleotide is formed of a heterocyclic base, a pentose sugar (ribose) and a phosphoric acid (H_3PO_4) (Watson and Crick, 1953).

The bases are classified according to their core structure in pyrimidines, with only one heterocyclic ring, and purines, with two heterocyclic structures. The pyrimidines present in RNA are uracil (U) and cytosine (C), while the purines are adenine (A) and guanine (G). Uracil is absent in the DNA and it is replaced by thymine (T).

The ribose of the RNA is also different from the sugar of the DNA (deoxyribose). The ribose contains a hydroxy group in the position 2', which is absent in the deoxyribose.

Different layers of structure exist inside the nucleotides. The glycoside binding between a base and the ribose forms a nucleoside. The nucleoside with the addition of a phosphate group becomes a nucleotide. The ordered concatenation of nucleotides defines the RNA primary structure. To form the primary structure, the nucleotides are connected by phosphodiester bonds through the oxygen on the 5' carbon of one and the 3' carbon of another (Figure 1). The primary structure is often considered a synonymous of 'sequence', usually represented as a consecutive list of nucleotide symbols (ACGU). The directionality is a fundamental characteristic of the primary sequence, and it is always defined from the 5' to the 3', till the end of the sequence.

The primary structure encodes the necessary properties for the formation of the secondary structure.

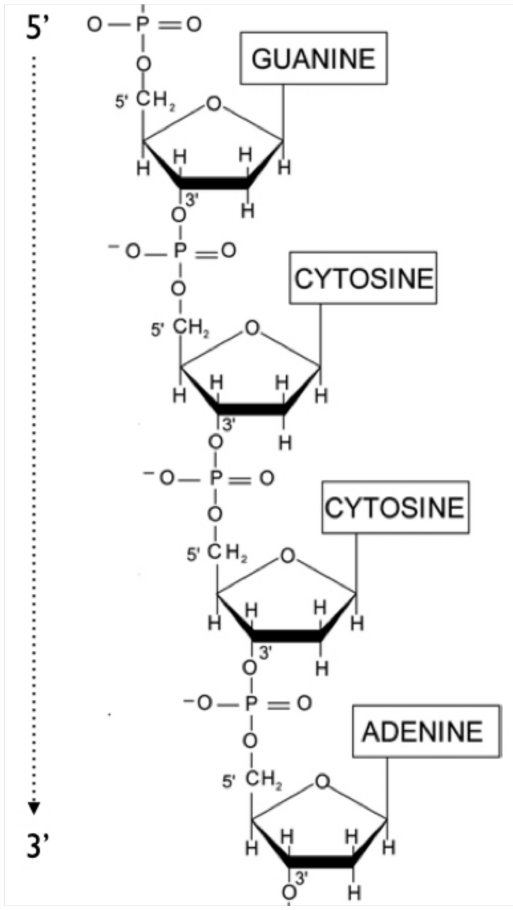


Figure 1: Primary sequence example with directionality highlighted (adapted from Schowen (1993)).

1.2 The secondary structure

1.2.1 Binding types

The secondary structure of nucleic acids is formed by base pairing interactions between nucleotides. In the DNA molecules, two strands of complementary nucleotides are interacting through strong hydrogen bonds between the Watson-Crick pairs Adenine-Thymine (A-T) and Cytosine-Guanine (C-G) to form the regular double helix (Watson and Crick, 1953).

Contrary to the DNA, the RNA in its native state exists in a single-stranded conformation. This feature allows the RNA to be more dynamic and with more degrees of freedom than the double-stranded DNA. The flexible native RNA in single-stranded conformation is able to form self-interactions along its primary structure. The pattern of the internal interactions for an entire RNA molecule is defined as its secondary structure (Doty et al., 1959).

The base-pairing interactions between the four RNA bases (Adenine, Cytosine, Guanine, Uracil) can be classified in different ways.

In the 1980s, with few experimentally determined structures available and the information coming only from the transfer RNA, short RNAs composed of 70 to 90 nucleotides and key elements for the translation (tRNA), the classification of the bonds was only based on the type of nucleotide involved (Saenger W. (1984) Principles of Nucleic Acid Structure. Springer-Verlag, New York, NY).

A more extensive classification from 1993 defined 28 possible pairings based on the type of the interacting bases (Tinoco, 1993). These base-pairing types can be grouped into 4 subclasses: purine-pyrimidine (10 pairings, including Watson-Crick, Wobble, and Hoogsteen interactions), homo purine-purine (7 pairings), hetero purine-purine (4 pairings) and pyrimidine-pyrimidine (7 pairings). A thorough description of these base pairs is beyond the scope of this thesis; it can be found in Tinoco, Jr. In Appendix 1 of: The RNA World, Cold Spring Harbor Laboratory Press, 1993, pp. 603-607.

Of the 28 possible base pairings, only six (AU, GU, GC, UA, UG, CG) interactions are stable. Accordingly, they are the most common interactions

within RNA molecules. The CG/GC pairs are the strongest and thus the most stable ones as they are formed by 3 hydrogen bonds. The other pairs are formed by only 2 hydrogen bonds.

More recently, crystallographic experiments showed that the majority of base pairs in structured RNAs show recurrent geometric patterns (Leontis and Westhof, 2001). These patterns arise because RNA nucleotides have three interactive edges that can form hydrogen bonds: the Watson-Crick edge, the Sugar edge (including the hydroxyl group), and the Hoogsteen edge for purines or 'CH' edge for pyrimidines. Interactions can thus be grouped into 12 classes characterized by the interactive edges involved in the hydrogen bonds and the relative orientation (cis/trans) of the glycosidic bonds of the two bases (Leontis and Westhof, 2001) (Figure 2).

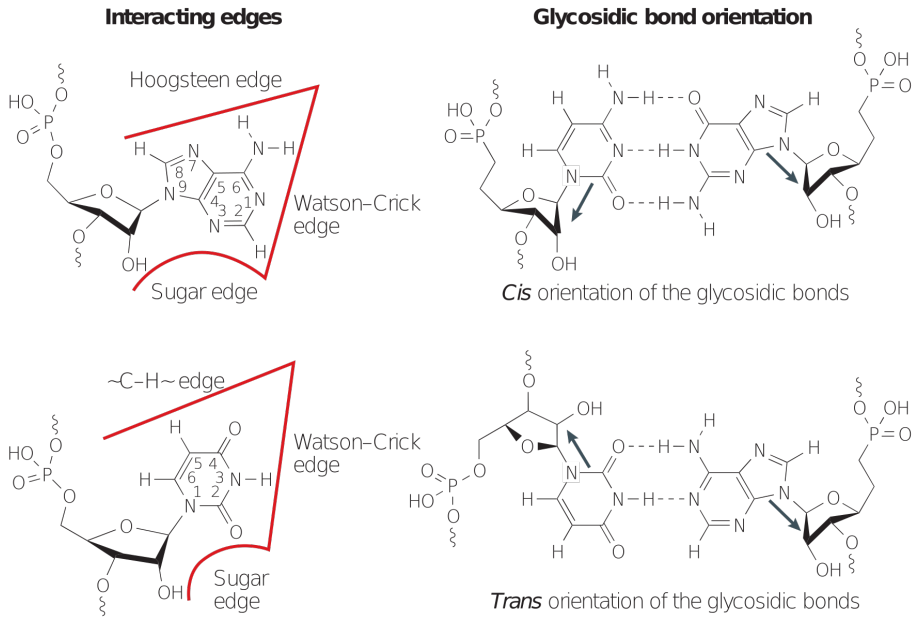


Figure 2: The three edges and the orientation of the bond (extracted from Schroeder et al. (2004)).

1.2.2 Structural elements

The RNA is able to form different structural elements of various lengths (Figure 3). Structural elements that can be present in the RSS are:

- Stem: double-stranded regions. The most stable RNA structural motif and usually the longest one.
- Hairpin-loop (H-loop): a very common structure. It is a combination of strong-complementary bases separated by unpaired nucleotides. For example, a common H-loop is AAAAACCCCUUUUU, where the complementary multi-A and multi-U segments form a stem, interrupted by a single-stranded region of multi-C.
- Internal-loop (I-loop): a loop that is internal to consecutive stems. The loop has the same number of nucleotides on the left (il) and on the right side (ir) (il = ir = n, where n $\in$ N).
- Bulge: a specific sub-class of I-loop, where only one side of the loop has unpaired nucleotides and the other is connected to the stem (e.g. il > 0, ir = 0).
- Multibranch-loop (M-loop): a complex structure composed of different sub-loop structures. The main architecture is usually composed of several branches combined together to a central loop.
- Pseudoknots: the most complex structure. A pseudoknot is formed when a loop region and bases outside of the loop interact. Due to their non-nested nature, the pseudoknots are an exception and for this reason cannot be predicted by dynamic programming algorithms (see Introduction 5.1 for more details).

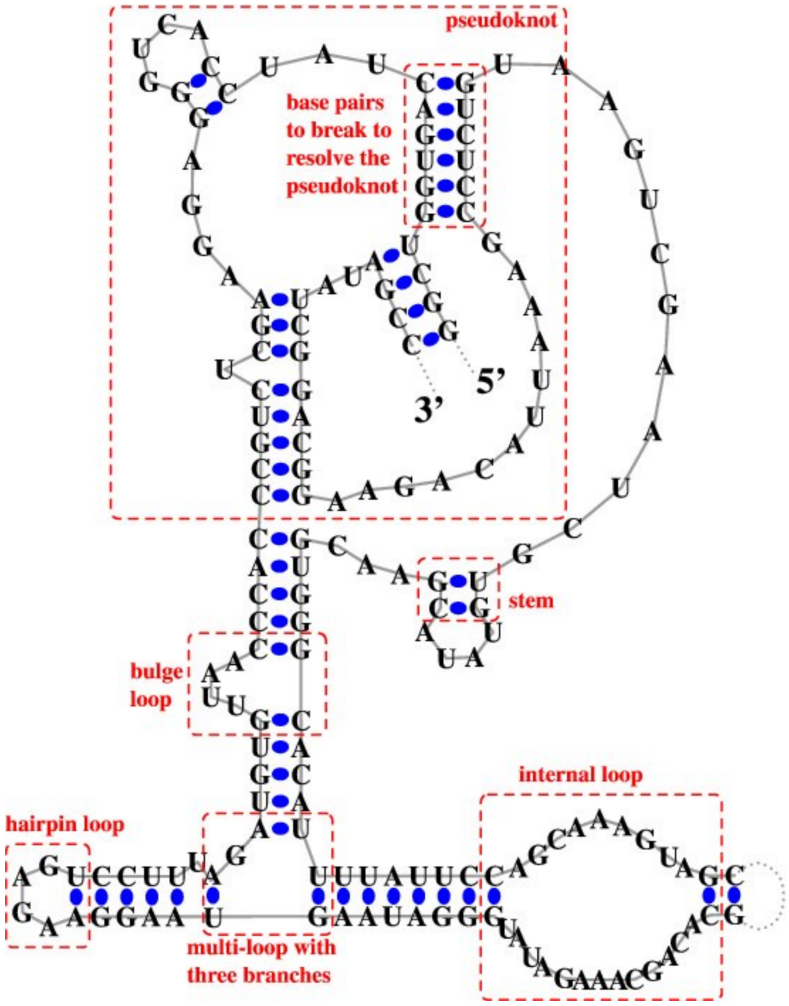


Figure 3: Structural motifs formed by the RNA secondary structure (extracted from RNAstrand webpage (Andronescu et al., 2008)).

1.3 The tertiary structure

The tertiary structure is the arrangement of the RNA in space where interactions between two-dimensional secondary structure elements create three-dimensional structural motifs such as helices (Batey et al., 1999).

During the folding of the tertiary structure, the secondary structure elements interact through *van der Waals* contacts, hydrogen bonds, and interactions between hairpin loops and bulges (Batey et al., 1999). The tertiary structure

interactions can be divided into 3 major categories: 1) interactions between two double-stranded helical regions (coaxial stacking, adenosine platform, and 2'-hydroxy-mediated helical interactions); 2) interactions between helical and unpaired regions (base triplex/triplexes, tetraloop motif, metal-core motif and the ribose zipper); 3) interactions between unpaired regions (loop-loop interactions and pseudoknots) (Batey et al., 1999).

The tertiary structure was thoroughly studied for tRNAs and ribozymes, such as the Tetrahymena self-splicing group I intron (Th-intron)(Rook et al., 1998; Lehnert et al., 1996). It was also discovered that specific ions have an effect on the RNA tertiary structure, such as magnesium (Mg2+), which is a fundamental element for the formation and the stabilization of the tertiary structure (Brion and Westhof, 1997). For example the Th-intron ribozyme needs to bind to at least 3 magnesium ions to be able to fold into an active tertiary structure (Batey et al., 1999).

More details regarding the RNA folding will be provided in Introduction 3.1.

2 The RNA secondary structure and its function in cellular processes

The secondary structure is fundamental for many aspects of the RNA biology such as correct functionality, but also for the interaction with proteins or other RNA molecules (Bellucci et al., 2011).

The ability of an RNA molecule to assume different secondary structures and its dynamicity are at the base of the theory of the 'RNA world' (Robertson and Joyce, 2012). This theory got more attention after the discovery of the ribozymes. The ribozymes are RNAs with enzymatic activity, something that was considered exclusive to proteins (Robertson and Joyce, 2012; Kruger et al., 1982). Interestingly, further theories speculate that a ribozyme of 40-60 nucleotides that form 3 stem-loops could be potentially enough to work as replicase, giving the base for a RNA-first world and highlighting the importance of the secondary structure for the functionality of the RNA (Robertson and Joyce, 2012).

The secondary structure is fundamental for the biological function of the

RNA, but it is also crucial for the activity of specific non-coding RNA classes. For example disrupting the secondary structure of a tRNA will affect translation, a basic step essential for life (Bernat and Disney, 2015).

The RSS is also crucial for the splicing of many pre-mRNAs, since the removal of the introns is also determined by the specificity of the secondary structure in proximity of the intron-exon junction (Shepard and Hertel, 2008; Buratti and Baralle, 2004).

The interaction of an RNA with proteins is one of the most important biological functions that can be influenced by the RNA secondary structure. Many RNA binding proteins (RBPs) contain specific domains that are able to bind the RNA, with selective specificity also for the RNA structure. For instance, RNA recognition motifs (RRM) (Auweter et al., 2006) and the K-homology domain are more prone to bind single-stranded RNA (ssRNA), while the double-stranded RNA binding domain (dsRBD) binds to dsRNA regions (Masliah et al., 2013). Furthermore, even in cases where the binding is only related to the sequence, the structure plays an important role since a specific sequence is only accessible to the binding when it is located at the bulge of a stem-loop structure (Lu et al., 2003). RBPs that are more promiscuous (i.e. able to bind many RNAs) are usually more prone to bind single-stranded regions (Dominguez et al., 2018).

Recent results also suggest that long and highly structured (i.e. enriched in double-stranded regions) RNA molecules can have an important role as scaffolding elements inside RNA granules (Maharana et al., 2018).

Due to its importance (see Introduction 1), the secondary structure is highly conserved (Pedersen et al., 2006; Washietl et al., 2005). The conservation of the RNA secondary structure is often associated to functionality (Ilyinskii et al., 2009; Ganot et al., 1997; Washietl et al., 2005; Pedersen et al., 2006), since the correct matches of the nucleotides add another level of complexity compared to the sequence conservation. The secondary structure is crucial for the function of many non-coding RNAs (ncRNAs) such as tRNAs, small nucleolar RNAs (snoRNAs) and microRNAs (miRNAs) (Vandivier et al., 2016; Ganot et al., 1997; Washietl et al., 2005). For these classes, the secondary structure is more likely to be conserved than the secondary structure of other

ncRNAs such as lncRNAs (Rivas et al., 2017; Delli Ponti et al., 2018).

The ribosomal RNAs (rRNAs) have an interesting evolutionary history regarding their structure. It was recently proposed that the small subunit (SSU) of the rRNAs evolves in a self-contained environment, where the secondary structure of the RNA of ancient species is conserved in the rRNA of the more evolved ones (Petrov et al., 2015). Regarding the long non-coding RNAs (lncRNAs), non-coding RNAs longer than 200 nucleotides, the secondary structure conservation is still a debated topic (Rivas et al., 2017; Somarowthu et al., 2015). The conservation of the RSS, especially for the lncRNAs, will be discussed in more details in Chapter II.

Disrupted structures or misfolded RNAs are also related to several pathologies (Bernat and Disney, 2015). Different studies suggested that SNPs are more prone to be related to a disease phenotype when the mutation is also affecting the secondary structure (Halvorsen et al., 2010; Bernat and Disney, 2015).

3 The RNA secondary structure *in vitro*

3.1 *In vitro* folding

The RNA folding is a hierarchical process (Brion and Westhof, 1997). The primary structure defines the RNA secondary structure (RSS), and the RSS is necessary to fold into tertiary structure.

The RNA folding *in vitro* follows a set of rules well studied and defined by the scientific community. The absence of a crowded cellular environment and the controlled conditions allow the study of a simplified model of the RSS folding, where the folding is guided prevalently by the sequence.

The *in vitro* folding of the RNA can be modeled as a stochastic search of the most energetically favorable and stable structure, passing through many other probable conformations, in a process similar to the protein folding (Wolynes et al., 1995)(Figure 4). The free energies of all possible conformations define an energy landscape, which is explored while searching the optimal structure that minimizes the free energy.

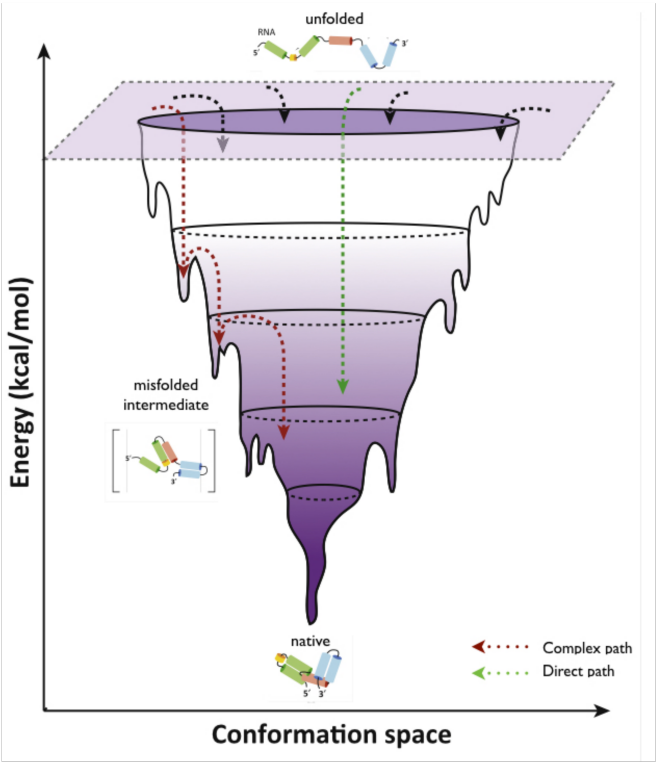


Figure 4: Representation of the RNA folding *in vitro*. The path toward the native structure could be complex or direct, and could lead to misfolded intermediates (adapted and modified from Incarnato et al. (2017)).

As previously explained, the RNA folding is considered a hierarchical process, since the secondary structure forms before the tertiary structure (Brion and Westhof, 1997). For this reason the two folding events can be treated as independent processes. The observation that the secondary structure is also faster to fold (μs to ms, depending on the length of the sequence) than the tertiary structure (ms to s) gives more support to the concept of a temporal hierarchy of the processes (Leamy et al., 2016).

The folding of an RNA into its secondary structure is also a hierarchical process where the stem-loops tend to form first. The folding dynamic of the early RNA stem-loops is also very similar to the beginning stage of protein folding, specifically the 'molten globule' state (Freisner and Gunn, 1996; Levitt et al., 1997). Hairpins with short loops have folding times between 10 and 100 μs (Crothers et al., 1974).

The folding of a transcript *in vitro* usually starts with the denaturation of its random coil conformation using different melting temperatures. The next folding step is in presence of different temperatures and salt concentrations, depending on the aim of the experiment (Baird et al., 2005). The presence of ions can also be fundamental for a fast folding. For example the presence of Mg2+ promotes the folding (London, 1991; Truong et al., 2013; Stein and Crothers, 1976) and the formation of tertiary interactions (Batey et al., 1999).

The majority of the *in vitro* folding studies were conducted on tRNAs (Crothers et al., 1974; Hilbers et al., 1976) and ribozymes (Banerjee and Turner, 1995; Mitchell et al., 2013; Rook et al., 1998) due to their known functionality and their moderate size ($<$ 100 nt).

The folding of long RNA molecules is a more complex procedure that can take up to minutes or hours (Chadalavada et al., 2002; Banerjee and Turner, 1995). This is mainly due to the formation of stable unfolded intermediates that have conformations very similar to the native structure (Treiber et al., 1998; Mitchell et al., 2013).

The *in vitro* folding was prevalently studied in conditions far from the ones present in the cellular environment, not only for the absence of the crowding effect (i.e. the presence of proteins and ligands) but also for the non-physiological salt concentrations (London, 1991; Truong et al., 2013)(Table 1).

Due to the many variables that are impossible to simulate, the study of the thermodynamics and the complete folding landscape of the RNA *in vivo* is still a challenge. To solve this problem, scientists are starting to work on an artificial cytoplasm that is able to mimic specific cellular characteristics. In the last years, several studies using artificial cytoplasm with crowding agents and physiological concentrations of ions and salt were developed (Desai et al., 2014; Dupuis et al., 2014; Nakano et al., 2014; Paudel and Rueda, 2014; Strulson et al., 2013; Tyrrell et al., 2015). I will explain *in vivo* effects such as the crowding in more detail in Introduction 4.1.

Table 1: Main differences of the *in vitro* and *in vivo* environment (obtained adapting the data from Leamy et al. (2016)).

Condition	*in vitro*	*in vivo*
Molecular crowding	0 %	20%-40%
Monovalent salt	0-1M	140mM K+
Divalent salt (total)	0-100 mM	20 mM
Ionic strength	0-1 M monovalent 0-3 M divalent	~0.142 M (eukaryotes) ~0.147 M (prokaryotes)

3.2 Experimental techniques *in vitro*

Until now, as for proteins and DNA, the most accurate way to assess the RNA structure is using NMR or X-ray techniques (Latham et al. 2005). However, there are very few crystals available for RNA, most of them for RNA of bacteria and synthetic organisms. For example, from the 1,059 RNA crystals validated with NMR/X-ray available in the RNAstrand database (Andronescu et al., 2008), 39% originate from synthetic constructs, while only 0.09% are from mouse (1 crystal) and 2.8% are from human (30 crystals). Moreover, the majority are rRNAs (26%; 276) and tRNAs (8%; 85). In addition, 84% of all available crystals are composed of complexes of the RNA with one or multiple ligands, while only 16% (175) are crystals of a single RNA, and 100% of these are from synthetic systems. Thus, the poor

availability of experimentally determined RNA structures, the complexity of the experiments and their high costs in terms of both time and money drove the scientific community toward the search for new techniques to assess the RSS.

In the last 10 years several experimental techniques were developed to study the RNA secondary structure at low-throughput and more recently also at high-throughput level (Mortimer et al., 2014; Strobel et al., 2018). The rapid evolution of omics techniques allowed the development of specific protocols able to profile the secondary structure landscape of complete transcriptomes of several organisms (Mortimer et al., 2014; Strobel et al., 2018)(Table 2). These experimental high-throughput techniques can be divided into chemical-based and enzymatic-based approaches.

Table 2: Main characteristics of the most known experimental techniques.

Methods	Type	Probe	RNA folding	Organisms	NGS platform
PARS and PARTE	enzymatic	RNase V1 (dsRNA) and S1 nuclease (ssRNA)	*In vitro*	*Saccharomyces cerevisiae*, *Homo sapiens* and *Danio rerio*	ABI SOLiD, Illumina HiSeq2000, Ion proton platform
ds/ssRNA-seq	enzymatic	RNase V1 (dsRNA) and RNase 1 (ssRNA)	*In vitro*	*Arabidopsis thaliana*, *Drosophila melanogaster* and *Caenorhabditis*	Illumina GAIIx and HiSeq2000
FragSeq	enzymatic	P1 nuclease (ssRNA)	*In vitro*	*Mus musculus*	ABI SOLiD3
SHAPE–seq	chemical	1M7	*In vitro*	*Bacillus subtilis*	Illumina GAIIx
DMS	chemical	DMS	*In vivo*	*A. Thaliana*, *S. cerevisiae*, *Homo sapiens* and *Oryza sativa*	Illumina HiSeq2000
MAP-seq	chemical	DMS, CMCT and 1M7	*In vitro*	Synthetic system	Illumina MiSeq
icSHAPE	chemical	NAI-N3	*In vitro* / *In vivo*	*Mus musculus*	Illumina HiSeq

3.2.1 Chemical-based techniques

Chemical-based techniques use small and highly reactive probes that are able to bind the RNA in order to obtain information on the RNA secondary structure. Peattie and Gilbert in 1980 performed the first chemical-probing experiment to study the structure of the tRNA (Peattie and Gilbert, 1980; Strobel et al., 2018). Since that year, several other techniques were developed, but the technologies became more popular with the discovery of SHAPE (Selective 2'-hydroxyl acylation analyzed by primer extension) chemistry in 2005 (Merino et al., 2005).

SHAPE is a technique based on chemical probes that are able to assess several features of a RNA molecule (Wilkinson et al., 2006; Merino et al., 2005). SHAPE chemistry is based on the activity of acylating agents, such as 1-methyl-7-nitroisatoic anhydride (1M7), which is able to react with flexible nucleotides forming a 2'- O-adduct. In contrast, nucleotides that are constrained by base pairing or tertiary interactions are unable to bind the chemical probe and appear as unreactive. Sites of 2'-O-adduct formation are then detected as stops to primer extension. The quantification of the reactivity helps to identify nucleotides that are in a double- or single-stranded conformation. Figure 5 gives an overview on the general active mechanism and provides more information regarding the probes.

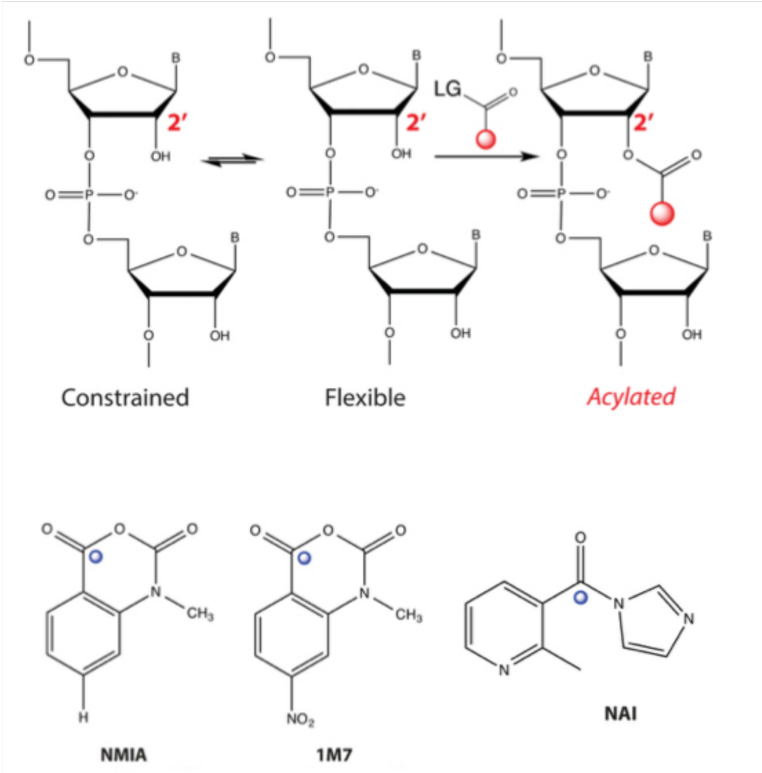


Figure 5: General mechanism and examples of probes for the SHAPE chemistry (adapted from Spitale et al. (2014)).

SHAPE was successfully applied to several RNA molecules (Rausch et al., 2017; Duncan and Weeks, 2008; Rice et al., 2014), including the complete HIV-1 genome (Wilkinson et al., 2008). The SHAPE protocol was modified and adapted for several tasks as the use of different probes allows the identification of several RNA properties. For example, SHAPE-seq is a variation of the technique that is optimized for NGS technologies (Lucks et al., 2011; Loughrey et al., 2014). SHAPE-map is based on the use of several probes (1M7, 1M6 and NMIA) to detect also long range interactions, stacked nucleotides, and pseudoknots (Siegfried et al., 2014; Smola et al., 2015). The Map-Seq protocol permits the quantitative probing of thousands of RNAs at once using Illumina technology (Seetin et al., 2014). The high-throughput protocol icSHAPE (*in vivo* click selected SHAPE) takes advantage of the probe NAI-N3 to profile the secondary structure of RNAs also *in vivo* (Spitale

et al., 2015).

SHAPE chemical probes are only able to bind to single-stranded nucleotides. A lack of signal is usually associated with double-stranded nucleotides, but with this lack of specificity it is impossible to distinguish between protected and interactive regions, especially *in vivo* where the protein-RNA interactions are more frequent.

There are other chemical-probing techniques that are not based on SHAPE chemistry. The most well-known and used one is probably DMS (dymethil sulfate). DMS chemical probing is often used *in vivo* (see Introduction 4.2 for details) due to the small size of the probe ($(CH_3O)_2SO_3$). However, this technique has a strong bias since the alkylating agent can only bind to adenine and cytosine and it is thus unable to provide structural information for all the nucleotides. Regardless of its limitation, DMS probing was applied to different organisms and conditions (Rouskin et al., 2014; Ding et al., 2014)

3.2.2 Enzyme-based techniques

The discovery of enzymes that are able to cut the RNA based on its local secondary structure allowed the development of several enzyme-based techniques. In this part I will focus on the most used techniques.

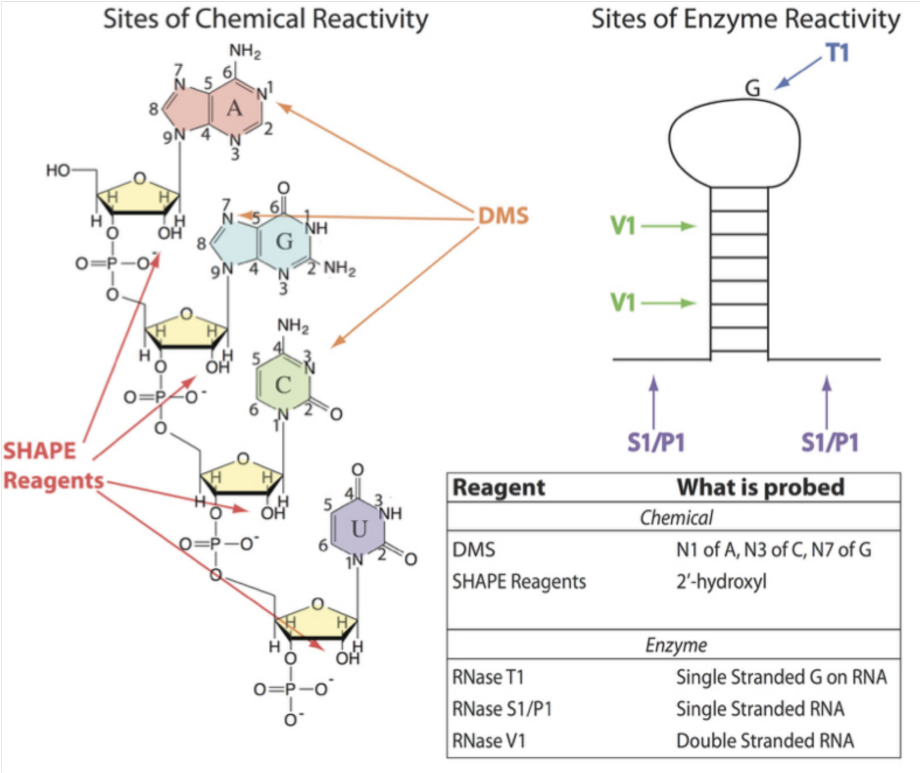
The RNase V1 is one of the first enzymes used to probe the RSS (Wyatt and Walker, 1989). It is able to cut the double-stranded nucleotides, with at least 3 nucleotides upstream and downstream of the cutting point (Wyatt and Walker, 1989).

PARS (parallel analysis of RNA structure) is one of the most used enzymatic techniques, able to distinguish double- and single-stranded regions using the catalytic activity of two enzymes: RNase V1 (able to cut double-stranded nucleotides, as previously specified) and S1 (able to cut single-stranded nucleotides). This technique was successfully applied high-throughput on entire transcriptomes (Kertesz et al., 2010; Wan et al., 2014). Recently, it was also modified for the use with Illumina technology (Saus et al., 2018). PARS is the only technique able to actually profile both single- and double-stranded nucleotides. This technique gives more coverage than SHAPE, which can

only probe flexible regions, as explained in the previous section. However, PARS also has limitations: since the size of the enzymes is large compared to the size of the small chemical probes, it could cause problems of resolution, and it still cannot be applied *in vivo*.

PARTE (parallel analysis of RNA structures with temperature elevation) is an extension of the PARS methodology. PARTE allows the genome-wide measurement of RNA folding energies. It was successfully used in *Saccharomyces cerevisiae* mRNAs by probing the secondary structure at temperatures ranging from $23°C$ to $75°C$ (Wan et al., 2012).

FragSeq (fragmentation sequencing) is a high-throughput enzyme-based RNA structure probing method. It uses fragments generated by digestion with the nuclease P1, which specifically cleaves only single-stranded nucleic acids (Underwood et al., 2010). This technique has huge limitations, since it can only cut single-stranded RNA and has thus the same limitations as SHAPE chemistry approaches, but without the advantage of using a small chemical probe. A summary of the general mechanisms of chemical-probes and enzymatic techniques is provided in Figure 6.



Reagent	What is probed
Chemical	
DMS	N1 of A, N3 of C, N7 of G
SHAPE Reagents	2'-hydroxyl
Enzyme	
RNase T1	Single Stranded G on RNA
RNase S1/P1	Single Stranded RNA
RNase V1	Double Stranded RNA

Figure 6: Modifications and sites of the interactions of the chemical-probes (SHAPE reagents and DMS) and enzymes (V1 and S1 for example) (adapted from Leamy et al. (2016)).

4 The RNA structure *in vivo*

4.1 *In vivo* folding

The RNA folding *in vivo* is a complex process, where many contributors could potentially affect the folding. While the RNA *in vitro* tends to fold incrementally by small conformation changes until it finds the structure that minimizes the free energy, (see Introduction 3.1), the folding *in vivo* is far more complicated and it can lead to a different structure. The *in vivo* folding is affected by different external contributors, such as general interactors or chaperones, which can actively influence the resulting structure (Figure 7). In general, the cellular environment has an effect on the RNA folding at a temporal (for example: co-transcriptional folding) and spatial level (for example: crowding effect).

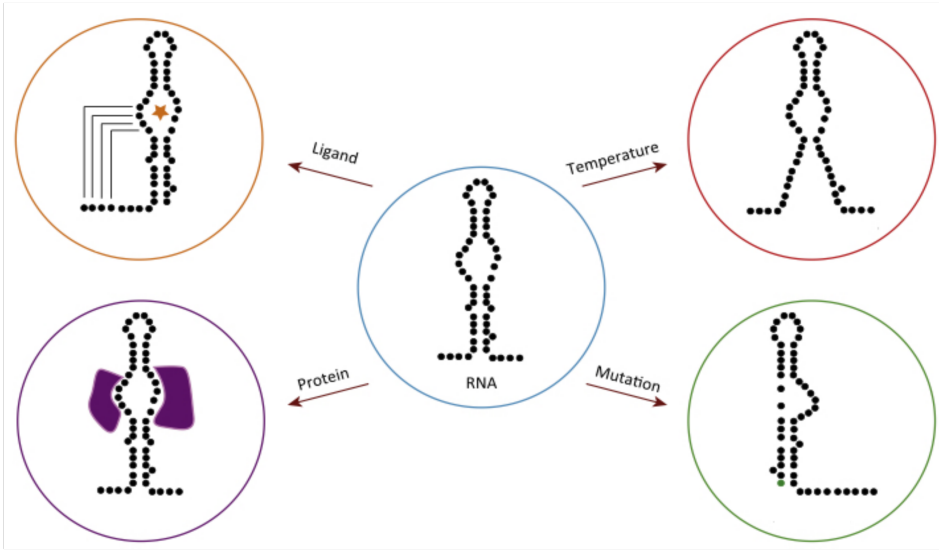


Figure 7: Simplified example of the possible external forces that can affect the RNA structure *in vivo* (adapted from Kwok et al. (2015)).

In contrast with the controlled *in vitro* environment, the cell is a crowded and complex environment where 30-40% of the cytosol is occupied by macromolecules (Minton, 2001; Zimmerman and Trach, 1991). This crowding effect is one of the most influential spatial contributors for the RNA folding.

The crowding effect of the cytoplasm has by itself an effect on the RNA folding *in vivo*. The presence of cosolutes and proteins, occupying a huge volume of the space available in the cytoplasm, reduces the degrees of freedom for macromolecule folding. This phenomenon is especially studied for the protein folding: having fewer degrees of freedom and thus a reduced number of possible intermediate structures available in the conformational space constrains the folding process (Zhou, 2004; Zhou et al., 2008). The crowding effect, or more precisely the excluded volume (i.e. the volume not available to be explored by flexible folding structures), was also studied for the thermodynamics of nucleic acids, including DNA duplex/hairpin motifs, RNA ribozymes, and telomerase pseudoknot RNA (Dupuis et al., 2014). A consensus of these studies is that highmolecular-weight polyethylene glycols (PEGs), used to simulate a crowding environment, increases the thermodynamic stability of the folded structures. The thermodynamics and kinetics of this phenomenon were studied at single molecule level in RNA for the folding of the GAAA tetra-loop receptor, were the effect of the PEGs promote a $>$60-fold increase in the folding equilibrium constant (Dupuis et al., 2014).

Regarding the temporal effects on the *in vivo* RNA folding, it was shown that active transcription processes can affect the secondary structure, since the *in vivo* folding can happen at the same time (Boyle et al., 1980; Brehm and Cech, 1983). This phenomenon affects the folding rate and thus leads to different folding times for different RNAs, and it can also influence the final and the intermediate structures (Pan et al., 1999; Heilman-Miller and Woodson, 2003). The first *in vivo* genome-wide data on the cotranscriptional folding of *E. coli* suggest as a general rule that the short-range interactions are fast and the related structures are formed early, while long-range interactions require intermediate structures to fold into their final structure (Incarnato et al., 2017).

As previously described in Introduction 3.1, one of the critical steps for the *in vitro* folding is the presence of misfolded structures that are energetically very similar to the native structure. These structures are very stable and can remain for hours, trapping the RNA in a wrong folding-pathway behind high energetic walls (Zemora and Waldsich, 2010). These misfolded

intermediates are also common *in vivo* (Jackson et al., 2006). However, it was shown that the folding of the RNA *in vivo* is generally faster than the folding *in vitro* (Mahen et al., 2010).

One of the possible explanations for the faster folding of the RNA *in vivo* is the presence of RNA chaperones, a class of proteins described in several contexts and with different main functions (Schroeder et al., 2004; Tompa and Csermely, 2004). Even though the proteins with chaperone activity are very different and difficult to catalog, they have two characteristics in common: 1) the ability to bind RNA; 2) the ability to destabilize RNA structures. The details of the mechanics behind the RNA chaperones are not completely understood, but few details are known or speculated (Zemora and Waldsich, 2010). The RNA chaperones are able to bind RNA, as previously stated, but with low affinity and in a promiscuous way (Herschlag, 1995). A weak interaction is necessary for the chaperones to be released after unfolding the RNA structure, allowing the RNA to fold again, and the non-selective binding is the key for the chaperone being functional for any kind of misfolded RNA (Zemora and Waldsich, 2010). It was also suggested that the RNA chaperones are enriched in disordered domains, which makes them more flexible and thus prone to bind RNAs promiscuously (Tompa and Csermely, 2004).

Similar to the chaperones, also the RNA helicases are active in unfolding the RNA structures to avoid kinetic traps. However, in contrast to the RNA chaperones, their activity is ATP-dependent (Bleichert and Baserga, 2007).

4.2 Experimental techniques *in vivo*

In contrast to the RNA in controlled *in vitro* conditions, the RNA *in vivo* is subject to a complex environment that can alter its structure by external forces. Several *in vitro* experimental techniques were also applied *in vivo*, but their applicability is not only subject to the limitations of the techniques themselves, but also to adverse influences by the complexity of the environment.

DMS was the first technique to be applied *in vivo* in 1988 (Climie and Friesen,

1988; Moazed et al., 1988). More recently it was also applied on the human transcriptome *in vivo* (Rouskin et al., 2014), and on *Arabidopsis* and yeast (Ding et al., 2014; Rouskin et al., 2014). Strikingly, the results suggested an almost complete lack of structure for the RNAs *in vivo*, especially in human (Rouskin et al., 2014). However, as previously highlighted, DMS chemical probing is only able to profile two nucleotides, with a partial coverage on the human transcriptome.

In contrast to DMS, icSHAPE is able to profile all nucleotides, and the small probe NAI-N3 was also successfully applied *in vivo* (Spitale et al., 2015). The results suggested a lack of structure in the coding RNAs, but this lack is not as drastic as was previously suggested by DMS studies. The icSHAPE technique was also used to show that especially the non-coding RNAs tend to be structured also *in vivo* (Spitale et al., 2015).

SHAPE-MaP was also successfully applied *in vivo*, but the measured structural profiles showed a low correlation with the profiles determined by icSHAPE (Smola et al., 2016). However, it was recently shown that the chemical probe 1M7, used by SHAPE-MaP, is not able to pass through a living cell membrane (Lee et al., 2017). This result suggests that the overall signal for SHAPE-MaP secondary structure measurements is coming from dead cells, where the structure is not informative since it could have been altered by post-mortem and stress activities.

PARS is still unable to profile the RNA genome-wide *in vivo*, but it was successfully applied under near-*in vivo* conditions to deproteinized natively folded RNAs extracted from lymphoblastoid cells (Wan et al., 2014).

In addition to the previously discussed limitations, all chemical probes are generally unable to distinguish between double-stranded regions and single-stranded regions that are bound to proteins, since both lead to a lack of signal (Mortimer et al., 2014). The experimental techniques are in general limited to the extent of establishing the interface and the effect of the proteins on the RNA structure. The complex conditions and the lack of a complete understanding of all the forces influencing the RNA *in vivo* are still a major problem, undermining the success of the RNA secondary structure probing *in vivo*.

A summary of the experimental techniques and their characteristics was provided in the Table 2.

5 The RNA structure *in silico*

Information on the secondary structure is crucial for understanding the function of an RNA and its role in cellular processes. However, structures determined by crystallographic experiments are available for only very few RNAs, among them only a small fraction from more evolved species such as human and mouse (see Introduction 3.2), and chemical- and enzymatic-based techniques alone are not able to provide the complete structural profile. In addition, these techniques must rely on computational approaches for the graphical overview, and SHAPE and PARS are not completely accurate when tested on RNAs for which crystallographic data are available (Delli Ponti et al., 2017; Wu et al., 2015). Thus, the development of *in silico* predictive models that can provide an alternative to expensive and time consuming experiments has always been of great importance (Zuker and Sankoff, 1984).

The two most widely used types of computational methods for predicting the secondary structure of an RNA are thermodynamics-based folding algorithms and comparative sequence analysis approaches. The main advantage of thermodynamics-based folding algorithms is that they can predict the structure based on the sequence only, without requiring any experimental data or homologous sequences. Comparative approaches can achieve higher performances, but they need information about homologous sequences. In addition to these approaches, there are thermodynamics-based algorithms that can use experimental constraints, such as SHAPE-profiles, to improve their predictive power. These hybrid methods will be discussed in Introduction 5.3.

5.1 Thermodynamic approaches

The RNA *in vitro* tends to spontaneously fold into the structure with the minimal free energy. However, an RNA can also fold into various

suboptimal structures with a low free energy, and it can assume different intermediate structures during the folding process (Treiber et al., 1998). Thermodynamics-based prediction applications are usually developed to search the minimum free energy (MFE) structure, although some can also provide suboptimal structures and their associated free energies (Gruber et al., 2008).

To compute the free energy of an RNA secondary structure, thermodynamics-based algorithms use a set of parameters that were first determined by optical melting experiments (Martin et al., 1971). The pioneering experiments described a set of 12 energetic parameters for Watson-Crick pairs (Xia et al., 1998). Other sets of parameters were measured for loops and GU pairs (Mathews et al., 2004, 1999). Thermodynamic parameters are available at the Nearest Neighbor Database (NNDB, https://rna.urmc.rochester.edu/NNDB/) (Turner and Mathews, 2010).

The most used computational approaches are based on dynamic programming algorithms (Nussinov and Jacobson, 1980). These algorithms sample every structure that can be obtained from the primary structure under a set folding rules (i.e. allowed nucleotide matches), searching for the conformation with the lowest free energy, which is considered the most probable native structure.

Three folding rules are at the core of any thermodynamics-based dynamic programming algorithm:

1. A nucleotide cannot participate in more than one base pairing interaction.
2. Based on sterical constraints, if two nucleotides are paired with each other, at least 3 unpaired bases should separate them.
3. For any two pairs of nucleotides (A, B) and (C, D) with position indices (a, b) and (c, d), the base pairing interactions are not allowed to cross and break the nested structure, i.e. if $a < c$, then $a < c < d < b$ has to hold.

The third rule also excludes pseudoknots, thus limiting the predictive power

of dynamic programming algorithms.

The most cited dynamic programming algorithms are RNAstructure (Reuter and Mathews, 2010) and RNAfold (Gruber et al., 2008), which are considered the gold standards of the field.

Thermodynamics-based dynamic programming algorithms are heavily limited by the length of the RNA molecule. On shorter molecules ($<$ 700 nucleotides), they can achieve accuracies of $\sim$0.70, but when the sequence is longer than 700nt, the performances drop drastically (Hajiaghayi et al., 2012; Lu et al., 2009). These algorithms are thus not well suited for predicting the secondary structure of long RNAs such as rRNAs or lncRNAs, which can have a total length of thousands of nucleotides. Moreover, as previously explained, the energy parameters of the thermodynamic models are based on *in vitro* data and thus unable to reflect the *in vivo* environment (Martin et al., 1971). The *in vivo* structure of an RNA molecule can be very different from the predicted one, since the presence of a cellular environment with protein interactions and external forces actively influences the folding (see Introduction 4.1). In general, thermodynamics-based folding algorithms are unable to consider these effects, since they treat the RNA molecule as isolated from any external forces.

Due to the limitations of free energy minimization algorithms, a number of other approaches were developed, among them an algorithm that maximizes the expected accuracy (MEA)(Lu et al., 2009),

Other approaches apply stochastic searching of the possible RNA structure generated using a Boltzmann distribution (Harmanci et al., 2009) or partition function based on probabilities (McCaskill, 1990).

Between the many tools based on different principle to predict the RNA structure only from the sequence, worth to mention: CONTRAfold (Do et al., 2006), Sfold (Ding et al., 2004), CentroidFold (Sato et al., 2009), Mfold (Zuker, 2003), RNAShapes (Steffen et al., 2006), GTFold (Swenson et al., 2012).

Dynamic programming algorithms can also be designed such that they can predict pseudoknots. However, to achieve this level of prediction, the tools have to sacrifice optimality and they usually require more

computational time than classic dynamic programming algorithms that exclude pseudoknots. An example is Pknots (Rivas and Eddy, 1999), which is part of the RNAstructure suite and one of the most used algorithms for pseudoknot prediction. Pknots is a thermodynamics-based dynamic programming algorithm that employs a diagram representation borrowed from quantum field theory (Rivas and Eddy, 1999). Due to the complexity of this approach, the computational time for predicting the secondary structure of a sequence of length n scales with n^6, and the required storage scales with n^4 (Rivas and Eddy, 1999; Jabbari et al., 2018).

The following programs are designed to also predict pseudoknots: KineFold (Xayaphoummine et al., 2005), CyloFold (Bindewald et al., 2010), pKiss (Janssen and Giegerich, 2015), and Knotty (Jabbari et al., 2018).

5.2 Comparative sequence analysis

As explained in Introduction 2 the secondary structure is conserved especially for specific non-coding RNAs. Therefore, information coming from homologous sequences can be exploited to build predictive algorithms with very high performances. This approach is well suited for RNAs that have many homologues, such as entire RNA families, and that have a known strong RSS conservation, for example the ancient rRNAs (Bokov and Steinberg, 2009). The basic approach is to align this large set of evolutionarily related RNA sequences and to scan them for sequence covariation. This approach is always combined with a predictive step. There are three different approaches for combining the alignment with the predictive step:

1. The sequences are first aligned and then the alignment is used as a constraint for the prediction. This approach is fast, but only when the sequence identity within the RNA family is high (75% or higher). Example programs are RNAalifold (Bernhart et al., 2008) and TurboFold (Harmanci et al., 2011).

2. The predictive step is done before the alignment. After all the predicted structures are generated, a consensus is selected as the best structure. This approach is ideal when the RSS is highly conserved, for example

in a family set. An example tool is RNAcast (Reeder and Giegerich, 2005).

3. The last approach is executing the predictive step and the alignment at the same time. Algorithms of this type are broadly applicable, even when there is low sequence or structural identity. Examples are Dynalign/Multilign (Fu et al., 2014; Xu and Mathews, 2011), Foldalign (Torarinsson et al., 2007), LocARNA (Will et al., 2007), PARTS (Harmanci et al., 2008), and RAF (Do et al., 2008).

Comparative sequence analysis is very limited by prior knowledge about the case-study RNA. Not only known homologues should be available in many different organisms to achieve the large number of sequences required, but they should also be characterized by a high sequence or structural similarity. With all these limitations, comparative sequence analysis-based approaches cannot be applied on novel discovered RNAs, or on debated and complex classes such as lncRNAs, where already defining the concept of homology is a challenge.

5.3 Integrative models: when the experiments meet the predictions

High-throughput experimental techniques such as SHAPE or PARS are able to profile the RNA at single nucleotide resolution. However, they can only provide a score for the propensity of each nucleotide to be in single- or double-stranded conformation, i.e. they cannot provide information on the actual base pairings. Since folding based on the structural profile alone does not have a unique solution, an algorithm is needed for predicting the actual base pairings and for visualizing the secondary structure.

Recently, it was shown that experimental data can be used as additional constraints for thermodynamics-based algorithms (Lorenz et al., 2016; Low and Weeks, 2010). Hard constraints force the algorithm to assign a specific conformation to a nucleotide and thus reduce the degrees of freedom of the tool. Soft constraints work as guideline, using the experiments as a continuous signal, where a weak signal allows more degree of freedom to the algorithm without imposing a structure. Due to their flexibility, soft

constraints are widely used and accepted by the majority of the algorithms. RNAstructure and RNAfold both accept DMS and SHAPE data as soft constraints (Lorenz et al., 2016; Hajdin et al., 2013). SHAPE data were successfully integrated into these thermodynamic algorithms to produce an improved visualization of the structure as well as to improve the performances of the algorithm. In some cases, incorporating experimental data improved the accuracy to up to 90% (Low and Weeks, 2010; Hajdin et al., 2013).

While the use of *in vitro* chemical-probed data to improve the predictive power of *in silico* prediction algorithms is well established (Delli Ponti et al., 2017; Wu et al., 2015), the use of *in vivo* data combined with predictive algorithms is still not well studied. At the time of the writing the thermodynamics-based approaches are still blind regarding the features guiding the *in vivo* RSS. A simple understanding of the structure *in vitro* could be not significant if the active *in vivo* structure has a different conformation prioritized by the environment. For example, machine-learning approaches could integrate multi-variables characterizing the RNA structure *in vivo*. This not only will be crucial to define the complex set of features influencing the RNA structure *in vivo*, but also would improve the predictive power of *in vivo* data. A piece of the *in vivo* puzzle will be provided in Chapter III and further analyzed in the General Discussion.

6 Toward the RNA Structurome

The genome-wide application of high-throughput profiling experiments (Introduction 3.2 and 4.2) is leading to the discovery of general structural properties of entire transcriptomes. Genome-wide analyses of RNA secondary structures can identify structural patterns and motifs in coding and non-coding RNAs from various organisms, and can thus help to understand how the structure is related to the function and to the expression of different types of RNA (Mortimer et al., 2014).

Experimental high-throughput techniques were applied to probe the structural profiles of the entire transcriptome of several different organisms and under different conditions. To date, the 'RNA Structurome' of the following

organisms is available: *Saccharomyces cerevisiae* (*in vitro*) (Kertesz et al., 2010), *Mus musculus* (*in vitro*/*in vivo*) (Spitale et al., 2015), *Homo sapiens* (*in vitro*) (Wan et al., 2014), *Danio rerio* (*in vitro*) (Kaushik et al., 2018), *Oryza sativa* (*in vivo*) (Deng et al., 2018), *Arabidopsis thaliana* (*in vivo*) (Ding et al., 2014), *Caenorhabditis elegans* (*in vitro*) (Li et al., 2012) and *Drosophila melanogaster* (*in vitro*) (Li et al., 2012).

Among the results of these studies there are interesting general findings. In human, *Drosophila*, zebrafish and *C. elegans* the UTRs are more structured (i.e. enriched in double-stranded nucleotides) than the coding regions, but this is not true for *Arabidopsis* and yeast. Also a three-nucleotide periodicity (unstructured followed by highly structured nucleotides; Kertesz et al. 2010) in the secondary structure data was detected in yeast, *Arabidopsis*, zebrafish, and human (Mortimer et al., 2014; Kaushik et al., 2018). This result suggests that, on average, the second nucleotide in a codon is the most likely to be involved in secondary structural interactions, since it is the highly-structured in the period (Mortimer et al., 2014; Kertesz et al., 2010). Moreover, since the ribosome density in translated sequences was shown to also have a periodicity of 3-nucleotides (Ingolia et al., 2009), the structural periodicity could be related to a facilitated translation (Kertesz et al., 2010). Consistently, in all the aforementioned organisms (yeast, human, *C. elegans*, *Arabidopsis*, *Drosophila*) there is a depletion of structure close to the start and the stop codon (Mortimer et al., 2014). This pattern was also found in yeast using the melting temperatures (see Introduction 3.2) (Wan et al., 2012).

In addition to high-throughput profiling experiments, also *in silico* methodologies were applied on multiple RNAs, offering an interesting and large collection of data that is available in public databases. The database compaRNA (Puton et al., 2013) offers RNA secondary structure predictions from 44 algorithms (http://iimcb.genesilico.pl/comparna/methods/) benchmarked against 265 RNA structures coming from PDB data. A past version of LNCipedia (Volders et al., 2013) contained information on 21,488 unique lncRNAs in human, including their secondary structure as predicted by RNAfold (Gruber et al., 2008). The current version of the database includes information on 107,039 lncRNAs (https://www.lncipedia.org/), but the predictions of the secondary structure were not yet available at the time of the writing.

The rapid increase of the amount of available high-throughput data, both experimental and predicted, is not only helping the scientific community to understand the folding rules and the influence of the secondary structure of an RNA on its function, but it is also opening novel and not yet completely explored possibilities for data based prediction approaches (e.g. machine learning) and large-scale analyses. In addition, applying genome-wide structure probing techniques both *in vitro* and *in vivo*, as was done for icSHAPE (Spitale et al., 2015), can be the next step to understand the structural differences between RNA *in vitro* and RNA *in vivo*.

CHAPTER I

A high-throughput approach to profile RNA structure

A high-throughput approach to profile RNA structure

The main goal when I started my PhD was to use high-throughput data from the yeast transcriptome to build a predictor of RNA secondary structure (Kertesz et al. 2010). At that time, a predictive tool trained on large-scale experimental data was still not available, as it was instead common for other fields, such as for instance protein-RNA interaction (Bellucci et al. 2011). The majority of the algorithms of RNA secondary structure (RSS) prediction were based on dynamic programming or on comparative analysis, as I explained in Introduction 5.

During the development of my algorithm, several new experimental techniques started to be published. A new PARS protocol was applied on the complete human transcriptome (Wan et al. 2014), while SHAPE, even if was only applied genome-wide on HIV-1, continued to be the standard in the field. icSHAPE was a change: a SHAPE chemistry-based technique applied high-throughput on mouse, also *in vivo* (Spitale et al. 2015).

All these new techniques, datasets and organisms were integrated as training set in my algorithm. CROSS (computational recognition of RNA secondary structure) is the first tool to predict the RNA secondary structure trained on high-throughput experimental data. Each technique was selected as independent set, and a Neural Network was trained to simulate each technique.

CROSS is able to profile an RNA molecule using only the sequence, at single nucleotide resolution and without sequence length restriction. This is especially important since the majority of the thermodynamics-based algorithms are restricted up to 1000 nucleotides.

CROSS simulates SHAPE data. As SHAPE, it can be used to increase the predicted power of thermodynamic softwares such as RNAstructure, generating *in silico* SHAPE-like constraints.

The ability of CROSS to profile long RNA molecules made it the perfect candidate to assess the RSS of lncRNAs. For this reason, several collaborators working on novel discovered or poorly characterized lncRNAs already employed the tool. Moreover, several tools of the laboratory, such for

example CROSSalign (see Chapter II) or a future version of catRAPID will rely on CROSS architecture. Moreover, CROSS helped several lines of research in the lab, for example it was employed to study the scaffolding properties of the lncRNAs in stress granules, showing how the secondary structure is an important component for their functionality (Botta et al. 2018).

CROSS was published in Nucleic acids research in 2017.

Riccardo Delli Ponti, Stefanie Marti, Alexandros Armaos & Gian Gaetano Tartaglia. A high-throughput approach to profile RNA structure. *Nucleic Acids Research*. 2017, Vol. 45, No. 5. doi:10.1093/nar/gkw1094. PMID: 27899588.

Delli Ponti R, Marti S, Armaos A, Tartaglia GG. A high-throughput approach to profile RNA structure. Nucleic Acids Res. 2017 Mar 17;45(5):e35–e35. DOI: 10.1093/nar/gkw1094

CHAPTER II

A method for RNA structure prediction shows evidence for structure in lncRNAs

A method for RNA structure prediction shows evidence for structure in lncRNAs

In this Chapter, I will introduce CROSSalign, a new tool developed as a further application of the CROSS algorithm. As specified in the previous Chapter, CROSS is a powerful tool, able to profile long and complex RNA molecules in short computational time. The profiles produced by CROSS are also an interesting resource, that can be exploited by other tools. Knowing this potential, in the laboratory we wanted to developed a new algorithm based on CROSS technology.

There are few algorithms available able to structural align RNAs of different length, especially for long molecules. The combination of the structural profiles of CROSS with a dynamic time warping algorithms (DTW), which allows comparison of profiles of different length, is the core of my new algorithm: CROSSalign. CROSSalign is able to compute the structural distance of one or many RNA sequences, regarding of their length, to assess the secondary structure similarity between them. The application of CROSSalign could shine new light on the similarities and the conservation of complex RNA molecules.

The structural conservation of lncRNAs is still debated. The structural conservation was supported for specific regions of lncRNAs such as *HOTAIR* (Somarowthu et al. 2015) and the RepA of *Xist* (Maenner et al. 2010). However, a recent statistical analysis suggested that the secondary structure conservation in the previous cases is not statistical significant (Rivas et al. 2017). To help unfold the mystery, I decided to apply CROSSalign to study structural conservation of lncRNAs.

The results of my analysis reveal a structural conservation between known lncRNA domains including *Xist* RepA and *HOTAIR* D2, supporting a structural conservation for lncRNAs. CROSSalign was also applied on single-stranded RNA (ssRNA) viruses, specifically on HIV. The results highlight regulatory regions with a similar structure in HIV and other ssRNA viruses, opening new questions regarding similar mechanisms mediated by the secondary structure. Worth to specify that RepA and D2 profiles were

accurately recognized in a pool of RNAs reverse engineered to have the same structure but different sequences.

CROSSalign is able to identify thousands of matches between long RNA molecules. The algorithm can be applied to build a 'structural homologome' among RNAs of different organisms (see General Discussion for more details). At the lab we are currently computing the structural homologome between all the long intergenic non-coding RNAs (lincRNAs) of human and mouse ($32x10^6$ interactions).

CROSSalign was recently submitted to Frontiers in Molecular Biosciences.

Riccardo Delli Ponti, Alexandros Armaos, Stefanie Marti & Gian Gaetano Tartaglia. A method for RNA structure prediction shows evidence for structure in lncRNAs. *Frontiers in Molecular Biosciences* (under review).

Riccardo Delli Ponti, Alexandros Armaos, Stefanie Marti & Gian Gaetano Tartaglia. A method for RNA structure prediction shows evidence for structure in lncRNAs. *bioRxiv*. 2018 July. doi:https://doi.org/10.1101/284869.

Delli Ponti R, Armaos A, Marti S, Tartaglia GG. A Method for RNA Structure Prediction Shows Evidence for Structure in lncRNAs. Front Mol Biosci. 2018 Dec 3;5. DOI: 10.3389/fmolb.2018.00111

CHAPTER III

Predicting the *in vivo* structure of RNA molecules

Predicting the *in vivo* structure of RNA molecules

As explained in Introduction 4.1, the RNA structure *in vivo* is a complex puzzle. If proteins are now accepted as pieces in the puzzle, their contribution to the *in vivo* folding is still poorly understood. At the moment of the writing, the use of *in vivo* data by computational approaches is still not explored.

Few trustable *in vivo* experimental data are available, and even less experiments provide both *in vivo* and *in vitro* data for the same dataset/condition. icSHAPE is one of the few techniques providing high-quality data, genome-wide and both *in vitro* and *in vivo* (Spitale et al. 2015).

For this project, I used icSHAPE data and high-throughput predictions to assess the contributions of the proteins and of the crowding effect on the RNA secondary structure *in vivo*. The idea behind the new approach is also to combine the contribute of the RNA sequence and the interacting proteins to achieve a better understanding of the RNA secondary structure *in vivo*. The sequence contribution is exploited using the same machine learning approach behind CROSS (see Chapter I), while to understand the protein contribution we used catRAPID, an algorithm developed in our lab to predict protein-RNA interactions (Bellucci et al. 2011), and a filtering procedure based on RNA binding domains (Ray et al. 2009).

While it is possible to predict *in vitro* data using only the sequence contribution, for the *in vivo* data to achieve better performances it is necessary to provide extra layers of information. The results shows a positive influence of the potential to form a crowded environment, where RNA fragments discriminated by the enrichment in protein binding domains are easier to be predicted than RNA fragments depleted of binding domains. Moreover, the implementation of protein data during the training of the algorithm shows a small improvement in the prediction of *in vivo* structural data.

My computational analysis is a small contribution to the understanding of the RNA secondary structure *in vivo*, but it is the first computational approach that uses the experimental *in vivo* data and that connects the

proteins and their effect on the RNA structure. During the next months I plan to improve and refine the analysis with the aim to build CROSSalive, the first predictor of structural *in vivo* data.

The algorithm could be combined with CROSS to provide the differences between the *in vivo* and *in vitro* structure of RNA molecules. The process behind the analysis can be refined and applied to more complex integrative approaches as Deep Learning procedure employing more contributors, not only the proteins, to achieve a comprehensive knowledge of the RNA structure *in vivo*. I will talk more extensively about this in the General Discussion.

The final manuscript will be submitted to *Bioinformatics*.

Riccardo Delli Ponti, Alexandros Armaos, Fernando Cid & Gian Gaetano Tartaglia. Predicting the *in vivo* structure of RNA molecules. *Bioinformatics* (in preparation).

Predicting the *in vivo* structure of RNA molecules

Riccardo Delli Ponti, Alexandros Armaos, Fernando Cid & Gian Gaetano Tartaglia

Introduction

The RNA secondary structure (RSS) is fundamental for its biological function, especially for the interaction with proteins (Bellucci et al., 2011). The secondary structure of the RNA is altered from *in vitro* to *in vivo* conditions due to the presence of a different environment and the action of external interactors, such as RNA binding proteins and ions (Minton, 2001).

Several experimental techniques probed the secondary structure of transcriptome of different organisms (Strobel et al., 2018). Specific techniques such as icSHAPE, DMS and SHAPE-MaP were applied *in vivo* (Spitale et al., 2015; Rouskin et al., 2013; Siegfried et al., 2014; Smola et al., 2015). However, the complex mechanisms contributing to the formation of the secondary structure *in vivo* are still poorly characterized. Previous analysis suggested a lack of structure for the RNA *in vivo* (Rouskin et al., 2013), while recent results proposed a structural conservation from *in vitro* to *in vivo*, especially for non-coding RNAs (Spitale et al., 2015).

If *in vitro* experimental data are well integrated in thermodynamic approaches, predicting the structure *in vivo* is very difficult (Delli Ponti et al., 2017). Indeed, computational methods cannot predict all the forces driving the RNA structure *in vivo* (Eddy, 2004) and the scientific community can rely only on few experimental techniques able to assess the RNA secondary structure *in vivo*. RNA undergoes a number of modifications in the cellular environment, including methylation (m6a), which is important for post-transcription regulation of gene expression (Wei and Moss, 1977). In the methyl-tranferase complex in mammals, *Mettl3* is the active component responsible of the majority of RNA methylation modifications (Meyer et al., 2012). Recently, icSHAPE *in vivo* data upon *Mettl3* knockdown indicated an

influence of m6a on the RNA secondary structure with a helicase-like activity, promoting a transition toward unpaired nucleotides (Spitale et al., 2015).

Through analysis of transcriptome-wide data we are building a method for the prediction of RNA secondary structure *in vivo*. One key element in our approach is the prediction of protein interactions (Cirillo et al., 2016) which allows us to mimic the complex cellular environment. For the first time we introduce a large-scale analysis of *in vivo* data and highlight a relationship between proteins and RNA structure, predicting *in vivo* data with an accuracy of 0.80 or higher.

Results and Discussion

DMS can study the contributions of adenine and cytosine to RNA structure (Mortimer et al., 2014) and the 1M7 and NMIA reagents of SHAPE-MaP have poor solubility and reactivity (Lee et al., 2017), icSHAPE is, at present, the most reliable *in vivo* technique. Following the strategy applied on our previous work (CROSS)(Delli Ponti et al., 2017), we selected 10^5 RNA regions encoding the highest icSHAPE signal for single- (reactivity=1) and double-stranded conformation (reactivity=0 or lack of signal). These fragments were used to assess the sequences contribution during the training process (see Methods: Training of the network).

Artificial neural networks (ANN) were initially trained only using sequence contributions: icSHAPE *in vitro* data and icSHAPE *in vivo* with and without *Mettl3* knockdown (m6a+/m6a-) (Spitale et al., 2015). Three ANNs (*in vitro*, *in vivo* m6a+, *in vivo* m6a-) were trained on the same conditions (training and testing sets of the same size), and cross-validated between each other (Figure 1a). The *in vitro* model is the one with the best performances in 10-fold validation during the training step, showing how it is easier to predict the RNA structure *in vitro* (0.86 accuracy or ACC; Figure 1b). However, during the cross-validation with the other datasets, we noticed that the *in vitro* model is not able to correctly predict the *in vivo* datasets. This is because of the complexity of the *in vivo* conditions, which cannot be predicted using the

sequence information only (Delli Ponti et al., 2017), since RNA structure is altered by the presence of proteins. To correctly predict the RNA secondary structure *in vivo*, we integrated additional contributions into the predicted model.

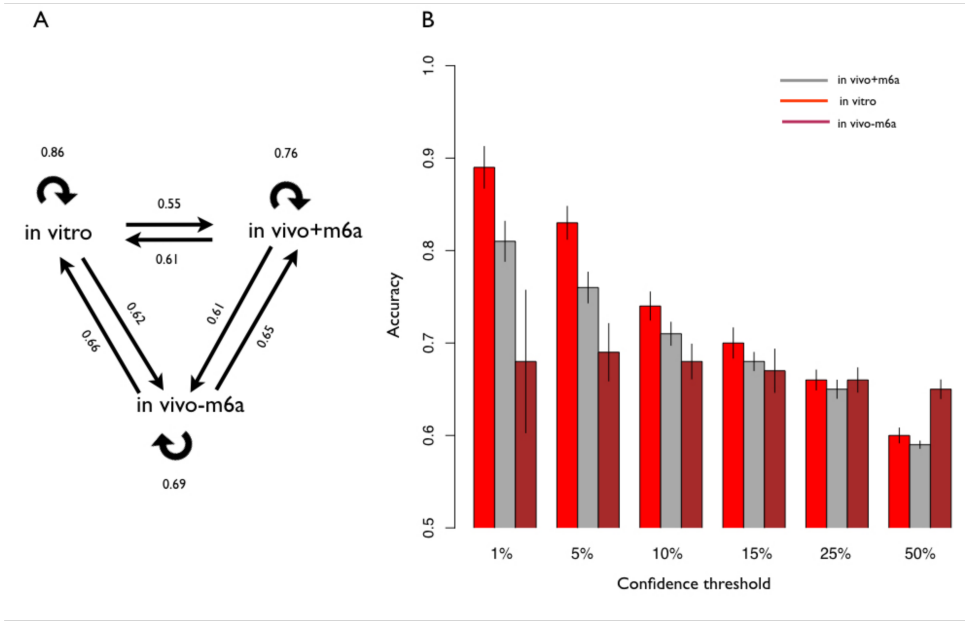


Figure 1: **(A)** Cross-validation between ANN trained on top/bottom 100'000 icSHAPE fragments. The accuracy is reported for the 5% of the testing dataset. Each comparison is done between the best 10 cross validated network against the other datasets. **(B)** Cross validation inside each specific (*in vitro*, *in vivo* m6a+, *in vivo* m6a-) dataset with the same training and testing conditions. The accuracies are reported for the top and bottom percentage of the testing set, where 50% is the complete set.

It is known that the crowding effect or presence of other molecules characterize the difference between *in vitro* and *in vivo* (Minton, 2001). The crowding environment, for instance, has a positive influence on the RNA folding, reducing the degree of freedom toward the structure taken in the cellular *milieu* (Dupuis et al., 2014). To study how the information coming from protein interactions is related to secondary structure predictions, we analyzed the RNA fragments enriched in RNA binding proteins (RBPs) domains (eg. ACACA for HNRNPL; see Methods: Motifs enrichment selection; Figure 2a). The presence of binding domains inside a RNA fragment could be intended as the potential binding and of a resulting crowded environment with more possible interactors. Our results indicate that it is easier to predict the secondary structure of RNA fragments enriched in RBPs motifs, while it is more difficult to predict the structure of fragments depleted of binding domains (Figure 2b). This finding suggests the important contribution of proteins for the predictions of RNA structure *in vivo*. The result also highlights the importance of the crowding effect, since the more a fragment is prone to bind proteins, the easier is to predict its structure. We note that the direct binding of proteins to an RNA could alter and affect the structure, especially if the binding is strong and specific, bringing to a huge difference between the native structure *in vitro* and *in vivo*.

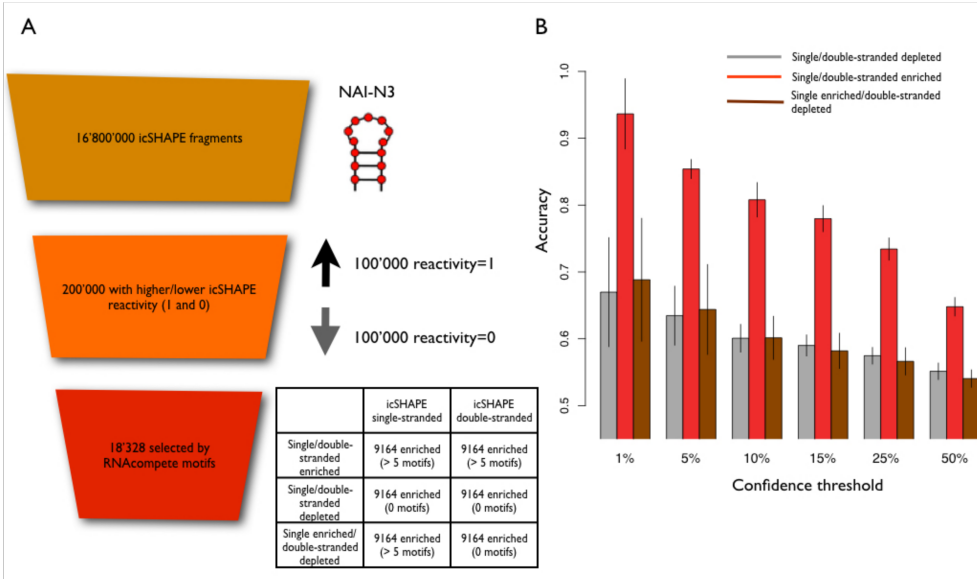


Figure 2: **(A)** Pipeline explaining the process of filtering for the RNA fragments according to their propensity to bind many RBPs (crowding effect). The RNAs are selected for the presence of known RBDs extracted from RNAcompete (Ray et al., 2009). Fragments with more than 5 motifs in their sequence are selected as highly contacted by proteins *in vivo* (enriched), while the ones with 0 motifs as poorly contacted (depleted). **(B)** Cross validation inside each specific dataset. The accuracies are reported for the top and bottom-ranked parts of the testing set using the predicted score as a sorting variable, where 50% is the complete set. Filtering the quality of the RNA fragments using as positive RNAs highly contacted and as negative RNA poorly contacted increases the predictive power of the network, comparing to use both classes enriched or depleted of contacts.

To further prove the protein contribution to predict the RNA secondary structure *in vivo* and how it is affected by the direct binding, we exploited RBPs prediction. This step is fundamental because it allows to directly compute binding of all proteins from first principles (Bellucci et al., 2011), without the use of binding domains retrieved from previous experiments (Agostini et al., 2013).

We used catRAPID to predict the interaction of several fragments of 640 proteins (described in Agostini et al. (2013) and Cirillo et al. (2013) if uniform) with 10^5 double/single-stranded *in vivo* RNA fragments (a total of 12.8×10^7 interactions; Methods: Selection of the protein contribution in the *in vivo* data). The protein fragments were firstly selected for their ability

to discriminate single and double-stranded RNA using icSHAPE data. We found that the dataset was enriched for proteins with GO terms associated to RNA structure, and further selected cases with strong signals (see Methods: Selection of the protein contribution in the *in vivo* data; Figure 3a). The 101 selected proteins were included into the training of the algorithm, to complement the information coming from the sequences. Using also the protein contributions leads to an improvement of the predictive power of the algorithm, up to an ACC of 0.88 during the 10-fold validation (Figure 3b).

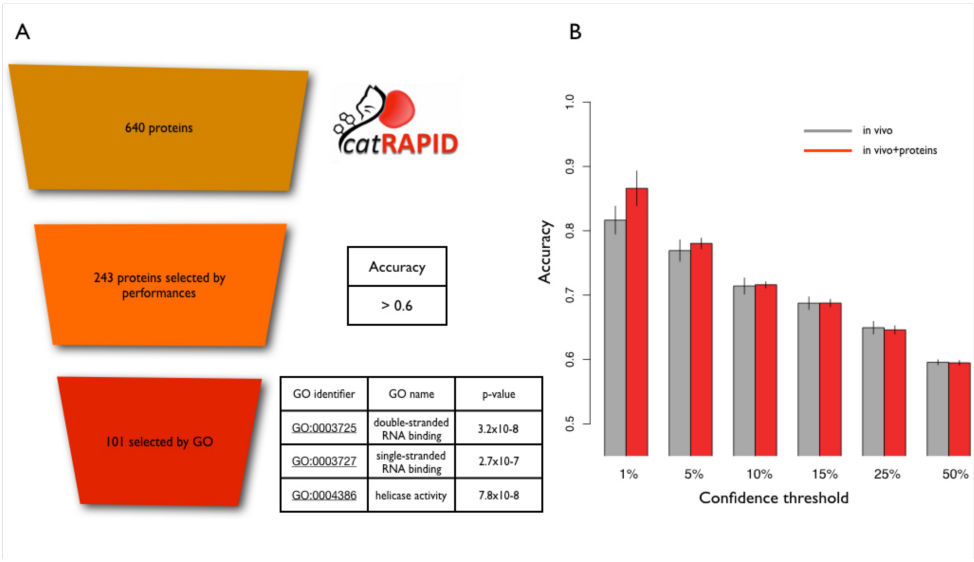


Figure 3: **(A)** Pipeline summarizing the process of filtering for protein contributions. **(B)** Cross validation inside each specific dataset (*in vivo* only sequence, *in vivo* sequence with proteins). The accuracies are reported for the top and bottom percentage of the testing set, where 50% is the complete set. Using the protein information coming from the catRAPID score of the best 101 proteins previously selected increases the predicting power *in vivo*.

With these novel approaches we plan to build CROSSalive, an algorithm able to improve the prediction of RNA secondary structure *in vivo* data. Although we fully understand the limits of our approach to completely understand the features influencing the RNA structure *in vivo*, we have shown for the first time a direct relation between protein interaction and RSS prediction. Future multi-features approaches, such as convolutional neural networks,

could integrate many properties of the *in vivo* environments, including the proteins, to provide a more complete understanding of the RNA secondary structure *in vivo*.

Our approach could be refined in several ways. The proteins could be further filtered or differently normalized to improve the predictive power of the ANN trained joining sequence and protein contributions. Moreover, the potential of crowding environment could be also integrated in the predictive process. The methodology could be tested on other *in vivo* structural data to further prove its predictive power.

The procedure reported for *in vivo* m6a+ data could be also applied for *in vivo* m6a- to predict structural changes related to RNA methylation. The combination of the two predictive approaches (m6a- and m6a+) would lead to a complete understanding of the RNA structure *in vivo* and the methylation pattern related to the structure.

Materials and Methods

Selection of the protein contribution in the *in vivo* data

We used catRAPID *omics* (Agostini et al., 2013) to select the proteins with a stronger binding with the RNA fragments having a higher propensity to be double or single-stranded, according to icSHAPE *in vivo* data (200'000 double- and single stranded fragments). From a starting pool of 630 proteins, we selected the fragments of proteins with a strong predictive power on RBPs domains. The fragments were then divided in enriched ($>$ 5 domains; at least 1/10 of the fragment length) or depleted (0 domains) of binding domains. Neural networks were trained, using the information coming from the sequence only, on different datasets using the combination of depleted/enriched fragments for the two classes (single-stranded and double-stranded secondary structure). The training sets were balanced and of the same size. Anyway, it is easier to predict the secondary structure of fragments when enriched *vs* depleted fragments define the two classes. On the contrary, when both classes are defined by depleted fragments (i.e. without any binding domains), it is more difficult to predict them.

Training of the network

For more information about the component of the ANN, check our previous paper Methods (CROSS; (Delli Ponti et al., 2017)). We selected the 100'000 fragments of 51 nucleotides with the middle nucleotides with a higher propensity to be single-stranded for icSHAPE reactivity, and the 100'000 with a higher propensity to be double-stranded. The window size is enough to capture the combinatorial complexity of icSHAPE data on mouse transcriptome (i.e., $4^{51} > 12x10^6$), which is also an accepted size for catRAPID algorithm. The sequence component was coded using a 'one hot encoding' procedure, where each nucleotide is converted in a 4mer notation: A = (1, 0, 0, 0), C = (0, 1, 0, 0), G = (0, 0, 1, 0) and U = (0, 0, 0, 1). This approach was used to train the ANNs using the sequence only, including the ones based on the RBP motifs enrichment.

To study the protein binding, the catRAPID scores of the 111 discriminative proteins were integrated in the training step. First, the youden cut-off was computed for each protein on the complete dataset of RNA fragments (single/double-stranded). Then the score of each protein was normalized using the cut-off, setting the scores higher to 1, and the lower to -1. The following 111 normalized scores were integrated for each RNA with the information coming from the sequence, for a total training complexity of 305 variables.

References

Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D., and Tartaglia, G. G. (2013). catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics (Oxford, England)*, 29(22):2928–2930.

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nature Methods*, 8(6):444–445.

Cirillo, D., Agostini, F., and Tartaglia, G. G. (2013). Predictions of protein-RNA interactions: Protein-RNA interactions. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(2):161–175.

Cirillo, D., Blanco, M., Armaos, A., Buness, A., Avner, P., Guttman, M., Cerase, A., and Tartaglia, G. G. (2016). Quantitative predictions of protein interactions with long noncoding RNAs. *Nature Methods*, 14(1):5–6.

Delli Ponti, R., Marti, S., Armaos, A., and Tartaglia, G. (2017). A high-throughput approach to profile RNA structure. *Nucleic Acids Research*, 45(5):e35–e35.

Dupuis, N. F., Holmstrom, E. D., and Nesbitt, D. J. (2014). Molecular-crowding effects on single-molecule RNA folding/unfolding thermodynamics and kinetics. *Proceedings of the National Academy of Sciences*, 111(23):8464–8469.

Eddy, S. R. (2004). How do RNA folding algorithms work? *Nature Biotechnology*, 22(11):1457–1458.

Lee, B., Flynn, R. A., Kadina, A., Guo, J. K., Kool, E. T., and Chang, H. Y. (2017). Comparison of SHAPE reagents for mapping RNA structures inside living cells. *RNA*, 23(2):169–174.

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell*, 149(7):1635–1646.

Minton, A. P. (2001). The Influence of Macromolecular Crowding and Macromolecular Confinement on Biochemical Reactions in Physiological Media. *Journal of Biological Chemistry*, 276(14):10577–10580.

Mortimer, S. A., Kidwell, M. A., and Doudna, J. A. (2014). Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics*, 15(7):469–479.

Ray, D., Kazan, H., Chan, E. T., Pea Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B. J., Morris, Q., and Hughes, T. R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology*, 27(7):667–670.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J. S. (2013). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, 505(7485):701–705.

Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E., and Weeks, K. M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods*, 11(9):959–965.

Smola, M. J., Calabrese, J. M., and Weeks, K. M. (2015). Detection of RNAProtein Interactions in Living Cells with SHAPE. *Biochemistry*, 54(46):6867–6875.

Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H. Y., Batista, P. J., Torre, E. A., Kool, E. T., and Chang, H. Y. (2015). Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, 519(7544):486–490.

Strobel, E. J., Yu, A. M., and Lucks, J. B. (2018). High-throughput determination of RNA structures. *Nature Reviews. Genetics*, 19(10):615–634.

Wei, C. M. and Moss, B. (1977). Nucleotide sequences at the N6-methyladenosine sites of HeLa cell messenger ribonucleic acid. *Biochemistry*, 16(8):1672–1676.

General discussion

Limits of the understanding of the RNA structure

The RNA secondary structure (RSS) is crucial for understanding the functionality, the role and the interactions of a RNA molecule (Bellucci et al., 2011). The folding process *in vitro* is controlled by primary structure alone while *in vivo* there is a complex environment in which multiple components may exert a strong influence (Rouskin et al., 2014). While computational approaches are still unable to understand the complex mechanisms of the folding *in vivo*, several thermodynamic algorithms are able to predict the RNA secondary structure *in vitro* using dynamic programming (for more details check Introduction). Regarding the computational approaches, RNAstructure (Reuter and Mathews, 2010) and Vienna (Gruber et al., 2008) are considered the golden standard for RNA secondary structure prediction. Even if they are constantly updated with new modules and features, the core of the thermodynamic approaches is still very similar to the first algorithm introduced by Zucker and Siegler almost four decades ago (Zuker and Sankoff, 1984; Stormo, 2006). Most of the biological rules of the thermodynamics-based algorithms are extracted from optical melting experiments, which provide energetic constraints for the parameterization of the free energies (Martin et al., 1971). These parameters, even if updated, are still based on few RNA structures, usually tRNAs, rRNAs and ribozymes (Martin et al., 1971).

NMR and X-ray are still considered the universal standards for the definition of a structure (Latham et al., 2005). If this is vastly true and accepted for proteins, the RNA crystals are instead few and limited, as I discussed in the Introduction 3.2 referring to the RNAstrand database. Analyzing the complete PDB database (https://www.rcsb.org/), only 1,276 RNA structures are available at the time of the writing. Of this pool, 808 crystals are of unknown taxonomy, with only 39 crystals coming from human. The crystals are limited in length, with only 47 RNA structures longer than 200 nucleotides, with a maximum of 2,880nt for 23S rRNA of *Deinococcus radiodurans* (ID 2O43).

Computational approaches are also limited by the length of the sequence, with an accuracy that drops for sequences larger than 700 nucleotides

(Hajiaghayi et al., 2012; Lu et al., 2009). Even if the algorithm can process large sequences, the computational times are huge and these approaches are usually not suited for high-throughput analysis (Agostini et al., 2013).

High-throughput characterization of the RNA structure

Limitations and restricted applicability of crystallographic techniques for the RNA polarized the scientific community toward the developing of new high-throughput technologies (for details check Introduction 3.2).

As I discussed in the Introduction 3.2 and 4.2, high-throughput techniques were applied on complete transcriptomes. The comparison of the results would lead to build an RNA structurome (see Introduction 6). This is achievable only using genome-wide data, which was impossible in the past having because only low-throughput crystallographic structural data were available. The advent of genome-wide techniques is bringing a major understanding of the universal role of RSS.

The majority of the predictive algorithms in different fields (for example protein RNA interaction) are based on machine learning approaches trained on experimental data, not on thermodynamic principles (Bellucci et al., 2011; Alipanahi et al., 2015; Cirillo et al., 2015; Livi et al., 2016; Danko et al., 2015; Mort et al., 2014). With the lack of experimental data regarding RNA structures, training a machine learning approach was an almost impossible task. However, the new flow of data coming from high-throughput techniques could be used to train machine-learning approaches. For example, the predictive power of icSHAPE data was already explored not to predict the RSS, but to improve the predictive power of protein-RNA interactions (Spitale et al., 2015).

SeqFold was one of the first tools based on experimental data (Ouyang et al., 2013). The algorithm is not a proper machine learning approach trained on the experimental data, but it uses the experimental profiles to select the best structure. Thermodynamic-based methodologies use experimental data, mainly to guide and improve their predictive power. These integrative approaches combine experimental data as constraints for thermodynamics-

based approaches, as done in other fields (Vendruscolo et al., 2003). These methods are able to improve performances with very good reliability (Low and Weeks, 2010; Hajdin et al., 2013; Lorenz et al., 2016), but the algorithms lose all their benefits since they would need an experiment to set the constraints.

Integrative approaches exploit experimental data to refine theoretical models. A direct result is an improvement in the predictive power of the algorithms, such as the use of SHAPE data inside thermodynamics-based approaches (Low and Weeks, 2010; Lorenz et al., 2016). Experimental data, used as a soft-constraint (see Introduction 5.3 for details), allow a reduction of the conformational space, with less folded intermediates to be sampled by the algorithm. In Chapter I I presented my algorithm, CROSS, a tool designed to high-throughput predict the RSS and trained on experimental data. One interesting feature of CROSS is the ability to simulate *in silico* SHAPE constraints, which can be used to improve the predictive power of RNAstructure and Vienna. This unique feature can lead to further developments in which algorithms can be used to generate new constraints to improve thermodynamic approaches. Alternatively, experimental techniques should be used to provide constraints to improve the predictive power of thermodynamic approaches. For instance, NMR chemical shift data were used as soft-constraints to improve the accuracy of secondary structure predictive approaches (Zhang and Frank, 2018).

SHAPE data are integrated in other computational approaches. The use of soft-constraint can be a guideline to improve different computational fields. For instance SHAPE score was used to discover structural domains that could not be identified using single-nucleotide resolution (Pollom et al., 2013). Moreover, SHAPE data were used to study structural conservation (Lavender et al., 2015b) and they were also combined with the sequence as guideline for multiple alignments (Lavender et al., 2015a). Since the structural score of CROSS can also be used as *in silico* alternative to SHAPE data to improve the predictive power of RNAstructure and Vienna, new applications of CROSS profiles can lead to novel improvements, as for SHAPE data. CROSS is the first algorithm to predict the RNA secondary structure trained on experimental data, without any sequence

length restriction and at single nucleotide resolution. The high-throughput application of CROSS can profile in few days entire transcriptomes, leading to the possibility of developing *in silico* structuromes.

During the developing of CROSS, five datasets of different techniques and organisms were trained using Neural Networks (see Chapter I). The models learned basic rules of the techniques and interesting characteristics arise from the cross-validation of the models.

The ability of CROSS to reproduce different experimental techniques is one of the most interesting feature that should be further exploited. Using more than one model on the same RNA allows to sample multiple conformations of the same structure, based on the characteristics learned from each technique. Sampling multiple conformations is a relevant point to understand the folding and the functionality of an RNA molecule. Indeed, SHAPE data were already used to sample multi structural conformations (Kutchko et al., 2015). Moreover, to sample long and complex RNA molecules, such as lncRNAs, different experimental techniques can be used to obtain multiple profiles, which can be combined to have a more complete view of the secondary structure (Novikova et al., 2012). The different modules of CROSS can be applied in the same way on long and complex RNA to extract useful structural information on regions with consensus (i.e. stable conformation) or disagreement (i.e. regions of variability) between the models trained on different techniques.

Worth to specify that the genome-wide techniques are not completely accurate (Delli Ponti et al., 2017; Wu et al., 2015). During the training of the ANNs (artificial neural networks) at the core of CROSS methodology, a fundamental step was the filtering and the assessment of high-quality data. Training fragments not consistently associated to a high-signal were discarded from the set. The quality of the data is a critical step to achieve better predictive power for an ANN (Najafabadi et al., 2015).

CROSS is not only a suggested application to profile the secondary structure of lncRNAs but, thanks to its high-throughput applicability, can be used to generate genome-wide *in silico* data that can be the base to a bigger understanding of the RNA structurome.

Toward the structural homologome

Having a powerful methodology to profile the RSS of very large molecules opens many possibilities and applications. To understand RNA folding, the simple prediction of secondary structure is not enough, but other related properties can be explored, such as for instance conservation. Indeed, the conservation of the secondary structure is an important property for the understanding of the functionality and the importance of a specific structure across the species.

CROSSalign was introduced in Chapter II and it represents the natural evolution of the CROSS method. The algorithm is not a new method, but a combination of CROSS and DTW (dynamic time warping) algorithm (Giorgino, 2009). The two approaches together allow the comparison of structural profiles providing a structural distance to evaluate the similarity. In essence, CROSSalign computes pairwise distances, evaluating structural homology only on the base of the predicted secondary structure of the two RNAs. This approach differs from the multiple-alignment, where it is necessary to use a large number of homologs to establish conservation (Tavares et al., 2018).

The ability of CROSSalign to provide pairwise alignments for large molecules and domains offers a huge versatility to understand structural conservation and similarity. Indeed, finding global structural similarities between different RNAs can lead to discover molecules that can share similar mechanisms or functionality, or even a class of molecules (Ganot et al., 1997).

Even if the overall sequence is not conserved for long and complex RNA molecules, specific important regions can be conserved in structure (Somarowthu et al., 2015). The search of conserved structural domains could be fundamental to discover regulatory regions (Lu et al., 2011). Given the importance of these domains, CROSSalign also allows the division of a profile into subdomains to search for specific structural conserved regions.

CROSSalign was build focusing on its high-throughput applicability. At the time of the writing, the structural profile of a target RNA can be compared with the complete set of long-intergenic non-coding RNAs (lincRNAs)

of several organisms (zebrafish, human, mouse, macaque and rat). The resulting compendium of interactions is a prime example of what can be called *in silico* 'structural homologome', based on the secondary structure similarities. This 'structural homologome' is a powerful instrument to identify RNAs with strong similarity in complete organisms. Indeed, novel RNAs could be clustered into structural families and their unknown functionality could be associated to their homologs in structure. The structural homology adds a new level of characterization, especially for molecules in which the sequence is not well conserved, such as for instance the lncRNAs. If CROSS profiles long and multiple RNAs to build an *in silico* 'structurome', CROSSalign lead to the high-throughput characterization of the complete set of a 'structural homologome'. For example, to gather further knowledge in how much human and mouse are truly similar, the complete set of human and mouse ncRNAs could be checked for structural homology using CROSSalign.

Rivas et al. in 2017 assessed that the secondary structure conservation previously reported for the lncRNAs (Somarowthu et al., 2015) was not statistically significant (Rivas et al., 2017). This assumption brought new skepticism regarding the already debated structural conservation for the lncRNAs. As I previously introduced in Chapter II, with the use of CROSSalign I found a strong structural homology among primates. Recent works supported the results, suggesting a structural conservation for a number of lncRNAs (Tavares et al., 2018). At present, more studies and analysis are needed to understand the complex pattern of the structural conservation of the lncRNAs.

In conclusion, CROSSalign is a useful and powerful approach that can establish in a high-throughput way the pairwise structural similarities of different RNAs. The algorithm can lead to novel discoveries and a deep understanding of the conservation of long and complex RNA molecules such as lncRNAs or ssRNA viruses, or to identify homologues in function through the homology in structure.

A piece toward the solution of the *in vivo* puzzle

If the rules behind the *in vitro* folding of the RSS could be considered a solved problem (Tinoco and Bustamante, 1999), the puzzle of the *in vivo* RNA structure and the complete set of cellular components influencing it is still a mystery (Leamy et al., 2016). As in a proper puzzle, many pieces can contribute to the formation of the *in vivo* structure of a specific RNA. The sequence is unconditionally a piece of the puzzle, but other components are still unknown.

Crowding component, chaperones and ligands have an effect on the correct *in vivo* folding (for details see Introduction 4.1). However, there is not a direct quantification of the contributions of these factors for the *in vivo* RSS folding. Moreover, computational approaches are not able to use *in vivo* constraints, and also at the time of the writing there are no algorithms to predict the RSS *in vivo*.

As introduced in Chapter III, the proteins and the crowding are an important component, and their computational contribution is fundamental to increase the performances of *in vivo* predictions. Combining sequence and protein information, and using icSHAPE *in vivo* data will lead to the building of CROSSalive, the first algorithm able to predict RSS *in vivo* data.

CROSS and CROSSalive could be used to provide profiles for the RSS both *in vivo* and *in vitro*. Comparing this information will lead to a major understanding of structural changes of the RNA, identifying regions with a different conformation *in vivo* and *in vitro*.

The results of Chapter III could be further improved with additional steps of proteins selection and a refined training. Moreover, the analysis could be applied also for *in vivo* data with an altered methylation pattern, trying to correctly predict for the first time structural modifications affected by methylation.

However, the complex puzzle of the *in vivo* folding is far to be solved. Multiple components can be necessary for the correct folding of the RNA, not only the proteins. A deep learning approach could be used to prioritize and integrate multiple components into a predictive approach. Deep learning

algorithms were successfully implemented in several biological analysis (LeCun et al., 2015). I will continue to talk about deep learning in the section of Future perspectives.

In summary, the algorithms that I developed during my PhD can be used to provide several layers of high-throughput information for the RSS. For example, using all the algorithms on a novel transcript it is possible to: 1) discover its secondary structure both *in vitro* and in future *in vivo*; 2) to obtain multiple structures based on multiple experimental techniques; 3) to discover structural homologs in different species.

Future perspectives

To understand and discover the characteristics of the RSS, it is necessary to have a continuous and abundant flow of experimental data. Indeed, genome-wide techniques will soon provide a complete understanding of the structural modifications for the RNA both *in vivo* and *in vitro*. Only in 2018 two new transcriptomes were profiled by an experimental technique (Deng et al., 2018; Kaushik et al., 2018). Yet, at the time of the writing, only icSHAPE was performed both on the *in vitro* and *in vivo* murine transcriptome (Spitale et al., 2015). Although *XIST* was profiled in both conditions using SHAPE reactives (Smola et al., 2016), many other transcripts have not yet been studied. To truly understand the *in vivo* 'structurome', we need more of these data. Information on the human transcriptome coming from the same technique, in the same conditions, and both *in vivo* and *in vitro*, is still missing.

The development of new experimental techniques is fundamental for the future of the field. Evolution of techniques *in vivo* is crucial and the accuracy of the measurements is essential, especially to train machine-learning algorithms. The search for new chemical-probes would be the key to understand RNA structural mechanics and to open the door to the developing of new techniques. The SHAPE probes already provide information on stacked nucleotides and long-range interactions (Siegfried et al., 2014). A recent study indicates a deeper complete understanding of the probing ability of SHAPE reactives, using molecular simulations to

unravel the binding mechanics (Mlýnský and Bussi, 2018). New probes will indeed provide important information for pseudoknots and tertiary structures. Importantly, SHAPE reactivities could be further investigated to discriminate *in vivo* between double-stranded nucleotides and regions inaccessible to the probe because bound to proteins. This discrimination of the data will also improve the predictive power of algorithms to predict the RSS *in vivo*.

I envisage that the scientific community will soon put more effort to solve the mystery regarding the structural conservation of lncRNAs. I speculate that the additional level of complexity coming from the RSS, compared to the sequence only, will be key to establish the conservation for the lncRNAs. To achieve this objective, new experiments will be done in order to obtain more structural data for the RNAs of different organisms. In this scenario, the high-throughput structural distances from CROSSalign could help to understand which RNAs are structurally similar in different species, reducing the number of samples for the experiments.

Importantly, fast computational approaches will be used to exploit information from *in silico* transcriptomes. High-throughput algorithms will be developed for this task, providing a new flow of *in silico* structural data that could be implemented in comprehensive databases. CROSS is one of the tools that is already deployed for this task and it could provide a big database of *in silico* structural profiles. A possible application of CROSS will lead to an improved dataset with the structural profiles for all the lncRNAs from LNCipedia (Volders et al., 2013) or for the recently curated FANTOM5 data (Hon et al., 2017).

With the recent increase of deep learning bioinformatic algorithms, a tool based on this architecture is still missing for the prediction of the RSS. Complex architectures, such as convolutional neural networks (CNN), would benefit from the huge amount of data coming from high-throughput experiments. Moreover, multiple features of the RNA structures can be converted into training data for the CNN to improve the predictive power on the RSS. Multiple features such as protein interactions, crowding effect, and ions concentration will improve the predictions, building an algorithm able to quantify the contributions of each external force on the RSS *in vivo*.

These algorithms could take advantage from meta-predictions, using *in silico* data for their training set using data coming from structural conservation (CROSSalign) or energetic values (RNAstructure).

The 'structural content' of an RNA defines important properties of the molecule that have not been investigated in this thesis. The concept was used in recent papers as discriminative feature especially in RNA granules (Botta-Orfila et al., 2018). While, the classification of lncRNAs is mainly based on their length, a novel structural classification will highlight their role in formation of biological condensates (Van Treeck and Parker, 2018).

Conclusions

During my PhD I focused on the discovery of structural properties of the RNA through developing novel algorithms and methodologies. The understanding of the predictive power of high-throughput experimental data was the basis for the development of my algorithms. The tools were all designed for high-throughput applicability, i.e. they are fast and without sequence length restriction. In the following points I will describe the main findings of my thesis:

- CROSS is the first algorithm trained on high-throughput experimental data (Chapter I) coming from different techniques (SHAPE, icSHAPE, PARS and also low-throughput NMR/X-ray data) and organisms (human, mouse, HIV and yeast). Five different models were built on these datasets to be applied as *in silico* alternative to experimental techniques or as a global consensus to profile the RSS.

- CROSS was successfully applied to profile the entire murine *Xist* and to understand the differences in structure in the complete set of human CDS and UTR regions. The high-throughput nature of CROSS makes it suitable to profile long and complex molecules such as lncRNAs or complete transcriptomes.

- CROSS was combined with a dynamic time warping algorithm (DTW) to build CROSSalign (Chapter II), a methodology able to assess the structure similarities of RNAs with different lengths. The algorithm identified a strong structural conservation for the RepA of *Xist* and the D2 of *HOTAIR*, especially for primates.

- The novel approach of CROSSalign and its high-throughput applicability allow assessing the structural similarities of thousands of RNAs by comparing them pairwise. The tool can be applied to understand and to build *in silico* 'structural homologome'.

- Interaction with other molecules alter RNA structure *in vivo*. The proteins and the crowded environment should indeed be taken into account for the prediction of RNA secondary structure *in vivo* data. The structure of RNAs where protein binding motifs are present is easier to predict, while the integration of protein prediction data using the

catRAPID method leads to an improvement in the predictions of the RSS *in vivo*. The approach is still being developed.

- The *in vivo* analysis could lead to build CROSSalive, a predictive approach to profile the RNA structure *in vivo*.

The suite of tools that I developed during my PhD has a wide applicability and can be employed to predict the structural profile of novel lncRNAs or to obtain entire '*in silico* structurome'. Combining the two algorithms on a novel RNA, it is possible to obtain the secondary structure profile *in vitro* and the structural homologs at transcriptome level. The application of the *in vivo* findings will lead to an algorithm that is able to predict the RNA secondary structure *in vivo*.

Appendix
List of publications

1. Federico Agostini, Davide Cirillo, **Riccardo Delli Ponti**, Gian Gaetano Tartaglia. SeAMotE: a method for high-throughput motif discovery in nucleic acid sequences. *BMC Genomics*. 2014; 15(1): 925. Published online 2014 October 23. doi: 10.1186/1471-2164-15-925 PMCID: PMC4223730

2. Federico Agostini, Davide Cirillo, Carmen Maria Livi, **Riccardo Delli Ponti**, Gian Gaetano Tartaglia. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*. *Bioinformatics*. 2014 October 15; 30(20): 29752977. Published online 2014 June 1. doi: 10.1093/bioinformatics/btu420 PMCID: PMC4184263

3. Livi CM, Klus P, **Delli Ponti R**, Tartaglia GG. catRAPID signature: identification of ribonucleoproteins and RNA-binding regions. *Bioinformatics*. 2015 Oct 31. pii: btv629. [Epub ahead of print] PubMed PMID: 26520853; PubMed Central PMCID: PMC4795616.

4. Klus P, **Ponti RD**, Livi CM, Tartaglia GG. Structural disorder and RNA-binding ability: a new approach for physico-chemical and gene ontology classification of multiple datasets. *BMC Genomics*. 2015 Dec 16;16:1071. doi: 10.1186/s12864-015-2280-z. PubMed PMID: 26673865; PubMed Central PMCID: PMC4681139.

5. **Delli Ponti R**, Marti S, Armaos A, Tartaglia GG. A high-throughput approach to profile RNA structure. *Nucleic Acids Res*. 2016 Nov 28. pii: gkw1094. PubMed PMID: 27899588.

6. Ribeiro DM, Zanzoni A, Cipriano A, **Delli Ponti R**, Spinelli L, Ballarino M, Bozzoni I, Tartaglia GG, Brun C. Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs. *Nucleic Acids Res*. 2018 Jan 25;46(2):917-928. doi: 10.1093/nar/gkx1169. PubMed PMID: 29165713; PubMed Central PMCID: PMC5778612.

7. Teresa Botta-Orfila, Fernando Cid-Samper, Iona Gelabert-Baldrich, Benjamin Lang, Nieves Lorenzo-Gotor, **Riccardo Delli Ponti**, Lies-Anne WFM Severijnen, Benedetta Bolognesi, Ellen Gelpi, Renate K Hukema-Felten, Gian Gaetano Tartaglia. Phase Separations Driven by

RNA Scaffolds and Protein Sequestration in FXTAS. *Cell Reports*, 2018 [under review].

8. E. Panatta, A. Lena, M. Mancini, A. Smirnov, **R Delli Ponti**, T. Botta-Orfila, G. Tartaglia, G. Calin, G. Melino, E. Candi. Ultra-conserved non-coding transcript T-UC291 controls keratinocyte differentiation by interfering with ACTL6A. *EMBO Reports*, 2018 [under review]

9. **Riccardo Delli Ponti**, Alexandros Armaos, Stefanie Marti & Gian Gaetano Tartaglia. A method for RNA structure prediction shows evidence for structure in lncRNAs. *Frontiers* [under submission]

10. Roberto Vendramin, Yvessa Verheyden, Hideaki Ishikawa, Lucas Goedert, Emilien Nicolas, Kritika Saraf, Alexandros Armaos, **Riccardo Delli Ponti**, Keiichi Izumikawa, Pieter Mestdagh, Prof. Denis Lafontaine, Gian Gaetano Tartaglia, Prof. Nobuhiro Takahashi, Jean-Christophe Marine. SAMMSON fosters cancer cell fitness by enhancing concertedly mitochondrial and cytosolic translation. *Nature Structural and Molecular Biology*, 2018 [ahead of publication]

11. Marion Alriquet, Adrían Martínez-Limón, **Riccardo Delli Ponti**, Martin Hengesbach, Gian Tartaglia, Giulia Calloni, and Martin Vabulas.Free RNA recruits TRMT6/61A methyltrasferase in the early phase of proteostasis stress. *Molecular Cell*, 2018 [under review]

Appendix
CHAPTER I

Delli Ponti R, Marti S, Armaos A, Tartaglia GG. A high-throughput approach to profile RNA structure. Nucleic Acids Res. 2017 Mar 17;45(5):e35–e35. DOI: 10.1093/nar/gkw1094

Appendix
CHAPTER II

Delli Ponti R, Armaos A, Marti S, Tartaglia GG. A Method for RNA Structure Prediction Shows Evidence for Structure in lncRNAs. Front Mol Biosci. 2018 Dec 3;5. DOI: 10.3389/fmolb.2018.00111

Appendix
Posters

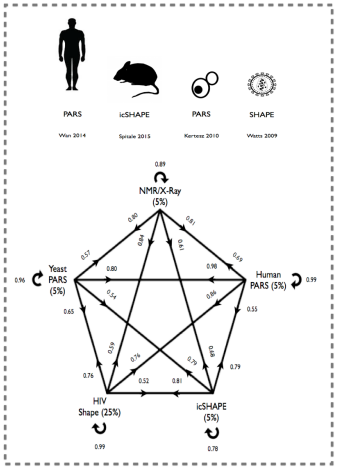
CROSS: A High-Throughput Approach to Profile RNA structure

Riccardo Delli Ponti[1,2], Stefanie Marti[1,2], Alexandros Armaos[1,2] and Gian Gaetano Tartaglia[1,2,3]

1) Gene Function and Evolution, Centre for Genomic Regulation (CRG)
2) Universitat Pompeu Fabra (UPF)
3) Institució Catalana de Recerca i Estudis Avançats (ICREA)

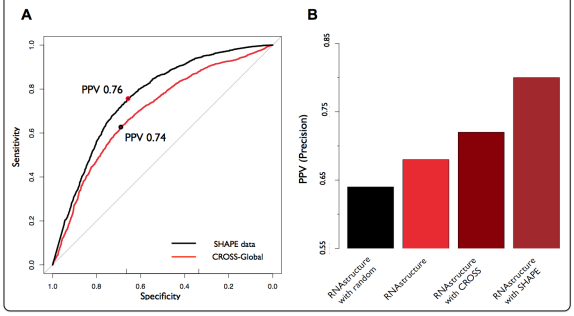
In addition to low-throughput techniques such as Nuclear Magnetic Resonance NMR and X-Ray crystallography [1], recent efforts have started to exploit biochemical reactions to perform high-throughput profiling of RNA structure. For instance, PARS (Parallel Analysis of RNA Structure) distinguishes double- and single-stranded regions using the catalytic activity of two enzymes, the RNase V1 (able to cut double-stranded nucleotides) and S1 (able to cut single-stranded nucleotides) [2,3], while SHAPE (Selective 2'-Hydroxyl Acylation analyzed by Primer Extension) [4,5] employs highly-reactive chemical probes such as NAI-N3 to characterize RNA backbone flexibility. Here we introduce the first computational approach - CROSS (Computational Recognition of Secondary Structure) - to predict the profile of an RNA structure using sequence information only trained on experimental high-quality data without any sequence length restriction. We validated CROSS on large-scale studies (PARS: yeast and human transcriptomes and icSHAPE: mouse transcriptome) as well as low-throughput (SHAPE: HIV RNA) and high-quality NMR/X-ray data.

Datasets



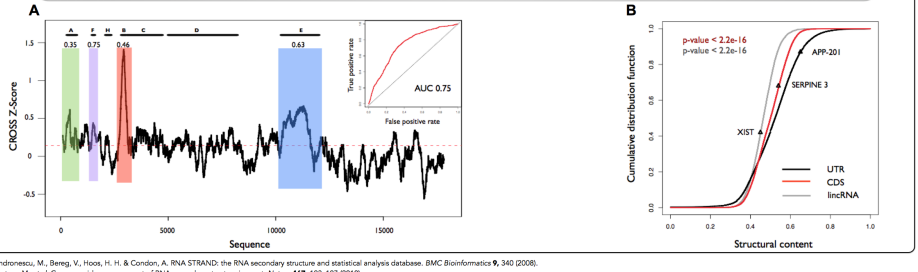
Validation

CROSS models are specific for individual techniques. Yet, each experimental approach has practical limitations and different range of applicability. For this reason, we integrated the five models into a new variable, *Global Score*, to provide the *consensus* signal. The algorithm was then tested on 22 independent structures. The performances are comparable with the experimental SHAPE data (**A**). *Global Score* was used as constraint to improve *RNAstructure* [6] (PPV from 0.64 to 0.72; up to 30% in some structures; **B**).



Long non-coding RNAs

In agreement with DMS experiments applied on *Xist* (AUC 0.75 between experimental and predicted high-signal profiles) [7], CROSS identifies the structural elements associated with repetitive regions Rep A, B, E and F and resolves their internal structures with correlations of 0.35, 0.46, 0.63 and 0.75, respectively (**A**). While the sequences of Rep A and B are conserved across species and show a high degree of structural content, the 3' region of *Xist* is variable and predicted by CROSS to be more single-stranded [8]. We employed CROSS to analyze the structural differences between human coding sequences CDSs, untranslated regions UTRs and long intergenic non-coding transcripts (lincRNAs; **B**). Long non-coding RNAs have complex regulatory abilities and their structure could be more flexible and less structured to provide a wide range of interactions. **Since CROSS is a fast algorithm able to profile the secondary structure of *Xist* in less than 2 minutes, a natural consequence is to apply the algorithm specifically on lncRNAs or genome-wide. analysis.**



1: Andronescu, M., Bereg, V., Hoos, H. H. & Condon, A. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics* **9**, 340 (2008).
2: Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103–107 (2010).
3: Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–709 (2014).
4: Spitale, R. C. *et al.* Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**, 486–490 (2015).
5: Wilkinson, K. A., Merino, E. J. & Weeks, K. M. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat Protoc* **1**, 1610–1616 (2006).
6: Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
7: Fang,R., Moss,W.N., Rutenberg-Schoenberg,M. and Simon,M.D. (2015) Probing Xist RNA Structure in Cells Using Targeted Structure-Seq. PLOS Genet, 11, e1005668.
8: Nesterova et al., Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. Genome Res., 11, 833–849.

References

Agostini, F., Cirillo, D., Bolognesi, B., and Tartaglia, G. G. (2013). X-inactivation: quantitative predictions of protein interactions in the Xist network. *Nucleic Acids Research*, 41(1):e31.

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838.

Andronescu, M., Bereg, V., Hoos, H. H., and Condon, A. (2008). RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC bioinformatics*, 9:340.

Auweter, S. D., Oberstrass, F. C., and Allain, F. H.-T. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, 34(17):4943–4959.

Baird, N. J., Westhof, E., Qin, H., Pan, T., and Sosnick, T. R. (2005). Structure of a folding intermediate reveals the interplay between core and peripheral elements in RNA folding. *Journal of Molecular Biology*, 352(3):712–722.

Banerjee, A. R. and Turner, D. H. (1995). The time dependence of chemical modification reveals slow steps in the folding of a group I ribozyme. *Biochemistry*, 34(19):6504–6512.

Batey, n., Rambo, n., and Doudna, n. (1999). Tertiary Motifs in RNA Structure and Folding. *Angewandte Chemie (International Ed. in English)*, 38(16):2326–2343.

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nature Methods*, 8(6):444–445.

Bernat, V. and Disney, M. D. (2015). RNA Structures as Mediators of Neurological Diseases and as Drug Targets. *Neuron*, 87(1):28–46.

Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R., and Stadler, P. F. (2008). RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9(1):474.

Bindewald, E., Kluth, T., and Shapiro, B. A. (2010). CyloFold: secondary structure prediction including pseudoknots. *Nucleic Acids Research*, 38(Web Server issue):W368–372.

Bleichert, F. and Baserga, S. J. (2007). The long unwinding road of RNA helicases. *Molecular Cell*, 27(3):339–352.

Bokov, K. and Steinberg, S. V. (2009). A hierarchical model for evolution of 23s ribosomal RNA. *Nature*, 457(7232):977–980.

Botta-Orfila, T., Cid-Samper, F., Gelabert-Baldrich, I., Lang, B., Lorenzo-Gotor, N., Delli Ponti, R., WFM Severijnen, L.-A., Bolognesi, B., Gelpi, E., Hukema-Felten, R. K., and Tartaglia, G. G. (2018). Phase Separations Driven by RNA Scaffolds and Protein Sequestration in FXTAS. *bioRxiv*.

Boyle, J., Robillard, G. T., and Kim, S. H. (1980). Sequential folding of transfer RNA. A nuclear magnetic resonance study of successively longer tRNA fragments with a common 5′ end. *Journal of Molecular Biology*, 139(4):601–625.

Brehm, S. L. and Cech, T. R. (1983). Fate of an intervening sequence ribonucleic acid: excision and cyclization of the Tetrahymena ribosomal ribonucleic acid intervening sequence in vivo. *Biochemistry*, 22(10):2390–2397.

Brion, P. and Westhof, E. (1997). Hierarchy and dynamics of RNA folding. *Annual Review of Biophysics and Biomolecular Structure*, 26:113–137.

Buratti, E. and Baralle, F. E. (2004). Influence of RNA secondary structure on the pre-mRNA splicing process. *Molecular and Cellular Biology*, 24(24):10505–10514.

Chadalavada, D. M., Senchak, S. E., and Bevilacqua, P. C. (2002). The folding pathway of the genomic hepatitis delta virus ribozyme is dominated by slow folding of the pseudoknots. *Journal of Molecular Biology*, 317(4):559–575.

Cirillo, D., Botta-Orfila, T., and Tartaglia, G. G. (2015). By the company they keep: interaction networks define the binding ability of transcription factors. *Nucleic Acids Research*, 43(19):e125.

Climie, S. C. and Friesen, J. D. (1988). In vivo and in vitro structural analysis of the rplJ mRNA leader of Escherichia coli. Protection by bound L10-L7/L12. *The Journal of Biological Chemistry*, 263(29):15166–15175.

Crothers, D. M., Cole, P. E., Hilbers, C. W., and Shulman, R. G. (1974). The molecular mechanism of thermal unfolding of Escherichia coli formylmethionine transfer RNA. *Journal of Molecular Biology*, 87(1):63–88.

Danko, C. G., Hyland, S. L., Core, L. J., Martins, A. L., Waters, C. T., Lee, H. W., Cheung, V. G., Kraus, W. L., Lis, J. T., and Siepel, A. (2015). Identification of active transcriptional regulatory elements from GRO-seq data. *Nature Methods*, 12(5):433–438.

Delli Ponti, R., Armaos, A., Marti, S., and Tartaglia, G. G. (2018). A method for RNA structure prediction shows evidence for structure in lncRNAs. *bioRxiv*.

Delli Ponti, R., Marti, S., Armaos, A., and Tartaglia, G. (2017). A high-throughput approach to profile RNA structure. *Nucleic Acids Research*, 45(5):e35–e35.

Deng, H., Cheema, J., Zhang, H., Woolfenden, H., Norris, M., Liu, Z., Liu, Q., Yang, X., Yang, M., Deng, X., Cao, X., and Ding, Y. (2018). Rice InVivo RNA Structurome Reveals RNA Secondary Structure Conservation and Divergence in Plants. *Molecular Plant*, 11(4):607–622.

Desai, R., Kilburn, D., Lee, H.-T., and Woodson, S. A. (2014). Increased ribozyme activity in crowded solutions. *The Journal of Biological Chemistry*, 289(5):2972–2977.

Ding, Y., Chan, C. Y., and Lawrence, C. E. (2004). Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Research*, 32(Web Server issue):W135–141.

Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., and Assmann, S. M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, 505(7485):696–700.

Do, C. B., Foo, C.-S., and Batzoglou, S. (2008). A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, 24(13):i68–i76.

Do, C. B., Woods, D. A., and Batzoglou, S. (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics (Oxford, England)*, 22(14):e90–98.

Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N. J., Van Nostrand, E. L., Pratt, G. A., Yeo, G. W., Graveley, B. R., and Burge, C. B. (2018). Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell*, 70(5):854–867.e9.

Doty, P., Boedtker, H., Fresco, J. R., Haselkorn, R., and Litt, M. (1959). Secondary structure in ribonucleic acids. *Proceedings of the National Academy of Sciences*, 45(4):482–499.

Duncan, C. D. S. and Weeks, K. M. (2008). SHAPE analysis of long-range interactions reveals extensive and thermodynamically preferred misfolding in a fragile group I intron RNA. *Biochemistry*, 47(33):8504–8513.

Dupuis, N. F., Holmstrom, E. D., and Nesbitt, D. J. (2014). Molecular-crowding effects on single-molecule RNA folding/unfolding thermodynamics and kinetics. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8464–8469.

Freisner, R. A. and Gunn, J. R. (1996). Computational Studies of Protein Folding. *Annual Review of Biophysics and Biomolecular Structure*, 25(1):315–342.

Fu, Y., Sharma, G., and Mathews, D. H. (2014). Dynalign II: common secondary structure prediction for RNA homologs with domain insertions. *Nucleic Acids Research*, 42(22):13939–13948.

Ganot, P., Caizergues-Ferrer, M., and Kiss, T. (1997). The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes & Development*, 11(7):941–956.

Giorgino, T. (2009). Computing and Visualizing Dynamic Time Warping Alignments in *R* : The **dtw** Package. *Journal of Statistical Software*, 31(7).

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R., and Hofacker, I. L. (2008). The Vienna RNA Websuite. *Nucleic Acids Research*, 36(Web Server):W70–W74.

Hajdin, C. E., Bellaousov, S., Huggins, W., Leonard, C. W., Mathews, D. H., and Weeks, K. M. (2013). Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proceedings of the National Academy of Sciences of the United States of America*, 110(14):5498–5503.

Hajiaghayi, M., Condon, A., and Hoos, H. H. (2012). Analysis of energy-based algorithms for RNA secondary structure prediction. *BMC Bioinformatics*, 13(1):22.

Halvorsen, M., Martin, J. S., Broadaway, S., and Laederach, A. (2010). Disease-Associated Mutations That Alter the RNA Structural Ensemble. *PLoS Genetics*, 6(8):e1001074.

Harmanci, A. O., Sharma, G., and Mathews, D. H. (2008). PARTS: probabilistic alignment for RNA joinT secondary structure prediction. *Nucleic Acids Research*, 36(7):2406–2417.

Harmanci, A. O., Sharma, G., and Mathews, D. H. (2009). Joint stochastic sampling for RNA secondary structure prediction. In *2009 IEEE International Workshop on Genomic Signal Processing and Statistics*, pages 1–4, Minneapolis, MN, USA. IEEE.

Harmanci, A. O., Sharma, G., and Mathews, D. H. (2011). TurboFold: Iterative probabilistic estimation of secondary structures for multiple RNA sequences. *BMC Bioinformatics*, 12(1):108.

Heilman-Miller, S. L. and Woodson, S. A. (2003). Effect of transcription on folding of the Tetrahymena ribozyme. *RNA (New York, N.Y.)*, 9(6):722–733.

Herschlag, D. (1995). RNA Chaperones and the RNA Folding Problem. *Journal of Biological Chemistry*, 270(36):20871–20874.

Hilbers, C. W., Robillard, G. T., Shulamn, R. G., Blake, R. D., Webb, P. K., Fresco, R., and Riesner, D. (1976). Thermal unfolding of yeast glycine transfer RNA. *Biochemistry*, 15(9):1874–1882.

Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J. L., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T. M., Severin, J., Lizio, M., Kawaji, H., Kasukawa, T., Itoh, M., Burroughs, A. M., Noma, S., Djebali, S., Alam, T., Medvedeva, Y. A., Testa, A. C., Lipovich, L., Yip, C.-W., Abugessaisa, I., Mendez, M., Hasegawa, A., Tang, D., Lassmann, T., Heutink, P., Babina, M., Wells, C. A., Kojima, S., Nakamura, Y., Suzuki, H., Daub, C. O., de Hoon, M. J. L., Arner, E., Hayashizaki, Y., Carninci, P., and Forrest, A. R. R. (2017). An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature*, 543(7644):199–204.

Ilyinskii, P. O., Schmidt, T., Lukashev, D., Meriin, A. B., Thoidis, G., Frishman, D., and Shneider, A. M. (2009). Importance of mRNA secondary structural elements for the expression of influenza virus genes. *Omics: A Journal of Integrative Biology*, 13(5):421–430.

Incarnato, D., Morandi, E., Anselmi, F., Simon, L. M., Basile, G., and Oliviero, S. (2017). In vivo probing of nascent RNA structures reveals principles of cotranscriptional folding. *Nucleic Acids Research*, 45(16):9716–9725.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)*, 324(5924):218–223.

Jabbari, H., Wark, I., Montemagno, C., and Will, S. (2018). Knotty: efficient and accurate prediction of complex RNA pseudoknot structures. *Bioinformatics*.

Jackson, S. A., Koduvayur, S., and Woodson, S. A. (2006). Self-splicing of a group I intron reveals partitioning of native and misfolded RNA populations in yeast. *RNA (New York, N.Y.)*, 12(12):2149–2159.

Janssen, S. and Giegerich, R. (2015). The RNA shapes studio. *Bioinformatics (Oxford, England)*, 31(3):423–425.

Kaushik, K., Sivadas, A., Vellarikkal, S. K., Verma, A., Jayarajan, R., Pandey, S., Sethi, T., Maiti, S., Scaria, V., and Sivasubbu, S. (2018). RNA secondary structure profiling in zebrafish reveals unique regulatory features. *BMC genomics*, 19(1):147.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107.

Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. (1982). Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, 31(1):147–157.

Kutchko, K. M., Sanders, W., Ziehr, B., Phillips, G., Solem, A., Halvorsen, M., Weeks, K. M., Moorman, N., and Laederach, A. (2015). Multiple conformations are a conserved and regulatory feature of the *RB1* 5 UTR. *RNA*, 21(7):1274–1285.

Kwok, C. K., Tang, Y., Assmann, S. M., and Bevilacqua, P. C. (2015). The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends in Biochemical Sciences*, 40(4):221–232.

Latham, M. P., Brown, D. J., McCallum, S. A., and Pardi, A. (2005). NMR Methods for Studying the Structure and Dynamics of RNA. *ChemBioChem*, 6(9):1492–1505.

Lavender, C. A., Gorelick, R. J., and Weeks, K. M. (2015a). Structure-Based Alignment and Consensus Secondary Structures for Three HIV-Related RNA Genomes. *PLOS Computational Biology*, 11(5):e1004230.

Lavender, C. A., Lorenz, R., Zhang, G., Tamayo, R., Hofacker, I. L., and Weeks, K. M. (2015b). Model-Free RNA Sequence and Structure Alignment Informed by SHAPE Probing Reveals a Conserved Alternate Secondary Structure for 16s rRNA. *PLOS Computational Biology*, 11(5):e1004126.

Leamy, K. A., Assmann, S. M., Mathews, D. H., and Bevilacqua, P. C. (2016). Bridging the gap between in vitro and in vivo RNA folding. *Quarterly Reviews of Biophysics*, 49.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lee, B., Flynn, R. A., Kadina, A., Guo, J. K., Kool, E. T., and Chang, H. Y. (2017). Comparison of SHAPE reagents for mapping RNA structures inside living cells. *RNA (New York, N.Y.)*, 23(2):169–174.

Lehnert, V., Jaeger, L., Michel, F., and Westhof, E. (1996). New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3d model of the Tetrahymena thermophila ribozyme. *Chemistry & Biology*, 3(12):993–1009.

Leontis, N. B. and Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. *RNA (New York, N.Y.)*, 7(4):499–512.

Levitt, M., Gerstein, M., Huang, E., Subbiah, S., and Tsai, J. (1997). Protein folding: the endgame. *Annual Review of Biochemistry*, 66:549–579.

Li, F., Zheng, Q., Ryvkin, P., Dragomir, I., Desai, Y., Aiyer, S., Valladares, O., Yang, J., Bambina, S., Sabin, L. R., Murray, J. I., Lamitina, T., Raj, A., Cherry, S., Wang, L.-S., and Gregory, B. D. (2012). Global analysis of RNA secondary structure in two metazoans. *Cell Reports*, 1(1):69–82.

Livi, C. M., Klus, P., Delli Ponti, R., and Tartaglia, G. G. (2016). *cat* RAPID *signature* : identification of ribonucleoproteins and RNA-binding regions. *Bioinformatics*, 32(5):773–775.

London, R. E. (1991). Methods for measurement of intracellular magnesium: NMR and fluorescence. *Annual Review of Physiology*, 53:241–258.

Lorenz, R., Luntzer, D., Hofacker, I. L., Stadler, P. F., and Wolfinger, M. T. (2016). SHAPE directed RNA folding. *Bioinformatics (Oxford, England)*, 32(1):145–147.

Loughrey, D., Watters, K. E., Settle, A. H., and Lucks, J. B. (2014). SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Research*, 42(21):e165–e165.

Low, J. T. and Weeks, K. M. (2010). SHAPE-directed RNA secondary structure prediction. *Methods (San Diego, Calif.)*, 52(2):150–158.

Lu, D., Searles, M. A., and Klug, A. (2003). Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition. *Nature*, 426(6962):96–100.

Lu, K., Heng, X., and Summers, M. F. (2011). Structural Determinants and Mechanism of HIV-1 Genome Packaging. *Journal of Molecular Biology*, 410(4):609–633.

Lu, Z. J., Gloor, J. W., and Mathews, D. H. (2009). Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, 15(10):1805–1813.

Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., and Arkin, A. P. (2011). Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences of the United States of America*, 108(27):11063–11068.

Maharana, S., Wang, J., Papadopoulos, D. K., Richter, D., Pozniakovsky, A., Poser, I., Bickle, M., Rizk, S., Guilln-Boixet, J., Franzmann, T. M., Jahnel, M., Marrone, L., Chang, Y.-T., Sterneckert, J., Tomancak, P., Hyman, A. A., and Alberti, S. (2018). RNA buffers the phase separation behavior of prion-like RNA binding proteins. *Science*, 360(6391):918–921.

Mahen, E. M., Watson, P. Y., Cottrell, J. W., and Fedor, M. J. (2010). mRNA Secondary Structures Fold Sequentially But Exchange Rapidly In Vivo. *PLoS Biology*, 8(2):e1000307.

Martin, F. H., Uhlenbeck, O. C., and Doty, P. (1971). Self-complementary oligoribonucleotides: adenylic acid-uridylic acid block copolymers. *Journal of Molecular Biology*, 57(2):201–215.

Masliah, G., Barraud, P., and Allain, F. H.-T. (2013). RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cellular and molecular life sciences: CMLS*, 70(11):1875–1895.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7287–7292.

Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940.

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

Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and Weeks, K. M. (2005). RNA structure analysis at single nucleotide resolution by selective 2′-hydroxyl acylation and primer extension (SHAPE). *Journal of the American Chemical Society*, 127(12):4223–4231.

Minton, A. P. (2001). The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media. *The Journal of Biological Chemistry*, 276(14):10577–10580.

Mitchell, D., Jarmoskaite, I., Seval, N., Seifert, S., and Russell, R. (2013). The long-range P3 helix of the Tetrahymena ribozyme is disrupted during folding between the native and misfolded conformations. *Journal of Molecular Biology*, 425(15):2670–2686.

Mlýnský, V. and Bussi, G. (2018). Molecular Dynamics Simulations Reveal an Interplay between SHAPE Reagent Binding and RNA Flexibility. *The Journal of Physical Chemistry Letters*, 9(2):313–318.

Moazed, D., Robertson, J. M., and Noller, H. F. (1988). Interaction of elongation factors EF-G and EF-Tu with a conserved loop in 23s RNA. *Nature*, 334(6180):362–364.

Mort, M., Sterne-Weiler, T., Li, B., Ball, E. V., Cooper, D. N., Radivojac, P., Sanford, J. R., and Mooney, S. D. (2014). MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biology*, 15(1):R19.

Mortimer, S. A., Kidwell, M. A., and Doudna, J. A. (2014). Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics*, 15(7):469–479.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1).

Nakano, S.-i., Miyoshi, D., and Sugimoto, N. (2014). Effects of molecular crowding on the structures, interactions, and functions of nucleic acids. *Chemical Reviews*, 114(5):2733–2758.

Novikova, I. V., Hennelly, S. P., and Sanbonmatsu, K. Y. (2012). Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Research*, 40(11):5034–5051.

Nussinov, R. and Jacobson, A. B. (1980). Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 77(11):6309–6313.

Ouyang, Z., Snyder, M. P., and Chang, H. Y. (2013). SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Research*, 23(2):377–387.

Pan, T., Fang, X., and Sosnick, T. (1999). Pathway modulation, circular permutation and rapid RNA folding under kinetic control. *Journal of Molecular Biology*, 286(3):721–731.

Paudel, B. P. and Rueda, D. (2014). Molecular crowding accelerates ribozyme docking and catalysis. *Journal of the American Chemical Society*, 136(48):16700–16703.

Peattie, D. A. and Gilbert, W. (1980). Chemical probes for higher-order structure in RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 77(8):4679–4682.

Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. (2006). Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Computational Biology*, 2(4):e33.

Petrov, A. S., Gulen, B., Norris, A. M., Kovacs, N. A., Bernier, C. R., Lanier, K. A., Fox, G. E., Harvey, S. C., Wartell, R. M., Hud, N. V., and Williams, L. D. (2015). History of the ribosome and the origin of translation. *Proceedings of the National Academy of Sciences*, 112(50):15396–15401.

Pollom, E., Dang, K. K., Potter, E. L., Gorelick, R. J., Burch, C. L., Weeks, K. M., and Swanstrom, R. (2013). Comparison of SIV and HIV-1 Genomic RNA Structures Reveals Impact of Sequence Evolution on Conserved and Non-Conserved Structural Motifs. *PLoS Pathogens*, 9(4):e1003294.

Puton, T., Kozlowski, L. P., Rother, K. M., and Bujnicki, J. M. (2013). CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Research*, 41(7):4307–4323.

Rausch, J. W., Sztuba-Solinska, J., and Le Grice, S. F. J. (2017). Probing the Structures of Viral RNA Regulatory Elements with SHAPE and Related Methodologies. *Frontiers in Microbiology*, 8:2634.

Reeder, J. and Giegerich, R. (2005). Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics (Oxford, England)*, 21(17):3516–3523.

Reuter, J. S. and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11(1):129.

Rice, G. M., Leonard, C. W., and Weeks, K. M. (2014). RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA (New York, N.Y.)*, 20(6):846–854.

Rivas, E., Clements, J., and Eddy, S. R. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature Methods*, 14(1):45–48.

Rivas, E. and Eddy, S. R. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots 1 1edited by I. Tinoco. *Journal of Molecular Biology*, 285(5):2053–2068.

Robertson, M. P. and Joyce, G. F. (2012). The Origins of the RNA World. *Cold Spring Harbor Perspectives in Biology*, 4(5):a003608–a003608.

Rook, M. S., Treiber, D. K., and Williamson, J. R. (1998). Fast folding mutants of the Tetrahymena group I ribozyme reveal a rugged folding energy landscape. *Journal of Molecular Biology*, 281(4):609–620.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J. S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, 505(7485):701–705.

Sato, K., Hamada, M., Asai, K., and Mituyama, T. (2009). CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Research*, 37(Web Server issue):W277–280.

Saus, E., Willis, J. R., Pryszcz, L. P., Hafez, A., Llorens, C., Himmelbauer, H., and Gabaldn, T. (2018). nextPARS: parallel probing of RNA structures in Illumina. *RNA*, 24(4):609–619.

Schowen, R. L. (1993). Principles of biochemistry 2nd ed. (Lehninger, Albert L.; Nelson, David L.; Cox, Michael M.). *Journal of Chemical Education*, 70(8):A223.

Schroeder, R., Barta, A., and Semrad, K. (2004). Strategies for RNA folding and assembly. *Nature Reviews. Molecular Cell Biology*, 5(11):908–919.

Seetin, M. G., Kladwang, W., Bida, J. P., and Das, R. (2014). Massively Parallel RNA Chemical Mapping with a Reduced Bias MAP-Seq Protocol. In *RNA Folding*, volume 1086, pages 95–117. Humana Press, Totowa, NJ.

Shepard, P. J. and Hertel, K. J. (2008). Conserved RNA secondary structures promote alternative splicing. *RNA (New York, N.Y.)*, 14(8):1463–1469.

Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E., and Weeks, K. M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods*, 11(9):959–965.

Smola, M. J., Christy, T. W., Inoue, K., Nicholson, C. O., Friedersdorf, M., Keene, J. D., Lee, D. M., Calabrese, J. M., and Weeks, K. M. (2016). SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proceedings of the National Academy of Sciences of the United States of America*, 113(37):10322–10327.

Smola, M. J., Rice, G. M., Busan, S., Siegfried, N. A., and Weeks, K. M. (2015). Selective 2-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nature Protocols*, 10(11):1643–1669.

Somarowthu, S., Legiewicz, M., Chilln, I., Marcia, M., Liu, F., and Pyle, A. M. (2015). HOTAIR forms an intricate and modular secondary structure. *Molecular Cell*, 58(2):353–361.

Spitale, R. C., Flynn, R. A., Torre, E. A., Kool, E. T., and Chang, H. Y. (2014). RNA structural analysis by evolving SHAPE chemistry. *Wiley interdisciplinary reviews. RNA*, 5(6):867–881.

Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H. Y., Batista, P. J., Torre, E. A., Kool, E. T., and Chang, H. Y. (2015). Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, 519(7544):486–490.

Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., and Giegerich, R. (2006). RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics (Oxford, England)*, 22(4):500–503.

Stein, A. and Crothers, D. M. (1976). Conformational changes of transfer RNA. The role of magnesium(II). *Biochemistry*, 15(1):160–168.

Stormo, G. D. (2006). An Overview of RNA Structure Prediction and Applications to RNA Gene Prediction and RNAi Design. *Current Protocols in Bioinformatics*, 13(1):12.1.1–12.1.3.

Strobel, E. J., Yu, A. M., and Lucks, J. B. (2018). High-throughput determination of RNA structures. *Nature Reviews. Genetics*.

Strulson, C. A., Yennawar, N. H., Rambo, R. P., and Bevilacqua, P. C. (2013). Molecular crowding favors reactivity of a human ribozyme under physiological ionic conditions. *Biochemistry*, 52(46):8187–8197.

Swenson, M. S., Anderson, J., Ash, A., Gaurav, P., Sksd, Z., Bader, D. A., Harvey, S. C., and Heitsch, C. E. (2012). GTfold: enabling parallel RNA secondary structure prediction on multi-core desktops. *BMC research notes*, 5:341.

Tavares, R. C. A., Pyle, A. M., and Somarowthu, S. (2018). Covariation analysis with improved parameters reveals conservation in lncRNA structures. *Bioxiv*.

Tinoco, I. (1993). APPENDIX 1: Structures of Base Pairs Involving at Least Two Hydrogen Bonds. In *The RNA world*, volume 24, pages 603–607. Cold Spring Harbor Laboratory Press.

Tinoco, I. and Bustamante, C. (1999). How RNA folds. *Journal of Molecular Biology*, 293(2):271–281.

Tompa, P. and Csermely, P. (2004). The role of structural disorder in the function of RNA and protein chaperones. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 18(11):1169–1175.

Torarinsson, E., Havgaard, J. H., and Gorodkin, J. (2007). Multiple structural alignment and clustering of RNA sequences. *Bioinformatics (Oxford, England)*, 23(8):926–932.

Treiber, D. K., Rook, M. S., Zarrinkar, P. P., and Williamson, J. R. (1998). Kinetic intermediates trapped by native interactions in RNA folding. *Science (New York, N.Y.)*, 279(5358):1943–1946.

Truong, D. M., Sidote, D. J., Russell, R., and Lambowitz, A. M. (2013). Enhanced group II intron retrohoming in magnesium-deficient Escherichia coli via selection of mutations in the ribozyme core. *Proceedings of the National Academy of Sciences of the United States of America*, 110(40):E3800–3809.

Turner, D. H. and Mathews, D. H. (2010). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 38(Database issue):D280–282.

Tyrrell, J., Weeks, K. M., and Pielak, G. J. (2015). Challenge of mimicking the influences of the cellular environment on RNA structure by PEG-induced macromolecular crowding. *Biochemistry*, 54(42):6447–6453.

Underwood, J. G., Uzilov, A. V., Katzman, S., Onodera, C. S., Mainzer, J. E., Mathews, D. H., Lowe, T. M., Salama, S. R., and Haussler, D. (2010). FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature Methods*, 7(12):995–1001.

Van Treeck, B. and Parker, R. (2018). Emerging Roles for Intermolecular RNA-RNA Interactions in RNP Assemblies. *Cell*, 174(4):791–802.

Vandivier, L. E., Anderson, S. J., Foley, S. W., and Gregory, B. D. (2016). The Conservation and Function of RNA Secondary Structure in Plants. *Annual Review of Plant Biology*, 67(1):463–488.

Vendruscolo, M., Paci, E., Karplus, M., and Dobson, C. M. (2003). Structures and relative free energies of partially folded states of proteins. *Proceedings of the National Academy of Sciences*, 100(25):14817–14821.

Volders, P.-J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., Vandesompele, J., and Mestdagh, P. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Research*, 41(Database issue):D246–251.

Wan, Y., Qu, K., Ouyang, Z., Kertesz, M., Li, J., Tibshirani, R., Makino, D. L., Nutter, R. C., Segal, E., and Chang, H. Y. (2012). Genome-wide measurement of RNA folding energies. *Molecular Cell*, 48(2):169–181.

Wan, Y., Qu, K., Zhang, Q. C., Flynn, R. A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R. C., Snyder, M. P., Segal, E., and Chang, H. Y. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, 505(7485):706–709.

Washietl, S., Hofacker, I. L., Lukasser, M., Httenhofer, A., and Stadler, P. F. (2005). Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nature Biotechnology*, 23(11):1383–1390.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Wilkinson, K. A., Gorelick, R. J., Vasa, S. M., Guex, N., Rein, A., Mathews, D. H., Giddings, M. C., and Weeks, K. M. (2008). High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states. *PLoS biology*, 6(4):e96.

Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2006). Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols*, 1(3):1610–1616.

Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R. (2007). Inferring Noncoding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering. *PLoS Computational Biology*, 3(4):e65.

Wolynes, P. G., Onuchic, J. N., and Thirumalai, D. (1995). Navigating the folding routes. *Science (New York, N.Y.)*, 267(5204):1619–1620.

Wu, Y., Shi, B., Ding, X., Liu, T., Hu, X., Yip, K. Y., Yang, Z. R., Mathews, D. H., and Lu, Z. J. (2015). Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic Acids Research*, 43(15):7247–7259.

Wyatt, J. R. and Walker, G. T. (1989). Deoxynucleotide-containing oligoribonucleotide duplexes: stability and susceptibility to RNase V1 and RNase H. *Nucleic Acids Research*, 17(19):7833–7842.

Xayaphoummine, A., Bucher, T., and Isambert, H. (2005). Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Research*, 33(Web Server issue):W605–610.

Xia, T., SantaLucia, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C., and Turner, D. H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37(42):14719–14735.

Xu, Z. and Mathews, D. H. (2011). Multilign: an algorithm to predict secondary structures conserved in multiple RNA sequences. *Bioinformatics (Oxford, England)*, 27(5):626–632.

Zemora, G. and Waldsich, C. (2010). RNA folding in living cells. *RNA biology*, 7(6):634–641.

Zhang, K. and Frank, A. (2018). RNA Secondary Structure Prediction Guided by Chemical Shifts. *Biophysical Journal*, 114(3):678a.

Zhou, H.-X. (2004). Protein folding and binding in confined spaces and in crowded solutions. *Journal of molecular recognition: JMR*, 17(5):368–375.

Zhou, H.-X., Rivas, G., and Minton, A. P. (2008). Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annual Review of Biophysics*, 37:375–397.

Zimmerman, S. B. and Trach, S. O. (1991). Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. *Journal of Molecular Biology*, 222(3):599–620.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415.

Zuker, M. and Sankoff, D. (1984). RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621.