UNIVERSITÉ DE STRASBOURG

Thèse présentée pour obtenir le grade
de Docteur de l'Université de Strasbourg

Dicipline: Aspects Moléculaires et Cellulaires de la Biologie

Par Denise Martinez Zapien

# *Approche structurale du rôle de l'ARN 7SK comme régulateur de la transcription eucaryote*

Soutenue publiquement le 22 Septembre 2011

Membres du Jury

| | |
|---|---|
| Dr. Olivier BENSAUDE | Rapporteur externe |
| Dr. Christiane BRANLANT | Rapporteur externe |
| Dr. Pascale ROMBY | Rapporteur interne |
| Dr. Hervé LE HIR | Examinateur |
| Dr. Anne-Catherine DOCK-BREGEON | Directeur de Thése |

# INDEX

# PREFACE

The study of the function of a macromolecule is related to the study of their structure. The function of the macromolecules depends on accurate recognition with other molecules and on their response upon this interaction. Some years ago the proteins were considered as the main functional macromolecules of the cells, given their physicochemical diversity, and nucleic acids as responsible for the storage and transmission of the genetic information. This point of view has been challenged with the discovery of RNAs participating in many cellular processes. The ability of RNA to fold into complex structures has been increasingly recognized as a source of diversity of RNA function. Contributing to this diversity is the ability of RNA to form complementary base pairs with other RNAs and with single-stranded DNA, and to interact with proteins as part of RNPs.

This manuscript describes the different studies aimed to the identification of the structural determinants in the interaction between the 7SK RNA and the HEXIM1 protein, two components of the 7SK snRNP that cooperate to inhibit the eukaryotic transcription by the RNA Polymerase II. It also intends to provide some insights into the structure of 7SK.

The Chapter I is an introduction to the eukaryotic transcription and its regulation by RNAs, highlighting the role of the 7SK snRNP.

The Chapter II is dedicated to a more detailed description of HEXIM1 to then address the characterization of the 7SK and HEXIM1 interaction by biochemical a biophysical methods in Chapters IV and V. The Chapter III describes the preparation of RNA and protein constructions, since the design and production of the target molecules with a high yield and purity is an essential part of the strategy for structural studies.

The Chapter VI presents some preliminary results of the study of the interaction of 7SK with an hnRNP, which were recently identified as major 7SK associated proteins.

The Chapters VII, VIII and IX are devoted to the structural study of 7SK snRNA. In the Chapter VII, a study of the secondary structure of 7SK using the recently developed SHAPE method is presented. Chapter VIII explains the strategy used for the characterization of the solution structure of 7SK by SAXS, emphasizing the confronted limitations and including some preliminary results. The Chapter IX describes the crystallization of a functional subdomain of 7SK and the problems encountered in our attempt to solve the structure.

Finally, the material and methods are given in the annexes.

# ACKNOWLEDGEMENTS

This work was conducted at the Laboratory of Expression of Genetic Information of the Integrated Structural Biology Department at the Institute de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), Illkirch, France, with a fellowship of the Consejo Nacional de Ciencia y Tecnologia (CONACyT), Mexico, and the Association pour la Recherche sur le Cancer (ARC), France.

I am heartily thankful to my supervisor, Anne-Catherine Dock-Bregeon, for her encouragement and strong support, for her confidence in my research, and for her invaluable advice and assistance during my work and the preparation of this manuscript. Her passion for the science inspired me and motivated me throughout my work.

I want to express my gratitude to Prof. Dino Moras for acceptance to his laboratory.

I gratefully thank the members of the jury, Dr. Olivier Bensaude, Dr. Pascale Romby, Dr. Christiane Branlant and Dr. Hervé Le Hir for giving me the honor of judging my work.

It is a pleasure to thank all our collaborators Bruno Kieffer and Isabelle Lebars (NMR), Sarah Sanglier, Jean-Michel Saliou and Cedric Atmanene (MS), Dimitri Svergun and Michal Gadja (SAXS), Frabrice Jossinet (HP1 model), Bruno Kaholz and Jean François Menetret (EM), and Yves Mély and Julien Godet (Fluorescence).

I want to thank my lab partners Emiko Uchikawa for the fruitful discussions and motivation and Alexandre Durand for his help in protein purification and in the optimization of several protein purification protocols. I want also to thank Meiggie Untrau, Nadège Muller and all the students who participated in the cloning of several constructions presented in this work.

I would also aknowledge Adam Ben Shem for his encouragement and invaluable discussions and advices for crystallization.

I want to thank the Structural Biology and Genomics technology platform, in particular Pierre Poussin for his excellent technical support and his advices for crystallization.

I am deeply grateful to Alastair McEwen for his enormous assistance for data collection and analysis of crystals and for his enthusiasm.

I would like to thank at each of the members of the Integrated Structural Biology Department, in particular to its director Patrick Schultz, and people with whom I shared the laboratory all these years and who contributed in many assorted ways to this thesis: Nada, Aline, Martin, Stephanie, Emiko, Emeline, Wassim, Serena, Pierre, Marie-Laure, Alexandre,

Justine, Laura, Massimo, Corinna, Angelita, Mari, Judith, Cédric, Heena, Vidhya, Sankar, Martin, Edouard, Loubna, Tiphaine, Benoit, Rita, Natacha, Maria, Pierre, Yann, Nicolas, Sergey, Julie, Didier Busso, Catherine Birck, James Stevenin, Natacha, Valerie, Jean-Claude Thierry...

Quiero agradecer también a mis amigos de México y de Francia porque han contribuido de diversas maneras a la realizacion de esta tesis.

Je tiens à remercier Luc pour d'avoir rendu ma vie plus joyeuse et interessante, pour sa confiance et son soutien qui m'ont été precieux. N'importe quel mot serait insufissant pour te remercier. Je veux aussi remercier à sa famille pour son accueil et pour tous ces dimanches partagés.

Finalmente, quiero expresar mi profundo agradecimiento a mis papàs y a mi hermana, porque sin su apoyo esto no seria posible, porque siempre han alentado mis ambiciones a pesar de lo que ello implica y por su confianza.

# CHAPTER I : INTRODUCTION

Until the middle of the 70s, the only known RNA molecules were the functional messenger (mRNA), the transfer (tRNA) and the ribosomal (rRNA). The mRNA was considered as a coding molecule for translating the genetic information contained in the DNA into proteins, while the tRNA and the rRNA, fulfilled generic roles during this translation. A landmark in the RNA biology was the discovery in 1982 of the ribozymes, demonstrating the RNA capacity to catalyze specific biochemical reactions. This prompted scientists to consider a «RNA world» and led to investigate other roles for RNAs. Today we know that the RNA plays a variety of structural, informational, catalytic and regulatory roles in the cells. Surprisingly, genome-wide analyses have shown that less than 2% of the human genome is translated into protein, yet more than 40% is thought to be transcribed into RNA (Matera et al. 2007; Mercer et al. 2009). This observation highlights the widespread roles and cellular processes in which RNAs would be involved, and raises the question of what are the functions of the cell delegated to RNAs instead to proteins, and why.

Recently, several studies have revealed that non-coding RNAs (ncRNAs), actively regulate mRNA transcription (Goodrich et al. 2006), which is a key point for the control of the gene expression since changes in the appropriate pattern of gene expression has profound effects on cellular function and underlies many diseases. The next chapter is an introduction to Transcription in Eukaryotes, then two textbook examples of ncRNAs in the regulation of transcription will be described, in order to introduce the main topic of this manuscripts, the role of the 7SK snRNP in the regulation of the transcription elongation.

# 1. The Eukaryotic Transcription

In all organisms, the transcription is carried out by DNA-dependent RNA polymerases (RNAP). While Bacteria and Archaea have only one RNAP, three nuclear RNAPs operate in all eukaryotes examined so far. RNAPI is specialized in the synthesis of the abundant pre-rRNA precursor of the large ribosomal rRNAs, RNAPII produces all mRNAs and some small nuclear RNAs (snRNA), and RNAPIII transcribes all tRNAs, the small ribosomal 5S rRNA, and an eclectic collection of genes whose main common features are that they encode structural or catalytic RNAs, generally shorter than 400 nucleotides (Werner et al. 2009).

## 1.1. Transcription by RNAPII

### a. RNAP II structural organization

Transcription of eukaryotic genes is a complex process requiring the action of a myriad of proteins. Central to the process is the RNAPII. The RNAPII is a multisubunit enzyme with 12 to 15 subunits, depending on the organism. The best characterized form of the enzyme, from *Saccharomyces cerevisiae*, comprises 12 different polypeptides (Cramer 2000; Cramer et al. 2001; Figure I.1). These subunits can be classified into three overlapping categories: subunits of the core domain having homologous counterparts in bacterial RNAP (Rpb1, 2, 3, and 11), subunits shared between all three nuclear polymerases (Rpb5, 6, 8, 10, and 12), and subunits specific to RNAPII but not essential for transcription elongation (Rpb4, 7, and 9).

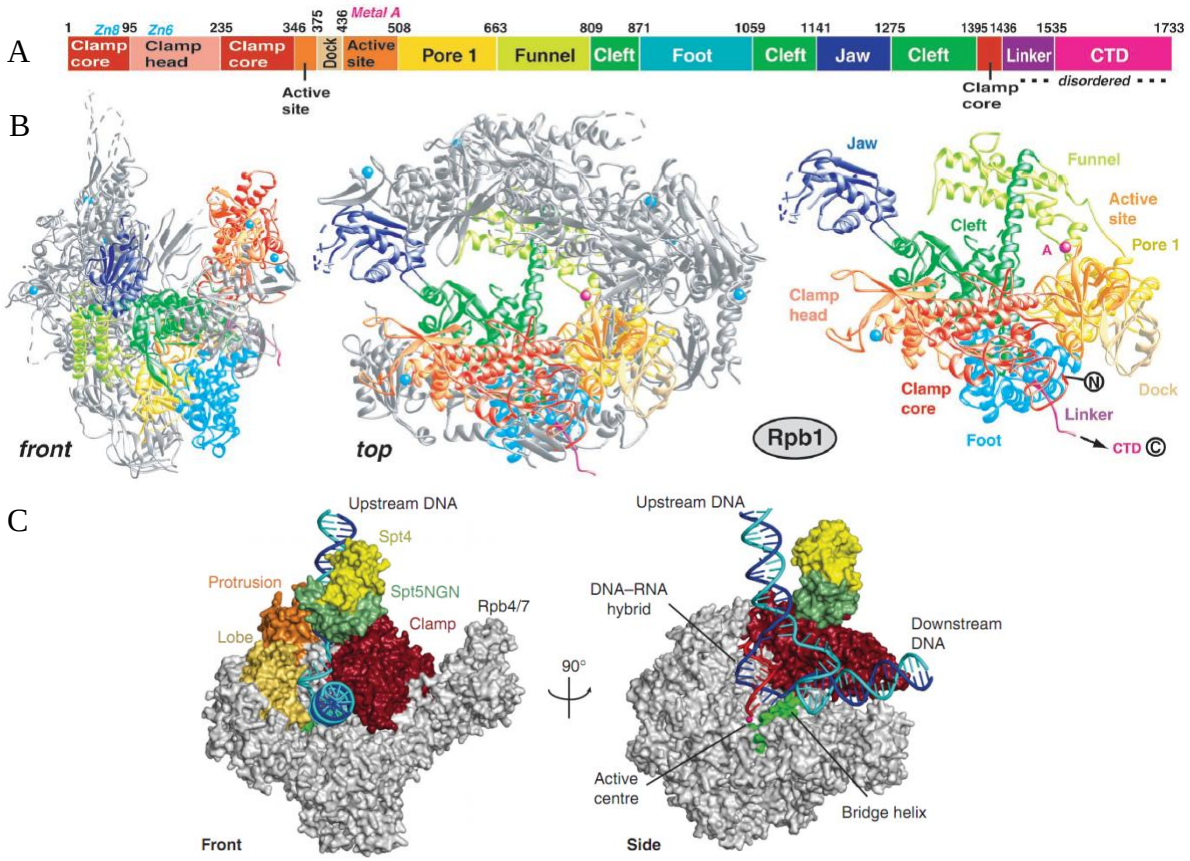



Figure I.1. Structural organization of RNAP II. A) Domains of the Rpb1. B) Front and top views of the RNAPII. The isolated Rpb1 is also shown and is colored as in A. C) Model of the complete yeast RNAPII elongation complex with bound Spt4/5. DNA template, DNA non-template, and RNA are in blue, cyan and red, respectively [from (Cramer et al. 2001) and (Martinez-Rucobo et al. 2011)].

### b. The "CTD"

A striking feature that distinguishes RNAPII from the other two eukaryotic RNAPs is the extended carboxy-terminal domain (CTD) of the largest subunit, Rbp1. The RNAPII CTD consists of heptapeptide repeats of the consensus sequence YSPTSPS. The CTD is thought to extend from the core of the polymerase, and is subject to modifications throughout the transcription cycle (Buratowski 2009). Modification of the CTD markedly affects its conformation and its ability to associate with factors that are involved in transcription initiation and elongation, RNA processing and termination (Meinhart et al. 2005; Phatnani et

al. 2006). Therefore, modification of the CTD is important for the coordination of transcription events, and different modification states of the CTD are characteristic of different transcriptional stages (Egloff et al. 2008; Figure I.2). Indeed, RNAPII changes from a hypophosphorylated (RNAPIIa) to a hyperphosphorylated (RNAPIIo) form during the transition from transcription initiation to transcription elongation. These phosphorylations occur in the CTD at Ser2 and Ser5. The level of Ser5 phosphorylation peaks early in the transcription cycle and remains constant or decreases towards the 3' end of the gene (Saunders et al. 2006). In contrast, Ser2 phosphorylation predominates during the transcription of the body and towards the 3' end of the gene and occurs concomitantly with productive elongation.

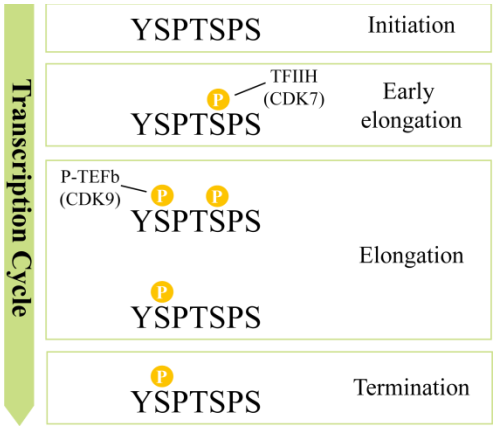



Figure I.2. The phosphorylation state of the CTD changes during the transcription. The phosphorylation positions at the heptad repeat of the CTD during the transcription cycle are indicated.

### c. The transcription cycle

RNAPII transcription cycle (Figure I.3) starts with the recognition of the core promoter, including the TATA box, the definition of the transcription start site (TSS), and the subsequent assembly of the preinitiation complex [PIC (Sikorski et al. 2009)]. In eukaryotic cells, DNA is wrapped around histone octamers to form nucleosomes, the primary unit of chromatin structure. Transcription requires the DNA to be accessible to sequence-specific transcription factors and to RNAPII, and requires the melting and reformation of the double helix throughout the length of the transcript (Li et al. 2007). Thus, the chromatin is a mechanistic player in transcription by creating either a stable, inaccessible or an accessible

chromatin structure. In this regard, chromatin-modifying enzymes have a key role in enhancing the access to the transcriptional machinery (Cairns 2009). In a process known as chromatin remodeling, covalent modifications of histones reduce protein-DNA interactions, or by using the energy of ATP hydrolysis, alter the histone-DNA contacts, thus leading to opening the chromatin structure surrounding the TATA box and other recognition sequences. These chromatin modifications allow the binding of sequence-specific transcriptional activators, typically composed of a DNA-binding domain (DBD) and an activation domain (AD), which can occur very far from the core promoter (Barberis et al. 2003). This event leads to the recruitment of the adaptor complexes such as SAGA or Mediator, which facilitate binding of general transcription factors (GTFs) at the promoter, to initiate transcription.

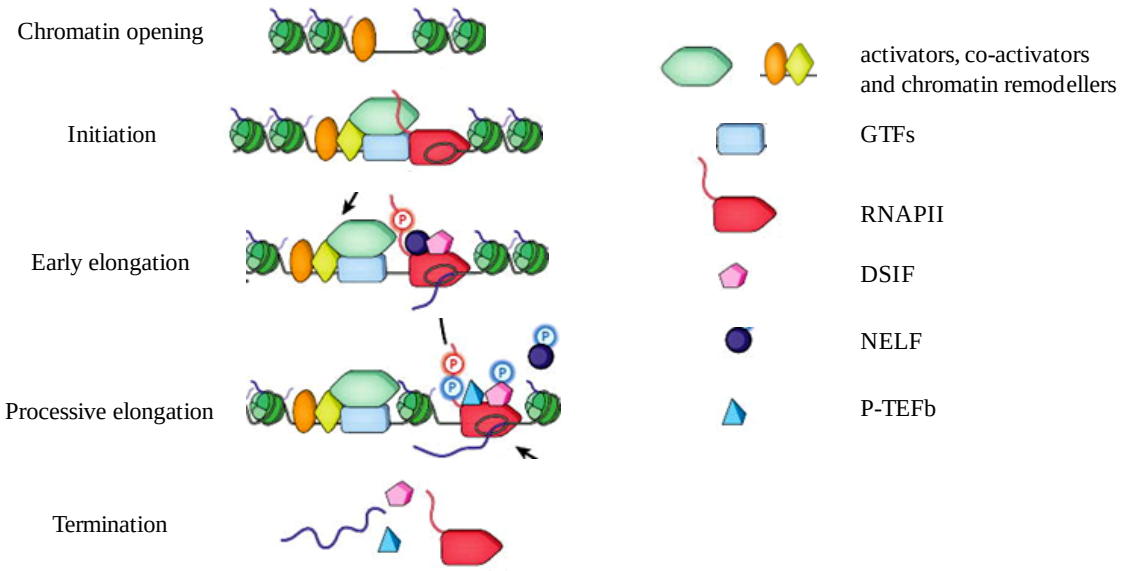



Figure I.3. The transcription cycle. Scheme showing the major steps of the eukaryotic transcription: initiation, elongation and termination. Adapted from (Fuda et al. 2009).

In the current model of RNAPII transcription, the first GTF to bind to the core promoter is TFIID (Orphanides et al. 1996; M. C. Thomas et al. 2006). TFIID recognizes the TATA element via the TATA-binding protein (TBP) subunit, and facilitates the positioning of the RNAPII at the core promoter. TFIIA and TFIIB also participate, by increasing the stability of TBP binding, and forming in this way the closed form of the PIC. Promoter loading onto RNAPII requires TFIIF, which forms a tight complex with the polymerase. TFIIB and TFIIF drive the binding of TFIIE that recruits TFIIH. TFIIH, which contains two helicases, then

melts 11 to 15 base pairs of DNA (transcription bubble) in order to position the single strand template in the RNAPII cleft (open complex) to initiate RNA synthesis. Following the establishment of the open complex, TFIIB is displaced as the nascent RNA extends beyond 4 nucleotides in length. The upstream region portion of transcription bubble collapses upon the formation of a stable RNA-DNA hybrid (8 to 9 base pairs), releasing accumulated energy that may help drive the promoter clearance. The promoter clearance coincides with the phosphorylation of Ser5 of the CTD, by the CAK kinase complex of TFIIH. This may destabilize contacts between the RNAPII and the GTFs. Then the RNAPII proceeds onto the elongation stage. The transcription elongation is the process by which the RNAPII moves through the coding region of the gene and incorporates nucleotides into the growing mRNA by phosphodiester bond formation in a template-directed manner (Armache et al. 2005). Finally, the termination of transcription involves release of the RNA transcript and the dissociation of the transcription complex from the DNA template.

The recruitment of the RNAPII to its promoter and its engagement into the elongation stage is not sufficient for an effective transcription, as previously believed, because RNAPII often pauses shortly after the TSS.

## 1.2. The Elongation Pausing

The transition from initiation to elongation is accompanied by changes in the phosphorylation state of the CTD. First, CTD is phosphorylated at Ser5 by the Cyclin-dependent Kinase 7 (CDK7) of TFIIH. This facilitates the binding of several factors involved in early elongation and modification of promoter-proximal histones (Phatnani et al. 2006). When the nascent transcript reaches about 20 nucleotides, its 5'end is modified through the addition of the 7-methyl-guanosine cap, which is critical for RNA stability, further RNA processing, export from nucleus and protein translation (Nechaev and Adelman 2010).

Then, RNAPII synthesizes 25 to 50 ribonucleotides before pausing, which corresponds to an abortive elongation (Nechaev and Adelman 2010). RNAPII pausing was first characterized on Drosophila *hsp* (heat-shock protein) genes (Gilmour et al. 1986). Indeed, the presence of RNAPII was observed in the promoter regions of uninduced *hsp70* genes, indicating the recruitment of RNAPII before transcription activation. The spontaneous release

of paused elongation complex was extremely slow, but became markedly faster upon heat shock, indicating that transcription output can be dramatically altered by regulating the efficiency of early elongation (Nechaev et al. 2008). This induction phenomenon was later described on several mammalian genes such as c-myc, c-fos and junB. With the advent of genome-wide approaches, it became evident that the pausing of the RNAPII is a widespread phenomenon (Nechaev and Adelman 2010).

The establishment of the paused RNAPII involves the coordinated actions of the transcription elongation factors DSIF [DRB (5,6-dichloro-1-4-D-ribofuranosyl benzimidazol) Sensitivity-Inducing Factor] and NELF [Negative Elongation Factor (Levine 2011; Nechaev et al. 2008)]. Interestingly, DSIF and NELF inhibit the production of long mRNA but not of short abortive transcripts. The exact mechanisms of action of DSIF and NELF in promoting RNAPII arrest are not well understood. However, it has been observed that DSIF interacts directly with transcribing polymerase and with the nascent transcripts via its Spt5 subunit. Then DSIF recruits NELF, whose RNA activity is required for the RNAPII arrest.

It was suggested that DNA elements sequence such as DPE (downstream promoter element) or its related motif called PB (Pause Button) are also implicated in the RNAPII pausing (Gilchrist et al. 2010; Levine 2011). The PB is a GC-rich sequence motif present in many paused genes. It has been proposed that these sequences confer an energy barrier for the melting of the double helix, impeding the RNAPII to move forward and allowing the binding of the DSIF and NELF.

Recently, it has been also suggested that the nucleosomes level of occupancy is also involved in RNAPII pausing (Gilchrist et al. 2010; Espinosa 2010). Indeed, paused genes [more than one-third of the genes in Drosophila cells are stalled (Nechaev, Fargo, et al. 2010)] display much lower nucleosome occupancy downstream of TSSs. Moreover, the loss of NELF leads to a decrease of RNAPII pausing, but also to an increase in nucleosome occupancy at promoters of paused genes. Interestingly, it leads also to a downregulation of many paused genes. These observations suggest that the pausing contributes to the formation of nucleosome free regions required for eventual gene activation.

The RNAPII pause seems to be an important “checkpoint” to ensure that only properly matured elongation complexes proceed through the gene (Levine 2011). Furthermore, paused RNAPII may foster synchronous and homogeneous patterns of gene expression (Espinosa 2010).

## 1.3. P-TEFb and the release of the RNAPII pausing

The Positive Elongation Transcription Factor b, P-TEFb, is responsible for the specific stimulation of the processivity of the RNAPII and of the release of the inhibitory effects of DSIF and NELF factors (Zhou et al. 2006; Peterlin et al. 2006; Kohoutek 2009). P-TEFb phosphorylates the CTD at the Ser2 position, yielding a RNAPIIo. The Spt5 subunit of DSIF and the RD subunit of NELF are also phosphorylated by P-TEFb. These phosphorylations lead to the dissociation of NELF and conversion of DSIF into a positive elongation factor. These events allow RNAPII to shift into the productive phase of transcriptional elongation. In addition, Ser2 phosphorylation of CTD provides a platform for assembly of complexes that travel with the RNAPII into the gene, including factors that regulate transcription elongation, RNA processing and termination, as well as the modification and remodelling of histones.

## 1.4. P-TEFb and HIV

Importantly, P-TEFb is a specific, cellular cofactor for efficient transcriptional elongation of the human immunodeficiency virus type 1 (HIV-1). In fact, the understanding of the general mechanism of P-TEFb stimulation of the RNAPII elongation has benefited tremendously from studies of the role of P-TEFb in regulating HIV-1 transcription.

HIV-1 encodes a small regulatory protein, Tat, which is essential for activating transcriptional elongation from the HIV-1 Long Terminal Repeat to produce the full-length viral transcripts (Falco et al. 2002; Karn 1999). To stimulate the processive transcription, Tat recruits P-TEFb to the 5'end of the nascent viral transcript (nucleotides +1 to +59), which forms a stem-loop structure called the Trans-Acting Response (TAR) RNA element (Figure I.4). Tat specifically binds to the TAR RNA just below the apical loop, at the level of a three-nucleotide bulge. Several flanking nucleotides in the stem participate also to the binding. The cooperative interactions among Tat, TAR and P-TEFb result in the recruitment of P-TEFb to the vicinity of the paused RNAPII (Price 2000; Zhou et al. 2006).

In the absence of Tat, HIV-1 uses an alternative strategy for the first rounds of viral transcription, where NF-κB (nuclear factor κ-light-chain-enhancer of activated B cells)

recruits P-TEFb to the HIV Long Terminal Repeats (LTR), which allows early production of Tat (Peterlin et al. 2006). Then the recruitment of P-TEFb via Tat and TAR RNA takes over, leading to higher levels of Tat.

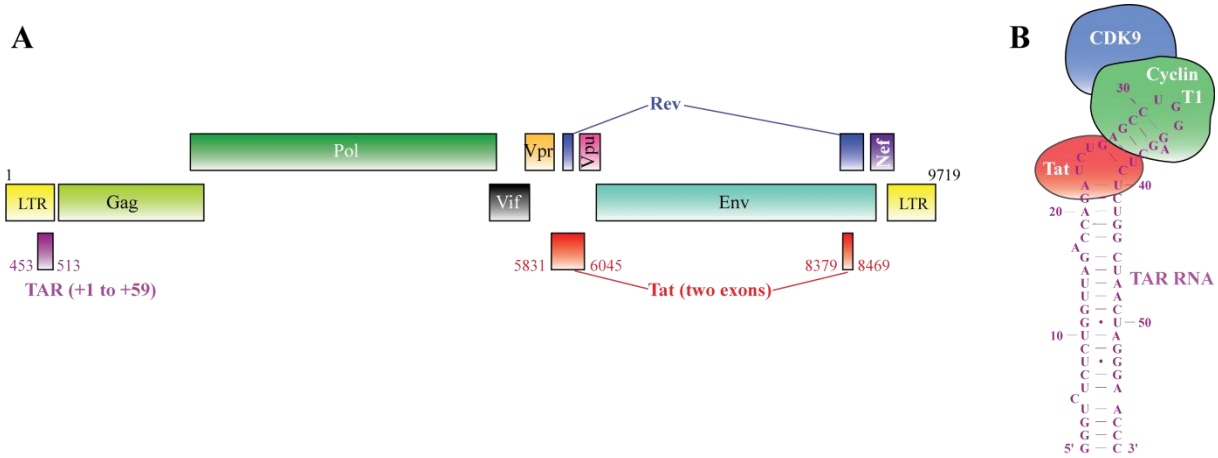



Figure I.4. HIV-1 TAR RNA and Tat protein. A) Structure of the HIV genome. Tat is encoded by two exons. Immediately downstream of the TSS is the transactivation response region, TAR. B) Interaction between TAR RNA, Tat and Cyclin T1. TAR encodes an RNA with a hairpin structure. Tat recognizes the bulge and the Cyclin T1 the apical loop (according to http://www.hiv.lanl.gov/).

Like NF-κB, others activators interact with P-TEFb such as c-Myc, the Class II Transactivator (CIITA), MyoD, HIC, B-Myb, GRIP1, MCEF, STAT3, steroid hormone receptors, and VP16 and can potentially recruit it to their respective promoters targets. Brd4, a general chromatin remodelling protein, also binds P-TEFb (Zhou et al. 2006). Brd4 is a double bromodomain-protein that binds to the Mediator complex and to acetylated chromatin, which is a hallmark of actively transcribed genes. Thus, Brd4 could represent a general recruitment factor of P-TEFb to RNAPII. Brd4 may also function in conjunction with NF-κB to recruit of P-TEFb to the HIV-1 promoter.

## 1.5. Structural organization of P-TEFb

P-TEFb is a heterodimer composed of the Cyclin-dependent kinase 9 (CDK9) and its regulatory partner Cyclin T1.

### a. CDK9

CDK9 is a serine/threonine kinase of 372 amino acids and a molecular mass of 42 kDa. This type of protein kinases transfers the γ-phosphate of ATP to the hydroxyl group of a serine, threonine or tyrosine residue on the target protein (Johnson 2009). Protein kinases share a common catalytic domain but there are a variety of different regulatory mechanisms. The fold of the catalytic kinase domain, comprising ~300 amino acids residues, consists of a small N-terminal lobe of about ~80 residues that contains a five-stranded β-sheet with one α-helix, the C-helix, which is important in regulation, and ~200 residues C-terminal lobe that is mostly α-helix with a few short β-strands (Figure I.5). A hinge region links the two lobes.

The catalytic pocket containing the ATP-binding site is formed at the interface between the two lobes. Non-polar side chains from both lobes enclose the adenine. The contacts also include three important regions: the flexible Gly-rich loop between strands β1 and β2 containing the motif GXGX(F/Y)G, the C-helix, and the hinge region between the lobes. The triphosphate moiety is stabilized by contacts to a metal ion bound by a DFG motif. An activation segment, also called T-loop (20 to 30 residues), involved in the substrate binding, is found between the DFG and an APE motifs in the C-terminal lobe. In CDKs, the catalytically competent conformation is promoted by phosphorylation of a Thr or a Tyr in the activation segment. In the non-phosphorylated state, the activation loop is supposed to block the catalytic site, interfering with ATP binding and preventing protein substrate binding.

The crystal structure of CDK9 within the P-TEFb complex exhibits a typical kinase fold (Figure I.5B), with the N-terminal lobe comprising the residues 16 to 108, and the C-terminal lobe comprising residues 109 to 330. In the structure reported by Baumli et al. (2008), CDK9 is in the active conformation, with the T- loop (activation segment) leaving free access for the substrate to the catalytic site. Indeed, the threonine of the activation segment, Thr186 (shown in pink in the Figure I.5B), is phosphorylated.

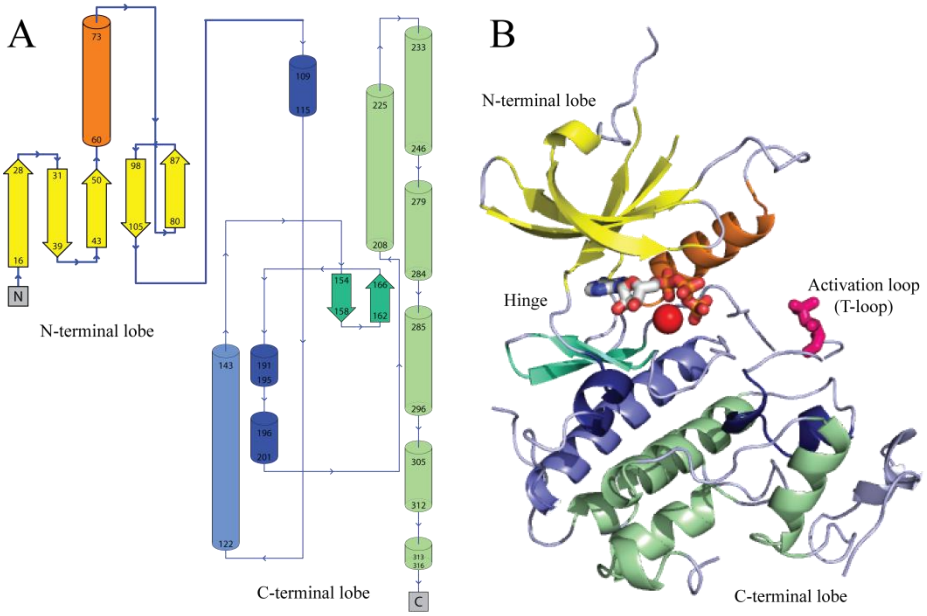



Figure I.5. Structural organization of CDK9. A) Topology diagram of CDK9 illustrating the structural domains of the kinase fold. The α-helices are represented by cylinders and β-strands by arrows (created with PDBsum). B) Crystal structure of CDK9 (PDB 3BLQ) colored as in A. In the active site, a Mg ion (red sphere) and the ATP (atom type coloured) are shown. The Thr186 at the T-loop is shown as "stick" model (pink).

Several post-translational modifications, such as phosphorylation and acetylation, regulate the activity of P-TEFb. The phosphorylation of Thr186 in the T-loop is considered essential for CDK9 activity. This phosphorylation seems to have a limited function as an organizational center (unlike other CDKs), its major role seeming to promote the active conformation of the T-loop for substrate recognition (Baumli et al. 2008). Another phosphorylation, at Ser175 promotes its interaction with Brd4 (Yang et al. 2005). Mutation at this serine inhibits the kinase activity and the binding to Brd4, and affects the in vivo association to the HIV-1 promoter. Autophosphorylation on the C-terminal region of CDK9 at Ser347, which is not conserved among other CDKs enzymes, enhances the binding of TAR RNA to the P-TEFb/Tat complex and ensures the nuclear localization of P-TEFb (Garber et al. 2000). In contrast, phosphorylation at Thr29 inhibits CDK9 kinase activity (Jiri Kohoutek 2009). Two acetylation sites have been also identified in CDK9. The acetylation of Lys44

promotes the P-TEFb activity, while the acetylation of Lys48 interferes with ATP binding and negatively regulates P-TEFb activity.

### b. Cyclin T1

As their name implies, all the CDKs require a Cyclin for activation. The main regulatory partner of CDK9 is the Cyclin T1. The Cyclin T2a, T2b and K are minor partners of CDK9, present at low concentrations in many cell types. The common structural feature of the family is the cyclin box motif (Pantano et al. 2005). This is a characteristic two-repeats folding motif of ~100 aminoacids long, connected by a linker peptide in extended conformation. The helices within each repeat are spatially disposed with the hydrophobic helix H3 surrounded by the other four. Cyclin boxes are usually inserted into a protein frame, with additional elements at the N and C termini of the cyclin box, which provide binding specificity for protein-protein interactions.

Cyclin T1 is a 726 amino acids residues long (87 kDa) protein, which can be divided in two major domains (Pantano et al. 2005). The N-terminus (1 to 300) contains the cyclin box and the TAR recognition motif (TRM). The C-terminus (301 to 726) is less characterized. It contains a putative coiled-coil region (379 to 430) and a His-rich domain (506 to 530). The C-terminal region is involved in the association with the CTD domain of RNAPII.

The crystal structure of Cyclin T1 in the P-TEFb complex (Figure I.6) shows, like in other cyclins, a canonical cyclin box, which comprises two bunches of five helices, framed by short N-terminal ($H_N$) and C-terminal ($H_C$) helices (Baumli et al. 2008). The two domains of the cyclin box are arranged around the central H3 and H3' helices (' marks the second domain). Despite their similarity in sequence and structure, cyclins involved in cell cycle control differ significantly from those involved in transcription by the length and orientation of the $H_N$ and $H_C$ helices. Contrarily to the cell cycle cyclins, where $H_C$ appears to have no function, it contributes to recognition of regulatory proteins in P-TEFb. A further difference lies in the "hydrophobic pocket", a recruitment site for CDK2 substrates, which is a not functional in Cyclin T1. Cyclin T1, in contrast to T2, contains the TRM at the C-terminal region of the cyclin box. The TRM interacts directly with the activation domain of Tat, and is important for the formation of the ternary complex Tat/TAR/P-TEFb. The Cys261 of Cyclin

T1, which is absent in Cyclin T2, also contributes to the interaction with Tat via a zinc ion. In line with this, only Cyclin T1 is involved in Tat-mediated transactivation.

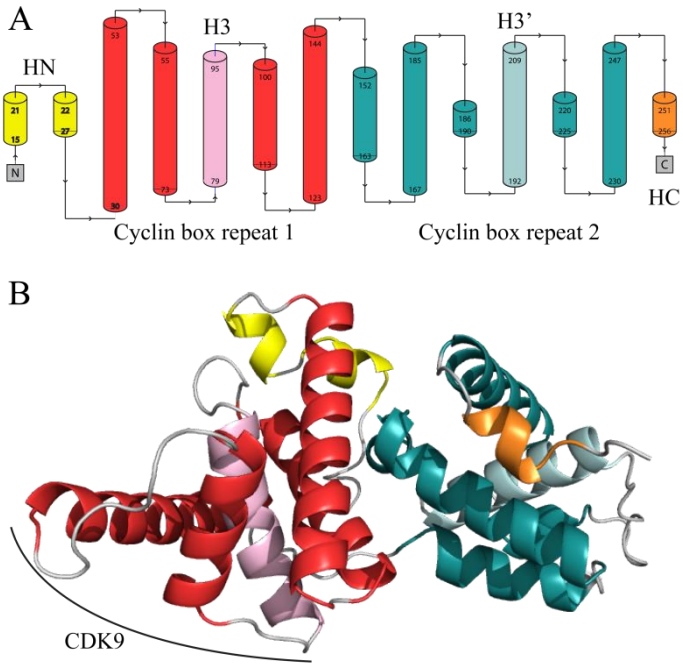



Figure I.6. Cyclin box domain of Cyclin T1. A) Topology diagram illustrating the structural organization of the cyclin box repeats of Cyclin T1 (α-helices are represented by cylinders; created with PDBsum). B) Crystal structure of Cyclin T1 (PDB 3BLQ) colored as in A. The surface interacting with CDK9 is indicated.

### c. CDK:Cyclin interaction

The crystal structure of P-TEFb (PDB 3BLQ) shows a rotation of Cyclin T1 with respect to the CDK9 when compared to cell cycle CDK/Cyclin complexes [such as CDK2/Cyclin A; (Baumli et al. 2008)]. This orientation results in a comparatively reduced number of contacts between CDK9 and Cyclin T1. The helix $H_N$ of Cyclin T1 is directed towards the solvent, as in the free Cyclin T1. Helices H3, H4 and H5 of the Cyclin T1 contact αC helix (in orange in Figure I.5) and β4 strand of the CDK9. The cyclin H5 helix runs parallel to the CDK αC helix and participates in the alignment of the active conformation.

The structure of P-TEFb co-crystallized with flavopiridol, a kinase inhibitor, shows that the binding of flavopiridol to the ATP-binding pocket of CDK9 induces a closing of the Gly-rich loop over the active site via van der Waals contacts with Ile25, Val33 and Phe30. The new position of the Gly-rich loop would exclude ATP binding.

The crystal structure of the P-TEFb/Tat (PDB 3MIA) has been recently published and shows that Tat acquires an extended conformation on the surface of the Cyclin T1 (Tahirov, Babayeva, Varzavand, Cooper, Sedore, et al. 2010). Interestingly, Tat inserts in a groove at the CDK9 and CyclinT1 interface, thus leading to a more stable complex, that is supposed to be more active (Figure I.7; see Chapter "Molecular description of HEXIM1" for further details and illustration).

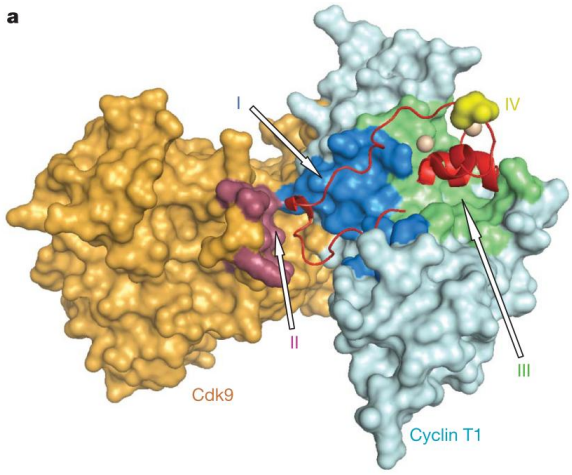



Figure I.7. Interactions between Tat and P-TEFb. CDK9 and Cyclin T1 are in orange and light blue, respectively. The interactions can be divided in: (1) the acidic/Pro-rich region and the Cys-rich and the core regions of Tat bind between the two domains of the cyclin box (blue and green regions, respectively); (2) A β-turn in the acidic/Pro-rich region of Tat contacts the T-loop of CDK9 (pink region); and (3) the Cys-rich region of Tat forms a second zinc finger, with the participation of Cys261 of Cyclin T1 (yellow region). Image from (Tahirov, Babayeva, Varzavand, Cooper, Sedore, et al. 2010).

The control of P-TEFb kinase activity is a pivotal point of regulation for the RNAPII transcription. The understanding of this regulation has also medical implications, since P-

TEFb is a specific cofactor for efficient transcriptional elongation during the HIV-1 gene expression

In 2001, two independent investigations showed that in human HeLa cells about half of P-TEFb was sequestered into a large kinase-inactive complex which includes a snRNA called 7SK (Nguyen et al. 2001; Yang et al. 2001). This discovery pointed out the crucial role of a ncRNA, the 7SK snRNA, in the regulation of the gene expression.

## 2. THE NON-CODING RNAS AS TRANSCRIPTIONAL REGULATORS

Several ncRNAs have been shown to actively participate in the regulation of transcription even by targeting directly the RNAP, or by targeting transcription factors. In this section two examples of ncRNAs regulators of the transcription, 6S and B2, will be presented before a more detailed description of 7SK snRNP.

### 2.1. Prokaryotic transcription regulation by the 6S RNA

6S is a ncRNA, conserved in many bacteria, that forms a specific complex with the RNAP associated to $\sigma^{70}$, the "housekeeping" $\sigma$ factor responsible of the transcription of most genes in growing cells (Wassarman et al. 2000; Wassarman 2007). In bacterial transcription, the association of the RNAP with a $\sigma$ factor is required for the initiation. The initiation involves the recognition of the promoter by the RNAP mediated by the specific interaction of the $\sigma$ factor to the conserved sequences at -10 and at -35. The DNA is then unwound between the -10 region and the start site of the transcription, what is called the "open complex", and the RNAP synthesizes some few ribonucleotides releasing the $\sigma$ factor, while the RNAP continues the elongation of the template.

6S RNA inhibits the $\sigma^{70}$-dependent transcription during the stationary phase, when 6S is most abundant, and activates several $\sigma^{S}$-dependent promoters in vivo (Trotochaud et al. 2004). $\sigma^{S}$ is the primary regulator of stationary phase and stress response genes. The phenotype of cells lacking of 6S RNA led to propos that 6S RNA would be important to balance nutrient utilization for long-term cell survival, in part by limiting stress responses to conserve energy (Trotochaud et al. 2006, 2004).

Even if the primary sequence of the different 6S RNA homologs is not particularly conserved, its secondary structure is conserved (Barrick et al. 2005). 6S RNA consists of a closing stem, a large central loop, and a terminal loop conserved domains separated by variable stems (Figure I.8). Interestingly, the secondary structure of 6S RNA is similar to the DNA conformation during transcription initiation in the "open complex" (Wassarman et al. 2000; Barrick et al. 2005). This similarity, suggesting the blocking of the active site of the

RNAP by 6S, could provide of an appealing mechanism for the inhibition of the $\sigma^{70}$-RNAP. Supporting this hypothesis, 6S RNA binds directly within the active site of the RNAP. Moreover, the region of the $\sigma^{70}$ factor that mediates the binding to the 6S RNA is also important for DNA interaction, so it has been suggested that RNA and DNA have overlapping sites on $\sigma^{70}$ that would result in a competition of both molecules for the binding (Klocko et al. 2009).

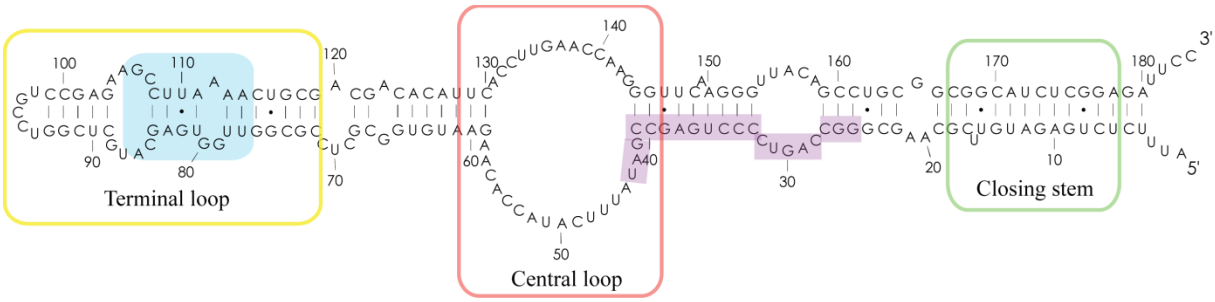



Figure I.8. The 6S RNA. Model of the secondary structure of 6S RNA showing its conserved structural elements. The sequence used as template for transcription is highlighted in pink, whereas the region predicted to contain the interaction site for $\sigma^{70}$ is in blue.

Surprisingly, 6S RNA acts as a template for synthesis of small RNAs, which liberates the RNAP from 6S RNA (Wassarman et al. 2006). Hence, the current biological model propose that 6S RNA binds the RNAP in the stationary phase and is able to repress transcription because nucleotide concentration is insufficient for transcription from the 6S RNA template. When the nucleotide concentrations increase, upon exit from stationary phase, the transcription from 6S RNA is able to proceed thereby relieving the repression by 6S RNA.

## 2.2. B2 and Alu RNAs in the eukaryotic transcription regulation

Mouse B2 and human Alu RNAs repress mRNA transcription by binding to the RNAPII during the cellular heat shock response (Espinoza et al. 2004; Mariner et al. 2008). Both are transcribed by the RNAPIII from short interspersed elements (SINEs), widely abundant in their respective genomes (Goodrich et al. 2010). Upon cellular stress, such as heat shock, the

level of B2 and Alu RNAS sharply increases. These ncRNAs bind to the RNAPII and repress transcription of several genes, while heat shock genes are stimulated. B2 and Alu RNAS do not block the assembly of the RNAPII and the GTFs complexes at the DNA promoter, but these complexes are transcriptionally inert since these ncRNAs disrupt contacts between the RNAPII and the DNA throughout the core promoter (Yakovchuk et al. 2009). In contrast, the interactions between the GTFs and the promoter are not affected and they probably allow the RNAPII to be held in the inactive complexes on the DNA.

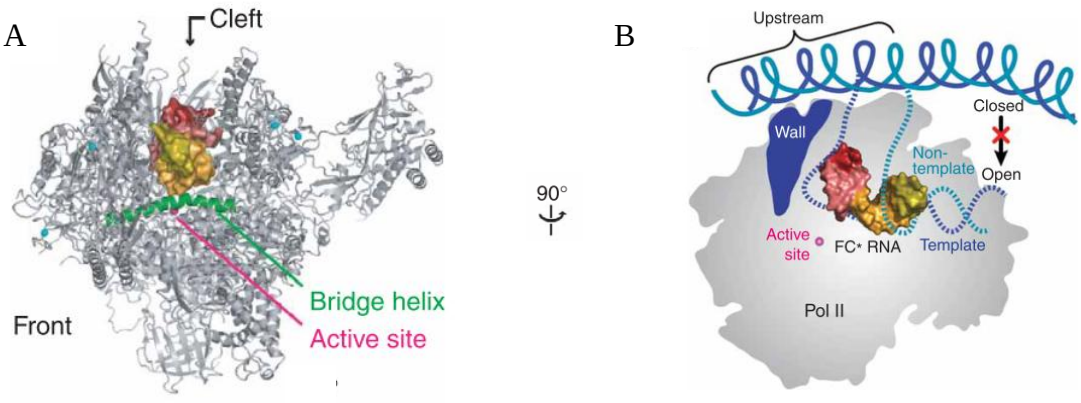



Figure I.9. Structure of the RNAPII/FC RNA complex. A) The RNAPII is shown in ribbon model in gray, and the FC RNA as a molecular surface in color. B) Model for inhibition of open complex formation by FC RNA. The promoter DNA is modeled as in the open complex. Adapted from (Kettenberger et al. 2006).

FC is an RNA aptamer that has been selected to bind yeast RNAPII. FC inhibits transcription initiation but not elongation, it competes with B2 for RNAPII binding suggesting that they may bind overlapping sites and use similar mechanism of inhibition. The crystal structure of the RNAPII in complex with a portion of FC has been solved (Kettenberger et al. 2006). The structure shows the RNA bound to a site that overlaps, but is not identical, with the binding site for nucleic acids in an elongation complex (Figure I.9). This suggests that the RNA inhibitor prevents entry of the promoter DNA during initiation, and that elongation complexes are not inhibited because pre-bound nucleic acids exclude the RNA inhibitor from the cleft.

The inhibition by these ncRNAs is reversible, but the mechanism by which the RNAPII is released is not known.

## 2.3. 7SK snRNP: a regulator of P-TEFb

7SK, an abundant snRNA (~2 × $10^5$ copies per cell), was identified on the 70s (Wassarman et al. 1991). However, its function, a riboregulator of a transcription elongation factor, remained unknown until 2001 (Nguyen et al. 2001; Yang et al. 2001). 7SK is found in metazoan organisms and is strongly conserved in higher vertebrates, with 332 nucleotides in human cells. 7SK gene is transcribed by the RNAPIII and has a type 3 promoter characterized by a Proximal Sequence Element (PSE), a TATA box located downstream of the PSE, and a Distal Sequence Element (DSE) upstream of the PSE (Schramm et al. 2002). 7SK snRNA contains a U-rich 3'end, as many RNAPIII transcripts, and a methylated γ-phosphate cap structure at the 5'end that protects it from degradation (Wassarman et al. 1991).

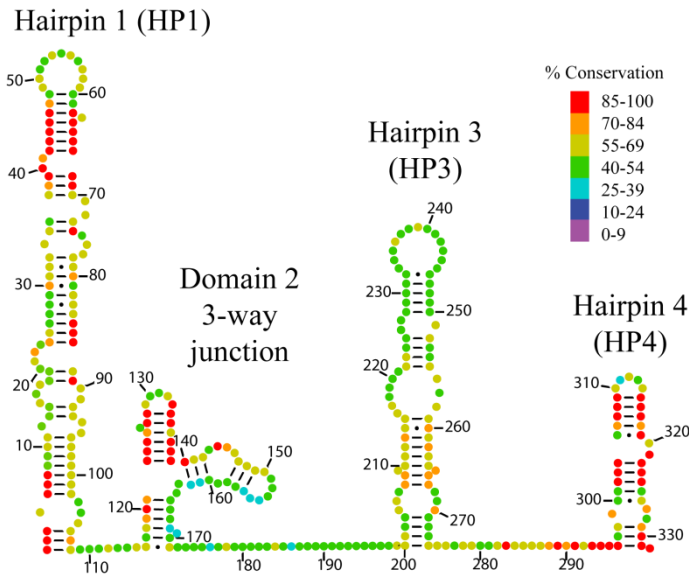



Figure I.10. Secondary structure and sequence conservation of 7SK. Model of the secondary structure of 7SK according to (Wassarman et al. 1991). The sequence is colored based on its conservation according to (Marz et al. 2009).

A secondary structure model has been proposed from probing experiments performed on 7SKsnRNP extracted from HeLa cells (Wassarman et al. 1991). In this model 7SK consists of three hairpin structures (comprising the nucleotides 1 to 108, 200 to 275, and 296 to 332, respectively) and a domain characterized by a three-way junction (nucleotides 116 to 171),

separated by stretches of single stranded regions (Figure I.10). More detailed information about the secondary structure of 7SK will be given in the Chapter VII "Probing the secondary structure of 7SK".

Sequence analyses have shown that two region, encompassing the 5' and 3'end hairpins respectively, are highly conserved across all organisms (Figure I.10), as well as one of the stem-loop elements of the three-way junction (Gruber, Koper-Emde, Marz, et al. 2008; Gruber, Kilgus, et al. 2008; Marz et al. 2009). In contrast, the region encompassing the hairpin 3 is only conserved in vertebrates.

7SK has been shown to associate to P-TEFb, and inhibit its kinase activity (Nguyen et al. 2001; Yang et al. 2001). But 7SK is not sufficient to inhibit P-TEFb. Indeed, in 2003 a novel component of the 7SK/P-TEFb snRNP was identified: HEXIM1 (Michels et al. 2003; Yik et al. 2003). HEXIM1 was able to specifically inhibit the kinase and transcriptional activities of P-TEFb in a 7SK snRNA-dependent fashion. HEXIM1 [Hexamethylene bisacetamide (HMBA)-inducible protein 1] was initially identified in human vascular smooth muscle cells as an up-regulated protein after treatment with the differentiating agent HMBA (Kusuhara et al. 1999). HEXIM1 will be described in more detail in the Chapter II "Molecular description of HEXIM1 protein".

Two regions of 7SK have been identified to be necessary for the binding of HEXIM1 and P-TEFb: the 5'end hairpin is essential for binding of both partners, while the 3' hairpin is essential for P-TEFb binding (Egloff et al. 2006; Bélanger et al. 2009). Indeed, a minimal 7SK containing only 5'end and 3'end hairpins can both bind and inactivate P-TEFb in HeLa cells (Egloff et al. 2006).

### a. The 7SK snRNP core

In 2008, two new partners of 7SK were indentified and showed to be stable components of the core 7SK snRNP: LaRP7 and MeCPE (Jeronimo et al. 2007; Markert et al. 2008; He et al. 2008; Krueger et al. 2008; Xue et al. 2010). In cells, most of 7SK is bound to LaRP7 (La-Related Protein 7) via its U-rich 3'end. MePCE (MethylPhosphate Capping Enzyme) is responsible for adding the cap structure at the 5'end. MePCE remains bound to 7SK after capping, stabilizing the interaction with LaRP7. Both, LaRP7 and MePCE act cooperatively to stabilize 7SK snRNA and maintain the integrity of 7SK snRNP in cells. It has been also

suggested that the interaction (and probably the specificity of recognition) between 7SK and HEXIM1 is enhanced by LaRP7 (Krueger et al. 2008; Markert et al. 2008; He et al. 2008). Direct interaction between HEXIM1 and LaRP7 that could explain this enhancement has been reported in vitro conditions, but in vivo requires 7SK (Krueger et al. 2008; Markert et al. 2008; He et al. 2008). Another explanation is that LaRP7 binding lead to a 7SK conformation that could facilitate the recognition by HEXIM1.

As HEXIM1 and 7SK, LaRP7 is only found in metazoan organisms, and the distribution of three proteins across the different animal clades coincides, suggesting a functional relationship among the three partners (Marz et al. 2009). This also suggests that this mechanism of transcription elongation regulation is a metazoan innovation. In contrast, MePCE homologs are also found in plants and fungi, suggesting alternative important functions.

### b. The 7SK/HEXIM1/P-TEFb complex

To date, the mechanism by which 7SK/HEXIM1 complex inhibits P-TEFb is not well understood but information of the elements involved in the interaction among 7SK, HEXIM1 and P-TEFb provides some insights about it. The stoichiometry of the 7SK/HEXIM1/P-TEFb complex is still controversial, particularly regarding how many P-TEFb are present. It most likely contains a single molecule of 7SK, a dimer of HEXIM1 and two copies of P-TEFb (Dulac et al. 2005; Li et al. 2005).

The interaction between P-TEFb and the 7SK/HEXIM1 complex seems to be mainly mediated by contacts between HEXIM1 and the Cyclin T1 (Dulac et al. 2005; Blazek et al. 2005; Yik et al. 2005; Schulte et al. 2005; Qintong Li et al. 2005). A detailed introduction of what is known about the regions of contact between HEXIM, 7SK and P-TEFb will be discussed in the following Chapter II "Molecular description of HEXIM protein". However, some in vitro experiments have shown that P-TEFb can bind specifically 7SK in a HEXIM1-independent fashion, even if P-TEFb inhibition requires HEXIM1 (Yik et al. 2003; Chen et al. 2004). Interestingly, this was shown to strogly depend on phosphorylation of the CDK9.

The modification state of the components is important for 7SK snRNP formation. For P-TEFb, the phosphorylation at the Thr186 in the T-loop of CDK9 is necessary for binding with 7SK:HEXIM1 complex (Li et al. 2005; Chen et al. 2004). Since the T-loop

phosphorylation is a hallmark of active P-TEFb, it seems that the 7SK/HEXIM1 complex sequesters activated P-TEFb that is ready for the kinase activity once released from the 7SK snRNP. These observation lead also to hypothesize that by binding CDK9 with the phosphorylated T-loop and consequently an "open" conformation, 7SK/HEXIM1 may physically block the access to the catalytic center of CDK9 kinase (Chen et al. 2004). In another hand, the dephosphorylation might also constitute a mechanism to release P-TEFb from the 7SK snRNP. In line with this, it has been shown that UV/HMBA initiates a $Ca^{2+}$/calmodulin-dependent signalling pathway to activate PP2B, a serine/threonine phosphatase that seems to facilitate the accessibility of the CDK9 T-loop to PP1$\alpha$ another phosphatase, which in turns dephosphorylate the Thr186 (Chen et al. 2008). This leads to the release of P-TEFb from 7SK snRNP.

The release of P-TEFb from the 7SK snRNP is also induced by stressful events that globally lead to the transcription inhibition, such as exposure of cells to Actinomycin D, a DNA-damaging agent, or to DRB, a kinase inhibitor.

### c. Remodelling 7SK snRNP

While LaRP7 and MePCE are permanent elements of the 7SK snRNP, this complex is essentially remodeled according to the transcription state. When HEXIM1 and P-TEFb are not bound, 7SK is found associated with other partners. A subset of heterogeneous ribonuclear proteins (hnRNP), such as Q, R, A1, A2, K as well as the RNA helicase A (RHA) have been identified as major 7SK snRNA-associated proteins (Barrandon et al. 2007; Van Herreweghe et al. 2007; Hogg et al. 2007). These hnRNPs are found mostly associated to 7SK under conditions in which 7SK is not associated to HEXIM1 and P-TEFb, this is after DRB or Actinomycin D cells treatment. Hence, it has been proposed that the nuclear level of active P-TEFb may be driven by the competitive interaction of 7SK with partners other than HEXIM1 (Figure I.11). Thus, 7SK snRNP would be a dynamic complex subjected to reversible remodeling, with 7SK snRNA playing a key role in the regulation of RNAPII.

The third hairpin of 7SK seems to be implicated in the interaction to some of the hnRNPs, and its suppression results in a 7SK/HEXIM1/P-TEFb resistant to stress-induced disassembly (Van Herreweghe et al. 2007).

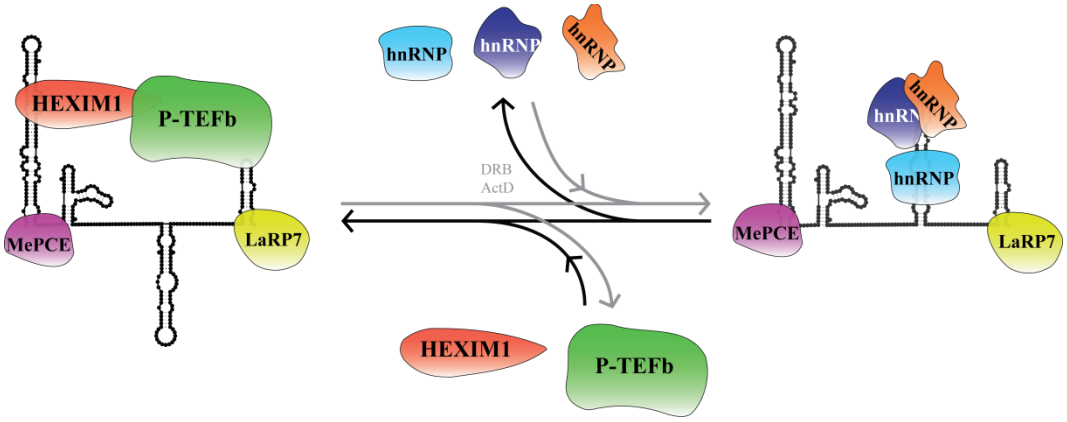



Figure I.11. Reversible remodeling of 7SK snRNPs. Transcription inhibition by DRB or Actinomycin D treatment induces dissociation of HEXIM1 and P-TEFb from 7SK snRNP, which then associates to the hnRNPs. The dynamic remodeling of the 7SK snRNP would modulate the level of active P-TEFb in the cell.

Hence, the different hairpins of 7SK seem to serve as platforms for various 7SK-binding proteins, and their evolutionary conservation might reflect the conservation of various interactions.

### d. A glance in the 7SK snRNPs and the cellular interaction network

Several additional roles of HEXIM1 have been reported since the discovery of the P-TEFb inhibitor function of HEXIM1. HEXIM1 suppresses the transcriptional activity of NF-κB, a nuclear factor that plays a pivotal role in regulating the expression of genes that influence cells differentiation, proliferation and inflammation (Ouchida et al. 2003). It has been shown that C-terminal domain of HEXIM1 is involved in this function. Similarly, a role for HEXIM1 as a potent inhibitor of CIITA (class II transactivator)-mediated transcription by sequestering P-TEFb from CIITA has been suggested (Jiri Kohoutek et al. 2006).

HEXIM1 also down-regulates the transcriptional activity of Estrogen Receptor alpha (ERα). This is both, by direct interaction through the HEXIM1 NLS, as well as by competition with ERα for binding to Cyclin T1 (Wittmann et al. 2005). In fact, increased HEXIM1 expression results in a decrease in estrogen-stimulated recruitment of ERα, P-TEFb, and Ser2 phosphorylated Pol II to promoter and coding regions of ERα-responsive genes

(Ogba et al. 2008). This may explain the HEXIM1-driven inhibition of breast cell growth since estrogens stimulates cell proliferation via ER (Wittmann et al. 2003). Also, it has been observed an estrogen-mediated down-regulation of HEXIM1 as well as a diminution of HEXIM1 in breast tumor cells (Wittmann et al. 2003). All these results point out an important role of HEXIM1 in breast cancer.

HEXIM1 can inhibit the Glucocorticoid Receptor (GR)-dependent transcription either by sequestration of P-TEFb, either by direct interaction (Shimizu et al. 2005; Yoshikawa et al. 2008). The interaction between HEXIM1 and GR requires the hinge region which links the DNA binding domain (DBD) and the Ligand Binding Domain (LBD) of the GR and the NLS domain of HEXIM1, and does not require 7SK (Shimizu et al. 2005; Yoshikawa et al. 2008).

Some evidences suggest that N-CoR/HDAC3 complex, involved in transcriptional repression, interacts with inactive large P-TEFb complex via direct binding between HEXIM1 and N-CoR. N-CoR/HDAC3 complex negatively regulates P-TEFb activity likely through deacetylation of CDK9 at Lys44, which was found to be important for the kinase activity (Fu et al. 2007). Hence, an interesting hypothesis emerged since the active form of P- TEFb is associated to Brd4, which in turns binds acetylated histones through its double bromodomain, a signature motif for binding of acetylated lysine (Jang et al. 2005; Yang et al. 2005). Hence, Brd4 may bind preferentially to the acetylated form of CDK9 and maintain its kinase activity in transcriptional elongation; on the contrary, the deacetylation of CDK9 by N-CoR/ HDAC3 or other HDACs may reduce the association of CDK9 with Brd4 and thus promote the interaction of P-TEFb with 7SK and HEXIM1 to form an inactive complex (Fu et al. 2007). Interestingly, Cyclin T1 is also acetylated in the active form of P-TEFb, and this acetylation triggers the dissociation of P-TEFb from the 7SK snRNP (Cho et al. 2009).

Nucleoplasmin (NPM), a nuclear phosphoprotein involved in ribosome biogenesis, cell growth and proliferation regulation, interacts with the 7SK/PTEFb free form of HEXIM1 via its BR and functions as a negative regulator of HEXIM1 (Gurumurthy et al. 2008). Over-expression of NPM decreases HEXIM1 proteins levels, but not HEXIM1 mRNA levels, through proteasome-dependent degradation and up-regulates P-TEFb-dependent transcription (Gurumurthy et al. 2008).

7SK has also been related to other functions. Depletion of 7SK in HeLa cells identified P-TEFb-independent regulatory phenomena. When 7SK concentrations are decreased to less than 5%, not only P-TEFb target genes are expressed at a higher rate, but

also genes not regulated by P-TEFb are both under or overexpressed (Eilebrecht et al. 2010). These observations point to a yet unidentified second function in transcription regulation exerted by 7SK snRNA. Indeed, it was shown recently that 7SK binds the chromatin factor and transcription regulatory hub HMGA1 (High Mobility Group protein) via the stem-loop encompassing the nucleotides 113 to 154, and promotes positive or negative regulatory activity. This suggests a role of 7SK in transcription initiation and in cell differentiation and proliferation by regulating HMGA1.

I-mfa (Inhibitor of MyoD family a) and HIC (Human I-mfa domain Containing) proteins interact with the His-rich domain and a Lys/Arg-rich motif (amino acids 250 to 275, overlapping the TRM) at the C-terminal region of Cyclin T1, leading to the inhibition of P-TEFb (Wang et al. 2007). This inhibition may involve the recruitment of P-TEFb to the transcription complex rather than a direct effect on its kinase activity. Since MyoD, a myogenic regulatory factor, was shown to bind P-TEFb, it has been suggested that I-mfa and HIC operate during development. Paradoxically, the 3'-UTR (UnTranslated Region) of HIC binds and activates P-TEFb by displacing 7SK (Young et al. 2007), probably in part by acquiring a structure that would mimic the 3'end hairpin of 7SK (nucleotides 4002 to 4030 of 3'UTR). These observations suggest a mechanism whereby gene expression can be controlled at the level of P-TEFb by mRNA as well as by protein modulators.

# CHAPTER II:
# MOLECULAR DESCRIPTION OF THE HEXIM1 PROTEIN

The human HEXIM1 protein is a 359 amino acids residues (41 kDa) protein that can be divided in three parts: a Proline-rich N-terminal half, a central region characterized by a cluster of basic amino acids (residues 150-165) and two acid clusters (residues 211-219 and 234-253), and a C-terminal domain rich in Glu and Leu.

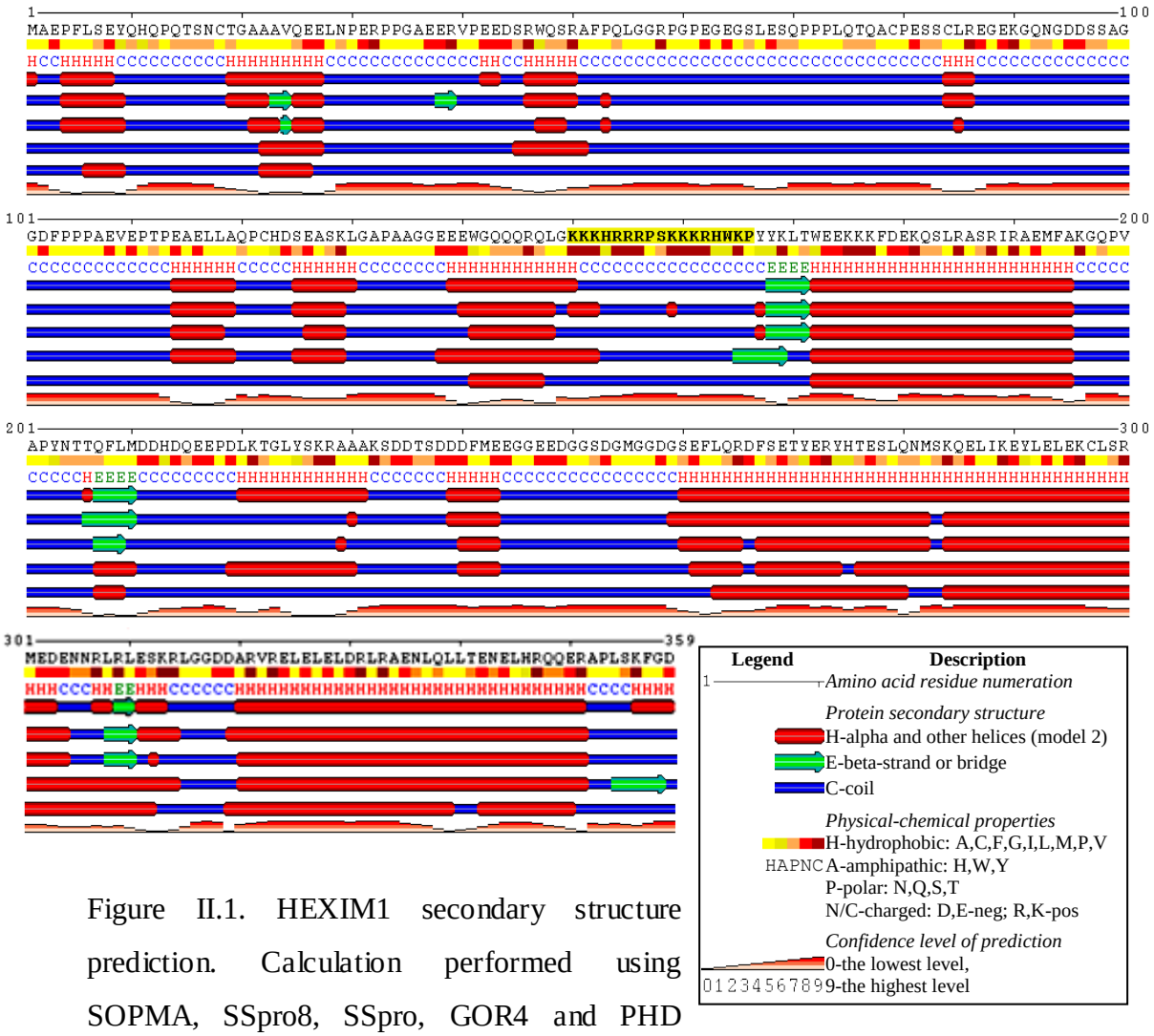



Figure II.1. HEXIM1 secondary structure prediction. Calculation performed using SOPMA, SSpro8, SSpro, GOR4 and PHD program respectively. ARM is highlighted.

Calculations from different secondary structure prediction programs (Figure II.1) propose a mainly unfolded N-terminal domain, as expected from its numerous Pro residues. A helical region is predicted in the central domain (residues 172-195) preceded by a short β sheet (nucleotides 168-171); the rest of the central domain is mostly unfolded. The C-terminal domain was predicted to consist of a long helical region (from around residue 260 to 351), which was confirmed by the structure determination of region 255 to 359 by NMR (Dames et al. 2007).

# 1. Functional Organization of HEXIM1

## 1.1. The N-terminal domain

The N-terminal domain (residues 1 to 149) of HEXIM1 is the less characterized one and has been shown to be dispensable for P-TEFb binding and inhibition, and for the association with 7SK (Michels et al. 2004; Yik et al. 2003; Dulac et al. 2005). Some evidences suggest that it could function as a self-inhibitory domain (see below).

## 1.2. The central domain

The central domain (residues 150 to 254) is characterized by a cluster of positively charged residues, the basic region (BR; residues 150 to 177), and a cluster of negatively charged residues (211 to 249) called the acid region [AR; (Michels et al. 2003; Barboric et al. 2005)]. A Nuclear Localization Signals (NLS) is found in this domain, which has been shown to direct HEXIM1 nuclear import, probably via the importin α-dependent pathway since association between HEXIM1 and importin α has been detected in vitro (Michels et al. 2003; Barboric et al. 2005). However, subcellular fractionation and immunofluorescence show that, while most HEXIM1 is found in the nucleus, a significant fraction is found in cytoplasm, both fractions being associated with RNA (Li, Cooper, et al. 2007).

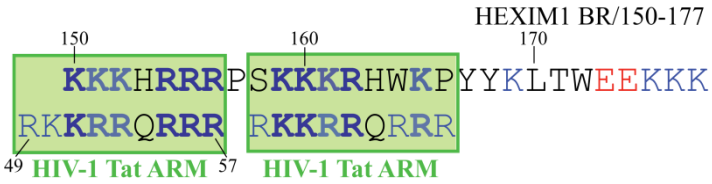



Figure II.2. Arginine-rich motif of HEXIM1 and Tat. The two HEXIM1 sequences in the BR similar to the ARM of Tat are boxed. Identical and conservative amino acids are highlighted. Positive and negative amino acids are colored in blue and red, respectively.

In cells, HEXIM1 binds specifically 7SK, and a conserved ARM (Arginine-Rich Motif; residues 150 to 165) in the BR is essential for this interaction (Michels et al. 2003; Barboric et al. 2005; Yik et al. 2004). Interestingly, this HEXIM' ARM is very similar to the arginine-rich TAR RNA-binding motif of the HIV-1 Tat protein (Figure II.2; Yik et al. 2004). It has been shown that a 28-residues peptide of HEXIM1 corresponding to the BR and expressed in HeLa cells is still able to recruit 7SK showing only a slightly reduced specificity (Yik et al. 2004). The crucial role of the ARM for 7SK interaction has been also proved by a mutant of the KHRRR (152 to 156) sequence, which is unable to bind 7SK (Michels et al. 2003).

It has been hypothesized that BR and AR would mediate an interaction between the N- and the C-terminal regions of HEXIM1 leading to an auto-inhibitory conformation of the protein. The binding of 7SK to the BR would disrupt this interaction yielding a conformational change of HEXIM1 which would unmask the P-TEFb binding site (Barboric et al. 2005). This model explains why 7SK is required for the binding and inhibition of P-TEFb by HEXIM1 (Yik et al. 2003; Michels et al. 2004; Yik et al. 2004; Haaland et al. 2005). Indeed, the removal of positive and negative charges of BR and AR regions alleviates the requirement of 7SK for the sequestration and inhibition of P-TEFb by HEXIM1 (Barboric et al. 2005). Accordingly, a HEXIM1 deleted of the 180 N-terminal residues exhibits constitutive binding and inhibition of P-TEFb in the absence of 7SK (Michels et al. 2004; Yik et al. 2003). Additional support comes from fluorescence studies and native gel electrophoresis analysis which suggest that HEXIM1 undergoes a conformational change upon RNA binding (Li, Cooper, et al. 2007). Importantly, these results also suggest that the C-terminal half of HEXIM1 is responsible for the repression of the CDK9 kinase activity and that 7SK does not participate in the inhibition.

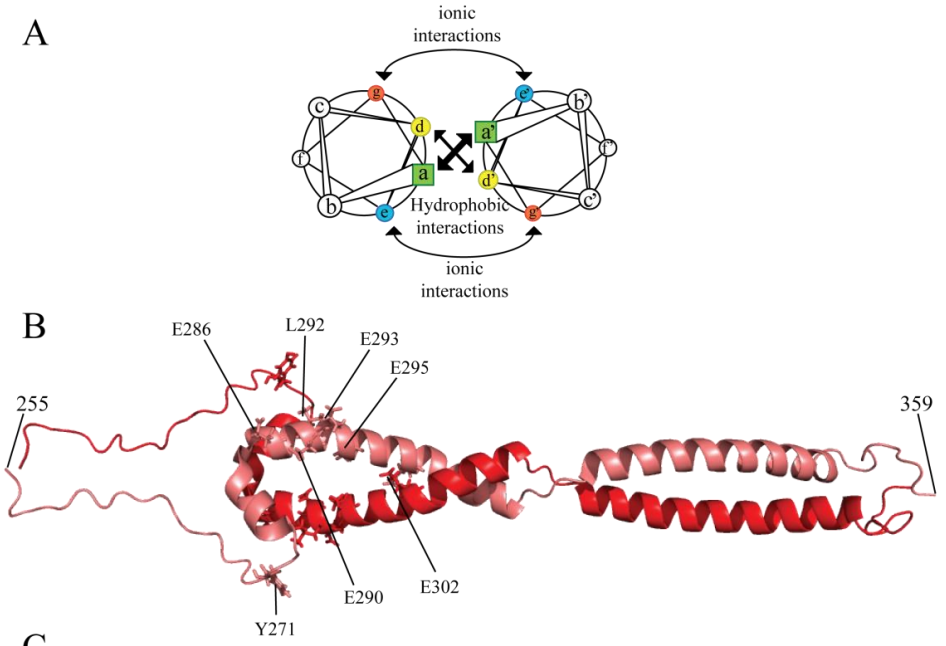
Crosslink experiments suggest that the region 210 to 220 of HEXIM1 also contacts 7SK, and the deletion of this region reduces the binding to 7SK (Bélanger et al. 2009). Next to this region, between BR and AR, is found another highly conserved region, the PYNT motif (residues 202 to 205), where Tyr203 and Thr205 were shown to be involved in P-TEFb binding and inhibition (Michels et al. 2004). Phe208 is critical for inhibition of P-TEFb kinase activity but not for its recruitment (Li et al. 2005).

## 1.3. The C-terminal domain

The dimerization of HEXIM1 seems to be a prerequisite for its incorporation into the large inactive complex of P-TEFb and therefore for binding and inhibition of P-TEFb in cells (Blazek et al. 2005).

In vivo and in vitro approaches have demonstrated that the C-terminal domain of HEXIM1 (residues 255 to 359) is responsible for this dimerization and for Cyclin T1 interaction, so it is also called the Cyclin T1 binding domain [TBD; (Dulac et al. 2005; Blazek et al. 2005; Yik et al. 2005; Schulte et al. 2005; Li et al. 2005)].

The three-dimensional structure of the TBD has been determined by NMR (Figure II.3). It forms a parallel bipartite homodimeric left-handed coiled coil comprising two segments (Lys284 to Lys313 and Asp319 to Gln348), preceded by a short $\alpha$ helix (residues Thr276 to Asn281; Dames et al. 2007). The composition of the second coiled-coil matches almost perfectly the canonical repeat motif that characterizes a left-handed coiled-coil. It consists of seven residues usually denoted (*a-b-c-d-e-f-g*) where *a* and *d* are typically nonpolar core residues found at the interface of the two helices, and *e* and *g* are solvent-exposed, polar residues that give specificity of interaction between the two helices through electrostatic interactions (Mason et al. 2004; Figure II.3A). The first coiled-coil segment diverges significantly from the consensus heptad repeat, exhibiting instead evolutionary conserved residues (Lys284, Ile288, Glu290, Tyr291, Leu292 and Glu295) in different positions of the heptad, which probably lead to a less tight $\alpha$-helix packing (Dames et al. 2007).




D

| Residue | Mutation affects: | Reference |
|---|---|---|
| Y203 | P-TEFb interaction and inhibition | Michels et al. 2004 |
| T205 | P-TEFb interaction and inhibition | Michels et al. 2004 |
| F208 | P-TEFb inhibition | Li et al. 2005 |
| Y271 | P-TEFb inhibition | Li et al. 2005 |
| E286/E290/E293 | P-TEFb interaction and inhibition | Dames et al. 2007 |
| E290 | P-TEFb inhibition | Dames et al. 2007 |
| L292 | P-TEFb interaction and inhibition | Dames et al. 2007 |
| E295 | P-TEFb inhibition | Dames et al. 2007 |
| E302 | P-TEFb interaction and inhibition | Dames et al. 2007 |

Figure II.3. TBD structure. A) Schematic representation of a parallel dimeric coiled-coil showing the heptad motif (the orientations is down the axis of the $\alpha$ helices, and from N-ter to C-ter and). B) Structure of dimeric TBD [PDB 2GD7 (Dames et al. 2007)]. Residues revealed important for P-TEFb binding and/or inhibition are indicated in sticks. C) Surface representation of TBD colored according to conservation, as based on the alignment from (Marz et al. 2009): variable (pink), average (gray) and conserved (blue). D) Table summarizing the residues involved in P-TEFb binding and/or inhibition.

As its name implies, the TBD mediates the HEXIM1 and Cyclin T1 interaction. The isolated TBD is able to directly bind the Cyclin T1 in a 7SK and CDK9-independent fashion (Michels et al. 2003; Yik et al. 2003; Schulte et al. 2005; Dulac et al. 2005; Dames et al. 2007). Cyclin T1 has been shown to interact with the first segment of the coiled coil and the short N-terminal $\alpha$-helix (Blazek et al. 2005; Schulte et al. 2005; Li et al. 2005; Dames et al. 2007). Dimeric HEXIM1 most probably binds two Cyclin T1 since it has been shown that two molecules of P-TEFb are present in the 7SK/HEXIM1/P-TEFb complex, even if P-TEFb is a monomer in the free form and only one molecule of 7SK is present (Dulac et al. 2005; Li et al. 2005). However, this is still under debate since characterization of the interaction between the TBD and the Cyclin box domain of Cyclin T1 by biophysical methods suggests that the dimeric coiled coil binds only one Cyclin T1 molecule (Schönichen et al. 2010).

Analysis of mutants has allowed identifying important residues for P-TEFb binding and inhibition. These are summarized in Figure II.3, including mutant in the PYNT region (Li et al. 2005; Dames et al. 2007; Michels et al. 2004).

## 2. HEXIM Family

HEXIM1 is closely related to HEXIM2 protein (Figure II.4). HEXIM2 has 286 residues and contains four exons whereas HEXIM1 has no introns (Michels et al. 2003). As HEXIM1, HEXIM2 also possesses the ability to inactivate P-TEFb to suppress transcription through a 7SK-mediated interaction with P- TEFb (Yik et al. 2005; Byers et al. 2005). HEXIM2 is able to functionally compensate HEXIM1 for its association with P-TEFb, when HEXIM1 is knocked down (Byers et al. 2005; Yik et al. 2005). Despite their similar functions, HEXIM1 and HEXIM2 exhibit distinct expression patterns in various human tissues and cell lines and since the N-terminal domain of HEXIM2 presents the major differences with HEXIM1, it has been suggested that whereas the two homologous HEXIM proteins are likely to have similar physiological functions and mechanisms of action, they may be regulated differently (Yik et al. 2005).

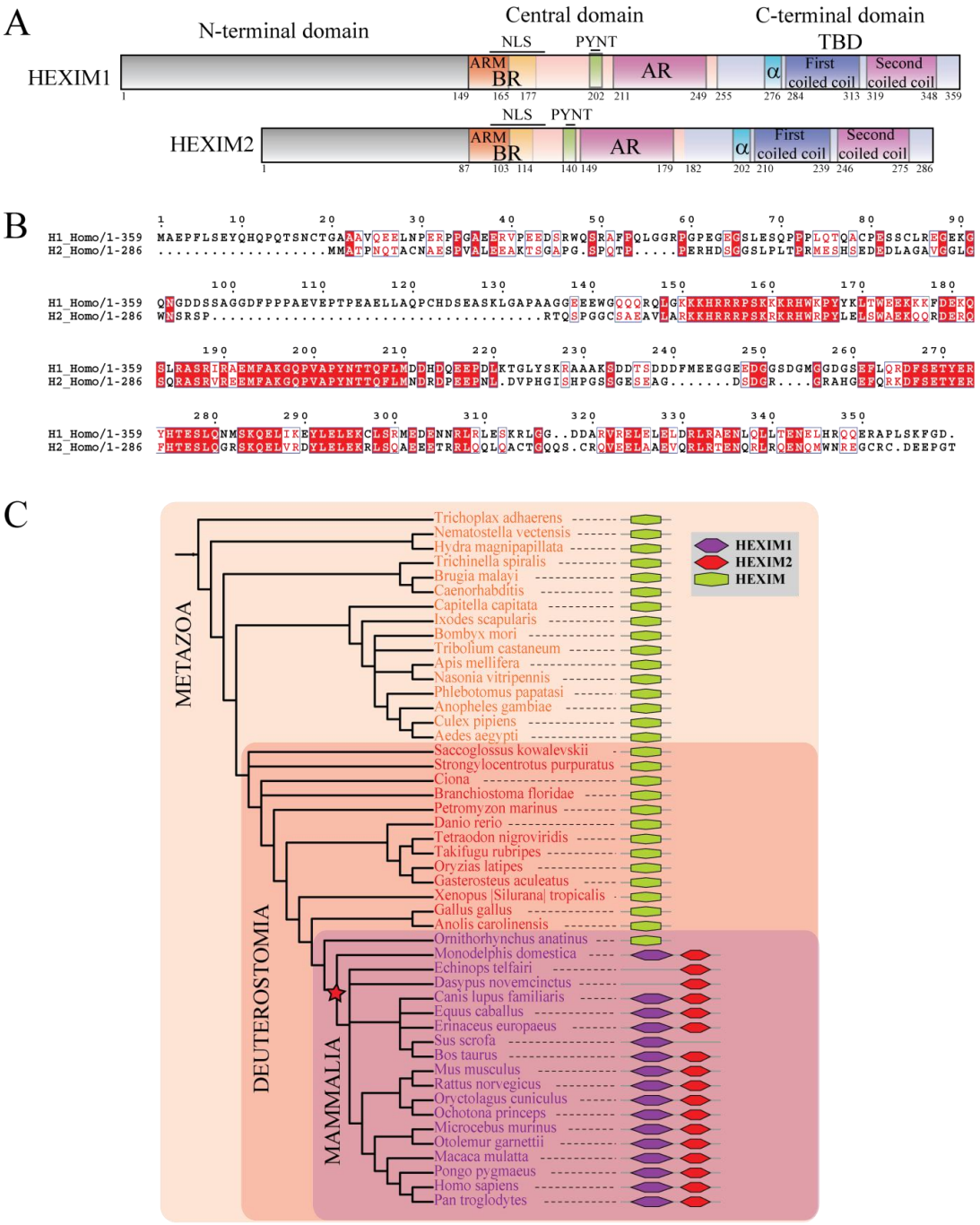



Figure II.4. HEXIM2. A) Schematic representation of HEXIM1 and HEXIM2. B) Sequence alignment of humans HEXIM1 and HEXIM2, identical residues are boxed in red and conservative mutations are boxed in white. C) Phylogenetic distribution of HEXIM protein based on (Marz et al. 2009). Red star indicated HEXIM duplication event. Phylogenetic tree built using iTOL (Letunic et al. 2007, 2011).

Homologs of HEXIM are found across nearly all metazoan (but not in flatworms). It has been proposed that HEXIM1 derived from the reverse transcription of HEXIM2 soon after the origin of mammals (only mammals contain two HEXIM proteins, see Figure II.4C; Marz et al. 2009). All HEXIM proteins present the three characteristic regions: the BR that includes the ARM and the NLS, the PYNT motif, and the coiled coil involved in the interaction with Cyclin T1 (Marz et al. 2009).

## 3. HEXIM1 and HIV1 Tat: Similar Mechanism for a Different Effect?

We saw (Figure II.2) that HEXIM1 contains an ARM (KKKHRRRP, residues 150 to 157) very similar to the ARM of HIV-1 Tat protein (RKKRRQRRR; Figure II.5), followed by a second positively charged sequence (KKKRHWKP) with partial resemblance to the ARM of Tat, and separated by a conserved Ser (Yik et al. 2004). In both proteins, the ARM is involved in the interaction with their RNA partner. Interestingly, the first ARM is not sufficient to mediate HEXIM1 binding to 7SK.

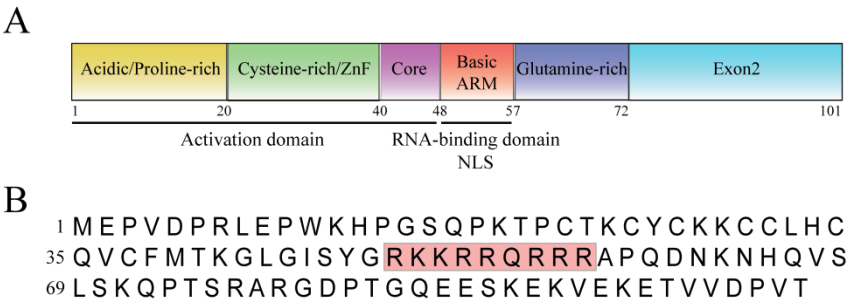



Figure II.5. The HIV Tat protein. A) Schematic representation of the functional domains of Tat. B) Sequence of HIV Tat protein. The ARM is highlighted.

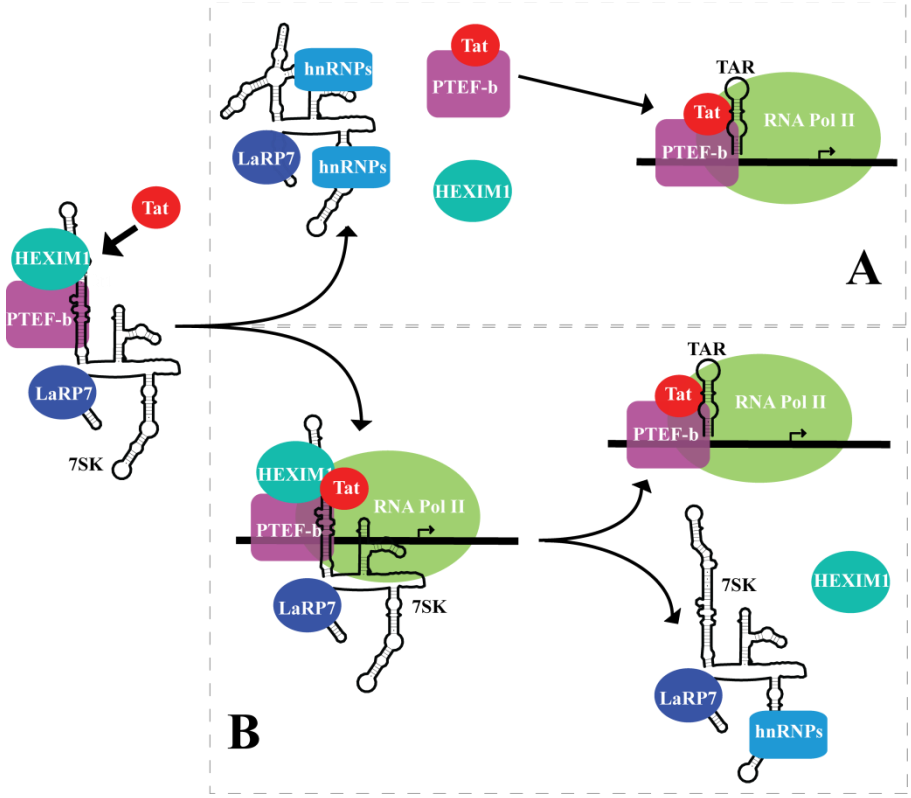



Figure II.6. Two models for the release of P-TEFb from 7SK snRNP mediated by Tat. A) Driven by competition: Tat competes with HEXIM1 for the binding to the Cyclin T1. 7SK undergoes a conformational change upon the loss of P-TEFb, which leads to the release of HEXIM1. The active P-TEFb is then recruited to the HIV-1 promoter by Tat. B) Driven by displacement: the large inactive P-TEFb complex is recruited by Tat to the HIV-1 promoter. As TAR is transcribed, Tat/TAR complex displaces HEXIM1/7SK complex from P-TEFb that turns into the active form.

However, Tat is able to compete with HEXIM1 for binding to 7SK and to disrupt the preformed HEXIM1/7SK complex (Sedore et al. 2007; Krueger et al. 2010; Muniz et al. 2010). Interestingly, Tat and HEXIM1 seem to bind the same region on 7SK (Muniz et al. 2010). Tat displaces HEXIM1 from Cyclin T1 probably because its higher affinity for Cyclin T1, preventing the formation of new HEXIM/7SK/P-TEFb complex, and suppressing the inhibitory effect of HEXIM1 on P-TEFb-dependent transcription (Barboric et al. 2007; Schulte et al. 2005; Sedore et al. 2007; Krueger et al. 2010). Since Tat and HEXIM1 bind the N-terminal domain of Cyclin T1 (Michels et al. 2003; Dulac et al. 2005) in a mutually

exclusive fashion, it has been proposed that both proteins compete for a similar binding surface on Cyclin T1 or that binding of one of them prevents simultaneous binding of the second protein due to steric hindrance or induced structural changes (Schulte et al. 2005; Figure II.6A). This idea is supported by evidences that suggest that the P-TEFb binding domain of Tat is essential and sufficient to cause the release of P-TEFb (Barboric et al. 2007; Krueger et al. 2010). It has been also proposed that 7SK undergoes a conformational change upon this release (Krueger et al. 2010).

However, this mechanism in which Tat drives the release of P-TEFb from 7SK snRNP has been recently challenged and remains controversial. Indeed, Tat, P-TEFb and the 7SKsnRNP, including HEXIM1, were all found to be recruited at the HIV promoter in the absence of TAR (D'Orso et al. 2010). Importantly, 7SK snRNP was released only when TAR was transcribed (Figure II.6B). This suggests an essential function of TAR to displace the 7SK/HEXIM1 from P-TEFb. Supporting this idea, it has been reported that the RNA binding domain of Tat, and not its Cyclin T1 binding domain, is required to release of P-TEFb (Muniz et al. 2010).

Consistent with both ideas, it has been observed that HEXIM1 overexpression inhibits the Tat-mediated transactivation of the HIV-1 LTR and decrease the level of Tat bound to P-TEFb (Yik et al. 2003; Fraldi et al. 2005; Sedore et al. 2007).

On the whole, the discovery of a strikingly similar ARM in Tat and HEXIM1 as well as the similarities of the protein binding sites of 7SK, strongly suggest that a similar architecture sustain both Tat/TAR/P-TEFb and HEXIM1/7SK/P-TEFb ternary complexes.

Recently, the crystal structure of the cyclin box domain of equine Cyclin T1 in complex with the Tat protein from the equine infectious anaemia virus (EIAV) and its corresponding TAR RNA (Anand et al. 2008) as well as the structure of human P-TEFb in complex with HIV-1 Tat (Tahirov et al. 2010) have been solved (Figure II.7). These structures provide structural basis to understand the recognition and specificity of the Tat/TAR/P-TEFb complex and some insights into how Tat functions in the recruitment and activation of P-TEFb. The CDK9/Cyclin T1 interaction surface is smaller than in other CDK/cyclin complexes (such as CDK2/Cyclin A). The crystal structure of Tat/P-TEFb shows that Tat fits into the groove at the heterodimer interface, stabilizing the interaction between CDK9 and Cyclin T1, and leading to a more active P-TEFb complex. The acidic/Pro-rich region of Tat binds as a random coil on the cyclin surface, whereas the Cys-rich and the core regions form a

random coil and two helices. These helices, along with the Cys261 of Cyclin T1, participate in the coordination of two Zn ions. Tat binding induces the disordering of the Cyclin T1 α-helix $H_C$, which exposes a buried surface in Cyclin T1 used for the binding of the Zn ion and the core region of Tat. Interestingly, since the helix $H_C$ is proposed to participate to the interaction with HEXIM1, the authors proposed that its unfolding may contribute to the release of P-TEFb from the 7SK snRNP. The comparison between the free and Tat-bound P-TEFb crystal structures shows that the first cyclin repeat of Cyclin T1 is shifted towards the CDK9 in the bound structure. As a consequence, and to avoid the possible resulting steric hindrance with the H5 helix of Cyclin T1, residues at the loop between β3 and αC of CDK9 change their conformation, pushing away and inducing a conformational change of residues of the β1β2-loop. These changes modify the substrate-binding surface of CDK9. This should affect both the specificity and the efficiency of phosphorylation.

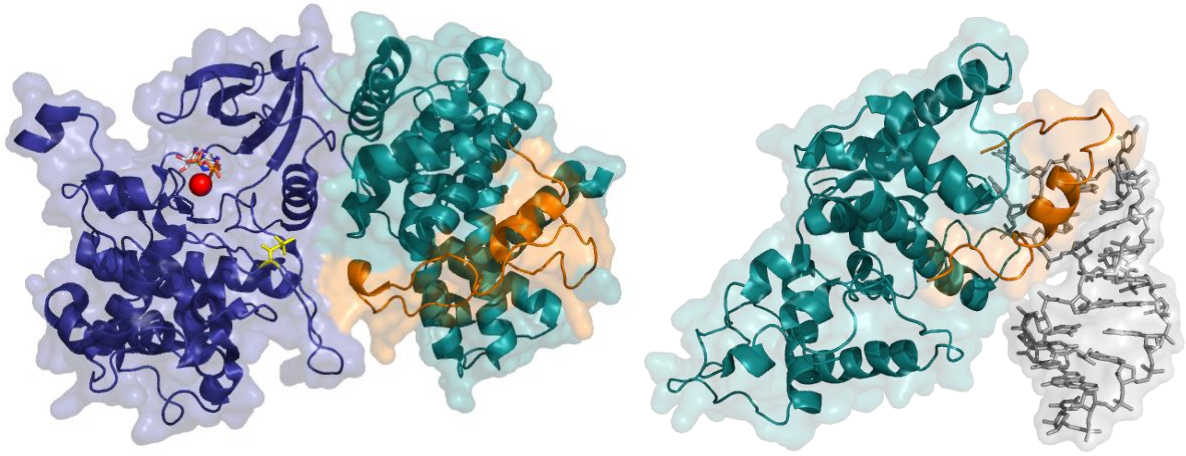


Figure II.7. Crystal structures of Tat complexes. Left, structure of HIV-1 Tat (residues 1 to 49; orange) in complex with human CDK9 (blue) and Cyclin T1 (green; PDB 3MIA). In the active site, a Mg ion (red sphere) and the ATP (coloured red) are shown. Right, crystal structure of the EIAV Tat (residues 41 to 69; orange) with its corresponding TAR RNA (nucleotides 3 to 24; gray) in complex with the equine Cyclin T1 (green; PDB 2W2H) showing how both Cyclin T1 and Tat participate to the recognition of TAR RNA at the level of the loop.

In the crystal structure of the equine Cyclin T1 in complex with the EIAV Tat and TAR RNA, Tat adopts a helical structure. This drives adjacent residues to interact with Cyclin T1. Interestingly, both Tat and Cyclin T1 participate to bind the stem loop structure of TAR. It is difficult however to draw conclusions about the TAR binding effect on the P-TEFb/Tat complex, since the residues involved in complex formation with TAR are different in EIAV Tat/equine Cyclin T1 and HIV-1 Tat/ human Cyclin T1 complexes (Tahirov et al. 2010). Additionally, the equine complex crystal was obtained with a fused construction Cyclin T1-Tat, and in the absence of CDK9. Also, the EIAV TAR RNA sequence is different from the HIV-1 one.

A main difference exists however between Tat/TAR/P-TEFb and HEXIM1/7SK/P-TEFb complexes. Whereas P-TEFb is inactive when bound to HEXIM1/7SK, its activity is induced when bound to Tat/TAR. There is still a lack of structural information of how that works. The three-dimensional structure of the 7SK/HEXIM1 complex should allow us to understand the basis of their recognition and specificity, which are their possible conformational changes, and what makes HEXIM1 an inhibitor of P-TEFb.

However, structural studies of 7SK/HEXIM1 complex represent a challenge because of the intrinsic flexibility of 7SK snRNA and the low content of secondary structure of HEXIM1. A characterization of a minimal but specific 7SK/HEXIM complex was necessary as a starting point for structural studies. This also implied the determination of the stoichiometry of the minimal 7SK/HEXIM complex.

Next chapters will describe the results of a biochemical and biophysical characterization of 7SK/HEXIM complex in the pursuit of a minimal complex for structural studies. It also outlines how the different molecules used for this work were designed; the choice of target molecule, cloning, expression, purity assessment in terms of chemical and conformational homogeneity will be discussed.

# CHAPTER III:
# MOLECULES PREPARATION

The choice of the target molecule is a key point for the structural study of biomacromolecules. For several techniques, full length 7SK and HEXIM1 were not appropriate or difficult to handle. Hence, different constructions of subdomains of these macromolecules were designed to overcome this problem. Furthermore, smaller and more compact substructures should be more suitable for successful crystallization. Each of the macromolecules designed was experimentally tested, and several parameters were considered for their selection. The most important of them was obviously the preservation of the biological function. Other fundamental aspects were the stability and the solubility.

Sample preparation is also an essential step since the quality of the biomacromolecules depends on the way they are prepared. In structural biology, large quantities of sample with a high purity degree are usually required.

In this work, different problems were faced for each of the biomacromolecule prepared. In general, several of our RNAs presented conformational heterogeneity. For proteins, some of the confronted problems were degradation, aggregation, and RNAses removal.

## 1. RNA PRODUCTION

### 1.1. T7 transcription

Most of the RNAs used for my pHD project were synthesized by T7 in vitro transcription. The T7 RNA polymerase is a single subunit DNA-dependent enzyme, capable of transcribing a complete gene without the need for additional proteins (Cheetham et al. 2000). Furthermore, T7 RNA polymerase has a stringent specificity for its own promoters, which contain a highly conserved sequence of 23 continuous base pairs including the start site for the RNA

(Davanloo et al. 1984). These features make it interesting and useful for biochemists. To design the sequence that will serve as a template for T7 RNA polymerase, it has to be taken into account that six of the conserved base pairs belong to the transcribed region and mutations in this region affect RNA synthesis (Kochetkov et al. 1998). A G at the position +1 and +2, generally ensures an optimal transcription. Since the human 7SK sequence starts at its 5' end by GG, the wild type human 7SK could be prepared by T7 in vitro transcription. However, to improve the performance of the transcription, most other of RNAs constructs (like stem-loop substructures) synthesized by T7 in vitro transcription were mutated to provide a 5'-GG at positions +1 and +2..

At the 3' end of the transcripts, T7 RNA polymerase can add non-template encoded nucleotides (Galperin et al. 2009). To overcome this problem we cloned most RNA constructs into the pHDV vector (Walker et al. 2003). Using this vector, a modified Hepatitis Delta Virus (HDV) ribozyme is transcribed at the 3'-end of the target RNA sequence. Since HDV ribozyme self-cleaves 5'end at the G +1 nucleotide, the transcripts are produced with homogeneous 3'-ends (see Figure A.1 in Annexes 1).

## 1.2. Templates

Three different sources of templates were used: plasmids, PCR products and synthetic oligonucleotides. Templates from linear plasmids were privileged for the production of currently used RNAs because once the correct sequence of a plasmid cloned has been confirmed, the DNA can be amplified by in vivo plasmid replication exploiting the high fidelity of bacterial DNA polymerase. A template generated by PCR was convenient as a quick way to test different substructures of 7SK deleted at its 3'end before cloning into a plasmid using a high fidelity DNA polymerase. Finally, the synthetic oligonucleotide (limited to ~120 nucleotides) template was generally used to generate the different versions of the stem-loop substructures of 7SK and test them also before cloning.

The transcription protocol adopted was adequate to produce several milligrams of RNA from 5 ml of transcription reaction, the standard volume used in the laboratory. RNAs produced from synthetic oligonucleotides were generally transcribed with a high yield by this protocol.

| Table III.1. 7SK constructions | | | | | |
|---|---|---|---|---|---|
| **RNA** | **Description** | **Template** | **System** | **Yield (1 step)*** | **Yield (2 steps)*** |
| **7SK** | Full length | Linear pHDV | in vitro | 0,7 mg | 0,2 mg |
| **7SKΔHP4** | Nucleotides 1 to 295 | Generated by PCR | in vitro | 0,4 mg | <0,1 mg |
| **7SKΔHP1** | Nucleotides 24 to 87 deleted | Linear pHDV | in vitro | 0,7 mg | 0,2 mg |
| **7SKΔ9** | Nucleotides 9 to 332 | Linear pHDV | in vitro | 0,6 mg | 0,1 mg |
| **7SK1-195** | Nucleotides 1 to 195 | Generated by PCR | in vitro | NE | |
| **7SKM1** | Nucleotides 290 to 295 mutated into UGUAGG | Linear pHDV | in vitro | 0,4 mg | <0,1 mg |
| **7SKΔD2** | Nucleotides 93 to 171 deleted | Generated by PCR | in vitro | NE | |
| **IL2** | Nucleotides 216 to 255 mutated into GAAA | Linear pHDV | in vitro | 0,5 mg | 0,2 mg |
| **IL3** | Nucleotides 206 to 269 mutated into A | Linear pHDV | in vitro | <0,1 mg | |
| **HP1L** | Nucleotides 1 to 108 | Generated by PCR | in vitro | 0,6 mg | |
| **HP1** | Nucleotides 24 to 87, nucleotides 25 and 26 mutated in G and 85 and 86 into C | Linear pHDV | in vitro | 0,35 mg | 0,3 mg |
| **HP1u** | HP1 with the apical loop mutated into UUCG | Linear pHDV | in vitro | 0,6 mg | 0,4 mg |
| **HP1a** | HP1 with the apical loop mutated into GAAA | Linear pHDV | in vitro | 0,5 mg | 0,35 mg |
| **KS1** | HP1 inserted in tRNA | pKSA | in vivo | 4 mg | |
| **KE1** | HP1 inserted in tRNA including a sephadex aptamer | pKSA | in vivo | 4 mg | |
| **KS1HP1u** | HP1u inserted in tRNA | pKSA | in vivo | 4 mg | |
| **KS1HP1a** | HP1a inserted in tRNA | pKSA | in vivo | 4 mg | |
| **LIL2** | Nucleotides 1 to 23 mutated into G and deletion of 195 to 332 | Generated by PCR | in vitro | <0,1 mg | |
| **L3L4** | Nucleotides 14 to 204 mutated into CUUG | Generated by PCR | in vitro | 0,25 mg | |
| **HP3** | Nucleotides 201 to 273; nucleotides 202 and 272 G and C, respectively | Linear pHDV | in vitro | 0,4 mg | 0,2 mg |
| **KS3** | HP3 inserted in tRNA | pKSA | in vivo | 4 mg | |
| **Dom2** | Nucleotides 88 to 190 and 88 to 90 mutated into GGG | Synthetic oligonucleotide | in vitro | 1 mg | |
| **L3** | Nucleotides 214 to 289 ; nucleotides 215 and 288 mutated into G and C, respectively | Synthetic oligonucleotide | in vitro | 1,4 mg | |

* For in vitro in mg/ml of transcription; for in vivo in mg/L of bacteria culture.

NE: not estimated. See annexes for secondary structures.

For all constructions but HP1 (see Table III.1), the self-cleavage of HDV ribozyme was 100% efficient. In the case of HP1, even upon longer incubations at 37°C with 40 mM $MgCl_2$ the efficiency was around 50% (Figure III.1). It has been shown that the self-cleavage efficiency of HDV ribozyme is sensitive to a number of factors, including its flanking regions (Chadalavada, Cerrone-szakal, & Bevilacqua, 2007; Chowrira et al., 1994). Flanking RNA sequences may pair nucleotides in the ribozyme sequence preventing its proper folding and therefore self-cleavage.

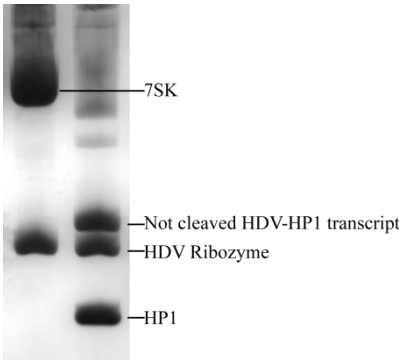



Figure III.1. Ribozyme self-cleavage. Comparison between 7SK and HP1 transcripts upon ribozyme self-cleavage treatment.

## 1.3. RNA purification

When I arrived to the laboratory, a protocol of RNA purification consisting of a desalting G25 sepharose chromatography, a DNAse treatment and a G200SW chromatography was used (Figure III.2). However, this protocol was long, and the RNA was more exposed to the RNAses activity. Furthermore, the DNAse treatment proved to be sometimes inefficient. For this reason we adopted the gel purification system, cheap and faster (several gels could be done in parallel), and convenient to purify the transcript product of 5 ml transcription reaction on one gel. Adjusting the percentage of acrylamide, RNAs of different lengths could be separated from the plasmid template, ribozyme and abortive transcripts. After electrophoresis, RNA band was identified by UV shadowing, without any staining, and excised. RNA was eluted passively, since this system preserved the RNA without degradation.

After gel purification some acrylamide was retained even after ethanol precipitation. The removal of acrylamide was essential to avoid damaging of the MonoQ column and for a good quality and better yield of the purification in general. For this reason, a double filtering system was adopted consisting of a glass wool and a Minisart filter.

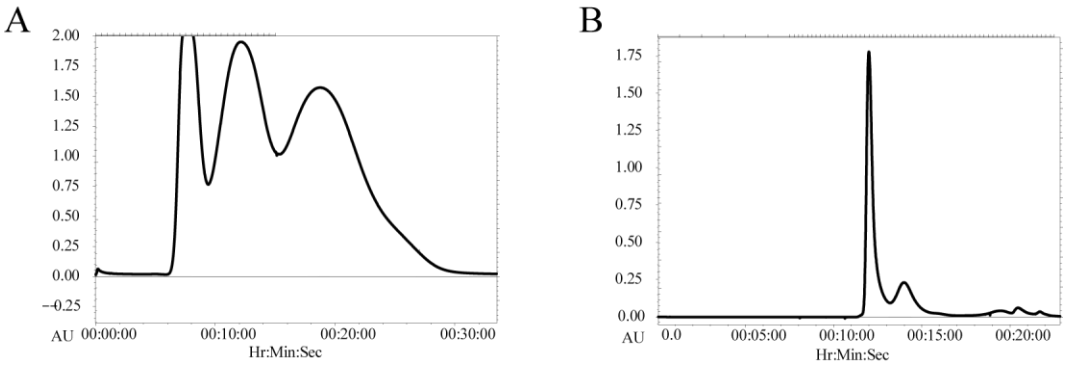



Figure III.2. 7SK purification by chromatography system. A) G25 chromatography removes rNTPs, and part of the plasmid template. B) G2000SW separates 7SK from the HDV ribozyme.

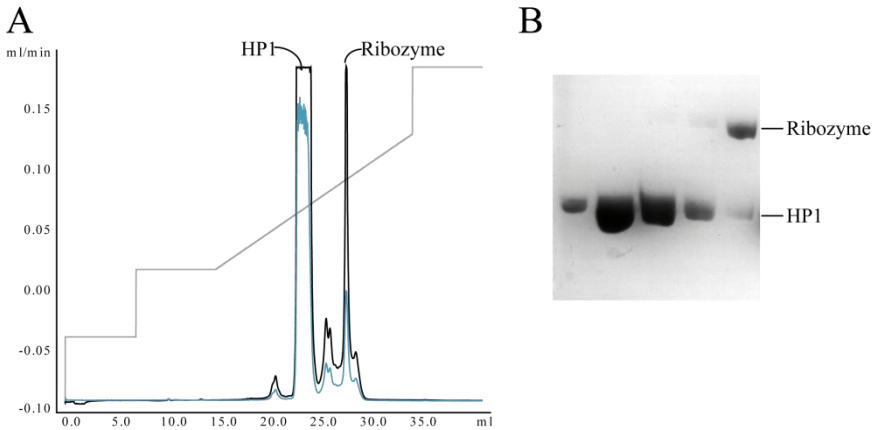



Figure III.3. MonoQ chromatography. A) Typical Mono Q chromatogram of HP1. Absorbances at 280nm (blue) and 260nm (black), and NaCl gradient (gray) are shown. B) Analytical polyacrylamide gel from the MonoQ peaks.

MonoQ is a high resolution ion exchange column. In ion exchange chromatography, the adsorption of the molecules to the solid support is driven by the ionic interaction between

the oppositely charged ionic groups in the sample molecule and the functional ligand in the support. Elution is achieved by increasing the ionic strength. Separation is obtained since different substances have different degrees of interaction with ion exchanger due to differences in their charges, charge densities and distribution of charge on their surfaces. RNAs are eluted depending on their overall (phosphate) negative charge per molecule, this is according to their size. The MonoQ chromatography was important to improve the quality and homogeneity of all our RNAs and an essential step for crystallization. For small RNAS, MonoQ chromatography also allowed to completely remove the ribozyme after gel purification (Figure III.3). However, for some RNAs, particularly full length 7SK and long constructs, the yield of this step was considerably poor, so MonoQ was avoided when possible.

## 1.4. Thermal treatment

7SK and many of its substructures showed different conformations. These conformations were detected by agarose gel, gel filtration, and even by MonoQ chromatography (see Figure A.7 in Annexes 1).

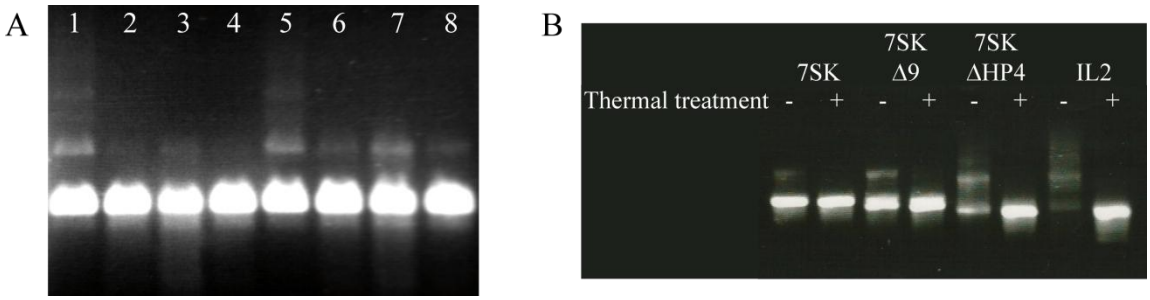



Figure III.4. Thermal treatment. A) Different thermal treatments were tested to obtain a single conformation of 7SK visualized by agarose gel: (1) not treated; (2) 2min at 85°C then 15min at 37°C; (3) 2min at 85°C then slow cooling to 20°C; (4) 2min at 85°C then fast cooling to 0°C for 5min; (5 to 10) dialysis overnight then treatment as 1 to 4, respectively. B) Agarose gel showing some 7SK constructs before and after thermal treatment.

Hence, a thermal treatment was required to achieve a single conformation. The conditions for the thermal treatment were experimentally established (see Annexes 1). The best treatment consisted in 1 minute heating at 85 °C, then fast cooling on ice for 5 minutes in presence of 2 mM $MgCl_2$. Treatments at higher temperatures or $MgCl_2$ concentrations led to RNA degradation (Figure III.4). This treatment was systematically performed after ethanol precipitation and after freezing of the longer constructs of 7SK.

## 1.5. RNAs produced in vivo

Because milligrams quantities of RNAs were required for SAXS and crystallization, in vivo production of RNA was an attractive, less costly way to obtain large quantities of RNA. This system allows producing recombinant RNA in *Escherichia coli* by exploiting a tRNA as a scaffold to disguise it as a natural RNA and thus to hijack the host machinery, escaping cellular RNAses. The method makes uses of a pBSTNAV vector which was initially used for tRNA overproduction in *Escherichia coli* (Meinnel et al. 1988). pBSTNAV includes a synthetic lipoprotein promoter upstream to the tRNA gene, and an rrnC transcription terminator downstream. This vector was modified to enable the insertion of a target RNA in the anticodon stem of tRNA (Ponchon et al. 2007). The tRNA chimera is recognized by cellular enzymes that precisely process the primary transcript, incorporate modified nucleotides and repair their 3' end.

Since the desired RNA is inserted in the anticodon stem, only stem-loop RNA structures are suitable for this system. Hence, HP1, HP1u, HP1a and HP3 (see Table III.1) were cloned and produced using this system, giving rise to a pKSA_HP or pKE1_HP.

This system allowed purifying several milligrams of RNA from standard cultures of *Escherichia coli* without induction. Furthermore, the tRNA scaffold could be used to orient HP1 constructs during SAXS analysis (see Chapter VIII) since tRNA structure is already know. For the same reason, these constructs were interesting for crystallization trials (see Chapter IX).

Table III.2 summarizes some advantages and disadvantages of the T7 in vitro and in vivo production systems.

| Table III.2. In vitro vs in vivo RNA production | | |
|---|---|---|
| **System** | **T7 in vitro transcription** | **In vivo production** |
| **Advantages** | • Simple and fast<br>• No restriction in length or conformation of RNA<br>• Adaptable for the incorporation of modified nucleotides | • High yield from 1L culture<br>• Less costly<br>• tRNA modification in vivo stabilizes the production |
| **Disadvantages** | • Yield depends on template<br>• More expensive | • Restricted to RNAs with 5' and 3' ends paired<br>• tRNA scaffold may interfere with RNA function<br>• Demands cloning |

## 2. Proteins Production

During my pHD, I mainly investigated the interaction between 7SK and the human HEXIM1, but I also explored the interaction of 7SK with other of its partners like UP1 and LaRP7, and conducted some kinase activity test of P-TEFb using a fragment of the CTD of PolII. Different constructs of HEXIM1 were designed, their stability during purification evaluated and tested for crystallization (see Table III.3).

### 2.1. Vectors

Human HEXIM1 gene cloned into a pET21 was a kind gift of Olivier Bensaude. The HEXIM1 sequence was then cloned in the laboratory into a pET28, the construction currently used. This construction includes a $(His)_6$ tag fused to HEXIM1 sequence at its C-terminal position linked by Lysine-Glutamate amino acid residues. Using this clone, different versions of HEXIM1 were designed and cloned into a pET-MCN vector (Diebold et al. 2011; see Figure III.5). This required mutation of the BamHI site of HEXIM1 at position 565 (amino acids 182).

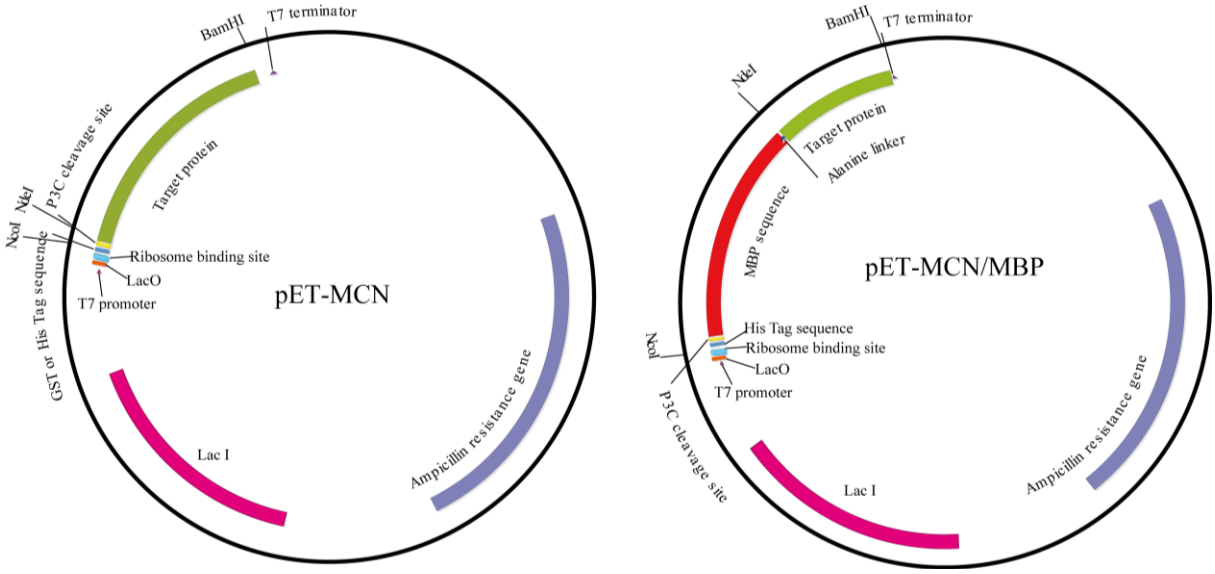



Figure III.5. Protein vectors. A circular map of pET-MCN and pET-MCN/MBP, showing their main features and the endonuclease restriction sites used to clone the different constructions.

The use of pET-MCN vector had several advantages:

1. Affinity purification tag position and/or nature are changeable
2. Protease cleavage site position and/or character are also changeable
3. Compatible for co-expression
4. NdeI and BamHI restriction sites act as universal restriction sites to clone the target sequence in any vector of the pET-MCN series.

## 2.2. Tags

Milligrams quantities of highly purified protein are typically required for structural biology studies, particularly for crystallization. The $(His)_6$ tag was privileged in our laboratory because high yields of purified protein were obtained after one chromatography step, and cleavage of small tags is often not required to grow suitable crystals. However, large-affinity tags like Maltose Binding Protein (MBP) or Glutathione-S-Transferase (GST), may provide an advantage by enhancing the solubility and expression of the target protein (Hammarström et al. 2002). These large affinity tags have been also used as a strategy to crystallize protein

not readily crystallizable (Smyth et al. 2003; Moon et al. 2010), since these carrier proteins can provide molecular surfaces that are favourable to crystal lattice formation. Recently, a MBP containing different surface mutations to reduce surface entropy (SER) and encourage crystal lattice formation has been designed (Moon et al. 2010). Three-dimensional structures of proteins fused to this MBP construction (MBP/SER) have been solved. Furthermore, MBP protein structure was used to solve the crystallographic phase problem by molecular replacement.

Since crystallization attempts of different versions of HEXIM1 failed, this MBP/SER system was used to improve our chances of a successful crystallization. The MBP/SER clone in the pETXM1 vector was kindly provided by Sebastian Charbonier and Yves Nominé. In a first period MBP/SER was fused to the N-terminus of the ARM of HEXIM1 by a linker consisting of three Alanine residues.

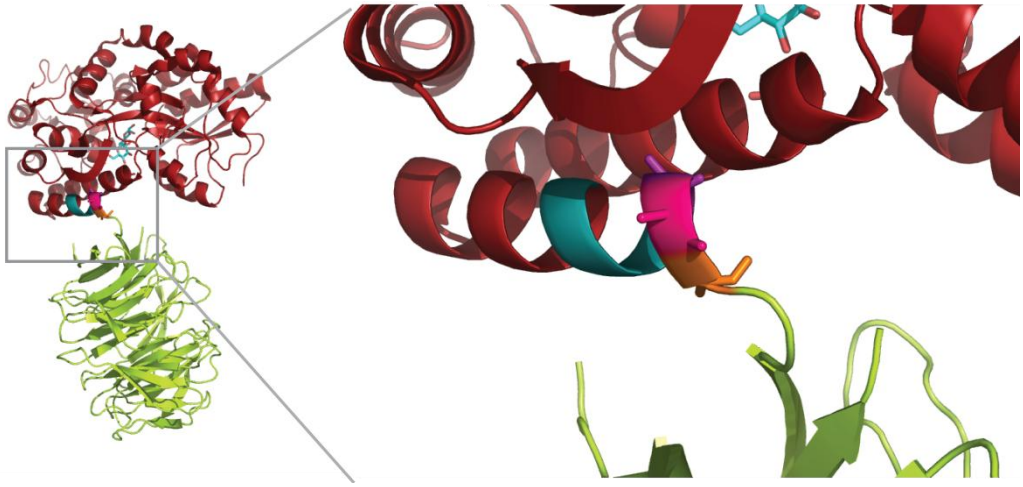


Figure III.6. Linker connecting MBP and target protein. At left, the three dimensional structure of MBP (red) fused to the Rack1 protein (green). At right, zoom of the linker region highlighting the three Ala (in violet, pink and orange, respectively) which are part of the C-terminal $\alpha$-helix of MBP (PDB 3M0).

Later, to benefit of the advantages of the pET-MCN system, MBP/SER was cloned into a pET-MCN vector between the NcoI and NdeI restriction sites (see Figure III.5). In this way, any protein could be cloned using the NdeI and BamHI restriction sites, resulting in a MBP protein fused at the N-terminus of the target protein. A $(His)_6$ tag linked by a 3C

protease cleavage site at the N-terminus of MBP allows purifying the whole construction by His-tag affinity chromatography, more efficient and less expensive than purification by amylose resin. The length of the linker region connecting the MBP to the target protein has been shown to be important for a successful crystallization (Moon et al. 2010). If the linker is too long, the fusion protein may have undesired conformational flexibility, but if it is not long enough the connection to the C-terminal $\alpha$-helix may disrupt the structure of the downstream target protein (Figure III.6). To test this parameter, two different linker length were generated consisting of one or two Ala residues followed by His-Met residues, product of the translation of the NdeI recognition site.

## 2.3. Design of constructions

The design of the boundaries for the different HEXIM1 constructions were based on the secondary structure predictions (see Chapter II "Molecular description of HEXIM1"), limited proteolysis experiments, characterization of physicochemical domains (Figure III.7), and from the knowledge of the functional regions of HEXIM1.

Hence, the N-terminal boundaries were:

- Q120, this mutant was a kind gift of Olivier Bensaude.
- E114, based on secondary structure predictions that suggest a $\alpha$ helix from 114 to 119.
- G136, based on the characterization of physicochemical domains of HEXIM1. The plot (Figure III.7) shows a kink at this position indicating the end of the N-terminal physicochemical domain. Also, G136 is located before a predicted $\alpha$ helix.
- G149, the beginning of the ARM of HEXIM1.

And the C-terminal boundaries were:

- G317, end of the first segment of the coiled coil of the TBD.
- R273, position before the short $\alpha$ helix of the TBD.
- M255, end of the central domain; G259, position just before the predicted coiled coil.
- Y225, a chymotrypsine cleavage site seen by limited proteolysis (Figure III.7).

- D179, end of the BR

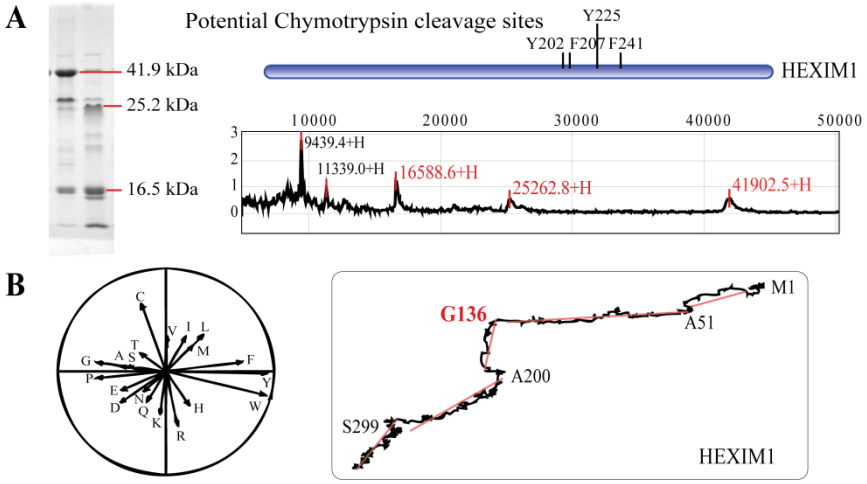



Figure III.7. Design of boundaries for HEXIM1 constructions. A) Analysis of limited chymotrypsine proteolysis by SELDI (Surface-enhanced laser desorption/ionization). Mild chymotrypsine digestion produces two peptides with MW of 25.2 and 16.5 kDa that could correspond to a cleavage at Y225. B) Characterization of physicochemical domains of HEXIM1 (by Olivier Poch). Left, vector diagram for the amino acids where the length and the direction depend on their physicochemical properties. Right, the resulting plot for HEXIM1.

The different HEXIM1 constructions used during this work are schematically summarized in the Figure III.8 (and see Table III.3).

## 2.4. Expression

With the exception of P-TEFb, these recombinant proteins were over expressed in BL21(DE3) strain of *Escherichia coli*. BL21 is a $Met^+$ derivative of B834 (a restriction-modification defective, galactose-negative, methionine auxotroph of *E. coli* B). DE3 lysogens contain a derivative of phage lambda that supplies T7 RNA polymerase by transcription from the lacUV5 promoter in the chromosome.

**Table III.3. Summary of purification protocol for protein constructions**

| Protein | MW (kDa) | Tag | Tag Loc | Purification 1 | | Purification 2 | | Purification 3 | | Purification 4 | | Purification 5 | | Yield |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Step | Buffer | Step | Buffer | Step | Buffer | Step | Buffer | Step | Buffer | |
| HEXIM FL | 41.7 | $(His)_6$ | C-Ter<br>N-Ter | Ni-NTA | 1 + 1.4 mM βM | S200 | 9 + 2mM DTT | SP | 16 + 2mM DTT 0.25 to 0.5M NaCl | | | | | 3-6 |
| HEXIM 120-359 | 28.9 | $(His)_6$ | C-Ter | Ni-NTA | 2 | S200 | 10 + 7mM βM | | | | | | | 3-5 |
| HEXIM 136-359 | 27.2 | $(His)_6$ | N-Ter | Ni-NTA | 3+1.4 mM βM | S200 | 9 + 2mM DTT | | | | | | | 3-6 |
| HEXIM 114-317 | 25.5 | $(His)_6$ | N-Ter | Ni-NTA | 3 + 1.4 mM βM | S75 | 11 | | | | | | | NS |
| | 52.1 | GST | N-Ter | GSH | | | | | | | | | | 1-2 |
| HEXIM 114-273 | 20 | $(His)_6$ | N-Ter | Ni-NTA | 3 | S75 | 11 | | | | | | | NS |
| | 46.6 | GST | N-Ter | GSH | | | | | | | | | | 1-2 |
| HEXIM 114-255 | 17.9 | $(His)_6$ | N-Ter | Ni-NTA | 2 | | | | | | | | | NS |
| | 44.5 | GST | N-Ter | GSH | | P3C | 12 + 7 mM βM | | | | | | | < 2.5 |
| HEXIM 114-225 | 14.8 | $(His)_6$ | N-Ter | Ni-NTA | 2 | | | | | | | | | NS |
| | 41.4 | GST | N-Ter | GSH | | P3C | 12 + 7 mM βM | | | | | | | < 1.5 |
| HEXIM 136-273 | 18.7 | $(His)_6$ | N-Ter | Ni-NTA | 4 | S200 | 4 | | | | | | | 3-6 |
| | 56.6 | $(His)_6$ MBP | N-Ter | Ni-NTA | 5 | P3C (Optional) | 13 | Phenyl | 17 1 to 0M $(NH_4)_2SO_4$ | Dialysis + SP | 19 0.1 to 1M NaCl | S200 | 16 | ~2 |
| HEXIM 136-179 | 48.4 | $(His)_6$ MBP | N-Ter | Ni-NTA | 6 | P3C (Optional) | 14 | Q | 15 0.1 to 1M NaCl | S75 | 16 | | | 3-5 |
| HEXIM 149-259 | 55.3 | $(His)_6$ MBP | N-Ter | Ni-NTA | 6 | P3C (Optional) | 14 | Q | 15 0.1 to 1M NaCl | S75 | 16 | | | 3-5 |
| HEXIM 149-179 | 44.4 | MBP | N-Ter | Amylose | 7 | Q | 15 + 20mM Maltose 0.1 to 1M NaCl | S200 | 16 | | | | | 1-2 |
| UP1 | 34.2 | GST | | GSH | 8 | Thrombin | 8 | S200 | 18 | | | | | ~10 |

MW including tag; Yield in mg/L of bacteria culture; GSH: Glutathione sepharose; βM: β–mercaptoethanol; NS: No soluble.
Buffers: **(1)** 50mM Tris pH 8, 500mM NaCl, 5mM $MgCl_2$; **(2)** 50mM Tris pH 8, 500mM KCl, 10mM $MgCl_2$; **(3)** 50mM Tris pH 8, 500mM NaCl, 10mM $MgCl_2$; **(4)** 50mM Tris pH 8, 500mM NaCl, 6mM $MgCl_2$; **(5)** 100mM Tris pH 8, 500mM NaCl; **(6)** 20mM Tris pH 8, 500mM NaCl; **(7)** 20mM Tris pH 8, 250mM NaCl; **(8)** 50mM Tris pH 7.6, 500mM KCl; **(9)** 50mM Tris pH 7.6, 500mM NaCl, 5mM $MgCl_2$; **(10)** 20mM Tris pH 8, 200mM KCl, 2mM $MgCl_2$; **(11)** 20mM Tris pH 8, 500mM KCl, 2mM $MgCl_2$; **(12)** 50mM Tris pH 8, 250mM KCl, 5mM $MgCl_2$, 0.5mM EDTA; **(13)** 100mM Tris pH 8, 100mM NaCl; **(14)** 20mM Tris pH 8, 100mM NaCl; **(15)** 20mM Tris pH 8.5; **(16)** 50mM Tris pH 7.6, 5mM $MgCl_2$; **(17)** 100mM Tris pH 8, 0.5mM EDTA; **(18)** 50mM Tris pH 7.6, 200mM KCl; (19) 20mM NaMES pH 6.

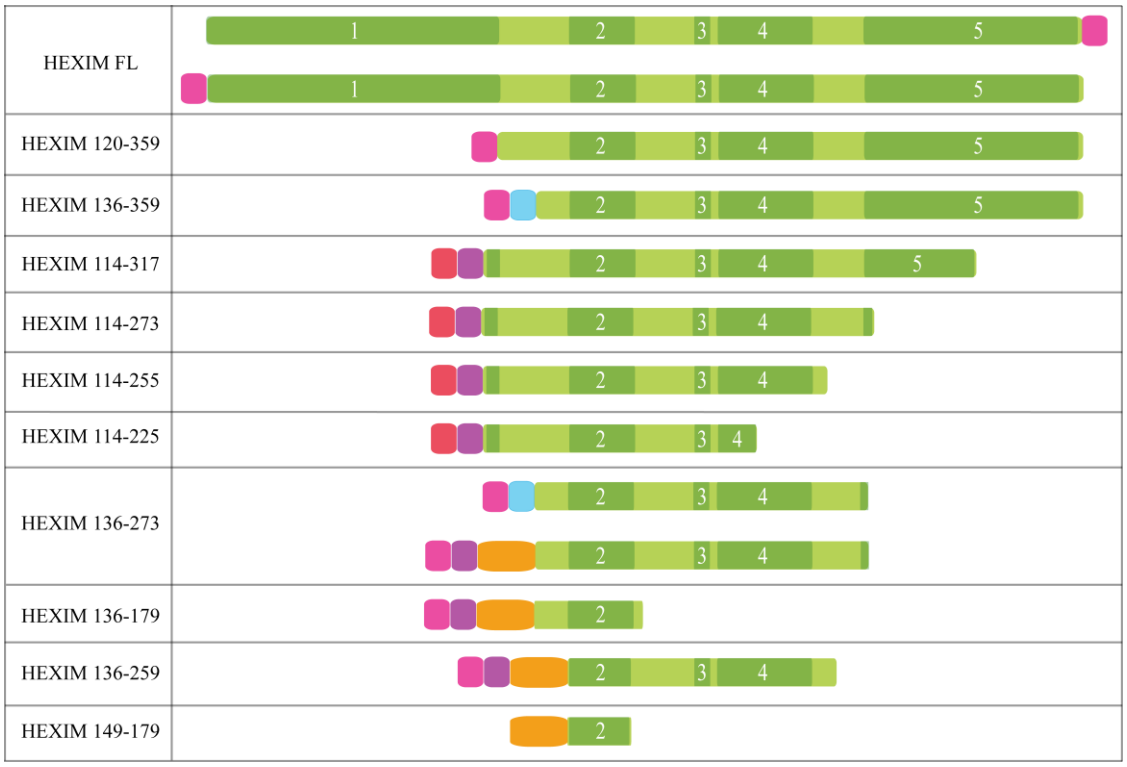



Figure III.8. HEXIM1 constructions. Schematic representation of all HEXIM constructions used during this work. The different HEXIM1 domains are highlighted: N-terminal domain (1), ARM (2), PYNT motif (3), AR (4), and TBD (5). $(His)_6$ tag (pink), $(His)_6$ or GST tag (red), MBP tag (orange), Thrombin cleavage site (blue), and 3C cleavage site (violet) are shown.

Since a functional P-TEFb depends on several post-translational modifications of the CDK9 and its Cyclin T1 associated (Kohoutek 2009), it was produced in insect cells infected by a Baculovirus vector.

The auto-induction protocol, originally developed by Studier (Studier 2005), was used to grow the cells. The auto-induction protocol provides for the expression of proteins without the need to add inducers such as IPTG during mid-log phase of cultures. The method is based upon a buffered medium that contains a mixture of carbon sources, including lactose in limited amount. The bacteria initially use glucose; when glucose is exhausted, lactose can enter the cell and induce expression of the T7 polymerase from the DE3 lambda lysogen. Then the translational machinery of bacteria is used to overexpress the recombinant protein.

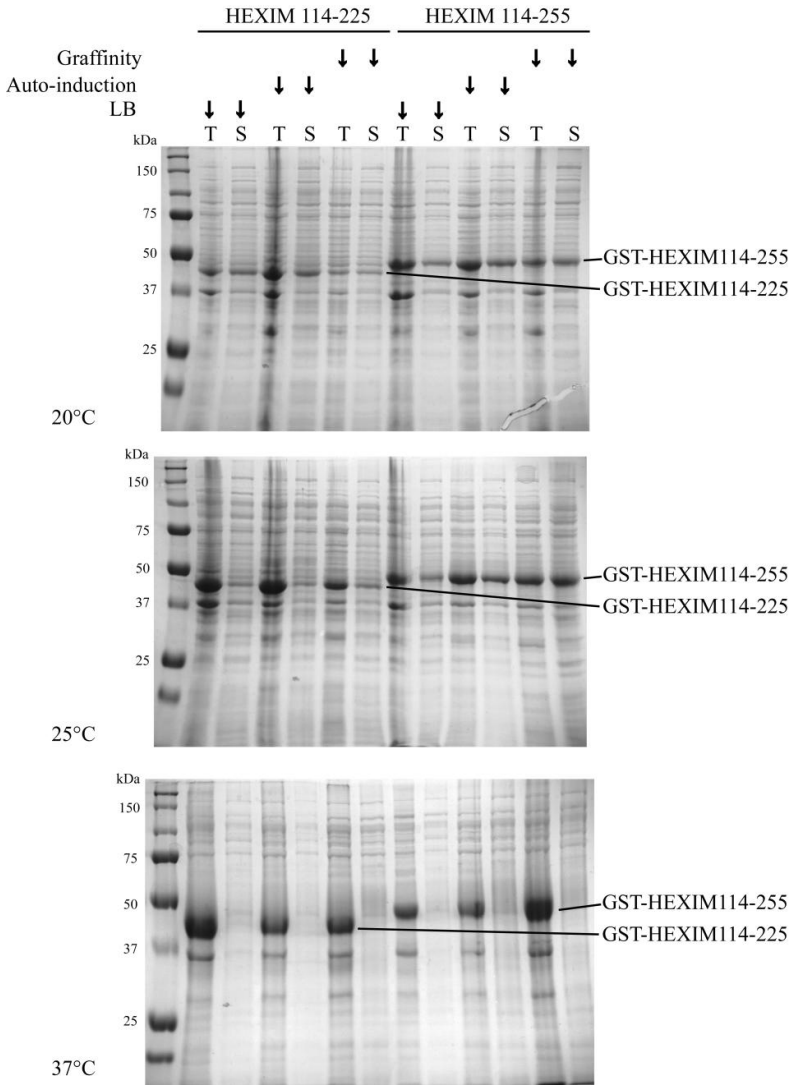



Figure III.9. An example of overexpression and solubility test of proteins. Denaturing gels were used to monitor the overexpression and solubility of proteins. Different media and temperatures were tested as indicated. The total (T) and soluble (S) fractions are shown.

The cultures were incubated at 25°C in a shaker for approximately 18 hours. Auto-induction medium and incubation at 25°C were the standard conditions used in the laboratory. For proteins with low level of overexpression or solubility, different media and incubation temperatures were tested. An example of the importance of the media selection and the temperature of growth for protein solubility is shown in the Figure III.9.

## 2.5. Purification

Proteins can be purified using purification techniques that separate according to specific properties. The Table III.4 shows the different techniques used during this work.

**Table III.4. Chomatography**

| Property | Technique |
|---|---|
| Biorecognition (ligand specificity) | Affinity chromatography |
| Charge | Ion exchange chromatography |
| Hydrophobicity | Hydrophobic interaction chromatography |
| Size | Gel filtration |

### a. Affinity chromatography

Affinity chromatography separates proteins on the basis of a reversible interaction between a protein and a specific ligand coupled to a chromatography matrix (Ad 2002a). In a single step, target molecules can be purified from complex biological mixtures. Proteins without substantial affinity for the ligand will pass directly through the column, whereas one that recognizes the ligand will be retarded in proportion to its affinity constant. Elution of the bound protein is achieved by changing such parameters as salt concentration or pH, or by addition of a competitor ligand in solution (Cuatrecasas et al. 1968).

To facilitate the purification by affinity chromatography, all the proteins used for this work were fused to an affinity tag, as previously described. The tag binds strongly and selectively to an immobilized ligand on a solid support, and contaminants are washed away. Affinity chromatography typically yields purities >90% in a single column step (Fong et al. 2010). Most of our constructions have a $(His)_6$ tag. The Histidine interacts with immobilized $Ni^{2+}$ ion in the matrix, as electron donor groups on Histidine imidazole ring readily form coordination bonds with the immobilized transition metal. The protein can be easily eluted by adding free imidazole. Because the relatively small size and charge of this tag, the protein activity is rarely affected. Besides, many proteins with $(His)_6$ tag have been deposited in the Protein Data Bank (Terpe 2003). The removal of this tag was a parameter tested only for crystallization.

Another tag used in this work is the Glutathione S-transferase (GST). The GST interacts with the immobilized glutathione in the matrix, and can be eluted with gluthatione in

solution. This was used because it can help to stabilize the recombinant protein (Terpe 2003). In contrast with $(His)_6$ tag, the GST tag (25 kDa) was systematically removed by protease cleavage. In general, the tag was removed by proteolysis on the affinity resin. Some proteins precipitated or were degraded after protease cleavage (see Table III.3). For some proteins, elution by protease cleavage resulted in a purified protein as shown in Figure III.10, since some unwanted proteins can interact with the affinity matrix. However, because protease cleavage was inefficient for some proteins and long incubations were required, we did not perform it systematically.

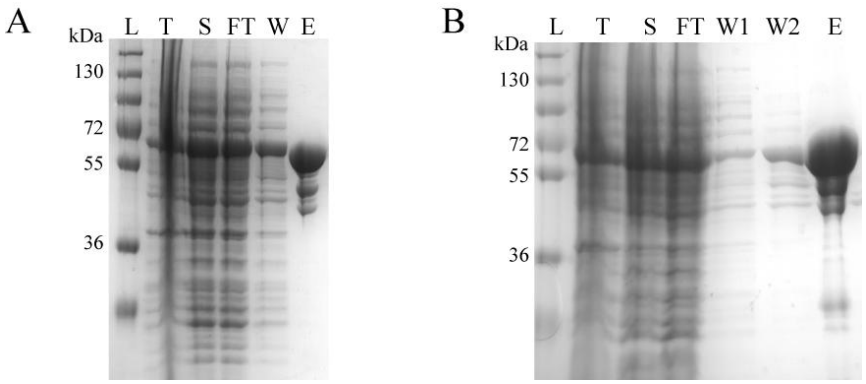



Figure III.10. Two methods for elution from affinity matrix. Gel showing His-MBP-HEXIM136-273 eluted by protease cleavage (A) or by addition of imidazole (B). Lanes correspond to ladder (L), total (T), soluble (S), flowthrough (FT), wash at high salt concentration (W or W1), wash at 20 mM Imidazole and elution (E) fractions.

Since HEXIM, LaRP7 and UP1 are RNA binding proteins, they bound bacterial RNAs non-specifically. To prevent this problem, cell lysis and affinity chromatography were typically performed at high ionic strength. A determinant step to remove most of the RNA non-specifically bound to the proteins, consisted in successive washes at low and high salt concentration before elution from the affinity matrix.

### b. Ionic exchange chromatography

Ion exchange chromatography separates proteins with differences in charge. The separation is based on the reversible interaction between a charged protein and an oppositely charged

chromatographic medium. Elution is performed by increasing salt concentration, and the proteins are eluted according to their ionic interaction with the charged support. Hence, the protein was eluted using a linear gradient of NaCl. This step was important either to remove the RNAses which seemed to elute in the flowthrough of the cationic exchange HiTrap SP, or to completely remove contaminant bacterial RNAs, as these were bound strongly to the anionic exchange HiTrap Q (Figure III.11).

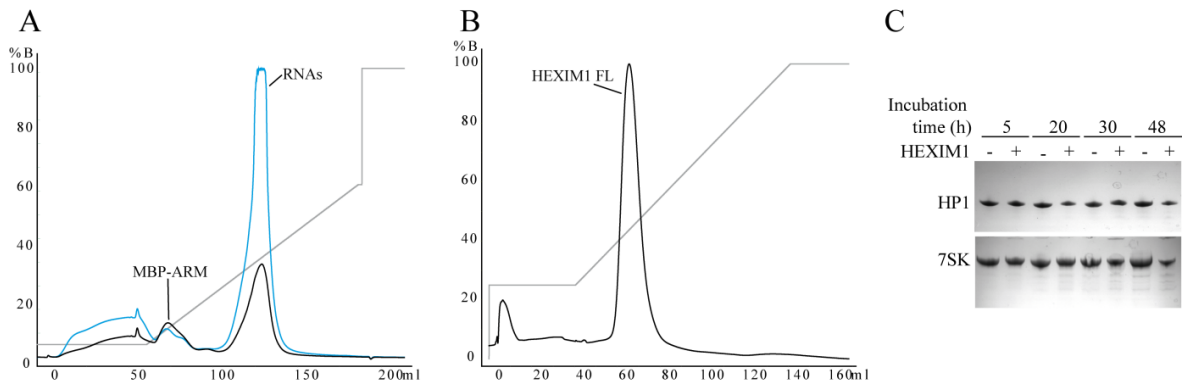



Figure III.11. Ion exchange chromatography. A) Typical anionic exchange chromatogram of MBP-ARM protein; the absorbance at 280nm (black line), 260nm (blue line), and salt gradient is shown. B) Typical cationic exchange chromatogram of HEXIM1 FL. C) RNases test of HEXIM1 analysed by denaturing polyacrylamide gel. Incubations times of HP1 or 7SK at 20°C in presence or not of purified HEXIM1 are indicated.

### c. Hydrophobic interaction chromatography

This technique takes advantage of the hydrophobic areas located on the surface of protein that interact with the hydrophobic groups attached to the stationary column. The hydrophobic interaction is favoured at high ionic strength. When the salt concentration is increased, the water molecules are sequestered by the salt ions, which decreases the number of water molecules available to interact with the charged part of the protein. As a result of the increased demand of solvent molecules, the proteins begin to interact with one another and with the resin via the hydrophobic patches on their surface. By gradually lowering the salt

concentration in the buffer, hydrophobic interactions are decreased and proteins elute from the matrix at different salt concentrations depending on the strength of their hydrophobic interactions.

This technique was used for MBP-HEXIM1136-273 construction since some proteolysis was observed even after ion exchange chromatography. Hence, to completely remove the proteases, a hydrophobic interaction chromatography was included before the ion exchange column. As well, this step greatly increased the purity of the protein (Figure III.12).

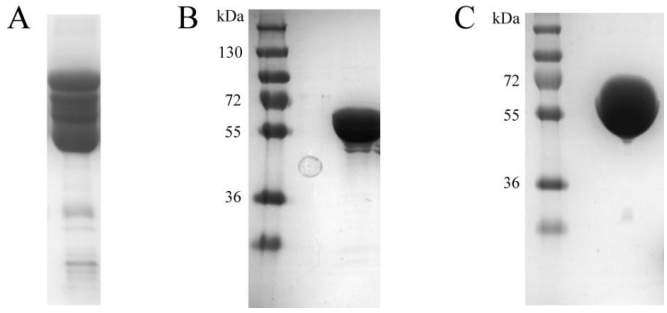



Figure III.12. Hydrophobic interaction chromatography. Degradation (A) of His- MBP-HEXIM136-273, and its purification including (B) or not (C) a hydrophobic interaction chromatography step.

### d. Gel filtration chromatography

The gel filtration chromatography separates proteins with differences in molecular size. Separation is achieved using a porous matrix to which the molecules have different degrees of access according to their size. Smaller molecules diffuse further into the pores and move through the beads more slowly, while larger molecules enter less or not at all and thus move through the solid phase faster. Hence, proteins are eluted in decreasing order of size.

Gel filtration chromatography was typically used as a last step in protein purification. For some proteins, and especially for full length HEXIM1, aggregates were removed at this step (Figure III.13). Gel filtration chromatography is also currently used for buffer exchange, so it was usually used to condition the protein in the suitable buffer for analysis, storage or crystallization.

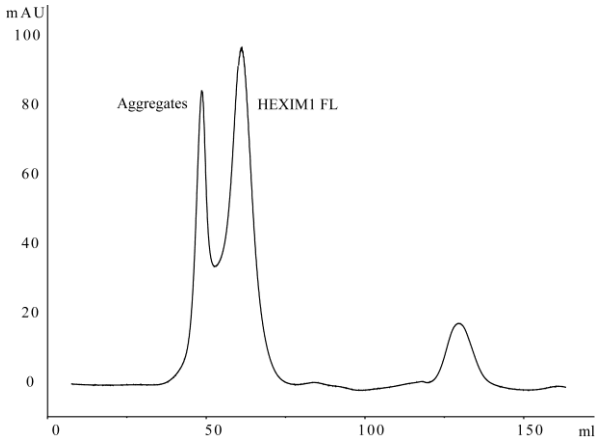



Figure III.13. Gel filtration chromatography. A typical gel filtration chromatogram of full length HEXIM1.

The Figure III.14 shows the MBP-HEXIM136-273 as an example of the purification protocol used in this work. The estimated yield of each protein purification is indicated in the Table III.3.

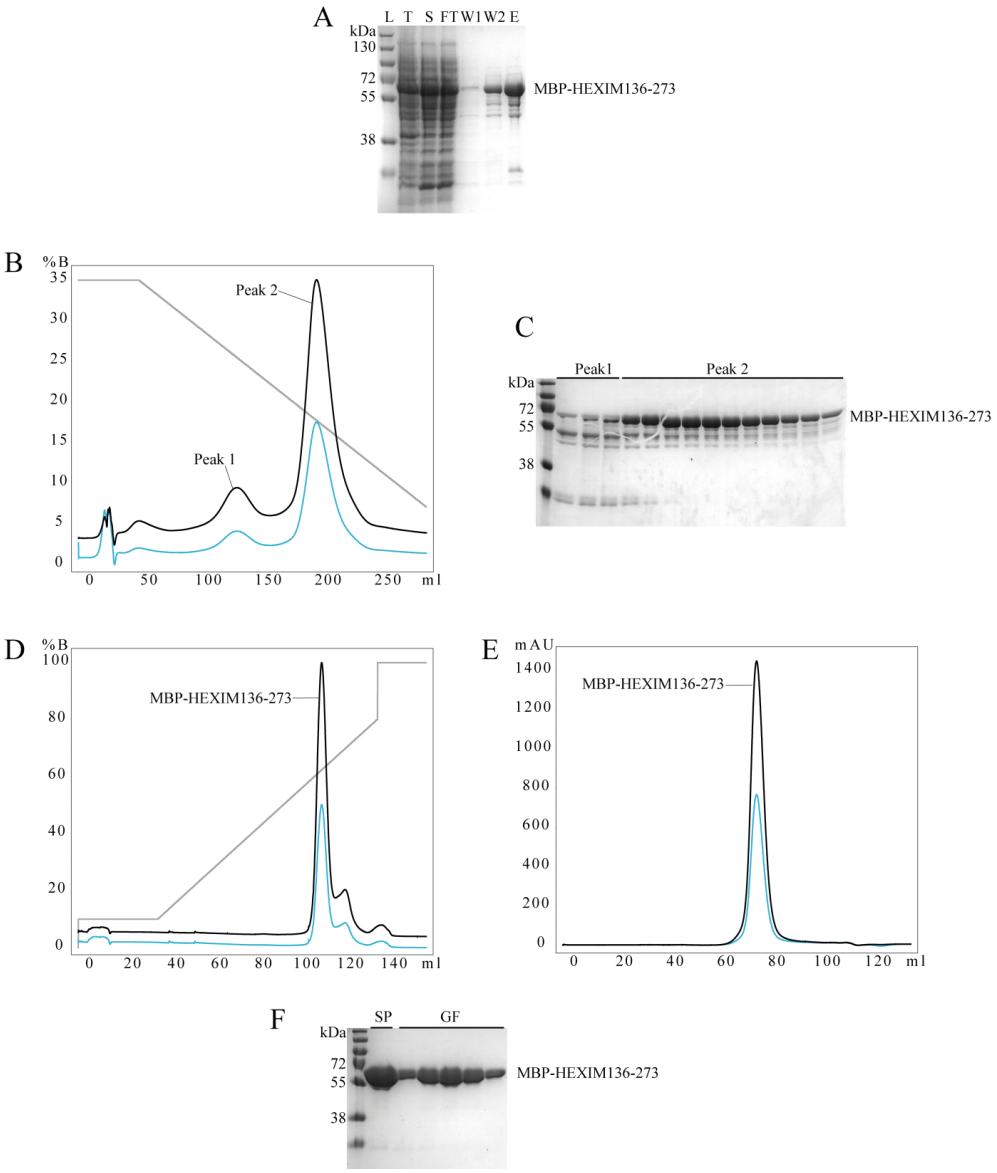



Figure III.14. MBP-HEXIM136-273 purification. A) Lysis and affinity chromatography monitored by SDS-PAGE: ladder (L), total (T), soluble (S), flowthrough (FT), wash at low and high salt concentration (W1), wash at 10 mM imidazole (W2) and elution (E). B) Hydrophobic interaction chromatogram, absorbance at 280 (black) and 260 nm (blue) are shown, as well as the ammonium sulphate gradient (gray). C) Gel analysis of the hydrophobic interaction chromatography. D) Cationic exchange chromatogram (SP). E) Gel filtration chromatogram. F) Gel analysis of SP and GF chromatography as indicated.

# CHAPTER IV: CHARACTERIZATION OF 7SK/HEXIM1 COMPLEX BY BIOCHEMICAL APPROACHES

## 1. SIZE EXCLUSION CHROMATOGRAPHY AND ELECTROPHORETIC MOBILITY SHIFT ASSAY

Size Exclusion Chromatography (SEC) provides a simple method for the characterization of RNA/protein complex. In theory the interaction between two or more macromolecules results in the formation of a complex with a larger Stokes radius than the isolated partners that therefore elutes faster. It could provide a qualitative estimation of the strength of the interaction, with stable complex eluting like a narrower peak than the less stable ones. This method also allows estimating the effect of the buffer condition on the stability of the complex.

The conception of the Electrophoretic Mobility Shift Assay (EMSA) was introduced in the study of the ternary complex of RNA polymerase II, DNA, and RNA (Chelm et al. 1979). This technique is based on the observation that protein-nucleic acid complexes migrate more slowly than free acid nucleic molecules when subjected to non-denaturing polyacrylamide or agarose gel electrophoresis. Nucleic acid migrates through an agarose or polyacryalamide gel matrix towards the anode upon application of an electric field due to the net negative charge of its sugar-phosphate backbone. Migration of nucleic acid through the gel is governed mainly by its molecular weight (smaller molecules travel faster), but also by its three-dimensional conformation. Interaction of a protein that modulates the nucleic acid conformation or substantially increases the molecular weight and changes the charge of the ribonucleoprotein particle, can lead to differential mobility in gel (Ryder et al. 2008). It should be considered that a "caging effect" stabilizes protein-nucleic acid complexes in the gel, meaning that the gel matrix impedes the partners' diffusion away, so concentrations remain locally high and promote prompt re-association (Fried et al. 1981). This approach can

produce information on the affinity of the RNA molecules and protein partners, and also on the conditions that favor the interactions.

Generally, EMSA experiments are performed in presence of an aspecific competitor (such as heparin or tRNA). Indeed, the "caging effect" and the nature of the protein are strong sources of unspecific effects. A competition experiment in the presence of a non-labelled specific competitor is also an important control to verify the specificity of the interaction (Fried et al. 1981). Furthermore, mutants can be characterized also using competition experiments. In the first case, the band of the complex should be eliminated, while in the second one it should not be affected.

Thus, in order to gain insights about the elements in 7SK and HEXIM1 that participate in the recognition and specificity between these molecules and to delineate the minimal 7SK/HEXIM1 complex for crystallization, SEC and EMSA analysis were performed.

## 1.1. 7SK/HEXIM1 complex

We tested first if the recombinant HEXIM1 is able to interact with the in vitro synthesized 7SK. As a reference, 20 µM of HEXIM1 were injected. As shown in the Figure IV.1, a single peak was observed corresponding to an apparent molecular weight (MW) of 476.9 kDa. The MW calculated by ProtParam of our construction is 41.7 kDa. Although it has been shown that HEXIM1 is a dimer (corresponding to 83.4 kDa for our construction), the apparent MW is still more than 5 times larger. This can be explained since estimation of MW by SEC makes two major assumptions: 1) molecules are spherical and 2) molecules do not interact with the gel material. Secondary structure predictions of HEXIM1 showed that it contains long non-structured regions, particularly at its N-terminal domain, and its C-terminal forms a long coiled coil (Dames et al. 2007; Schulte et al. 2005). This should explain its apparent high MW.

Next, 10 µM 7SK was injected. A single peak was observed which corresponds to an apparent MW of 562.4 kDa; the calculated MW of 7SK is 101.4 kDa. However, the calibration was done for proteins and is not applicable to nucleic acids.

Then, a 7SK/HEXIM1 complex with a ratio 1:2 (hereafter the ratio will be referred as RNA:Protein monomer) was injected. The SEC profile showed two peaks, one corresponding to an excess of free 7SK and the other corresponding to the complex. We controlled that with

increasing the ratio to 1:4, some free HEXIM1 could be observed. We conclude that the recombinant HEXIM1 was able to interact with our preparation of 7SK.

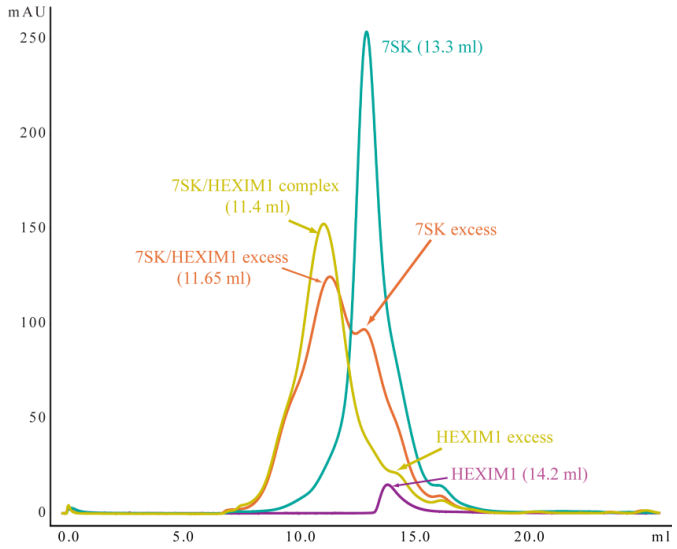



Figure IV.1. SEC analysis of 7SK/HEXIM1 complex. 7SK (blue), HEXIM1 (violet), 7SK/HEXIM1 complex (1:2, orange) and 7SK/HEXIM1 (1:4, yellow) profiles at 280 nm absorbance are shown. Elution volumes are indicated.

Then, the binding interaction between in vitro synthesized 7SK and the recombinant HEXIM1 was tested by EMSA. EMSA was performed using $^{32}$P-labelled RNAs by in vitro transcription in the presence of $^{32}$P-CTP. Since it has been reported that HEXIM1 is able to bind dsRNA (Li, Cooper, et al. 2007), a final concentration of 4 µM of total tRNA, which is in large excess to the labelled RNA, was added to each reaction to minimize the non specific interactions. All complexes were incubated 30 min before loading into a native gel.

Increasing concentrations of HEXIM1 resulted in the formation of one or two 7SK/HEXIM1 complexes (Figure IV.2), confirming that HEXIM1 is able to interact with 7SK in vitro, as previously reported (Michels et al. 2004; Barboric et al. 2005; Byers et al. 2005). A second band of 7SK/HEXIM1 complex was visualized at 1 µM or higher HEXIM1 concentrations and only when most of 7SK had shifted into the first complex. It has been sometimes interpreted as the binding of a second HEXIM1 on 7SK (Muniz et al. 2010; Byers et al. 2005). However, we noted that the apparition of the second band was more obvious when the protein was ageing, so we hypothesized that the second band could be due to protein aggregation or oxidation (HEXIM1 required the presence of a reducing agent during

purification and storage). To discern if the bands corresponded to multimers or even dissociation of the dimer of HEXIM, further analyses were required (see Chapter V.2 "Mass Spectrometry").

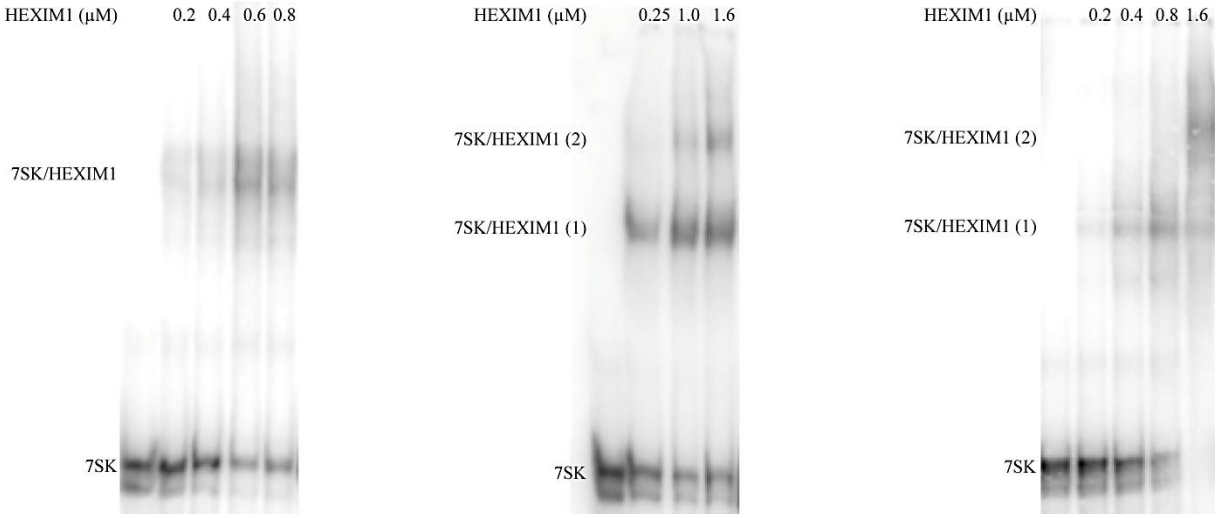



Figure IV.2. EMSA analysis of 7SK/HEXIM1 interaction. EMSAs performed with different ranges of concentrations (as indicated) of HEXIM1.

## 1.2. HP1/HEXIM1 complex

Base upon mutational analysis in vivo, it had been reported that the 5'end hairpin of 7SK is important for HEXIM1 binding (Egloff et al. 2006). Hence, the isolated hairpin encompassing the nucleotides 24 to 87 of 7SK, HP1, was tested for its interaction with HEXIM1 (Figure IV.3). HP1 was injected at 20 µM and showed a sharp peak. When the HP1:HEXIM1 complex (ratio 1:2) was injected a sharp peak was observed with an elution volume of 13.5 ml, followed by a small peak corresponding to the free HP1.

To test if this complex was stable, the fractions from the HP1/HEXIM1 complex were collected and injected again into the column. No dissociation was observed. Thus, HP1 with HEXIM1 forms a stable complex. That makes it a good candidate for crystallization.

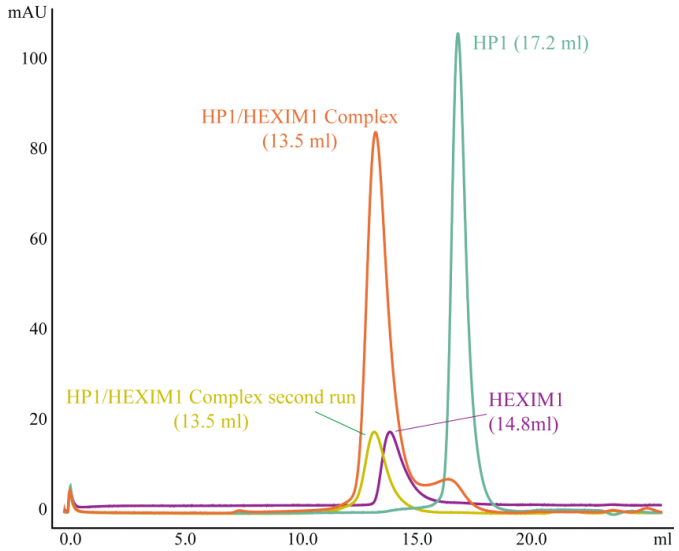



Figure IV.3. SEC analysis of HP1/HEXIM1 complex. HP1 (blue), HEXIM1 (violet), HP1/HEXIM1 complex (1:2, orange) and HP1/HEXIM1 second injection (yellow) profiles at 280 nm absorbance are shown. Elution volumes are indicated.

Next, the interaction between HP1 and HEXIM1 was tested by EMSA (Figure IV.4). HP1 was shifted by increasing concentrations of HEXIM1. The band was sharper than for 7SK/HEXIM1 complex in agreement with the results of SEC experiments. The specificity of the interaction observed was confirmed by the effective competition of 50 nM of non-labelled HP1. This result indicated that the elements of 7SK necessary for a specific binding to HEXIM1 are contained in HP1 as previously reported. Indeed, during my pHD work, the in vitro interaction of HP1 with HEXIM1 was reported by another team, and a crosslink between the U30 of HP1 and the region encompassing the aminoacids 210 to 220 of HEXIM1 was identified (Bélanger et al. 2009).

EMSA showed that the interaction of HP1 and HEXIM1 was specific in those experimental conditions. The advantage of EMSA is that large concentration of a competitor can be used to mask the non specific interaction, since the RNA of interest is labelled. In SEC experiments, it was not possible to use tRNA to ensure only specific interactions. To check that the interactions observed by SEC are specific, the interaction between HEXIM1 and HP3, a 7SK hairpin which does not participate in the binding to HEXIM1, was tested.

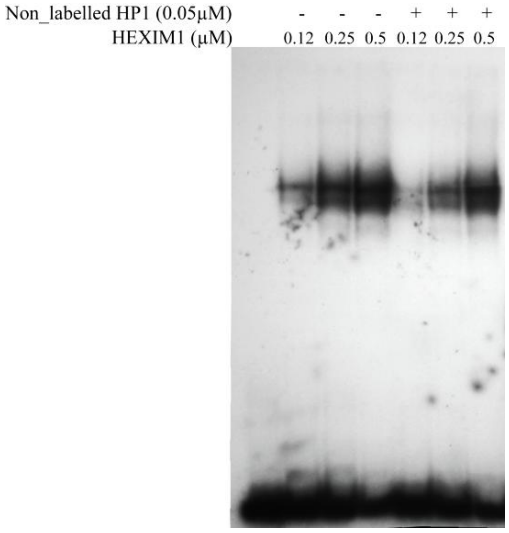



Figure IV.4. EMSA analysis of HP1/HEXIM1 interaction. Increasing concentrations of HEXIM1 were incubated with $^{32}$P-labelled HP1 in the absence or presence of non-labelled HP1.

HP3 eluted in a sharp peak. When HP3 and HEXIM1 were injected with a ratio 1:2 after incubation at 20°C, no complex was observed (Figure IV.5).

These experiments confirmed that SEC is a suitable technique.

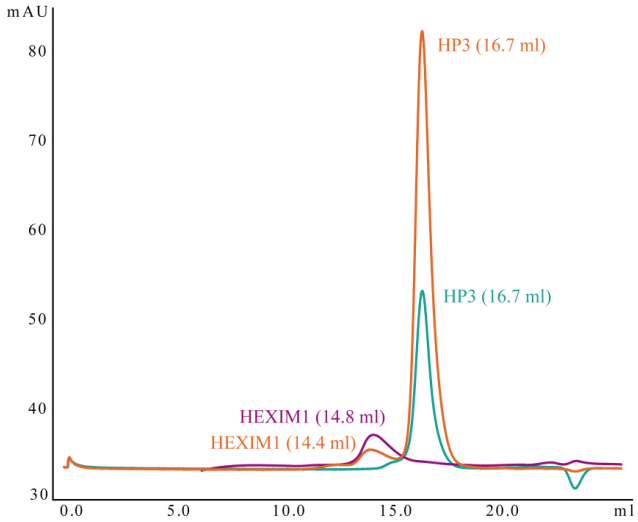



Figure IV.5. SEC analysis of HP3/HEXIM1 complex. HP3 (blue), HEXIM1 (violet), HP3+HEXIM1 (1:2, orange) profiles at 280 nm absorbance are shown. Elution volumes are indicated.

Many attempts of crystallization of HP1/HEXIM1 complex were performed, but were unfortunately unsuccesful. This prompted us to design HP1 variants in order to improve our

chances for crystallization. HP1 contains a large apical loop consisting of eleven nucleotides, most probably flexible that may prevent the crystallization. Hence, the apical loop was mutated into UUCG (construction called HP1u), which forms a highly stable tetraloop. A different strategy was to graft HP1 in the anticodon arm of a tRNA, thinking that HP1 should be stabilized and that the tRNA scaffold may favourably contribute to the crystallization. Two constructions were designed: KS1 and KE1, where KS1 contains a sephadex aptamer (see Chapter III "Molecules Preparation"). These RNA constructions were interesting because they can be over-expressed in *Escherichia coli* with high yields. Also, KS1 and KE1 may provide information on whether HEXIM1 is able to recognize HP1 in a different context than 7SK. Finally, HP1L consisting in the whole 5'end hairpin was also tested since a longer RNA may stabilize the protein interaction.

Variants of HP1 mentioned above were tested for HEXIM1 interaction. All RNAs were able to interact with HEXIM1 (Figure IV.6). These results confirmed that the sequence and the size of the apical loop is not determinant for the binding to HEXIM1 and suggested that HEXIM1 is able to recognize HP1 even out of the 7SK context.

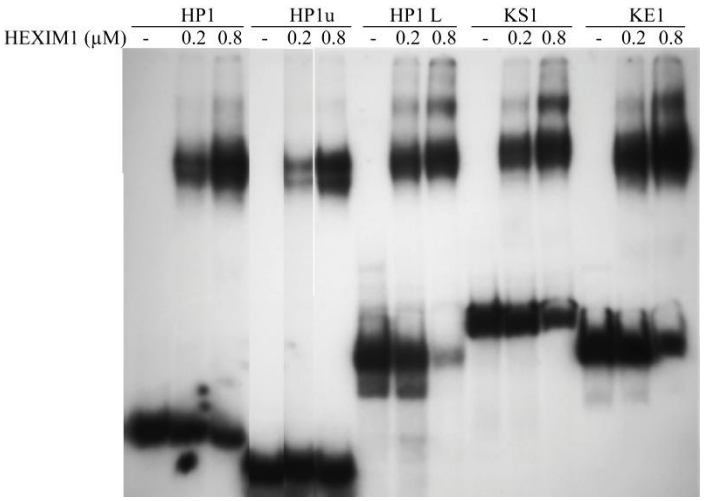



Figure IV.6. EMSA analysis of HP1 variants for interaction with HEXIM1. The different RNAs and the concentrations of HEXIM1 used are indicated.

To analyse in detail the binding of these different variants of HP1, a more subtle range of concentrations was used (Figure IV.7). HP1 and HP1L bound HEXIM1 with apparent similar affinities, whereas KE1 showed a slightly lower one. A possible explanation is that

HEXIM1 binding requires some flexibility. The anticodon arm of tRNA, which contains several GC basepairs, probably affords some rigidity.

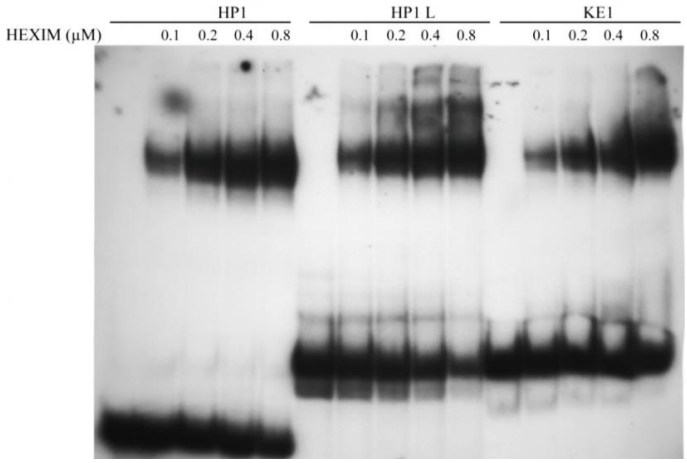



Figure IV.7. Further EMSA analysis of HP1 variants. The different RNAs and the concentrations of HEXIM1 used are indicated.

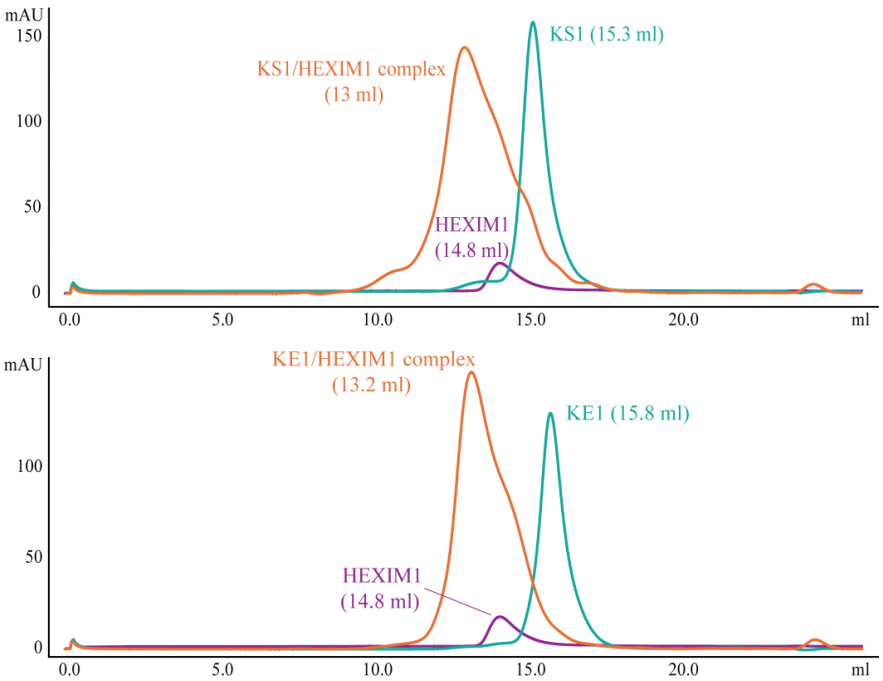



Figure IV.8. SEC analysis of KS1/HEXIM1 (upper panel) and KE1/HEXIM1 complexes (lower panel). KS1 or KE1 (blue), HEXIM1 (violet), KS1/HEXIM1 or KE1/HEXIM1 complexes (1:2, orange) profiles at 280 nm absorbance are shown. Elution volumes are indicated.

SEC analyses were also performed to test KS1 and KE1 interaction to HEXIM1 (Figure IV.8). Both eluted as a single peak and were able to interact with HEXIM1 (ratio 1:2), KE1 showing a slightly narrower peak than KS1. KS1 showed a small peak of free RNA. These results confirm that HEXIM1 is able to recognize HP1 regardless of the context.

This indicates that HP1 is probably forming an independent, autonomous domain in 7SK. All these complexes were tested for crystallization but no crystals were obtained.

### 1.3. Shorter HEXIM1 constructions

In order to further reduce the size of the complex expecting to improve the chances for crystallization, shorter constructions of HEXIM1 were produced. The N-terminal domain of HEXIM1 does not participate in 7SK binding (Yik et al. 2003; Michels et al. 2004), and secondary structure prediction showed that it is a mainly unstructured region, and therefore certainly no crystallizable. Also in denaturing conditions as SDS-PAGE, this domain generates an aberrant migration of HEXIM1 (Ouchida et al. 2003; Michels et al. 2003; Yik et al. 2003) most likely due to its high Proline content which decrease the electrophoretic mobility as a result of kinks and structural rigidity. Hence, the 120 N-terminal residues of HEXIM1 were deleted in HEXIM1 120-359. Then, we used the SEC to test the binding of HEXIM1 120-359 to HP1 (Figure IV.9).

HEXIM1 120-359 (MW 28.9 kDa, 57.8 kDa for a dimer) profile showed two peaks, the principal one eluted at 15.9 ml. The volume corresponded to an apparent MW of 80.5 kDa. This observation suggested that the N-terminal domain mainly accounted for the high elution volumes observed for the full length HEXIM1 (see above and Table IV.I).

The HP1/HEXIM1 120-359 (ratio 1:2) showed that the complex elutes at 14.5 ml. The presence of another peak eluting as the free HP1 indicates an underestimation of the concentration of HEXIM1 120-359. Thus, these results confirmed that the N-terminus of HEXIM1 is not involved in the interaction with HP1

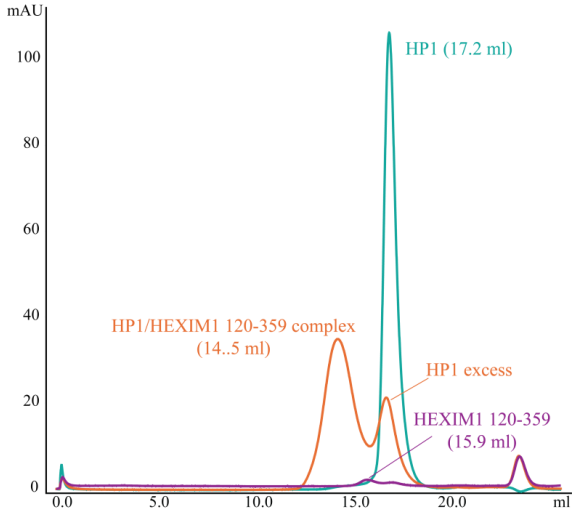



Figure IV.9. SEC analysis of HP1/HEXIM1 120-359 complex. HP1 (blue), HEXIM1 120-359 (violet), HP1/ HEXIM1 120-359 (1:2, orange) profiles at 280 nm absorbance are shown. Elution volumes are indicated.

This result was further confirmed by EMSA analysis (Figure IV.10). As expected, HEXIM1 1-120 was able to bind HP1 and gave rise to a sharper complex band than wild type HEXIM1. Unfortunately no crystals were obtained for HP1/HEXIM1 1-120 complex.

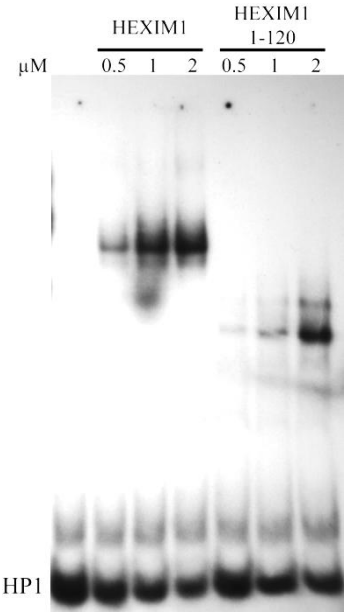



Figure IV.10. EMSA analysis of HP1/HEXIM1 1-120 complex. Comparison between wild type HEXIM1 and HEXIM1 120-359 for their HP1 binding. The different HEXIMs and their concentrations are indicated.

In order to further minimize the construction of HEXIM1 domain for the interaction with HP1, we designed constructions of HEXIM1 deleted from the N-terminal and C-terminal domains. A short α helix from residues 114 to 119 is predicted by most of the secondary structure prediction programs (see Chapter II), which marks the beginning of a more structured region of the protein, so the inclusion of this region may help to stabilize HEXIM1. In another hand, the TBD is essential for Cyclin T1 binding and for HEXIM1 dimerization but not for 7SK binding. Hence, HEXIM1 114-255 was produced, which is deprived of the 113 N-terminal residues and of the TBD, and was expected to be a monomer.

The SEC profile of HEXIM1 114-255 showed a sharp peak which corresponded to an apparent MW of 18.6 kDa in agreement with the 16.2 kDa calculated for a monomer (Figure IV.11). This suggests that the central region of HEXIM1 probably forms a globular domain. The complex HP1/HEXIM1 114-255 (ratio 1:2) formed a sharp peak with only a small peak of free protein.

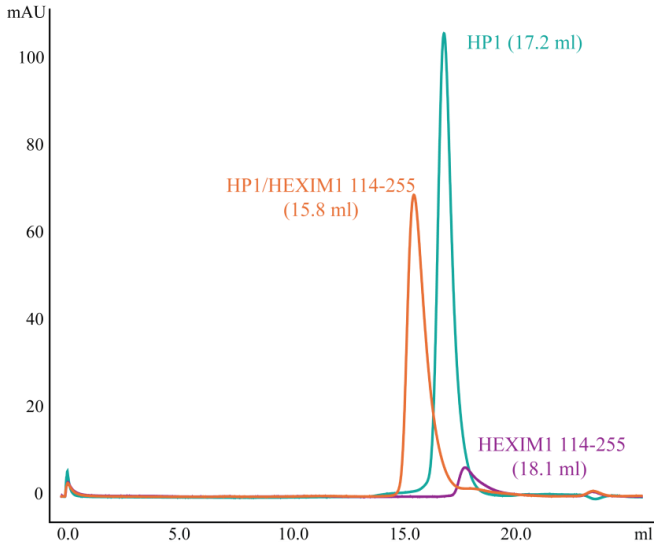



Figure IV.11. SEC analysis of HP1/HEXIM1 114-255 complex. HP1 (blue), HEXIM1 114-255 (violet), HP1/ HEXIM1 114-255 (1:2, orange) profiles at 280 nm absorbance are shown. Elution volumes are indicated.

With crystallization in mind, we analyzed also the KS1 and KE1 complexes with HEXIM1 114-255. The SEC profiles of the complexes showed narrow peaks (Figure IV.12). Thus, HEXIM1 114-255 was still able to recognize HP1 in the context of KS1 and KE1. The

narrowness of the peaks confirmed that these molecules represent good candidates for crystallization. Unfortunately, this protein showed a very low solubility and attempts to concentrate it for crystallization trials failed (see Figure III.9 and Table III.3).

The elution volumes of all the isolated RNA and protein constructions, their apparent and calculated MW are summarized in Table IV.1. Also, in Table IV.2, are summarized the elution volumes of all RNA/protein complexes observed; the apparent MW of these complex are only indication, since the estimation was very inaccurate.

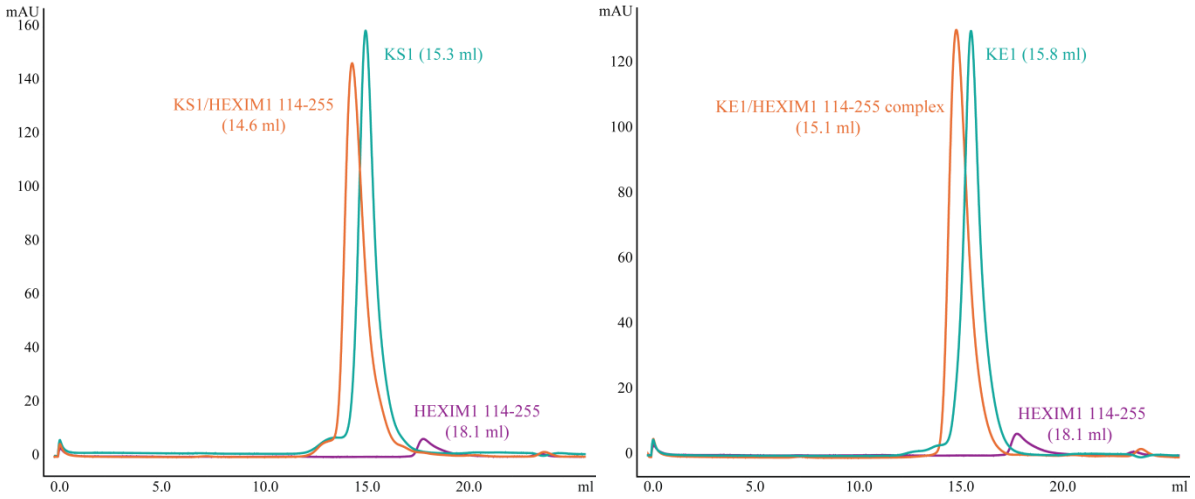



Figure IV.12. SEC analysis of KS1/HEXIM1 114-255 (left) and KE1/HEXIM1 114-255 complexes (rigth). KS1 or KE1 (blue), HEXIM1 144-255 (violet), KS1/HEXIM1 114-255 or KE1/HEXIM1 114-255 complexes (1:2, orange) profiles at 280 nm absorbance are shown. Elution volumes are indicated.

**Table IV.1. SEC (Superose 6) parameters for RNA and proteins**

| Molecule | | Elution Volume | Apparent MW | Calculated MW |
|---|---|---|---|---|
| RNA | 7SK | 13.3 ml | 562.4 kDa | 101.4 kDa |
| | HP1 | 17.2 ml | 36.0 kDa | 18.6 kDa |
| | KS1 | 15.3 ml | 140.6 kDa | 48.5 kDa |
| | KE1 | 15.8 ml | 98.1 kDa | 39.9 kDa |
| | HP3 | 16.7 ml | 51.3 kDa | 22.3 kDa |
| Protein | HEXIM1 | 14.8 ml | 476.9 kDa | 41.7 kDa |
| | HEXIM1 120-359 | 15.9 ml | 80.5 kDa | 28.9 kDa |
| | HEXIM1 114-255 | 18.1 ml | 18.6 kDa | 16.2 kDa |

| Table IV.2. SEC (Superose 6) parameters for RNA/protein complexes | | |
|---|---|---|
| **Complex** | **Elution Volume** | **Apparent MW** |
| 7SK/HEXIM1 | 11.4 ml | 2,331 kDa |
| HP1/HEXIM1 | 13.5 ml | 514 kDa |
| KS1/HEXIM1 | 13.0 ml | 737 kDa |
| KE1/HEXIM1 | 13.2 ml | 638 kDa |
| HP1/ HEXIM1 120-359 | 14.5 ml | 250 kDa |
| HP1/ HEXIM1 114-255 | 15.8 ml | 98 kDa |
| KS1/ HEXIM1 114-255 | 14.6 ml | 233 kDa |
| KE1/ HEXIM1 114-255 | 15.1 ml | 162 kDa |

To further delimit a functional HEXIM1 domain for RNA binding for crystallography, the interaction of different versions of HEXIM1 and HEXIM2 were tested by EMSA for the binding to HP1L. All the secondary structure predictions showed a $\alpha$ helix beginning around the residue 138, just before the RNA binding region of HEXIM1 (see above). Hence we decided to expand the N-terminal deletion to residue 135. Two deletion boundaries at C-terminus were chosen, one at 317 and therefore including the first segment of the coiled coil of TBD, and a second one at 273, just before the short $\alpha$ helix at the N-terminus of the TBD. This C-terminal region of HEXIM1 was included because crosslink experiments coupled to LC-MS (Liquid Chromatography-Mass Spectrometry; O'Gorman et al. 2005) suggested an interaction with 7SK, which may strengthen the complex (personal communication with C. Barrandon). The corresponding C-terminal deletions in HEXIM2 were also performed. Thus, HEXIM1 136-317 and HEXIM1 136-273, as well as HEXIM2 1-242 and HEXIM2 1-199 were produced. All these proteins were able to interact with HP1 L (Figure IV.13).

HEXIM1 136-317 and HEXIM2 1-242 show two bands of similar intensities at low concentration of protein. This may be explained by a mixture of monomers and dimers with similar affinity for RNA. Indeed, the constructions ends after the first segment of the coiled coil, which is not sufficient for a stable dimerization (Schönichen et al. 2010). The gels suggest that monomeric HEXIM1 is able to interact with HP1L, probably with similar affinity than dimeric HEXIM1.

In another hand, HEXIM1 136-273 and HEXIM2 1-199 are probably monomers. HP1L/HEXIM1 136-273 complexes were difficult to analyse since the complex band with higher mobility migrated at the same level than bands present in the absence of protein. HEXIM1 136-273 seemed to form a first complex with HP1L, then a second complex appeared at higher protein concentrations. The situation was similar for the equivalent

construction HEXIM2 1-199. This experiment suggess that HP1L contains two binding sites for HEXIM1 monomers.

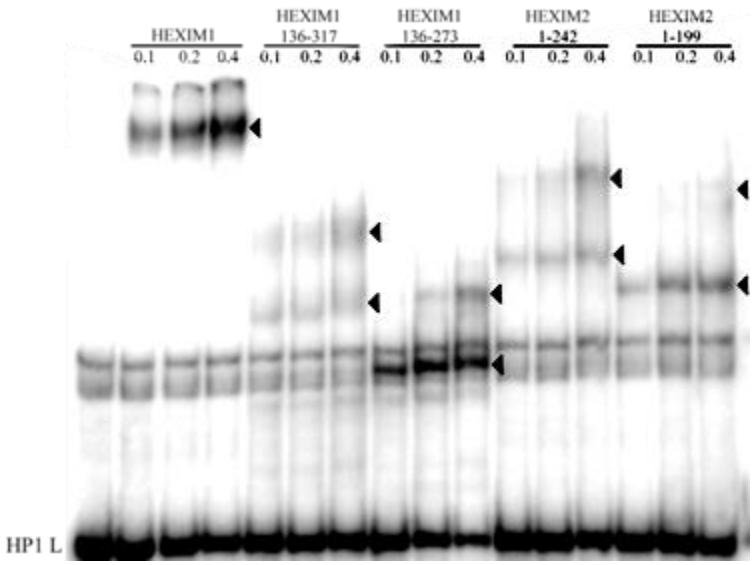



Figure IV.13. EMSA analysis of HEXIM1 and HEXIM2 constructions for interaction with HP1L. The different HEXIMs and their concentrations are indicated. Bands of complex are indicated with arrows.

These questions were further investigated later (see Chapter V.2 “Mass Spectrometry”).

## 2. Footprinting Assays on 7SK/HEXIM1 Complex

Some RNA footprinting assays were performed in order to identify the specific 7SK sequence that binds HEXIM1 and to investigate potential new HEXIM1 binding regions outside HP1. The basic principle of RNA footprinting is the measurement of accesibility of RNA using a probe (enzyme or chemical reagent) that is able to cut or modify the RNA. If a protein interact with the RNA, sites on the RNA where the protein is bound are inaccesible to the probe. After electrophoretic separation, the inaccesible sites appear as blanks in an otherwise regular RNA cleavage or modification pattern, thus revealing the footprint of the binding protein. As probe we used a hydroxyl-selective electrophile, the 1-methyl-7-nitroisatoic anhydride (1M7), and the method called Selective 2′-Hydroxyl Acylation analyzed by Primer Extension (SHAPE). The details of this method and the probe will be discused on the Chapter VII. Briefly, IM7 measures the local flexibility of every nucleotide within the RNA by reacting with its ribose 2'-hydroxyl position when unconstrained, in single stranded regions (Mortimer et al. 2008; Merino et al. 2005). The reaction generates a 2'-O-adduct and modification sites are revealed by primer extension. Hence, if HEXIM1 binds single stranded regions on 7SK or leads to changes on the flexibility of 7SK (due to conformational changes for instance), modification on the flexibility profile of 7SK should be seen. By using different primers (see Annexes 1 Figure A.9), nearly the full 7SK sequence was explored.

The flexibility profiles of 7SK (3µM) in the absence or the presence of HEXIM1 (6 µM) are summarized in Figure IV.14. Both profiles were very similar and no protection was observed. The nucleotides that showed the highest decrease in flexibility were G130, U129 and C150. Some nucleotides showed a higher flexibility in the presence of HEXIM1. These were C97, C173, A164, A270, U275 and A278 nucleotides, all of them, apart C97, predicted in single stranded regions of 7SK. Previous footprinting experiments on 7SK snRNP extracted from HeLa cells and using enzymatic and chemical probes showed protection of nucleotides included in the region 100 to 196 of 7SK (see Figure VII.2; Wassarman et al. 1991), region where we observed also most of the flexibility changes. It must be noted that Wassarman et al. observed more nucleotides protected on this region than us, but they may account for all the protein partners of 7SK.

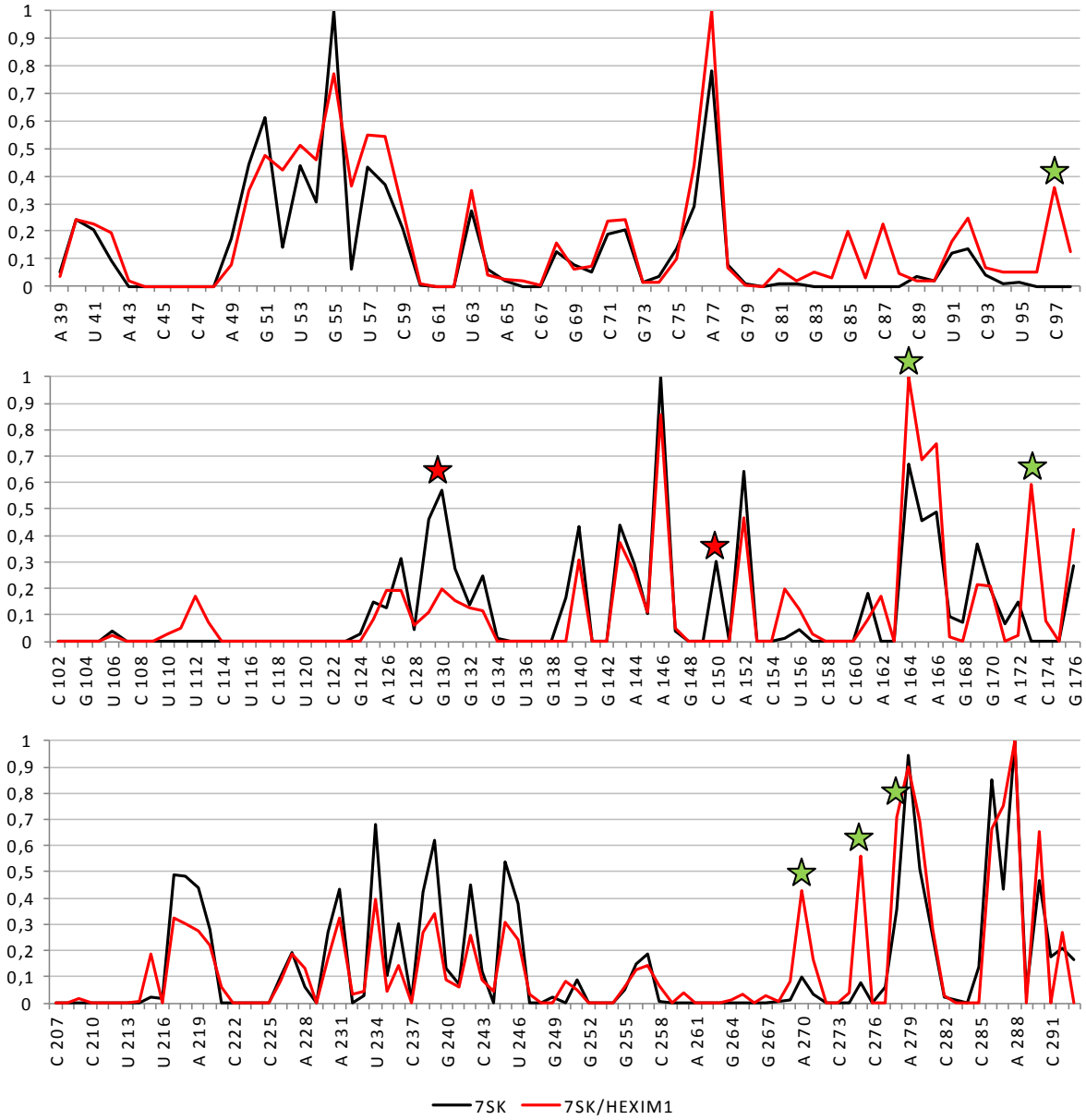



Figure IV.14. 7SK/HEXIM1 footprinting. The flexibility profiles of 7SK in the absence or the presence of HEXIM1 are shown. Red and green stars indicate decrease and increase of flexibility in the presence of HEXIM1, respectively. Briefly, the bands intensities on gels were normalized using noRNAlize (Vicens et al. 2007; see details Chapter VII), and then normalized to the unity.

Our results suggested that HEXIM1 most probably binds a double stranded region and therefore IM7 is not a suitable probe to identify the 7SK interaction region for HEXIM1. Indeed, HEXIM1 has been shown to be a dsRNA binding protein (Li, Cooper, et al. 2007).

Therefore, probes such as the RNAse V1 or MPE (Methidiumpropyl-EDTA)-Fe(II) should be suitable. These footprinting experiments, intended to reveal conformational changes on 7SK, need further optimization of the protocol. Indeed, we now know that particular attention must be paid to the tendency of HEXIM1 to form non specific interactions to obtain a significant. Nevertheless, other tools were more suitable to delineate the site on 7SK for HEXIM1 interaction. These are described in the next chapter.

## 3. Discussion and Conclusions

So far we showed that the isolated HP1 element (nucleotides 24 to 87) of 7SK was able to interact with HEXIM1. Also, HP1 grafted into the anticodon arm of a tRNA was recognized by HEXIM1, suggesting that the context in which is embedded is not determinant for the interaction. This also implies that HP1 should acquire its correct fold in an independent way in vitro. Indeed, we observed by SHAPE experiments (for details see Chapter VII) that this region of 7SK has the same profile either isolated or in the 7SK-context (Figure IV.15). Thus, HP1 is an autonomous, functional element of 7SK. We also observed that the length and the sequence of the apical loop of HP1 was not determinant for the interaction.

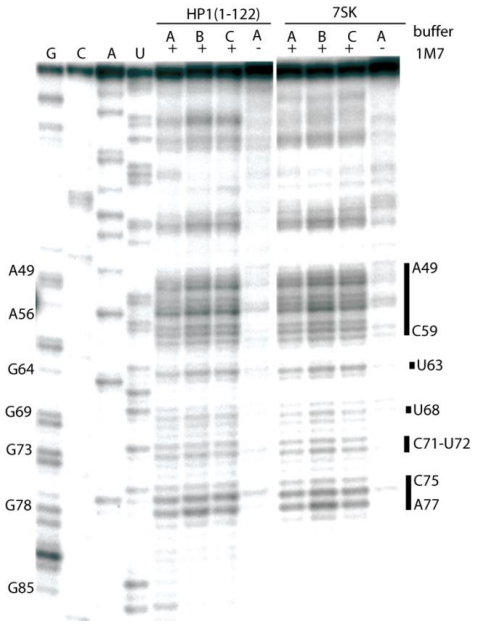



Figure IV.15. HP1 preserves its functional secondary structure as seen by SHAPE. SHAPE profiles of HP1 isolated or in 7SK-context are shown in parallel with sequencing reactions (see Chapter VII for details). Flexible regions are identified on the right.

HP1 (nucleotides 24 to 87) is then a suitable construction for studying in detail the interaction with HEXIM1.

In summary, footprinting experiments confirmed that HEXIM1 binds only double-stranded regions on 7SK, but suggested flexibility changes upon the interaction. Also, they

underlie the region encompassing the domain 2 of 7SK as interesting for further investigations.

# CHAPTER V: CHARACTERIZATION OF 7SK/HEXIM1 COMPLEX BY BIOPHYSICAL APPROACHES

## 1. NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY

In order to investigate the interaction between 7SK/HEXIM at higher resolution and understand what characterize the 7SK site for HEXIM1, Nuclear Magnetic Resonance (NMR) footprinting studies were performed by Isabelle Lebars and Bruno Kieffer from the Biomolecular NMR Laboratory at the IGBMC, and results were published in (Lebars et al., 2010, and see Annexe 4).

The NMR is a powerful tool for studying the structure of and the interaction between biomacromolecules such as proteins and RNAs in solution (Clarkson et al. 2003). NMR spectroscopy exploits the magnetic properties of atomic nuclei to obtain information about their physical environment such as chemical and electronic states. This gives insights into the structure and dynamics of the molecules. Interactions between molecules can be also studied by NMR spectroscopy since changes on the environment of an atom upon interaction with a second molecule will affect its NMR properties.

Within the size limits of NMR measurements on RNA (upper limit of around 50 to 100 nucleotides depending on the complexity of the spectrum), information about base-pairing pattern, conformational equilibria, secondary structure motifs (such as hairpins and bulges), and mapping of interaction surfaces of RNA with small proteins or other ligands, etc., can be derived (Fürtig et al. 2003). Thus, NMR spectroscopy provides information about changes in base paired nucleotides (see below) making it a suitable method for studying the interaction between 7SK and HEXIM1.

NMR studies were performed on HP1 that we had shown to be an autonomous structure of 7SK, specifically recognized by HEXIM1.

## 1.1. Determinants for a specific 7SK/HEXIM1 interaction

### a. HP1 sequence assignment

The region of imino proton resonances of G and U contain valuable information about base pairing in the RNA molecule. These signals are only observable when imino protons are protected from exchange with protons from the solvent and are therefore involved in H-bonds.

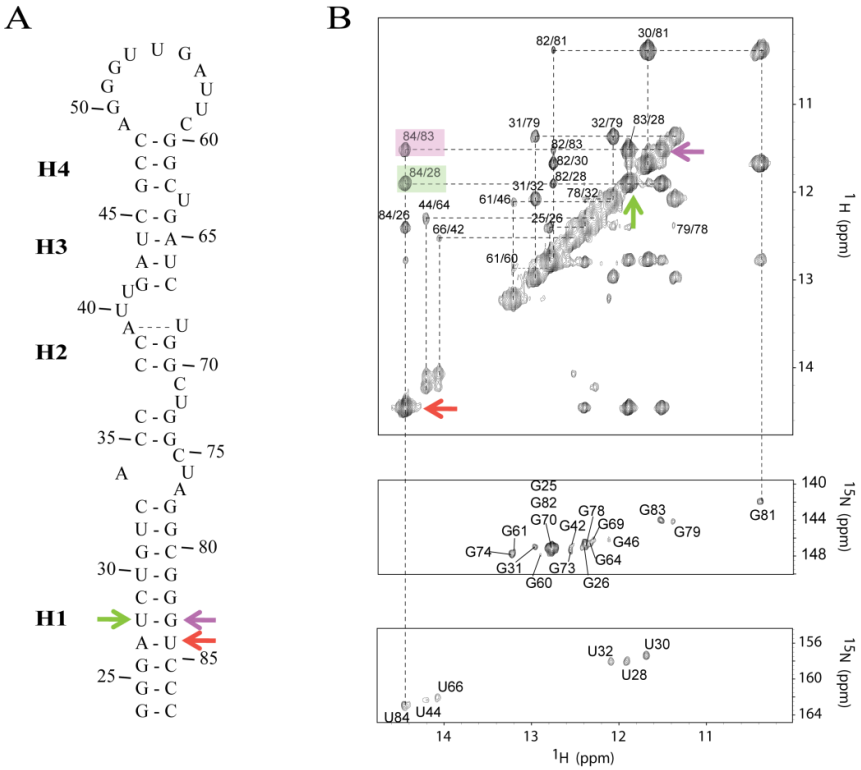



Figure V.1. Determination of the secondary structure HP1. A) Model of the secondary structure of HP1 as seen by NMR studies. B) Imino/imino protons region of NOESY spectrum recorded at 10°C in 90/10 $H_2O/D_2O$ with a mixing time of 300 ms. Dashed lines represent NH/NH sequential assignment (top). Regions SOFAST-HMQC, recorded at natural abundance of $^{15}N$ G imino protons (middle) and $^{15}N$ U imino groups (bottom). The starting point for helix assignment was based upon identification of A27-U84 pair (red arrow points U imino proton), the only A-U exhibiting connectivity (highlighted with a color square) with a G-U pair (green and violet arrows point the U28 and G83 imino protons, respectively).

The starting point for the determination of H-bond pattern is the sequential assignment of imino proton resonances in a 2D NOESY (Nuclear Overhauser Effect SpectroscopY) experiment which correlates all protons within a distance of 5 Å (Fürtig et al. 2003). By using NOE information, it is possible to distinguish G:C, A:U and G:U base pairs. NOE signals between imino protons of neighboring base pairs are observable and facilitate the sequential assignment of the imino proton signal. Hence, the first step was to determine the base pair pattern of HP1 by NOESY analysis. This supported the existence of the four helical segments shown in Figure V.1, numbered from the basal to the apical end of the hairpin. Also, it revealed an open conformation of A39-U68, usually represented as a base pair, in agreement with the U68 flexibility observed by SHAPE experiment (Figure IV.15). The assignments were verified by $^{1}H$-$^{15}N$ SOFAST-HMQC (Selective Optimized Flip-Angle Short-Transient Heteronuclear Multiple Quantum Coherence) analysis recorded at natural abundance of $^{15}N$.

### b. HP1 titration with HEXIM1 ARM

Mapping with the full length HEXIM1 was initially tried, but the spectra were not interpretable. It has been previously reported that the HEXIM1 ARM was fully functional for 7SK binding (Yik et al. 2004). Hence, the interaction of the ARM peptide of HEXIM1 (residues 149 to 165) with HP1 was mapped. The chemical shift of imino protons was monitored as a function of peptide concentration; the observed spectral changes are summarized in the Figure V.2. These changes included:

- observation of non-averaged signals from free and bound state that is characteristic of tight binding;
- frequency shifts that indicate specific binding;
- changes in peak intensity that reveal modification of solvent accessibility;
- appearance or disappearance of a resonance that indicate the formation or melting of a base pair, respectively;
- uniform broadening of all resonances that indicates a non-specific binding.

It is important to note that titration beyond a HP1/ARM ratio of 1:1.3 resulted in an overall broadening of the resonances, so no precise information can be obtained upon further addition of peptide.

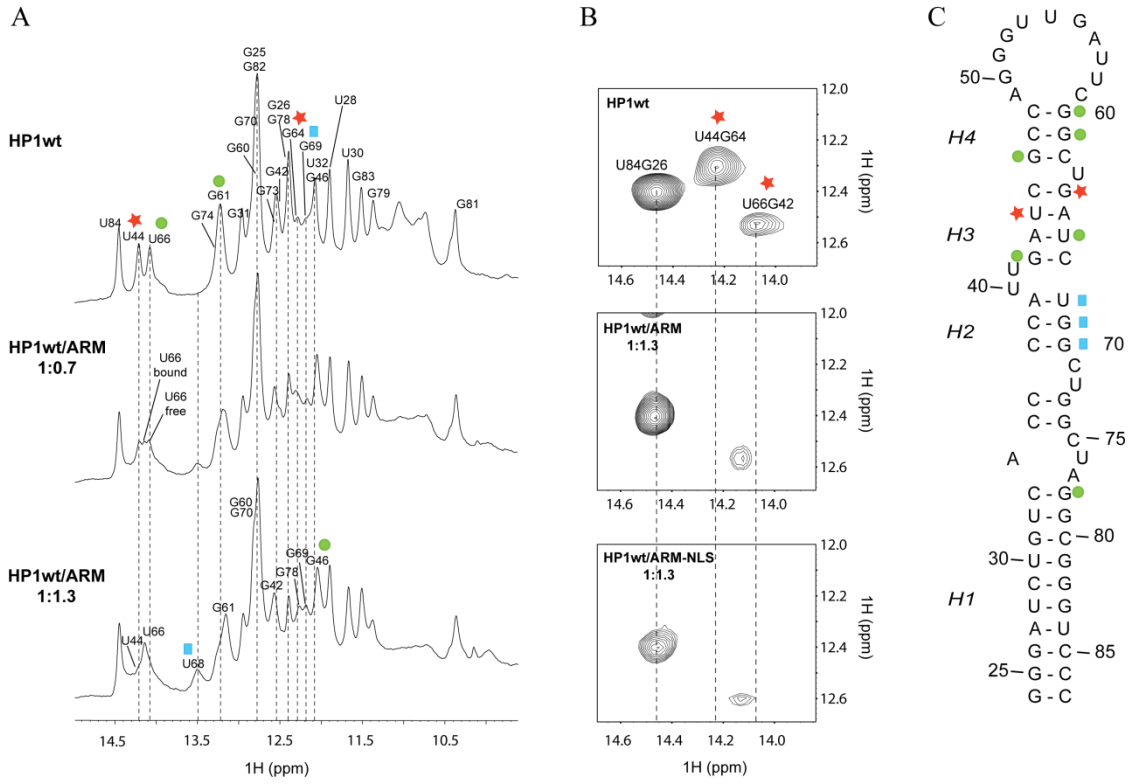



Figure V.2. Mapping of HP1 interaction with ARM. A) Titration of HP1 by ARM as indicated. Dashed lines indicate imino protons that undergo chemical shift changes. B) Imino/imino protons region of NOESY spectrum recorded at 10°C in 90/10 $H_2O/D_2O$ with a mixing time of 300 ms. Observable U84G26, U44G64 and U66G42 correlations in free HP1 (top) disappeared when bound to ARM (middle) or ARM-NLS (bottom), whereas U66G42 was shifted. C) Summary of effect observed on HP1 upon interaction. Red stars indicate opening of base pairs, blue squares indicate stabilization of base pairs and green circles indicates specific binding.

The results shown in Figure V.2 suggest that ARM binds specifically the G42, G46, G60, G61, G78 and U66, which seemed to be bound tighter. The spectra also suggested that the region encompassing the conserved GAUC motifs (stem 42-45/64-67) is melted upon addition of ARM, whereas A39-U68 base pair is formed. Also, H2 stem seemed to be stabilized. Importantly, all these nucleotides, except G78, are located in the apical region of HP1 pointing out the main role of this region for HEXIM1 interaction.

These interpretation were strengthened by $^{1}H$-$^{13}C$ HSQC (Heteronuclear Single Quantum Coherence) analyses on HP1 selectively labeled at A and U with $^{13}C$-$^{15}N$ enriched nucleotides. Additionally, observed connectivities probably corresponding to an H-bond network between the bulged Us disappeared, suggesting a more open conformation upon ARM binding. Similar results were obtained when the HEXIM1 ARM-NLS motif (residues 149-179) was used to map the interaction with HP1, suggesting that HEXIM1 ARM contained the determinant elements for HP1 recognition.

## c. ARM peptide specificity

The singular organization of ARM, with two ARM sequences separated for a Proline and a conserved Serine, let us to ask the role of these residues. Thus, titrations of HP1 with peptides ARM-S158C, ARM-P157G and ARM-P157K mutants were performed. NMR spectra showed a uniform broadening of all imino protons resonances upon mutant ARM binding, suggesting a loss of specificity (Figure V.3).

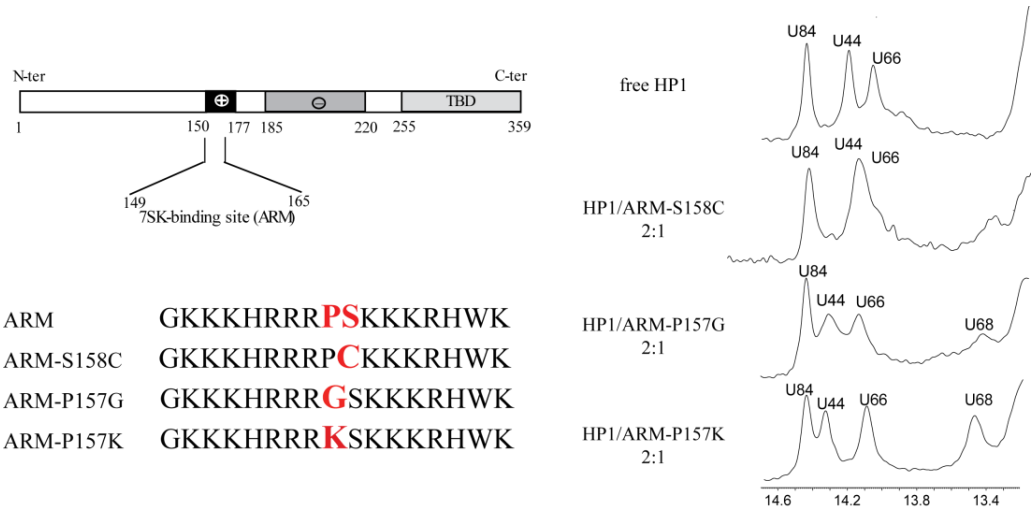



Figure V.3. Effect of Pro157 and Ser158 mutations on HP1 binding. Left, a scheme of HEXIM1 indicating the location of ARM (top), and the sequences of wild type and mutants ARM used for this study are shown. Right, regions corresponding to imino protons involved in AU pairs of HP1 upon interaction with ARM mutants as indicated are shown.

Interestingly, the GAUC motifs stem did not open, and the A38-U68 base pair formation was less evident (except for ARM-P157K). These results suggested an important role of Pro157 and Ser158 for a specific binding, which promotes the opening of the GAUC motifs stem.

Importantly, these results with point mutation of the peptide show that the change observed with ARM are specific and reflect the binding of HEXIM1 and not other effects such as electrostatic compensation of charges (positive amino acids on backbone phosphate) or annealing that has been observed with Tat-like peptides and a wide range of RNAs (Doetsch et al. 2011).

### d. HP1 specificity

Next, we further investigated the importance of the GAUC motifs and the participation of the bulged Us. Bulges are important as recognitions signal, but also favor the opening of grooves allowing insertion of protein elements (Weeks et al. 1993). Thus, three HP1 mutants were tested for their interaction with ARM: HP1ΔU4041 deleted of U40U41 bulge, HP1ΔU63 deleted of U63 and HP1dm with both GAUC motifs mutated into GGCC (but still with U bulges). The secondary structures of these HP1 mutants were determined by NMR (Figure V.4).

Interestingly, A39-U68 base pair is formed in the free HP1dm (GAUC mutated into GGCC), suggesting that the stabilization of H3 promotes a stabilization of the adjacent H2 helix. Titration of HP1 mutants with ARM resulted in an uniform broadening of resonances without significant changes on chemical shifts, and no opening of the H3 stem was observed. These results suggested an important role of bulged Us for specific recognition and for enabling the opening of the GAUC motifs stem. The importance of the GAUC sequence for a specific interaction with HEXIM1 was thus confirmed. Indeed, the presence of two GAUC separated by 10 to 30 nucleotides so that they can form a hairpin is a highly conserved pattern in 7SK and it has been used as a feature to identify 7SK in different species (Marz et al. 2009).

Again, the disappearance of specific effects with the mutated RNAs showed that the method is sensitive and adapted to our purpose.

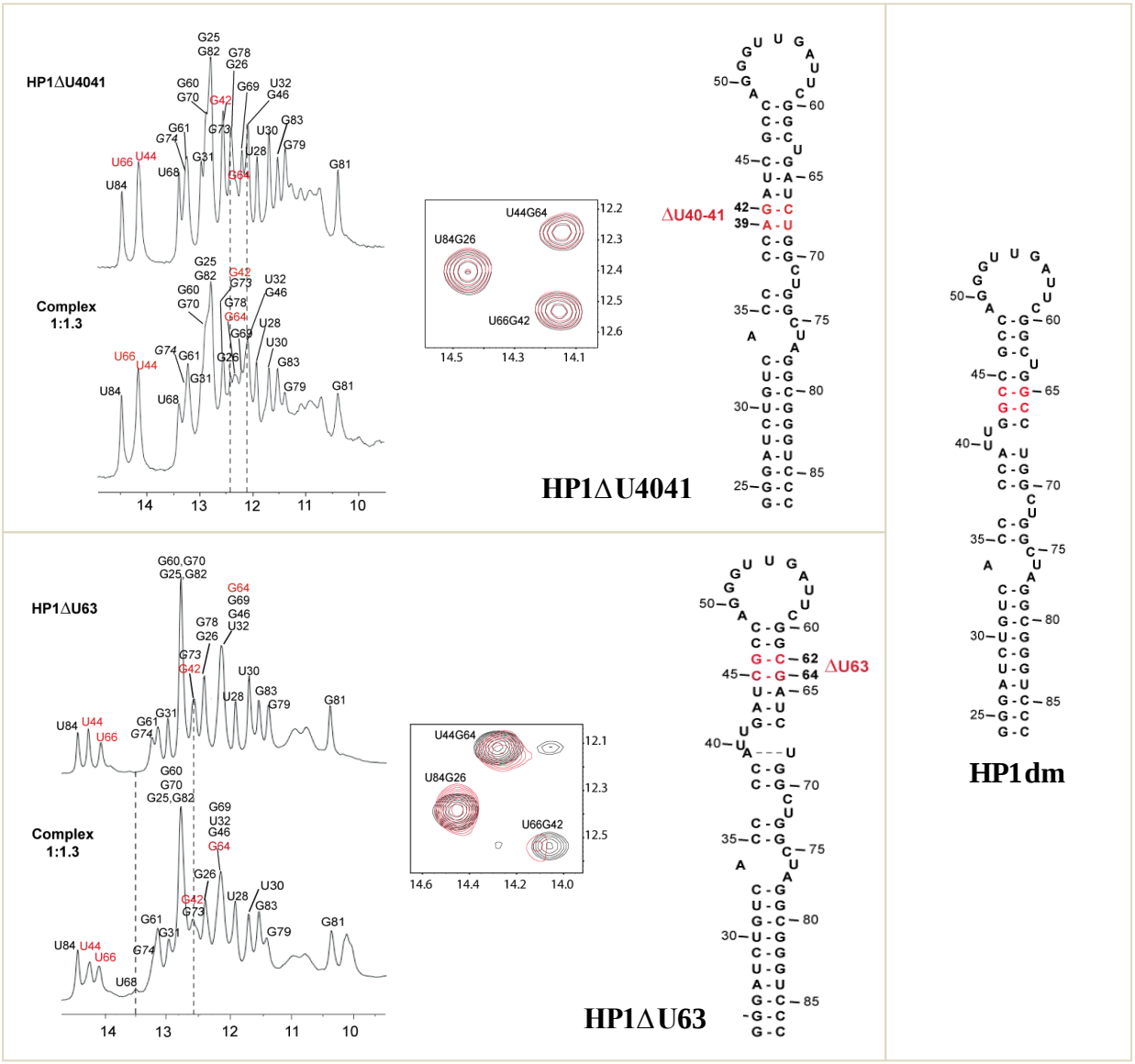



Figure V.4. Role of bulged Us on HP1 interaction with ARM. Left, imino protons region of NOESY experiments of free and ARM-bound HP1 mutants (as indicated). Dashed lines indicate imino proton that undergo chemical shift changes. Middle, imino/imino protons region of NOESY spectrum recorded at 10°C in 90/10 $H_2O/D_2O$ with a mixing time of 300 ms of free (black) and ARM-bound HP1 (red) mutants (as indicated) showing unchanged U84G26, U44G64 and U66G42 correlations. Right, models of the secondary structures of HP1 mutants as determined by NMR, with mutations highlighted in red.

## 1.2. Validation of determinants by EMSA

However, since ARM peptide is only a small region of HEXIM1 and HP1 one subdomain of 7SK, we next investigated the effect of these mutations in the context of full length 7SK and HEXIM1. Hence, the capacities of 7SKΔU4041, 7SKΔU63, 7SKdm, 7SKAU4344GC, 7SKU30C and 7SKG81A mutants for HEXIM1 interaction were tested by EMSA (Figure V.5). We chose to analyse the role of the base pair U30/G81, since a crosslink has been previously identified between U30 and amino acids 210 to 220 of HEXIM1 (Bélanger et al. 2009). Because this region was not included in the ARM peptide, this interaction could not be observed in our NMR experiment. As shown with previous EMSA experiments, HEXIM1 binding was mainly affected by mutations in the GAUC motifs stem and in the bulged Us confirming the NMR results, but not by mutations in U30 and G81 located in the H1 stem of the 5'end hairpin.

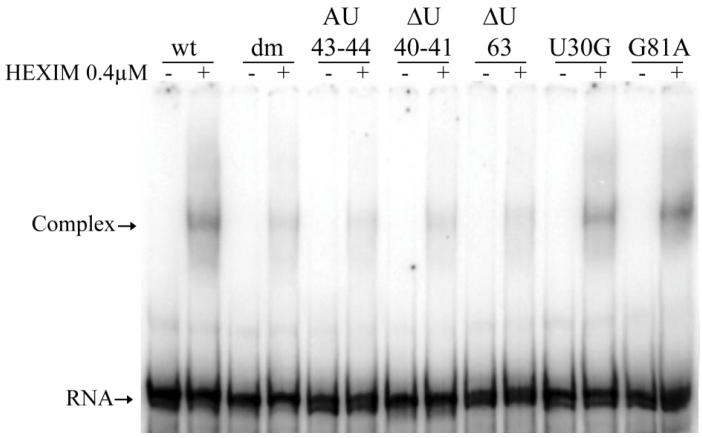



Figure V.5. Survey of the 7SK mutants' capacities for binding to HEXIM1. Different 7SK mutants were tested for binding to HEXIM1 as indicated.

In order to investigate more precisely the effects of the mutations, we decided to use HP1L, which encompasses the whole 5'end hairpin of 7SK. The use of HP1L has several advantages: it was easier to handle, the yield of the radioactive labelling was higher, and the bands in native gels were clearer allowing a more accurate estimation of the effect of the mutations. HP1L also allowed to investigate the basal region of the 5'end hairpin.

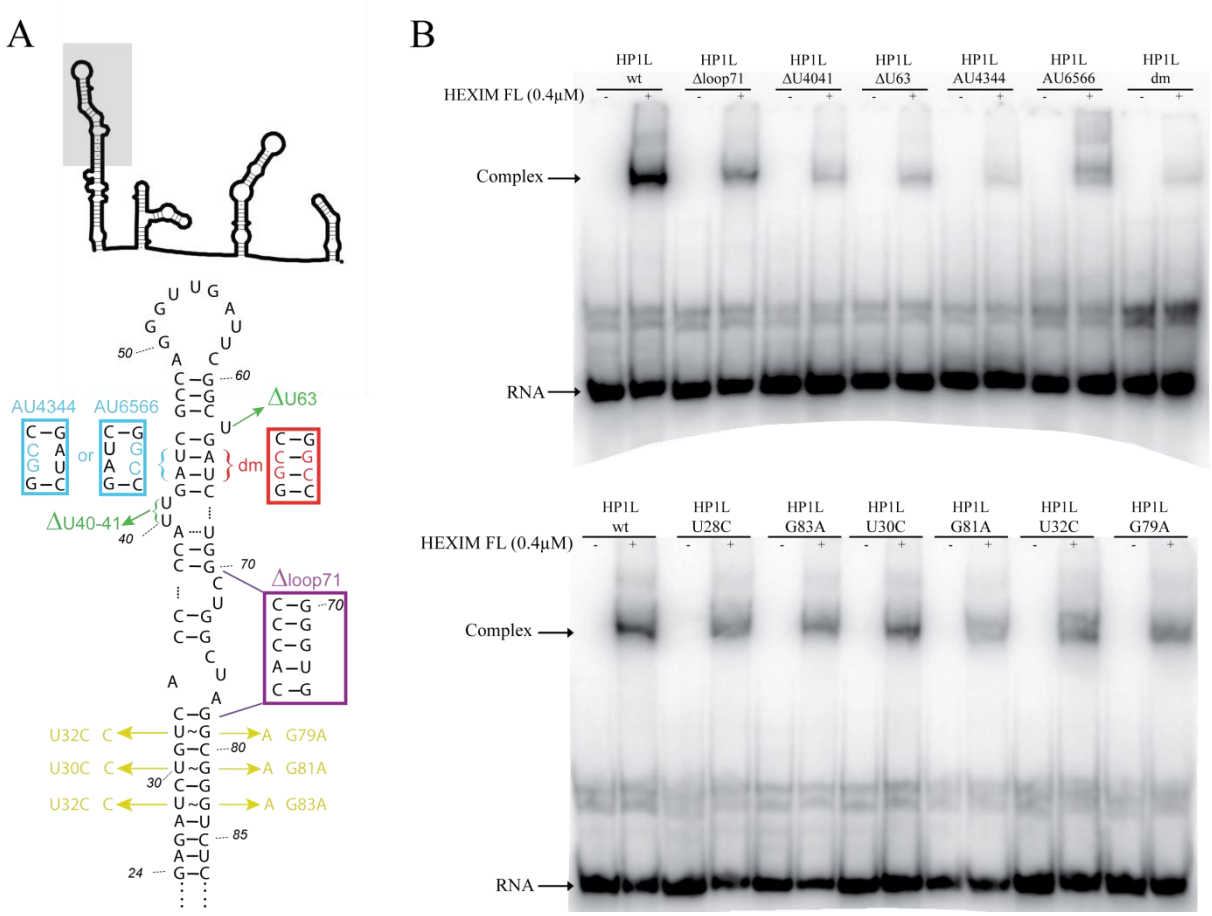



Figure V.6. Survey of the HEXIM1 binding capacities of HP1L mutants. A) Model of the 5'end hairpin of 7SK showing the different mutations tested in this study. B) Effect of mutations in the distal (top) and the proximal (bottom) regions of HP1L in their interaction to HEXIM1 tested by EMSA.

Thus the capacities of the different mutants of HP1L shown in Figure V.6A were tested for their interaction to HEXIM1 by EMSA (Figure V.6B). A general survey showed that, as seen by NMR, mutations in one or both GAUC motifs of H3 or in the bulged Us all affected the interaction with HEXIM1. The suppression of the nucleotides looped out in the region 71 to 77 also decreased the binding to HEXIM1, but in a lesser extent. Three GU base pairs are present in the H1 stem, probably conferring some flexibility to this stem, so we asked if mutations by more stable base pairs could affect the interaction to HEXIM1. Only the mutant G81A seemed to affect slightly the binding to HEXIM1.

When we investigated in more detail the effect of these mutations using a wide range of concentrations of HEXIM1, it was clear that mutations on the apical region of HP1L, but

not in the basal one, impaired the binding to HEXIM1 (Figure V.7). Mutants in the GAUC motif and in the bulged U40U41 had the higher effect, followed by mutants in U63 and in the region 70 to 77, whereas mutants in the basal stem showed a binding comparable to the wild type. Because these observations contradicted the previously reported participation of U30 in the HEXIM1 interaction, competition EMSA experiments were carried out to further confirm them (Figure V.8). While the capacity of HP1Ldm to compete was strongly impaired, the competition capacities of U30C and G81A were similar to the wild type HP1.

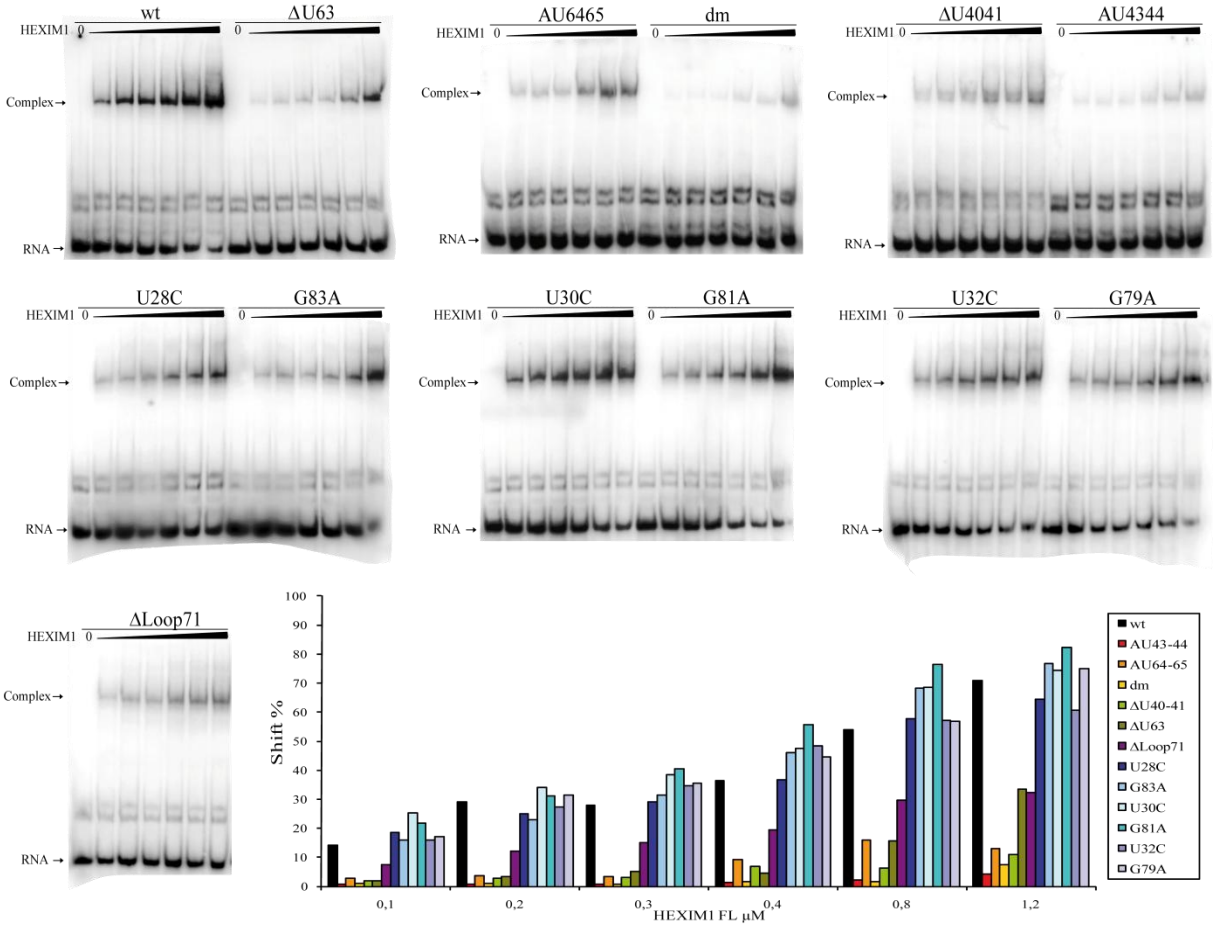



Figure V.7. Analysis of different HP1L mutants by EMSA. Different mutants of HP1L (as indicated) were incubated in absence or presence of increasing concentrations of HEXIM1 (0.1, 0.2, 0.3, 0.4, 0.8 and 1.2 μM) and analysed by EMSA. A summary of the results is shown (right bottom), the percentage of shifted RNA (calculated using Image Quant 5.2 software) is plotted against HEXIM1 concentration.

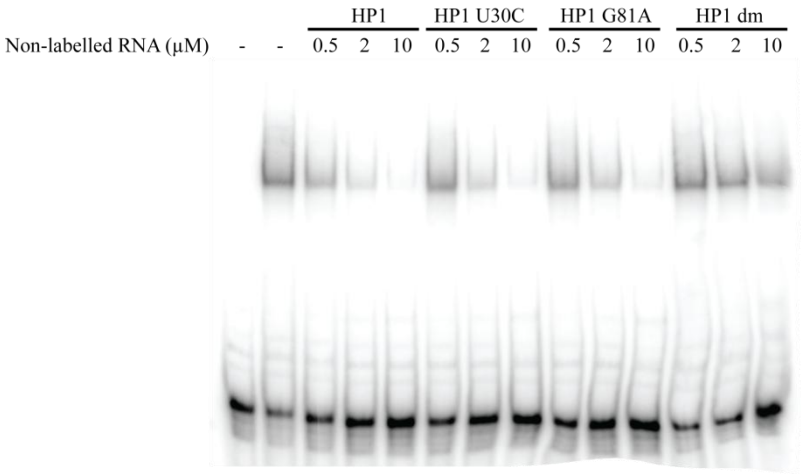



Figure V.8. Competition experiment with different HP1 mutants. Reactions were performed in absence (line 1) or presence of 1 μM HEXIM1 (lines 1-14) and in absence (line 2) or presence of increasing concentrations of the non-labelled RNAs (lines 3-14), as indicated.

We have shown that the ARM of HEXIM1 is able to bind specifically the apical region of HP1, and that the GAUC motif and the bulged Us have an essential role in the recognition. In the full length context of 7SK and HEXIM1, we confirmed the critical role of the GAUC motifs for the interaction. Mutations of the sequence of first GAUC motif (nucleotides 43 to 45) impaired in more extent the binding than the second GAUC motif (nucleotides 64 to 67). A detailed inspection of the mutants in the H1 helix showed a negligible effect of mutations in the GU wobble pairs, and suggested that they are not determinants for the HEXIM1 binding. The suppression of bulged Us compromised the binding, with a stronger impact for U40U41. Analysis of supplementary mutants, for instance the mutation but not deletion of the U bulges, should provide further information about the role of these bulges. In another hand, the internal loops in the middle region of the 5'end hairpin also impact the binding, probably by contributing to the flexibility of the distal region. Some flexibility of 7SK seem to be important as we observed a rearrangement of the apical region of HP1, with the opening of the GAUC motif stem and the base pair A39/G68 formation upon HEXIM1 binding. This rearrangement may be viewed as an induced fit of the 7SK structure to interact with HEXIM1. It may lead to larger conformational changes of the 7SK structure. Indeed, it has been proposed that conformational changes would remodel the

7SK snRNP to allow the recognition and interaction with P-TEFb. This remodelling would be key for the 7SKsnRNP regulation and therefore for transcription control.

In another hand, we showed that the ARM of HEXIM1 contains the determinants for a specific recognition of the GAUC motifs of HP1. This could be surprising, given the strong positive charge of the peptide and the lack of structure. However, we showed that Pro157, and particularly the Ser158 were important for a specfic binding since mutations of any of these aminoacids lead to a loss of the opening of the GAUC motif upon the interaction. It would be interesting to investigate if the mutations in these amino acids, and therefore the loss of the opening of the GAUC motif, affects the P-TEFb inhibitor function of HEXIM1.

HEXIM1 interacts with 7SK through an ARM. Interestingly, ARM are generally found in viral and phage proteins involved in RNA recognition (Patel 1999). It has been observed that although related at the primary sequences, ARMs from different proteins adopt different conformation depending on the RNA site recognized. It has been also shown that a single ARM can specifically recognize different RNA sites with different binding strategies (Smith et al. 2000). The ARM of HEXIM1 and its mainly unstructured nature may allow its interaction with multiples partners. Little is known about HEXIM1 when not associated to 7SK. However, only around 25% is found within the 7SK snRNP (Byers et al. 2005), and it has been reported also associated with other RNAs (Li, Cooper, et al. 2007). Also, it has been observed that HEXIM1 is also able to bind TAR and then bind and inhibit P-TEFb in vitro (Sedore et al. 2007), however it is not clear if this interaction exits in vivo. Indeed, we observed that HEXIM1 binds easily to bacterial RNAs during protein extraction.

These results opened several questions about the interaction of 7SK and HEXIM1. We observed that one ARM interacts with the apical region of HP1, but actually HEXIM1 is a dimer. Several possibilities can be imagined to account for the role of the second monomer. A simple explanation is that only one monomer of HEXIM1 interacts with 7SK. Another is that both monomers interact with 7SK, both at the GAUC motif or the second monomer with a different region. Other open question is if another region of HEXIM1 also participates, such as the amino acids 210 to 220 previously reported. Hence, we performed further investigations to try answering these questions.

Recently, it has been shown that Tat and HEXIM1 are able to bind 7SK in mutually exclusive way (Sedore et al. 2007; Muniz et al. 2010). This suggests that both proteins interact with an identical or overlapping region of 7SK or that they bind different conformations of 7SK. Also, 7SK has been observed associated to Tat in cells (Sobhian et al.

2010; D'Orso et al. 2010). Thus, HP1 may be also important in the HIV transactivation mediated by Tat. It would be interesting to know if the binding of Tat leads to the same conformational effects in HP1.

## 2. Characterization of 7SK/HEXIM complex by Native Mass Spectrometry

Our NMR approach revealed that one ARM peptide was enough to recognize the GAUC sequence of HP1, and pointed it out as a main binding site for HEXIM1. Specific effects on NMR spectra were observed, at least for addition up to a ratio of 1.3 peptide/HP1, but further addition of peptide lead to resonance broadening. However, HEXIM1 is able to dimerize via its C-terminal coiled coil, and several publications report that more than one HEXIM1 and P-TEFb molecule bind to one single 7SK (Blazek et al. 2005; Byers et al. 2005; Dulac et al. 2005; Li et al. 2005; Dames et al. 2007). Moreover, our gel shift experiments showed the appearance of a second band of complex with some RNA and protein constructions. These bands may be interpreted as the binding of a second protein, but EMSA experiments do not provide information about the stoichiometry of the observed complexes. Hence, a main issue was to understand what happens with the second monomer of HEXIM1, which also contains the RNA-binding ARM. Are two HEXIM1 monomers necessary for the binding of one 7SK? If yes, do both monomers bind together to the same RNA region? Or, is there a second binding site, and where is it? We wanted also to assess the oligomeric state of our HEXIM1 constructions.

Recently, a second HEXIM1 binding site located in the basal part of the 5' end hairpin has been proposed (Muniz et al. 2010). It comprises the region 18 to 27 and 84 to 95 (Figure V.9A) and was suggested to work in an interdependent fashion with the apical HEXIM1 binding site. This leads to hypothesize that each HEXIM1 of the dimer binds a different site on the 5' end hairpin (Figure V.9B). We wanted to gain some insights about the stoichiometry of 7SK/HEXIM1 complex since this is a prelude to more detailed structural studies, and essential for determining the minimal complex for crystallization trials. For this purpose, we carried out native mass spectrometry (MS) measurements, in collaboration with Sarah

Sanglier, Jean-Michel Saliou and Cédric Atmanene from the Laboratoire de Spectrométrie de Masse Bio-Organique (LSMBO), at the Institut Pluridisciplinaire Hubert Curien (IPHC).

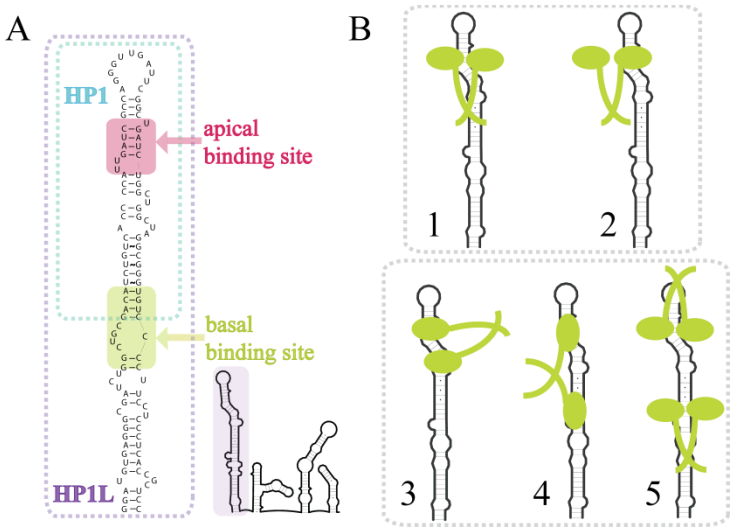



Figure V.9. The HEXIM1 binding sites on 7SK. A) Both, HP1 and HP1L with one and two sites, respectively, are indicated. B) Schematic representation of different models of interaction between HEXIM1 and 7SK: one site (upper panel), symmetric (left) or asymmetric (right) interaction; two sites (bottom panel), one monomer on adjacent (left) or remote (middle) sites, or one dimer on each site (right).

## 2.1. HEXIM1 binds HP1 as a dimer

The quality of the samples was controlled by MS in denaturing conditions in $H_20/CH_3CN/HCOOH$ (50/50/1). Then, the spectra of HP1 and HEXIM1 were analyzed in native conditions in 250 mM ammonium acetate pH 7.5 (Figure V.10). HP1 presented a population of 20.8 kDa in agreement with a monomer state. Surprisingly, the native spectrum of isolated HEXIM1 revealed two populations (41.5 and 83.3 kDa) consistent with monomers and dimers of HEXIM1, respectively. An explanation may be that the dimer was disrupted during the ionization process. Indeed, we observed that the ratio between the monomers and dimers also depended on voltage conditions of the measurement. The better efficiency of detection of smaller molecules led also to an overestimation of monomer population. We shall

see later that a monomer population was an interesting opportunity to compare the affinities of the monomer and the dimer for HP1. These results confirmed however that HEXIM1 forms a dimer even in the absence of RNA.

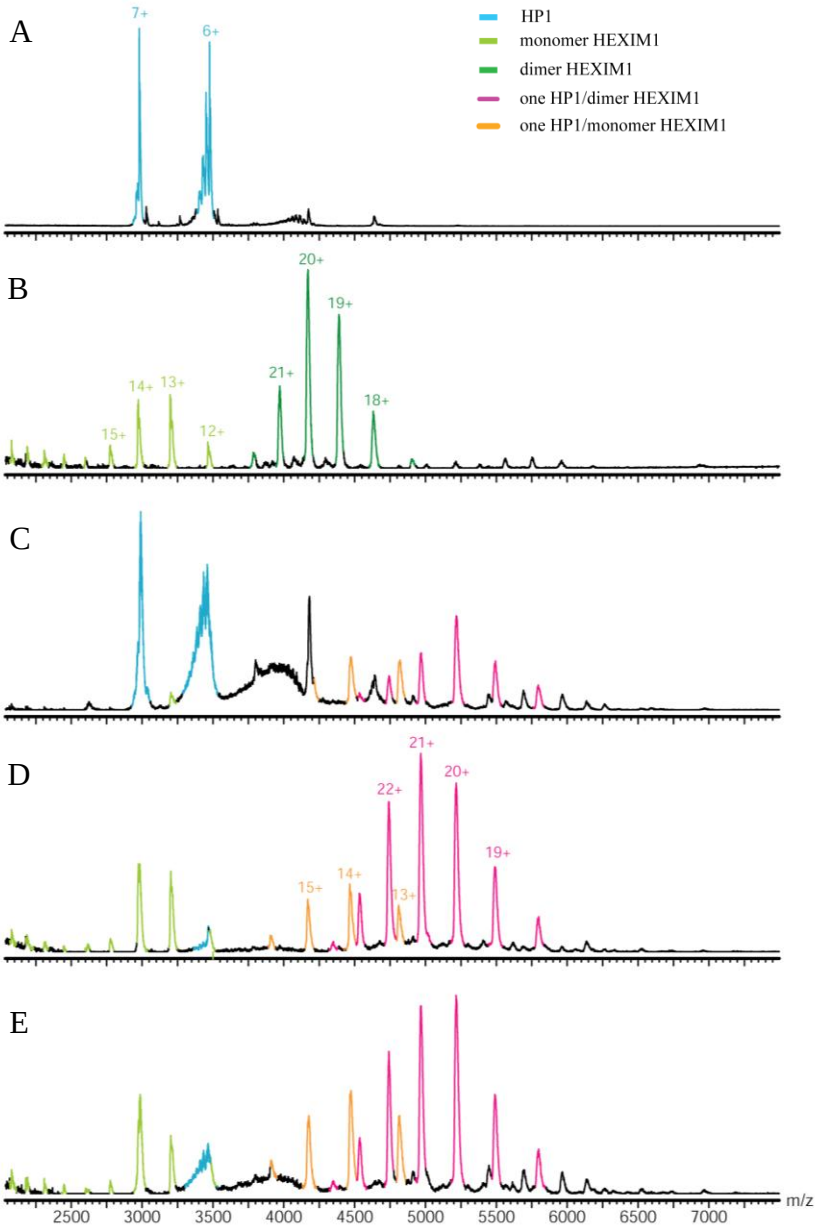



Figure V.10. HEXIM1 binds preferentially as a dimer. A) HP1 in native conditions. B) HEXIM1 in native conditions showing populations consistent with monomers and dimers. C, D, and E) HP1 and HEXIM1 mixed (ratio RNA/Protein of 1:1, 1:2 and 1:4, respectively) showing complex formation.

When HP1 was titrated with HEXIM1, we observed a population consistent with a complex of one HP1/dimer HEXIM1 (104.5 kDa). Only when the concentration of HEXIM1 was further increased, a small population consistent with a complex of one HP1/monomer HEXIM1 (62.6 kDa) appeared. Interestingly, a population of free HEXIM1 monomer was permanently present. On the whole, this experiment showed that HEXIM1 binds HP1 as a dimer and suggested that the dimer HEXIM1 has a higher affinity than the monomer HEXIM1 for HP1.

## 2.2. Control of the specificity of the interaction

In order to verify the specificity of the observed interaction in our experimental conditions, the binding between HP3 and HEXIM1 was tested (Figure V.11).

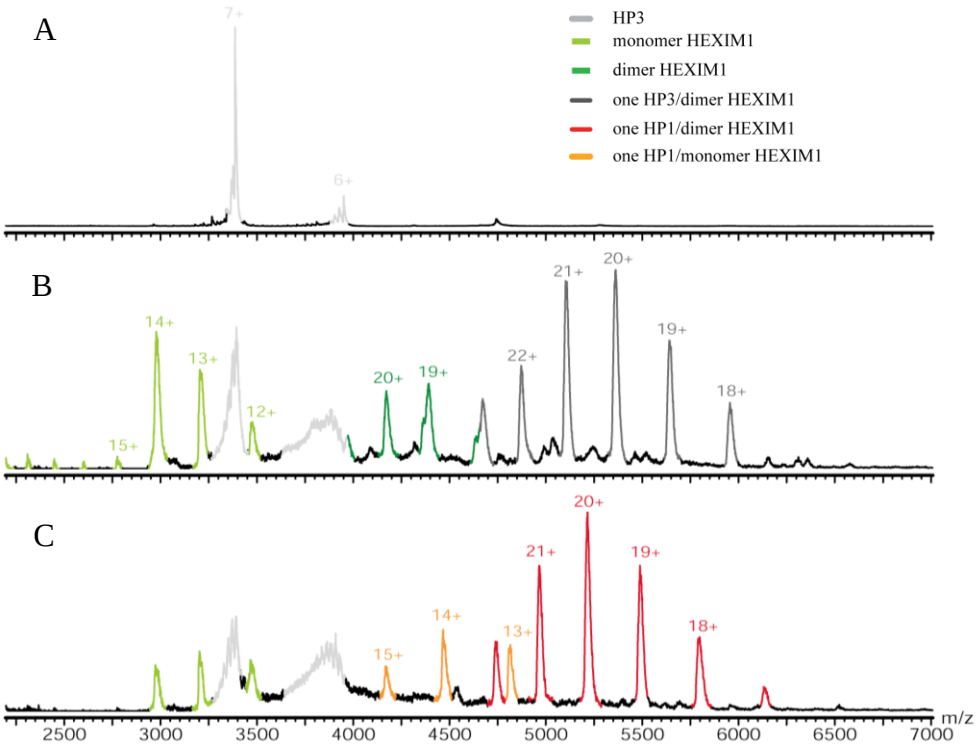



Figure V.11. HEXIM1 binds preferentially HP1 than HP3. A) HP3 in native conditions. B) HP3 and HEXIM1 mixed (1:1) showing complex formation. C) HP1 and HP3 competition experiment for HEXIM1 binding (ratio HP1:HP3:HEXIM1 of 1:1:2). Only a population consistent with a complex HP1/HEXIM1 is observed.

HP3 showed a population consistent with a monomer (23.7 kDa). When HP3 and HEXIM1 were mixed, a population consistent with one HP3/dimer HEXIM1 (107 kDa) was observed. This was actually not so surprising, since it has been reported that HEXIM1 is a promiscuous dsRNA binding protein (Li, Cooper, et al. 2007). However, SEC and EMSA experiments showed that HEXIM1 binds specifically HP1 but not HP3. Thus, we tested if competition experiments in the MS conditions allowed distinguishing the specific interaction of HEXIM1 for HP1 from the non specific ones (with HP3). Indeed, when HEXIM1 was mixed with identical ratio of both HP1 and HP3, only a population consistent with HP1/dimer HEXIM1 (104.1 kDa) was observed (Figure V.11C). These results suggest that HEXIM1 is able to bind a hairpin different than HP1, but that it has a higher affinity for HP1, as previously reported (Czudnochowski et al. 2010). Also, this showed that competitions experiments are a useful tool to distinguish specific interactions.

## 2.3. HP1L or HP1 bind one HEXIM1 dimer

We then hypothesized that one monomer in the dimer of HEXIM1 could interact with the apical binding site of 7SK (as seen by NMR) and the other monomer with the basal binding site, proposed by (Muniz et al. 2010). Each of the sites would contribute to the affinity of 7SK/HEXIM1 complex [model B4 in Figure V.9]. Consequently, HP1L that contains both sites, should be bound with higher affinity by the HEXIM1 dimer than HP1, which contains only one site (Figure V.9A). A different possibility offered by the extended HP1 is that two dimers could bind (model B5 in Figure V.9). This would explain the appearance of slow-migrating band on EMSA gels but would contradict all evidences suggesting that one dimer HEXIM1 binds one molecule of 7SK (Blazek et al. 2005; Byers et al. 2005; Dulac et al. 2005; Qintong Li et al. 2005).

The interaction between HP1L and HEXIM1 was analysed (Figure V.12). HP1L was present in a population consistent with a monomer state (34.8 kDa). When HP1 L was titrated with HEXIM1, a population consistent with one HP1L/dimer HEXIM1 complex (112.7 kDa) was observed. A minor population consistent with a HP1L/monomer HEXIM1 complex (76.9 kDa) was also detected, but disappeared when HEXIM1 concentration was increased. No

population consistent with a HP1L/tetramer HEXIM1 was observed, supporting the previous reports showing that only one dimer of HEXIM1 interacts with 7SK.

Next, we tested if HEXIM1 has a higher affinity for HP1L than for HP1 (Figure V.13). For this, a competition experiment between HP1 and HP1L for the binding to HEXIM1 was performed. When HEXIM1 was in presence of both of HP1 and HP1L at identical ratio, both populations consistent with a HP1/dimer HEXIM1 and a HP1L/dimer HEXIM1 (104.5 and 112.7 kDa, respectively) were detected. Surprisingly, no evidences that would suggest a higher affinity for HP1L was observed. These results suggested that the extension of HP1L to the second proposed HEXIM1 binding site does not provide any advantage for the interaction.

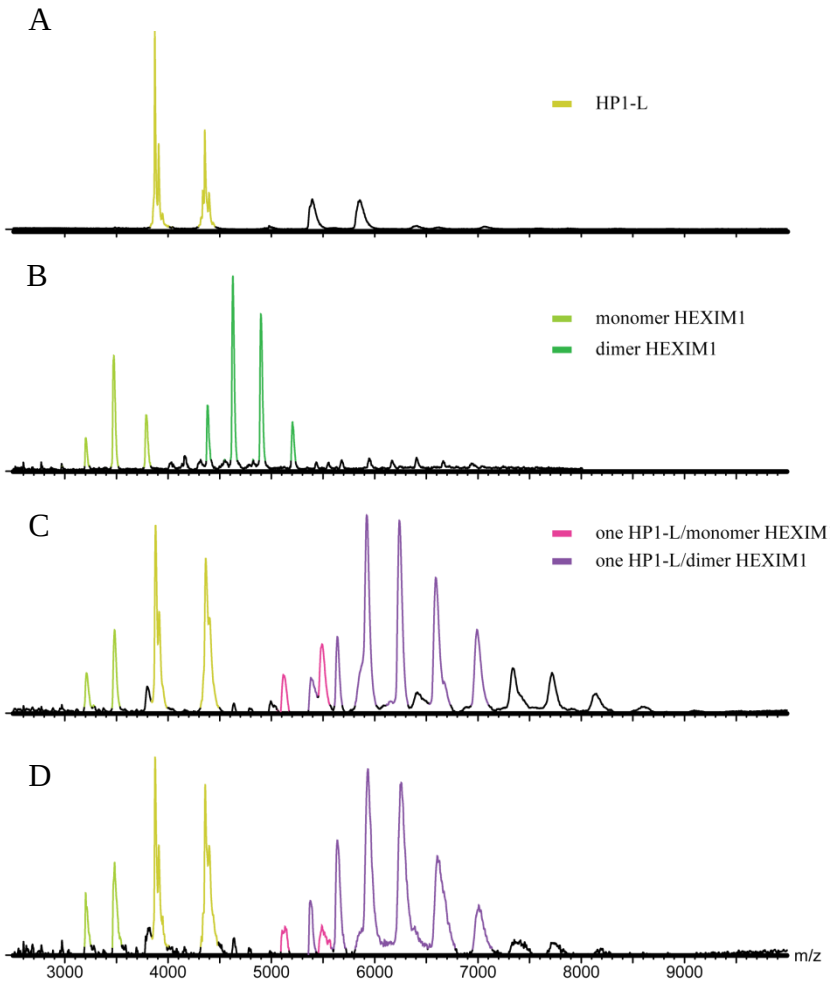



Figure V.12. HEXIM1 binds HP1 L. A) HP1L in native conditions. B) HEXIM1 in native conditions. C, and D) HP1L and HEXIM1 mixed (1:1, and 1:2, respectively) showing complex formation.

At this stage, we were left with several options to understand the participation, if any, of the second monomer of HEXIM1. The second monomer could be binding near or on the GAUC site, giving rise to a semi-symmetrical model, with HEXIM1 closing on the RNA. Alternatively, the second monomer could be not binding the RNA, and instead could be involved in recruitment of some other partner, for instance. To test these models, we used monomeric HEXIM1.

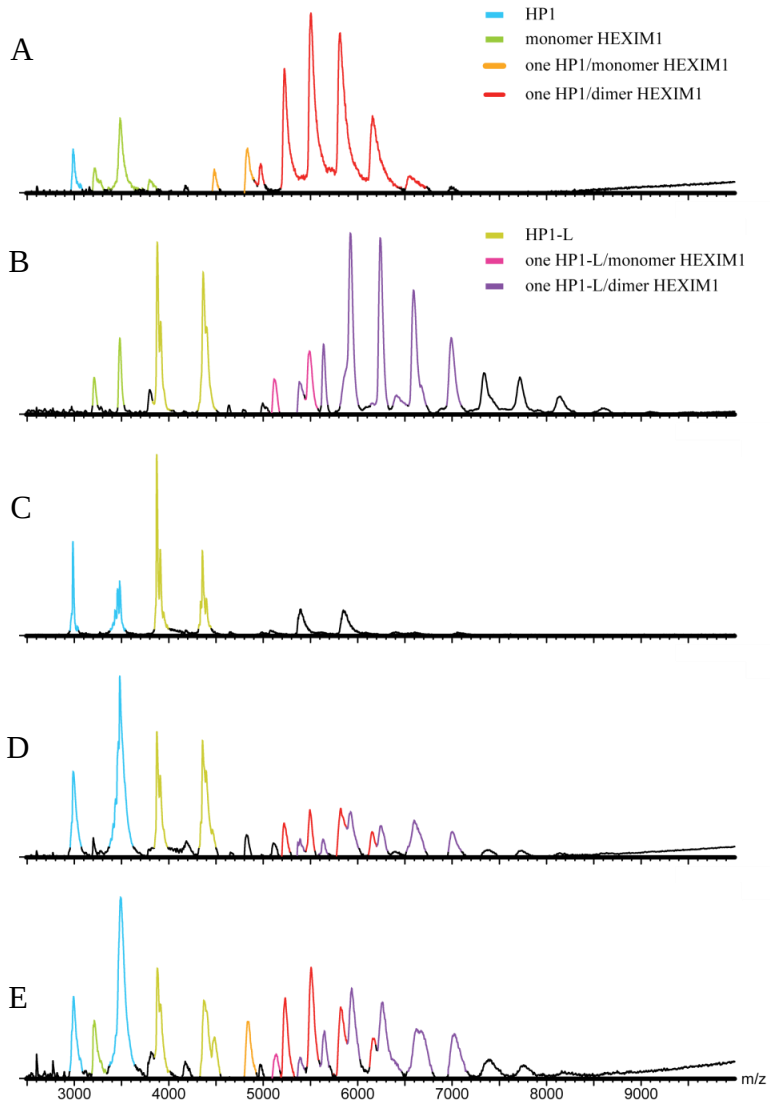



Figure V.13. HEXIM1 binds both HP1 and HP1L in competition experiment. A) HP1 and HEXIM1 were mixed (1:1), a population consistent with a complex is observed. B) HP1L and HEXIM1 mixed (1:1) showing complex formation. C) HP1 and HP1 L (1:1) mix. D and E) HP1 and HP1L competition experiment for HEXIM1 binding (1:1:2, and 1:1:3, respectively). Both HP1/HEXIM1 and HP1L/HEXIM1 complexes are observed.

## 2.4. Oligomeric state of HEXIM1 deleted of the coiled coil

To analyse in more explicit way the existence of a second HEXIM1 binding site on 7SK, we used the constructions HEXIM1 114-317, and HEXIM1 136-273. EMSA suggested that HEXIM1 114-317 is in a dimer/monomer equilibrium, while HEXIM1 136-273 is probably a monomer, following the extent of deletion of their C-terminal coiled coil. This was confirmed by the MS analysis of HEXIM1 114-317, which revealed two populations consistent with monomers and dimers (23.7 and 47.4 kDa, respectively; data not shown). The ratio between the monomer and the dimer populations was more important than that observed for wild type HEXIM1 suggesting that the shortening of the coiled coil destabilizes the dimer, as previously reported in a study showing the importance of the two segments in the TBD (Schönichen et al. 2010). When HEXIM1 114-317 was titrated with HP1, only a population consistent with a monomer HP1/dimer HEXIM1 114-317 complex (68.3 kDa) was observed. This result confirmed that the dimer HEXIM1 binds HP1 with higher affinity than the monomer HEXIM1. It is also possible that the presence of HP1 stabilizes the dimer HEXIM1 114-317 as previously proposed (Blazek et al. 2005).

## 2.5. HP1L binds two HEXIM1 monomers

HEXIM1 136-273, as expected, showed only one population consistent with a monomer protein (18.7 kDa). This offered an opportunity to inquire more explicitly into the participation of the second monomer, and see whether it binds RNA, or not. Since the publication by (Muniz et al. 2010) proposed a second binding site on the region 18 to 27 and 84 to 95 we measured the mass of complexes with the monomeric HEXIM1 and HP1L, as shown in Figure V.14.

When HP1L was titrated with HEXIM1 136-273, a population consistent with one HP1L/monomer HEXIM1 136-273 complex (53.5 kDa) appeared first (Figure V.14). Increasing HEXIM1 136-273 concentration led to a new population consistent with two monomers bound to HP1L in a complex of 72.2 kDa, supporting the existence of a second protein binding site on the 5'end hairpin of 7SK. At higher HEXIM1 136-273 concentration, a small population of complex with three monomers on one RNA could be detected.

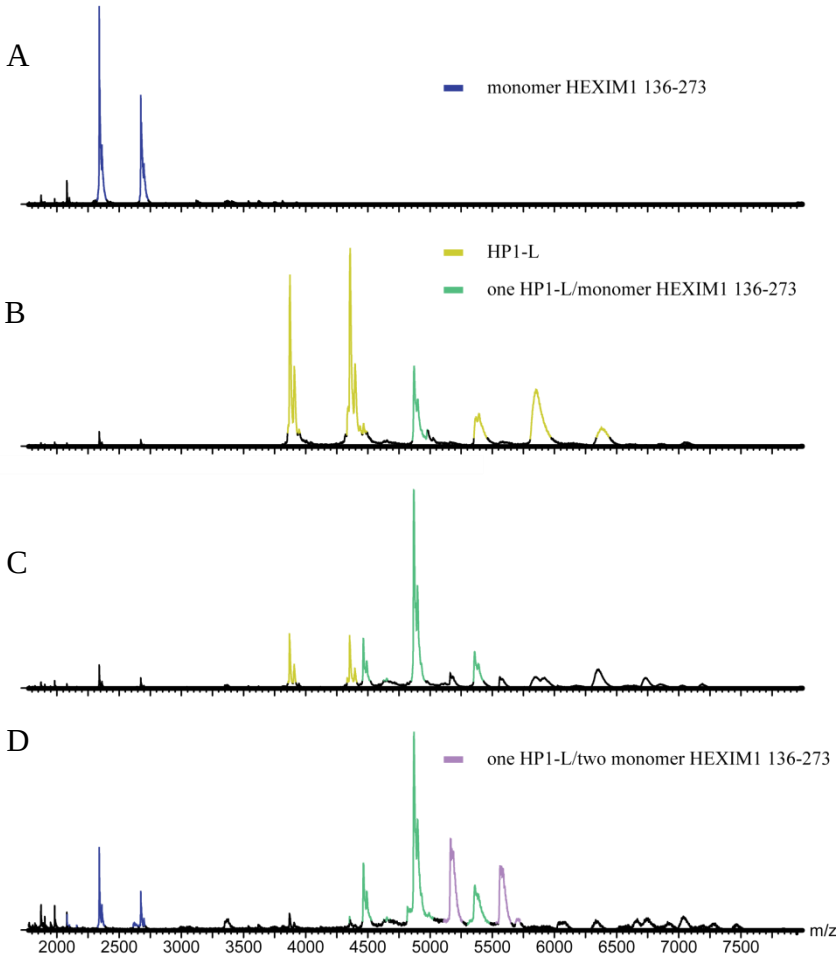



Figure V.14. Monomeric HEXIM1 136-273 binds HP1L. A) HEXIM1 136-273 in native conditions. A population consistent with a monomer is observed. B, C, and D) HP1L and HEXIM1 136-273 mixed (1:1, 1:2, and 1:1:4, respectively). The formation of one HP1L/one HEXIM1 136-273, one HP1L/ two HEXIM1 136-273, and one HP1L/three HEXIM1 136-273 complexes are observed.

When HP3 was titrated with HEXIM1 136-273, only a small population of complex 1:1 was observed, even at high monomer concentrations. This suggested that HEXIM1 136-273 has a higher affinity for HP1. We also concluded that the observed third site on HP1 is non-specific, and that the correct stoichiometry is 1:2, meaning that the second monomer participates to the RNA binding. This is consistent with, and explains a previous report, which concluded to a dimerization of HEXIM1 mediated by 7SK. Indeed, a mutant Flag-HEXIM11-

278 expressed in HeLa cells was able to immunoprecipitate HEXIM1 only in the presence of 7SK, but not a Flag-HEXIM11-150 lacking the RNA binding site (Blazek et al. 2005).

## 2.6. HP1 contains two HEXIM1 binding sites

To further question the localization of the binding site of the second monomer, we analyzed the interaction between HEXIM1 136-273 and HP1. When HP1 was titrated with HEXIM1 136-273, a population consistent with a HP1/monomer HEXIM1 136-273 complex (39.5 kDa) was observed (Figure V.15). Surprisingly, when the concentration of HEXIM1 136-273 was further increased, a population consistent with one HP1/two monomers of HEXIM1 136-273 (58.3 kDa) appeared. Actually, the profiles of the titration of both RNAs, HP1 and HP1L, were very similar.

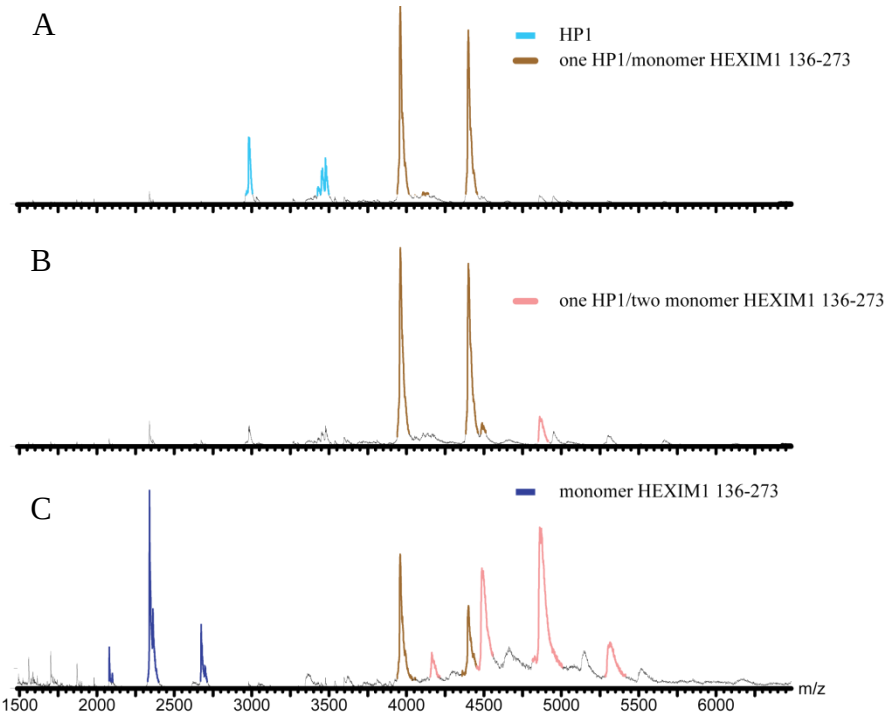



Figure V.15. HP1 binds two monomeric HEXIM1 136-273. A, B, and C) HP1 and HEXIM1 136-273 mixed (1:1, 1:2, and 1:1:4, respectively). The formation of one HP1/one HEXIM1 136-273, one HP1/ two HEXIM1 136-273, and one HP1/three HEXIM1 136-273 complexes are observed, as for HP1L.

So far, these results show that HEXIM1 binds two sites both on HP1, but that one site shows a lower affinity.

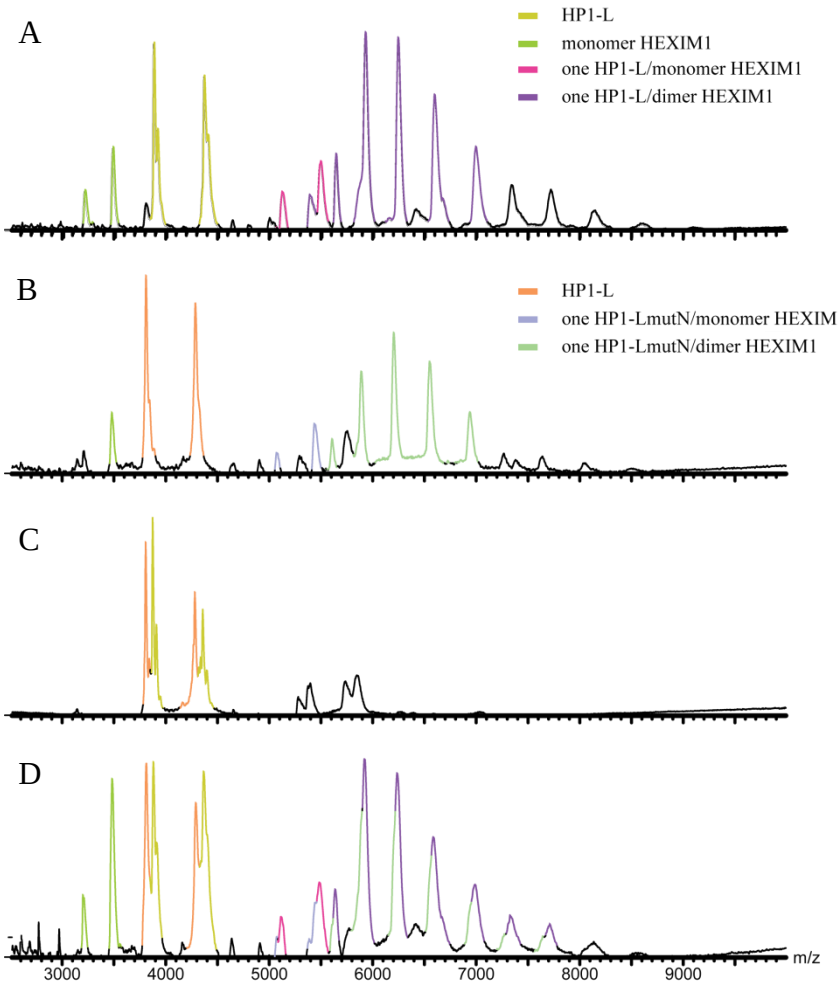



Figure V.16. HP1L and HP1L-mutN competition for HEXIM1 binding. A) HPIL and HEXIM1 were mixed (1:1), a population consistent with a complex is observed. B) HP1L-mutN and HEXIM1 were mixed (1:1), a population consistent with a complex is observed. C) HP1L and HP1L-mutN (1:1) mix. D) HP1L and HP1L-mutN competition experiment for HEXIM1 binding (1:1:2). Both, HP1L/HEXIM1 and HP1Lmut-N/HEXIM1 complexes are observed.

Given that our observation contradicted the location of the second binding site at the basal stem of HP1L, we decided to test directly this site by mutation analysis (Figure V.16). We used the HP1L-mutN, in which the sequence CUA (nucleotides 15 to 17) has been

mutated into A, resulting in the suppression of an internal loop. It was previously reported that truncation or alteration of the region C12-A27/U84-U95 abolished the interaction of HEXIM1 with this binding site. However, competition experiments for HEXIM1 binding between HP1L and HP1L-mutN showed the coexistence of both complexes, suggesting that HP1 L and HP1 L-mutN have a similar affinity for HEXIM1. In line with this, the populations of both free RNAs were equivalent in the presence of HEXIM1.

## 2.7. Towards a localization of the second binding site

The observation of sequential binding during the titration can reflect two extreme situations (but an intermediate model is also possible). (1) The second binding site is of weak affinity, and is bound only when the main protein binding site is saturated with the first monomer. (2) The second binding site is formed only upon binding of the first ARM. This is an appealing hypothesis favoured by to the observation of conformational changes of the GAUC site upon ARM binding. Indeed the GAUC motif stem opens, but the adjacent stem is stabilized, with the concomitant closing of A39-U68 base-pair.

In order to see whether the first binding was decisive for the second one, we used a mutant of the main binding site, called HP1L-mutU. This mutant has been completely deprived of the determinants for HEXIM1 interaction, with the deletion of the two bulges and the mutation of both GAUC into GGCC sequence. The resulting mass difference with HP1L is clearly measurable by MS ($\Delta$ = ~5 kDa). The experiment with this mutant was initially intended to assess the presence of a site in the basal region of the 5' hairpin (Muniz' basal site).

When HP1L-mutU was titrated with HEXIM1 136-273 (Figure V.17), a population consistent with one HP1L-mutU/one monomer HEXIM1 136-273 complex (52.6 kDa) was detected only at very high concentrations of HEXIM1 136-273, much higher than for HP1L and comparable to those at which we observed the binding to HP3. This let us to think that this interaction could be non-specific. To test the specificity of this interaction, we performed competition experiments using HP3. When HEXIM1 136-273 is mixed with HP1L-mutU and HP3, both in the same ratio, a small population of one HP1L-mutU/one monomer HEXIM1 136-273 was still observed. Concomitantly, a decrease of the population of free HP1L-mutU was clearly observed, but no of the free HP3. Interestingly, the peaks

corresponding to the HP1L-mutU/monomer HEXIM1 136-273 complex appeared at the same concentration to which the second monomer of HEXIM1 136-273 bound HP1.

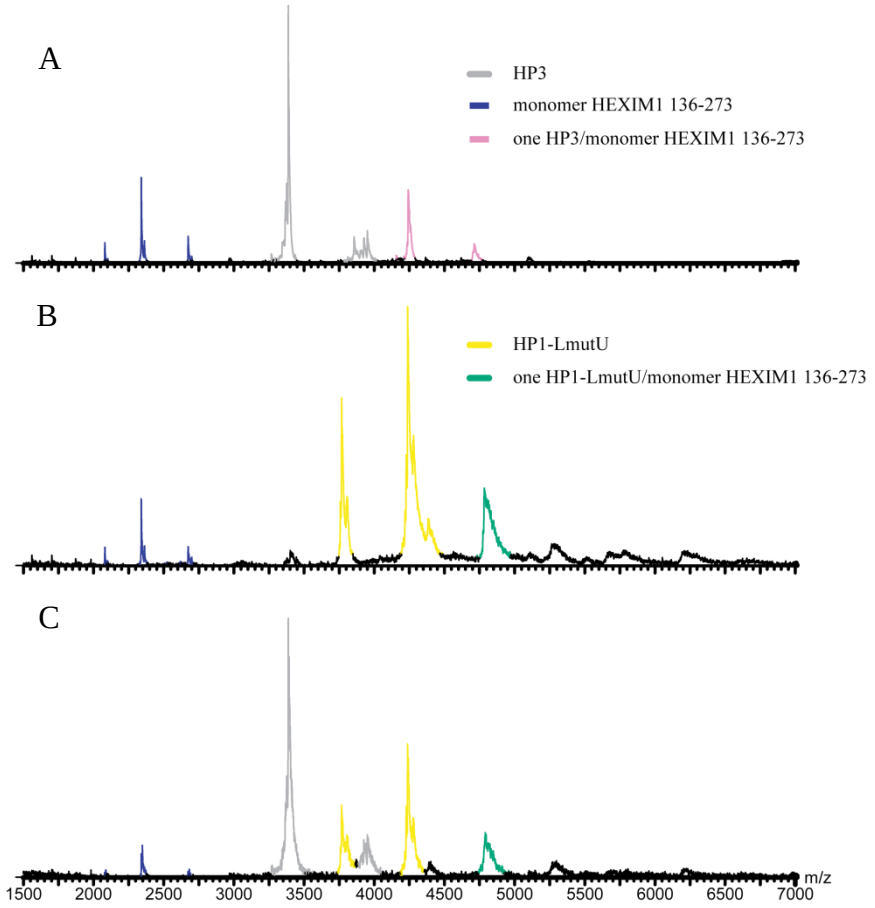



Figure V.17. HEXIM1 binds preferentially HP1L-mutU than HP3 in competition experiment. A) HP3 and HEXIM1 136-273 were mixed (1:2), a population consistent with a complex is observed. B) HP1L-mutU and HEXIM1 136-273 were mixed (1:2), a population consistent with a complex is observed. C) HP3 and HP1L-mutU competition experiment for HEXIM1 binding (1:1:2). Only HP1L-mutU/HEXIM1 1-273 complexes are observed.

These results suggested that HP1L-mutU still contains the second, lower affinity site. However, if both sites are interdependent, the low affinity may be also due to the loss of the main binding site. Thus, further investigations should be done to test this hypothesis.

## 3. Discussion and Conclusions

On the whole, the MS study of 7SK/HEXIM1 complexes shows that both HEXIM1 monomers participate to the binding, the dimer HEXIM1 binding preferentially the RNA. A first binding, most probably to the specific GAUC site described by our NMR study is followed by a second binding event, which is less strong but still specific, to an adjacent site comprised in HP1 (nucleotides 24 to 87) in contradiction with a previous report (Muniz et al. 2010). However, the precise location of the second binding site remains an open question.

Our studies with MS confirmed the weak specificity already reported for HEXIM1, and not surprisingly in view of the sequence of the ARM RNA binding site, rich in positively charged amino acids, and of the unstructured nature of HEXIM1 in the region responsible for RNA binding. This lack of specificity blurred our research, and prompts us now to turn in other techniques to further delineate the sequence of the RNA involved in the binding of the second monomer. EMSA seems a good technique for that, provided that it is performed in the presence of high concentration of tRNA, acting as non-specific competitor. However, EMSA also showed some bands that were not easy to interpret. These appeared with ageing protein, and were therefore hypothesized to reflect aggregates. The MS analysis suggests that, while these bands do not reflect multiple HEXIM1 dimers binding to the same RNA (never observed), they could however correspond to HEXIM1 dimers binding two RNAs (a very small amount of that type of complexes was seen in some experiments). This could also be due to artefacts linked to the high concentration (several micromolar) of each component, or to the absence of the other 7SK partners, such as LaRP7, which can be imagined to decrease the available region of 7SK accessible to HEXIM1.

The confrontation of these recent results with NMR mapping leaves open several questions. A strong concern is why the second binding event was not seen by NMR. This may be explained by the shortness of the ARM peptide used (amino acids 149 to 179). Thus, the second binding site could be ascribed to another region of HEXIM1, such as the region 210 to 220 pointed out by (Bélanger et al. 2009). In order to inquire into that, we analyzed the interaction between HP1 and a MBP-ARM protein (ARM was fused to the MBP in order to gain a more significant shift) by EMSA. Surprisingly, two complexes were formed (Figure V.18). Both bands are specific and are strongly reduced with a mutated HP1, deprived of one of the U bulges. First, this result suggested that the ARM of the second monomer should be

involved in the interaction to the second binding site. Second, it suggested that both binding sites would be interdependent since the loss of the main binding site leads to the loss of the second one. A better affinity for the second site upon binding on the first one is an appealing idea, in the line with the observed strengthening of the H2 stem of HP1 upon ARM binding by NMR. Thus, the interdependency of the sites is still another question to solve. The fact that the second binding site was not seen by NMR may be then explained if it does not involve changes in the imino-protons.

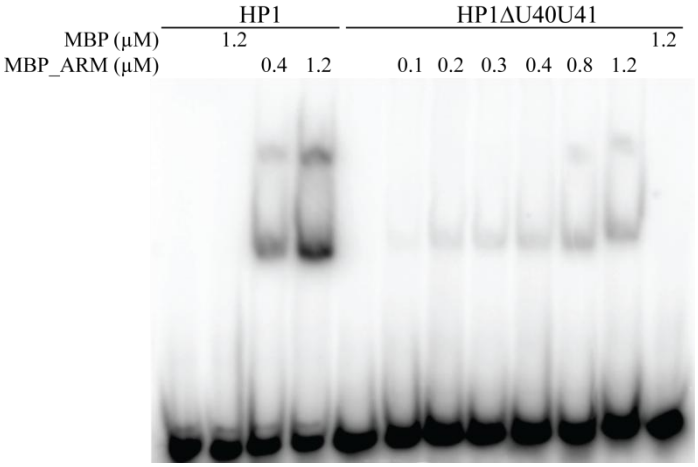



Figure V.18. EMSA analysis of the binding of MBP_ARM-NLS to HP1. Reactions were performed in absence or presence of MBP (control) or MBP_ARM-NLS proteins as indicated. Wild type and ΔU4041 HP1s were tested.

NMR mapping provided useful information that can be used as starting point for searching the second binding site. Indeed, it showed some stabilization at the base of the loop G70, and we described with EMSA analysis of mutants that this loop was contributing to the binding, so the loop at G70 should be investigated carefully. In addition, a crosslink between U30 and HEXIM1 has been reported (Bélanger et al. 2009), so this region should be also explored. Thus, the precise localization of the second site will require further experiments.

# CHAPTER VI :
# OTHER PROTEINS OF THE 7SK SNRNPS

When I started my thesis project in 2007, new partners of 7SK consisting of a subset of heterogeneous ribonuclear proteins (hnRNP) Q, R, A1, A2, K as well as the RNA helicase A (RHA) were identified by three different teams (Barrandon et al. 2007; Van Herreweghe et al. 2007; Hogg et al. 2007). It has been found that 7SK association to these proteins corresponded to a state where 7SK was release from P-TEFb and HEXIM1 (Barrandon et al. 2007; Van Herreweghe et al. 2007). Hence, it has been proposed that these hnRNP participate in the dissociation of the inactive 7SK/HEXIM1/P-TEFb complex by remodeling the 7SK conformation or stabilize the P-TEFb free 7SK. They are therefore important for the control of the 7SK.

The hnRNPs are RNA-binding proteins which are very diverse in structure, abundance, tissue specificity, and function. They participate in transcription regulation, telomere-length maintenance, splicing, RNA 3'end processing or mRNA nucleo-cytoplasmic transport, etc. They all contain RNA-binding motifs such as RRM (RNA-recognition motif), KH (K homology) domains and RGG (ArgGlyGly) boxes and have modular structure (Dreyfuss et al. 2002). We were interested in this regulation and hoping to be able to obtain structural insights into the binding specificity, we started working on the hnRNP K and A1. The hnRNP K contains three KH domains that mediate RNA and DNA binding, and is associated to multiple processes like chromatin remodeling, transcription, splicing, translation and mRNA stability (Bomsztyk et al. 2004). The hnRNP A1 participates in splicing, transport, stability and translation of mRNA and consists of an N-terminal domain called UP1 (unwinding protein 1) which contains two RRMs, and a C-terminal domain containing several RGG boxes (Xu et al. 1997). Since the recombinant hnRNP A1 was difficult to manipulate in the laboratory conditions, we chose to work with UP1 that has been shown to bind ssDNA or ssRNA (Xu et al. 1997).

We first studied the binding of hnRNP K and UP1 proteins to 7SK by EMSA (Figure VI.1). Both proteins binding to 7SK resulted in diffuse complexes. HnRNP K shifted 7SK only at high concentrations. UP1 produced complexes with a progressively lower mobility when increasing the protein concentrations. One possible explanation for this observation is that UP1 may undergo multimerization on 7SK. These results confirmed in vitro the

interaction between 7SK and hnRNP A1 and K reported before in vivo (Barrandon et al. 2007; Van Herreweghe et al. 2007).

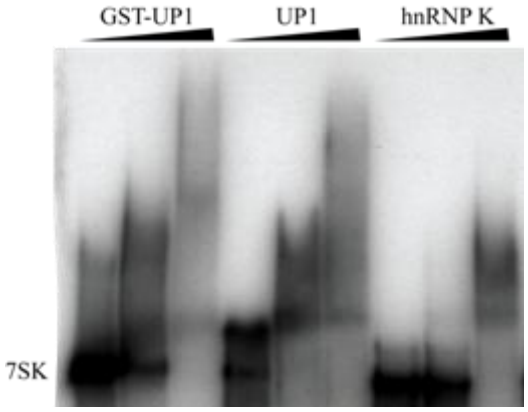



Figure VI.1. EMSA analysis of 7SK interaction to hnRNP K and UP1. Increasing concentration (0.125, 0.5 and 2 μM) of GST-UP1, UP1 or hnRNP K (as indicated) were incubated with 7SK and analysed by EMSA.

Previously, it had been proposed that HP3 of 7SK is the domain responsible for the interaction to hnRNP R, Q1, A2 and A1 (Van Herreweghe et al. 2007), but there was no information about the 7SK region binding hnRNP K. In order to confirm in vitro the interaction between UP1 and HP3 and to identify the 7SK domain determinant for hnRNP K interaction, we tested several RNAs constructions (HP1, HP3 and HP4) for interaction by EMSA (Figure VI.2). UP1 shifted HP3, but not HP1 or HP4. Surprisingly, HP3 showed two bands in native gels. This is probably due to two different conformations, since only one band was observed for HP3 in denaturing gel. Interestingly, UP1 only shifted the lower mobility HP3 band. Unlike 7SK binding, HP3/UP1 complex produced only one band. These results confirmed that HP3 is the 7SK domain for the interaction with hnRNP A1. HnRNP K failed to bind HP1, HP3 or HP4. One explanation is that hnRNP K requires more than one 7SK domain for the interaction or that it binds a different region of 7SK than those explored. However, hnRNP K was also difficult to manipulate, so we cannot rule out that our recombinant protein was not completely functional. This was further indicated by the high concentration (2 μM) that was required to shift 7SK. For this reason, only the HP3/UP1 complex was pursued for further characterization.

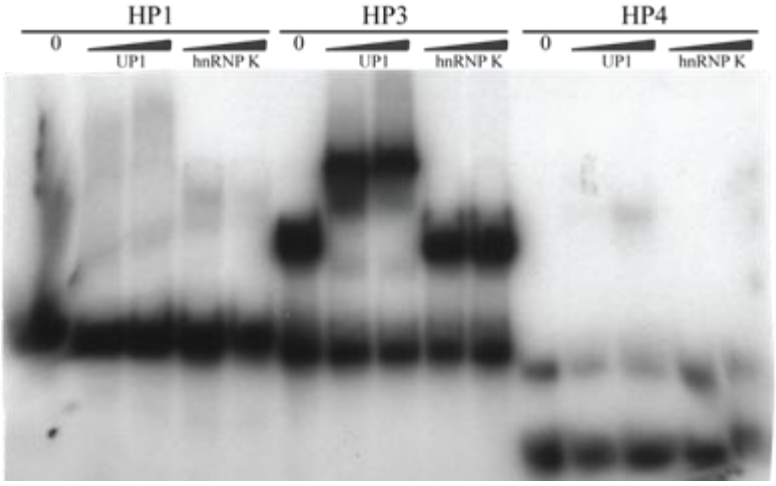



Figure VI.2. EMSA analysis of interaction between the different 7SK hairpins and hnRNP K or UP1. Increasing concentration (0.5 and 1 µM) of UP1 or hnRNP K (as indicated) were incubated with the isolated hairpins of 7SK and analysed by EMSA.

Then, we wanted to verify that the HP3/UP1 interaction is specific. Thus, non-labelled HP3 or 7SK were added to the interaction reaction and analysed by EMSA (Figure VI.3). Both, HP3 and 7SK effectively competed with labelled HP3 (compare the lines with 0.05 and 0.1 µM UP1 for each condition). Additionally, we also controlled if HEXIM1 was interacting with HP3, confirming that our EMSA conditions reflected specific interactions.

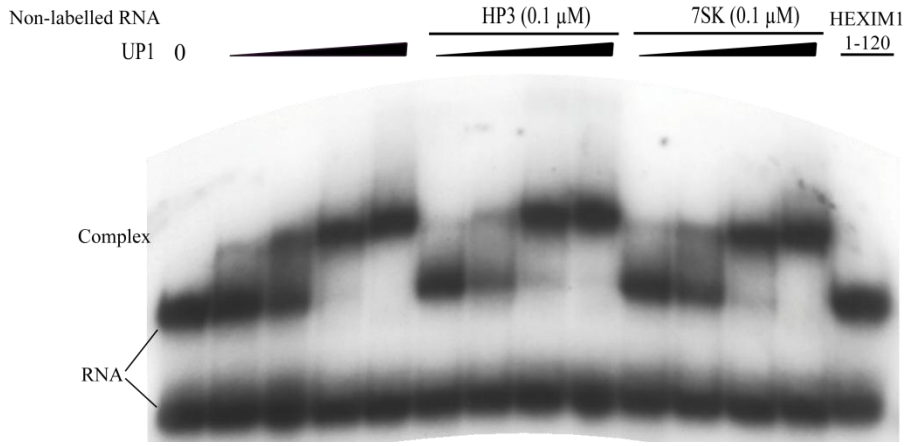



Figure VI.3. Analysis of the specificity of the HP3/UP1 interaction. Non-labelled HP3 or 7SK (as indicated) was incubated in presence of increasing concentration (0.05, 0.1, 0.25 and 0.5 µM) of UP1 and radiolabelled HP3, and analysed by EMSA. As a control, incubation with 1 µM of HEXIM1 1-120 is also shown.

Next, gel filtration chromatography analysis was performed to further characterize the HP3/UP1 complex. HP3 was purified using a gel filtration TSK-G2000 SW column to isolate the functional HP3 conformation. Purified HP3 showed a unique sharp peak in Superose 6 column with an elution volume of 16.3 ml (Figure VI.4). UP1 also presented a narrow peak profile with an elution volume of 18 ml, corresponding to an apparent MW of 18 kDa, slightly lower than the estimated MW by ProtParam, 22.2 kDa, and in agreement with a monomer protein. When a HP3/UP1 complex with a ratio of 1:1 was loaded into the column, after 30 minutes incubation at 20°C, two peaks were observed, one most probably corresponding to the free protein, and the other one most probably corresponding to the HP3/UP1 complex (elution volume of 15.2 ml). The free HP3 peak disappeared. Because free UP1 was observed, a HP3/UP1 complex at a ratio of 2:1 was then analysed. Three peaks were observed, one corresponding to the free protein, a main second peak corresponding to the first HP3/UP1 complex, and a new peak with a smaller elution volume of 14 ml. Since HP3 is in excess, this new complex probably corresponded to a UP1 binding two HP3 molecules. This is feasible since UP1 consists of two RRM. Unlike gel filtration chromatography, EMSA reactions contained a high excess of tRNA to mask non-specific interactions that may explain the differences observed. The fractions corresponding to the complex were reloaded in the column, and two peaks profile was observed, corresponding to free HP3 and UP1, respectively, showing that the complex dissociated easily.

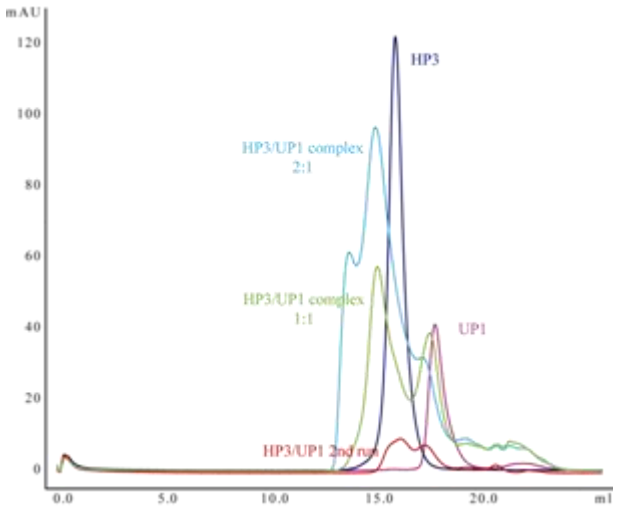



Figure VI.4. Analytical gel filtration chromatography of HP3/UP1 complex. HP1 (blue), UP1 (pink), HP3/UP1 complex (1:1, green), HP3/UP1 complex (2:1, light blue), and HP3/UP1 complex second injection (red) profiles at 280 nm absorbance are shown.

Finally, we also carried out preliminary studies of HP3/UP1 complex by SAXS and a first solution envelope was obtained (Figure VI.5). Further details about SAXS experiments are described in chapter VIII.

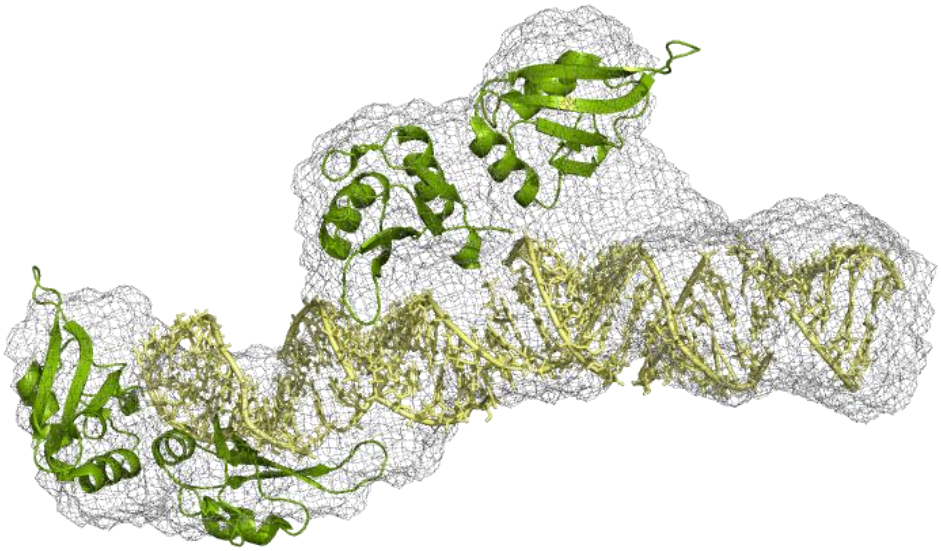


Figure VI.5. Solution envelope of HP3/UP1 complex. The crystal structure of UP1 [PDB 1L3K (Vitali et al. 2002)] and a model of HP3 calculated by MC-Sym (Parisien et al. 2008) were manually fitted in the solution envelope of HP3/UP1. A visible bulge at the middle of the hairpin could correspond to UP1 (two RRM) volume. Another large volume at the tip of HP3 might accommodate UP1 or reflect dynamics of the loop.

The enquiry about HP3 binding by UP1 was stopped there because we chose to focus on the HEXIM1/7SK interaction. However, we hope to have given some contribution for future studies in the team.

# CHAPTER VII :
# PROBING THE SECONDARY STRUCTURE OF 7SK

## 1. THE SECONDARY STRUCTURE OF THE RNAS

RNAs adopt complex three-dimensional folds for the precise presentation of chemical moieties that are essential for their functions as a biological catalyst, translator of genetic information, or structural scaffold (Batey et al. 1999).

Structural studies and sequences analysis have suggested that biological RNAs are composed of recurrent modular motifs that play specific functional roles (Holbrook, 2005; Nasalean et al. 2009). Some motifs direct the folding of the RNA or stabilize the folded structure through tertiary interactions. Others bind ligands or proteins or catalyze chemical reactions. Many of these structural motifs have been already identified and characterized (Leontis et al. 2003; Batey et al. 1999; Moore 1999; Hendrix et al. 2005; Nasalean et al. 2009).

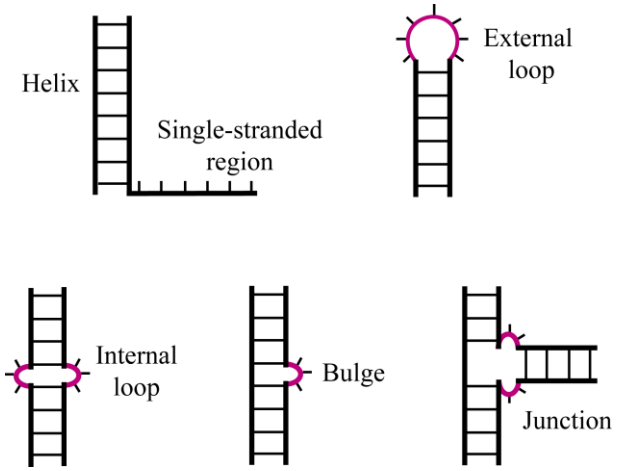



Figure VII.1. RNA secondary structure. Schematical representation of common RNA secondary structural elements [modified from (Batey et al. 1999)].

The principles that underlie the formation of such structural motifs are essentially found in the RNA sequence (most of denatured RNAs renature spontaneously in vitro). Indeed, single-stranded RNAs have a strong tendency to fold back and form Watson-Crick pairs, leading to hairpins or stem-loops of various length and complexities (Westhof et al. 2010). This is usually referred as the secondary structure of RNA. The secondary structure elements commonly described are helix, single stranded regions, bulges, external loops, internal loops and junctions (Figure VII.1). The hairpins defining the secondary structure can further assemble into intricate three-dimensional architectures.

Thus, the knowledge of RNA secondary structure is the first necessary step toward understanding the activity of the RNA (Westhof et al. 2010). There are different methods to determine the RNA secondary structure: prediction programs, comparative sequence analysis and experimental techniques.

**Table VII.1 RNA secondary structure prediction tools**

| Method | Program | Website | Webserver | Reference |
|---|---|---|---|---|
| **Free energy minimization** | RNAstructure | http://rna.urmc.rochester.edu/RNAstructure.html | No | (Reuter et al. 2010) |
| | Mfold | http://mfold.rna.albany.edu/?q=mfold/RNA-Folding-Form | Yes | (M. Zuker 2003) |
| | MCfold | http://www.major.iric.ca/MC-Fold/ | Yes | (Parisien et al. 2008) |
| | CONTRAfold | http://contra.stanford.edu/contrafold/server.html | Yes | (Do et al. 2006) |
| | RNAfold | http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi | Yes | (Gruber et al. 2008c) |
| | KineFold | http://kinefold.curie.fr/cgi-bin/form.pl | Yes | (Xayaphoummine et al. 2005) |
| **Comparative sequence analysis** | RNAalifold | http://rna.tbi.univie.ac.at/cgi-bin/RNAalifold.cgi | Yes | (Bernhart et al. 2008) |
| | LocaRNA | http://rna.informatik.uni-freiburg.de:8080/LocARNA.jsp | Yes | (Smith et al. 2010) |
| | MASTR | http://servers.binf.ku.dk/mastr/index.php | Yes | (Lindgreen et al. 2007) |
| | PETfold | http://rth.dk/resources/petfold/submit.php | Yes | (Seemann et al. 2008) |

The software tools for RNA secondary structure prediction calculate the free energy of a number of base-pairing schemes of a nucleotide sequence, and proposes the lowest energy potential secondary structure as the most probable (Reuter et al. 2010; Mathews et al. 2006). Comparative sequence analysis exploits the tendency for the global architecture of biological RNAs to be conserved (Bernhart et al. 2008; Hofacker et al. 2002). The phylogenetic covariance of two or more nucleotides that are distant in the primary sequence implies that

they interact at some level. Many tools for RNA secondary structure prediction using these methods have been published and are freely available online; some of them are presented in the Table VII.1.

Biochemical techniques allow determining experimentally the solution structure of RNA. Chemical and enzymatic probing is one of the most popular approaches for experimentally mapping the conformation of RNA molecules under defined conditions. In chemical probing, reagents modify specific functional groups either on the bases of different nucleotides or on the sugar-phosphate backbone and these modification sites can be shielded by base-pair hydrogen bonding, solvent inaccessibility or low flexibility (Weeks 2010). On the other hand, enzymatic probing exploits the existence of nucleases with a distinct preference either for unpaired or paired regions (Ehresmann et al. 1987). Hence, the reactivity of each nucleotide towards enzymes or chemicals can be used to differentiate single from double-stranded regions, and to obtain some information about the tertiary structure of the RNA. These data can then be used to constrain modelling of secondary structure of RNA.

## 2. 7SK Secondary Structure Models

Different models of the secondary structure of 7SK have been proposed from different approaches. In the next section three 7SK models will be discussed.

### 2.1. 7SK Wassarman and Steitz' s model

In 1991 Wassarman and Steitz published a model of the secondary structure of 7SK based on data from chemical and enzymatic probing (Wassarman et al. 1991). 7SK was extracted from HeLa cells in native conditions, and probing was performed in the deproteinized 7SK snRNA and in the 7SK RNP. The Wassarman and Steitz model is presented in the Figure VII.2.

In this model 7SK comprises four stem-loop or hairpins structures, separated by single stranded regions of different lengths. The 5' end hairpin and the hairpin 3 have big apical loops, as well as several internal loops. The 3' end hairpin (HP4) is the smallest one and has an apical pentaloop. The domain 2 consists of a three-way junction element. According to

Wassarman and Steitz, the major differences between the isolated 7SK and the 7SK snRNP occur in the region including G115 to G196 which encompasses the domain 2 and the region between domain 2 and hairpin 3. This region showed to be more accessible for chemical modification and enzymatic cleavage when the RNA is deproteinized.

Nowadays, this is the most generally accepted model for 7SK. However, the data is not fully accounted for the model. The inconsistencies are especially noteworthy in the hairpin 3 where the data does not fit with the model and in the single-strand linking this hairpin to the domain 2 for which data suggest a more structured region (see Figure VII.12). Besides, it represents 7SK as an open, extended RNA. Our SAXS data, however, suggested a more compact structure, not compatible with a RNA where the different hairpin structures are tethered in a long single strand (see Chapter VIII).

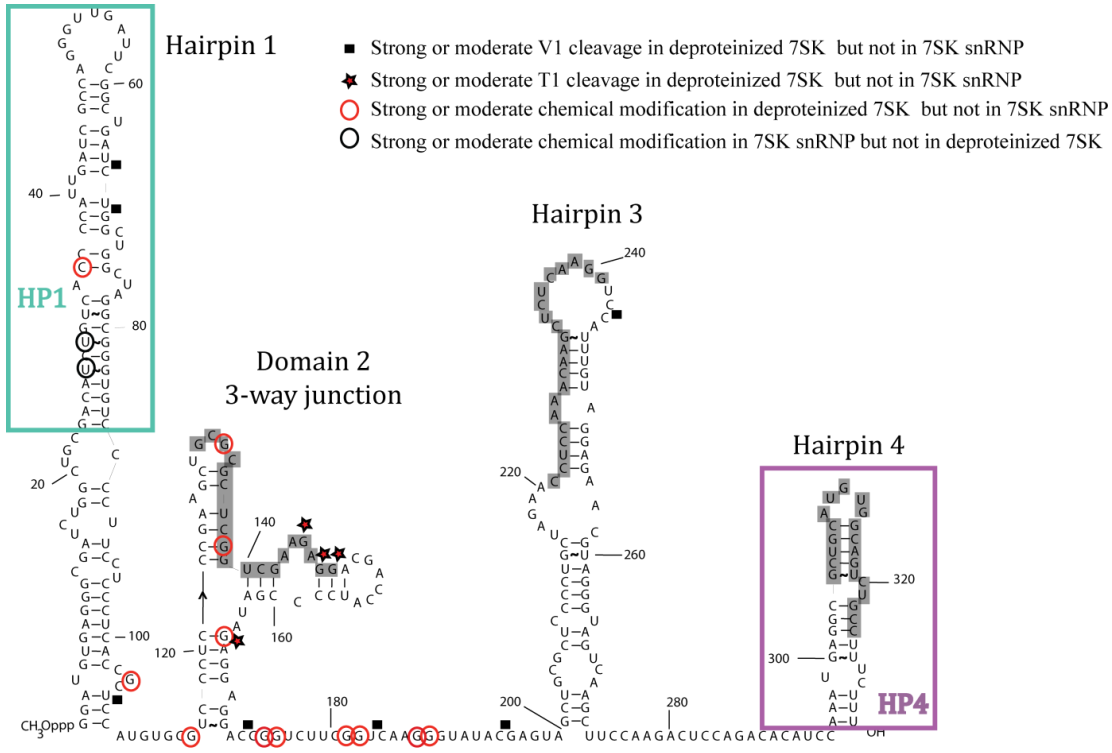



Figure VII.2. Wassarman and Steitz 7SK model. Model constructed from experimental data (with chemical probes DMS, kethoxal, and CMCT, and enzymatic probes RNAse T1, V1, H, and micrococcal nuclease) with 7SK extracted from HeLa cells (adapted from Wassarman and Steitz 1991). The hybridization sequences to the primers used for primer extension analysis are shadowed. HP1 and HP4 are also indicated.

## 2.2. 7SK Marz' s model

In 2009 Marz and collaborators proposed another secondary structure model of 7SK (Marz et al. 2009). Through a bioinformatics analysis approach, they identified 7SK in lower eukaryotes. They extended the collection of 7SK RNAs using a specialized automaton. Several features of previously known 7SKs were used to identify new 7SK sequences, in particular, a Pol III promoter sequence, a highly conserved GATC pattern (repeated twice), and a poly-T stretch of five thymidines within seven nucleotides as a termination signal. Thus, 7SK sequences were found across all animal phyla, with the exception of Platyhelminthes (flatworms). Marz and col. performed a structural alignment, analysed the patterns of base co-variation and constructed a consensus secondary structure model (Figure VII.3).

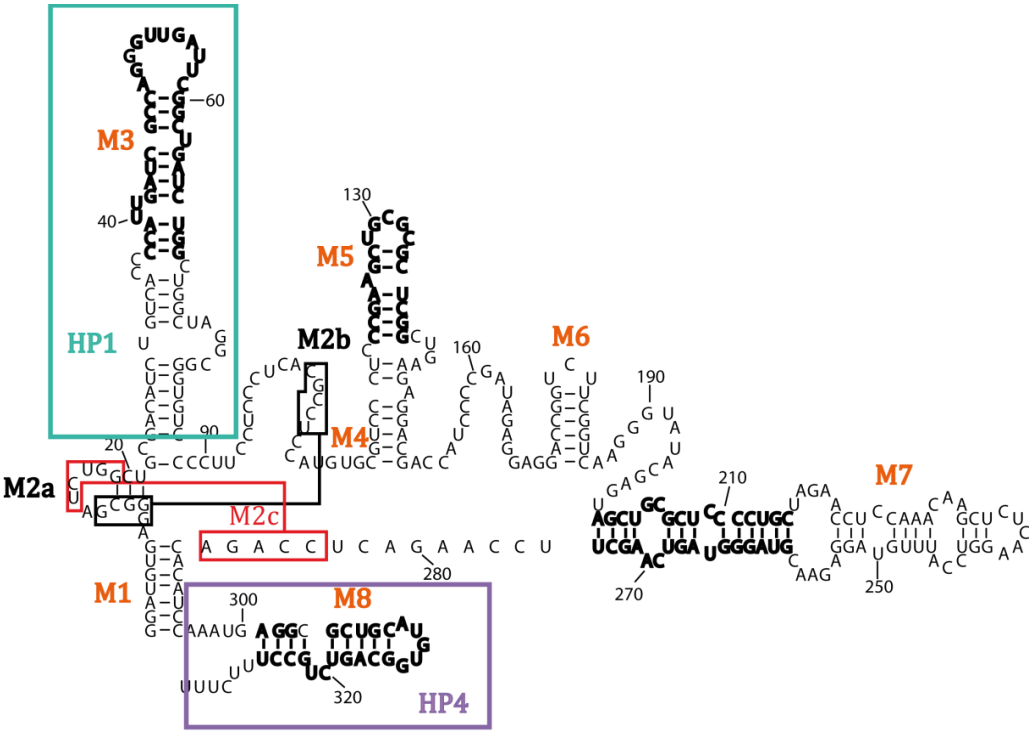



Figure VII.3. Marz's 7SK model. Model obtained from comparative sequence analysis (Marz et al. 2009). Constant elements in both Wassarmand and Steitz's and Marz's models are highlighted in bold. HP1 and HP4 are indicated.

One of the striking differences with the previous model is that in Marz's model the 5' and the 3' end of 7SK are gathered together by the M1 stem. This results in a circular, more

compact molecule. Most of the domain 2 region also shows a different secondary structure, including an additional stem-loop structure, M6. An interesting feature of this model is region M2, which is proposed to form three distinct structural alternatives (M2a, M2b or M2c as shown in Figure VII.3) suggesting the refolding of M2 as part of the core functionality of 7SK. The apical regions of the 5' end hairpin, HP4, as well as the small stem-loop structure encompassing C122-G139 (here called M5), are the same than in the Wassarman and Steitz model.

## 2.3. 7SK Eilenbrecht's model

Several models of secondary structure of 7SK based upon energy minimization programs have been presented in different research papers. One of them introduced in (Luo et al. 1997) and revisited in (Eilebrecht et al, 2010) presents 7SK as a circular molecule closed similarly by the M1 stem like in Marz's model, but in an extended version which includes the 3'end U-rich "tail" (L4; Figure VII.4).

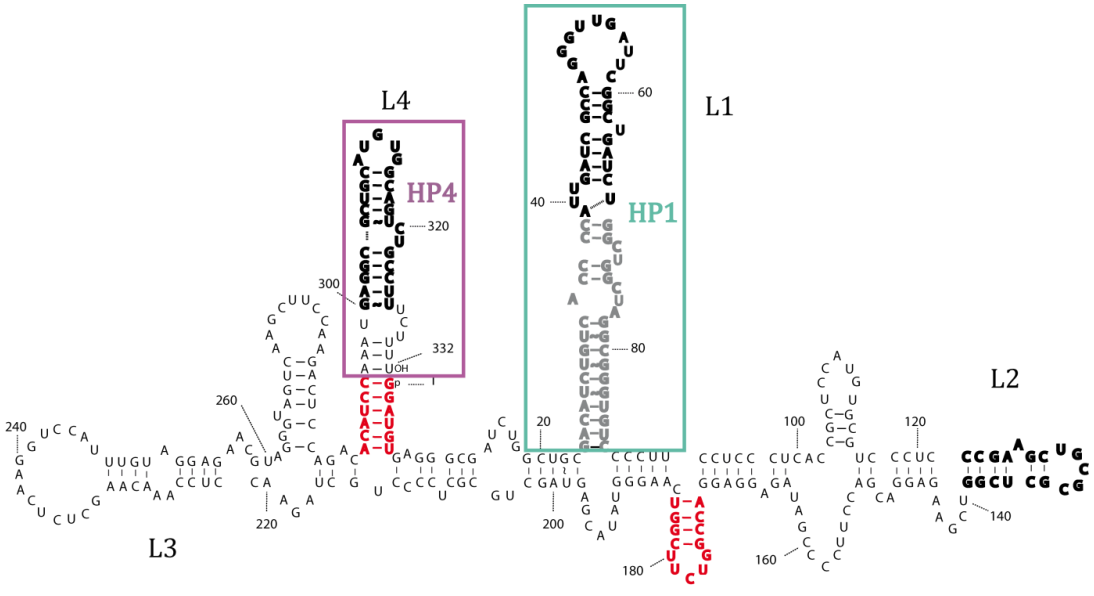



Figure VII.4. Eilebrecht's 7SK model. Model obtained from free energy minimization tools (DNASIS Pro Software according to Luo et al., 1997), but calculation details were poorly documented. The elements that coincide with both (black), only with Wassarman and Steitz (gray), or only with Marz's (red) are indicated [modified from (Eilebrecht et al, 2010)].

This model shows a 7SK subdivided into four main stem-loop structures with various junctions. The small hairpin called M6 (nucleotides 172 to 184) in Marz model is present. In contrast, the region A200-U275 presented as long stem-loops in the previous models (named HP3 or M7) shows a very different arrangement here, with two smaller hairpins connected by a junction element.

Interestingly, some structural elements are recurrent. These are HP1, HP4, and the small hairpin in the apical region of L2 (called M5 in Marz model), showed in black in the Figure VII.4.

## 3. The Secondary Structure of Synthesized Full Length 7SK

The goal of my project was the structural characterisation and analysis of the recognition mechanism between 7SK and its protein partners, in particular the HEXIM1 protein. Structural approaches like X-ray crystallography, SAXS, RMN, Cryo-Electron Microscopy, etc., require large quantities of material highly purified and homogeneous. T7 in vitro transcription is a generally used method to obtain sufficient amount of RNA for structural studies. Hence, 7SK was synthesized by T7 in vitro transcription and the conditions for its folding in a unique conformation were determined experimentally. To verify if this conformation corresponded to the conformation of 7SK extracted from HeLa cells, chemical and enzymatic probing experiments were performed.

Another objective for probing the secondary structure of 7SK is to create useful models for developing hypotheses regarding RNA function. Hence, these experiments were intended to allow us to get information about hinges in the 7SK structure, and to design functional subdomains for protein interaction. These functional subdomains constructions were important for several techniques, including crystallization.

In a first period of my pHD work, the RNAse T1 and the RNAse V1 were used with the intention of performing a quick analysis of the secondary structure of 7SK.

## 3.1. Enzymatic probes

### a. Ribonuclease T1

RNA T1 is a fungal endonuclease that specifically cleaves internucleotides bond adjacent to the 3'-phosphate of unpaired guanosine residues in RNA, with the intermediary formation of guanosine 2'-3'-cyclic phosphate. The hydrolysis generates fragments with a 3'-phosphate (Ehresmann et al. 1987).

### b. Ribonuclease V1

RNAse V1, from cobra venom, is a non-sequence specific endonuclease that preferentially cuts double-stranded or structured regions, generating fragments with a 5'-phosphate. The minimum size of the RNA substrate is 4 to 6 nucleotides. It also cleaves single-stranded region in stacked conformation (Ehresmann et al. 1987).

Probes were used under conditions were less than one cleavage occurred per RNA molecule with a statistical distribution. However, since the cleavages may introduce conformational rearrangements in RNA that potentially provides new targets (secondary cuts) to RNAse, interpretation was sometimes difficult. In particular, a very high T1 reactivity in the position G55 of the HP1 apical loop was observed. For this reason a recently developed method called Selective 2′-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) was used in a second period of my work. Also, we expected from SHAPE a more generic method, independent of the sequence.

## 3.2. Selective 2′-Hydroxyl Acylation analyzed by Primer Extension (SHAPE)

SHAPE chemistry uses a hydroxyl-selective electrophile, the 1-methyl-7-nitroisatoic anhydride (1M7), to map the local flexibility of each nucleotide within the RNA (Mortimer et al. 2008). When a nucleotide is unconstrained (flexible), its ribose 2'-hydroxyl position

preferentially adopts conformations that react with 1M7 to form 2'-O-ester adducts. Conversely, base paired or otherwise conformationally constrained nucleotides (by tertiary interactions, for instance) are unreactive (Merino, Wilkinson, Coughlan, & Weeks, 2005; Figure VII.5).

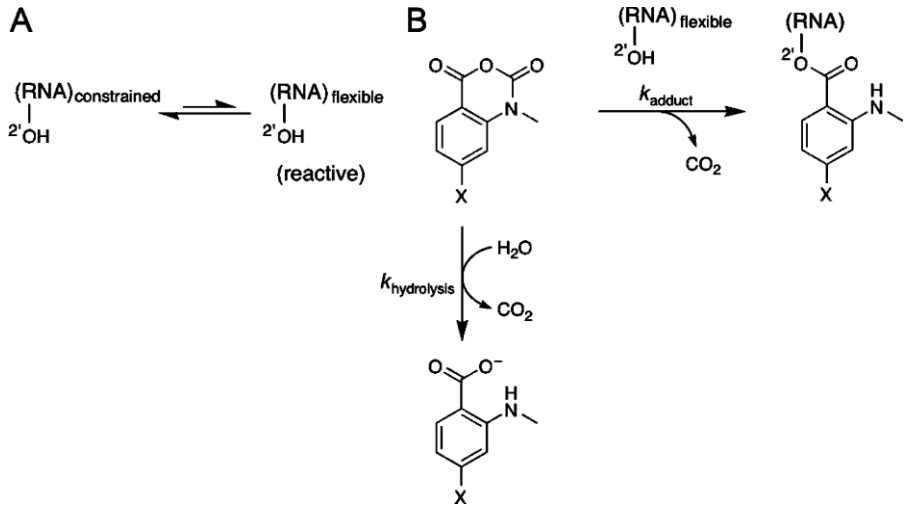



Figure VII.5. Mechanism of RNA SHAPE chemistry. A) The nucleophilic reactivity of 2'hydroxyl group is selectively enhanced at flexible positions. B) Parallel reaction of M-methylsatoic anhydre derivatives with RNA 2'-hydroxyl groups and with water (modified from Mortimer & Weeks, 2007).

Since every nucleotide has a 2'-hydroxyl, 1M7 reacts generically with all four nucleotides. Furthermore, as an electrophile capable of reacting with a hydroxyl group in aqueous solution will face competition from analogous hydrolysis reaction, 1M7 undergoes a parallel, self-inactivating, hydrolysis reaction with a half-life of 14 seconds (Mortimer et al. 2007; Merino et al. 2005). This short reactivity time implies that time-resolved analysis can be performed to map conformational states of the RNA.

### 3.3. Analyses of the modified sequence

The identification of the cleavages or modification can be done by two different methods depending on the nature of the modification and the length of the RNA molecule. The first

approach, which uses end-labeled RNA, only detects scissions and is limited to RNA containing several tens of nucleotides. After enzymatic digestion, the generated RNA fragments are sized by electrophoresis in a denaturing gel and bands are visualized by autoradiography. The length of the labeled fragment indicates the distance between the 5'labeled-end and the cleaved position. Determination of the position of the cleavages is facilitated by an alkaline hydrolysis ladder and a sequencing reaction electrophoresed in parallel. The second approach uses primer extension. In primer extension, a specific DNA labeled primer is extended with a Reverse Transcriptase, which can synthesize cDNA from a RNA template. When the primer anneals downstream of a modified or cleaved nucleotide, the reverse transcription is stopped at such position. The resulting labeled cDNA chains are sized by denaturing electrophoresis. To identify this position, a parallel dideoxyribonucleotide sequencing reactions are carried out on unmodified RNA using the same primer. Bands are visualized by autoradiography. Because the primer can be designed to hybridize any sequence within RNA, this method can be adapted to RNAs of any length (Ehresmann et al. 1987).

Hence, enzymatic cleavage can be detected by both methods, while SHAPE, as its name implies, uses primer extension as the presence of a 2'-O-adduct causes the reverse transcriptase to stop exactly one nucleotide prior to the modified base.

## 3.4 Unravelling the secondary structure of synthesized full length 7SK

### a. Probing experiments: global results

The goal of this work was to develop a model for the global architecture of 7SK. In order to assess buffer effects, probing experiments were performed in three different contexts: without monovalent salt and 2mM $MgCl_2$ (the standard buffer condition used in the laboratory to store the RNAs); in presence of 200mM KCl and 10mM $MgCl_2$ (the buffer currently used for HEXIM1 and binding assays); and in the presence of 100mM NaCl and 6mM $MgCl_2$ (the buffer advised in the standard SHAPE protocol (Wilkinson et al. 2006), and used for SAXS measurements). The products of subsequent reverse transcriptions were analysed using sequencing gels.

To cover the whole 7SK sequence, the use of multiple primers complementary to various regions of 7SK was needed (see Figure A.9 in Annexes 1). To explore the 3' end of 7SK a construction including an extension at the 3' end to allow the hybridization of a primer (called RT) was produced. This 3' extension was designed to acquire a stable independent structure, and should not interfere with the 7SK structure. This construction was only used to probe the 3' end using the RT primer. The rest of the experiments were carried out with the unmodified 7SK.

Examples of the gels after reverse transcription from each primer used are presented in Figures VII.6 to VII.11. Many flexible segments are visible as regions of increased IM7 modifications in the (+) lanes as compared to the (-) control lanes. The regions that clearly present high 1M7 reactivity correspond to 49-59, 129-134, 139-145, 161-166, 176-181, 191-199, 225-250, and 310-314 sequences. In contrast, some regions present low 1M7 reactivity suggesting regions of secondary structure like 80-100, 205-215, 259-268, and 300-309 sequences.

These results, summarized in the Figure VII.12, are globally consistent with the results obtained by (Wassarman et al. 1991) suggesting that the overall secondary structure of in vitro synthesized 7SK and the secondary structure of the HeLa cells extracted 7SK are similar. We observed also that the 3' extension to probe the 3' end of 7SK did not interfere with the 7SK structure.

The data fit especially well with the 5' end hairpin, HP4, and with the small hairpin in the apical region of domain 2. The existence of the M6 hairpin in Marz model is supported, as well as the hairpin structure encompassing C103-G115 in the Eilenbrecht model. However, none of the models reflects completely the data. The hairpin 3 was the region that presented more striking differences with the Wasserman and Steitz results. Also, some constrained nucleotides are unexplained by this model, notably those found in the single stranded regions linking the hairpins. To discriminate between the different models, further analysis was required.

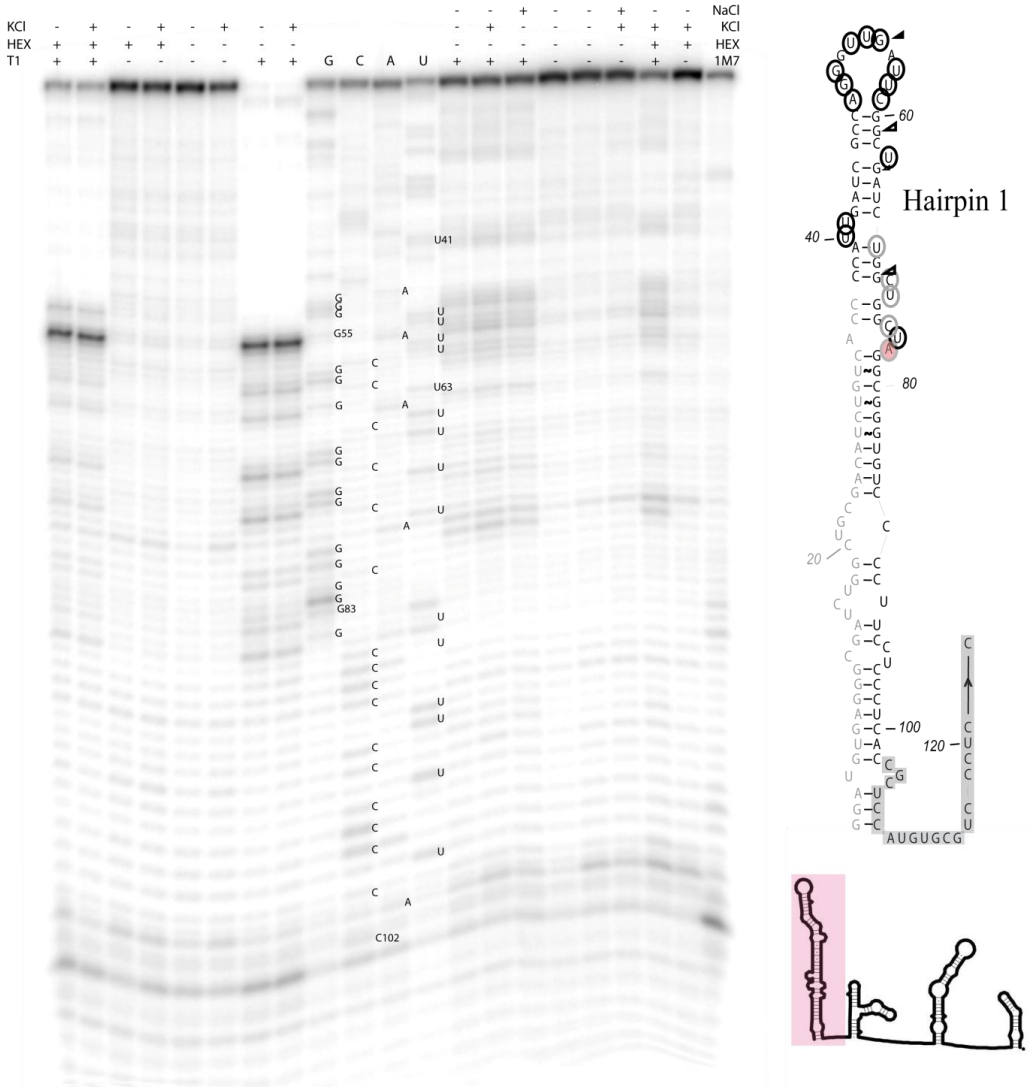



Figure VII.6. Analysis of primer C extension. Left, a probing experiment performed under different buffer conditions as indicated and in presence (+) or absence (-) of enzymatic (T1) or chemical probes (1M7) visualized by electrophoresis. Sequencing reactions were loaded in parallel and sequence is indicated. Right, summary of the results in the region of 7SK analyzed, shadowed in pink in a global 7SK model. Primer C hybridation sequence is highlighted, high IM7 reactive nucleotides are indicated with a black open circle, moderated 1M7 reactive nucleotides are indicated with a gray open circle, and T1 cleavage positions are indicated with a triangle. Nucleotides corresponding to either local degradation either pauses of the reverse transcriptase are indicated with a light red closed circle. Not analysed nucleotides are in gray. The HEXIM1 footprinting is discussed in Chapter IV.

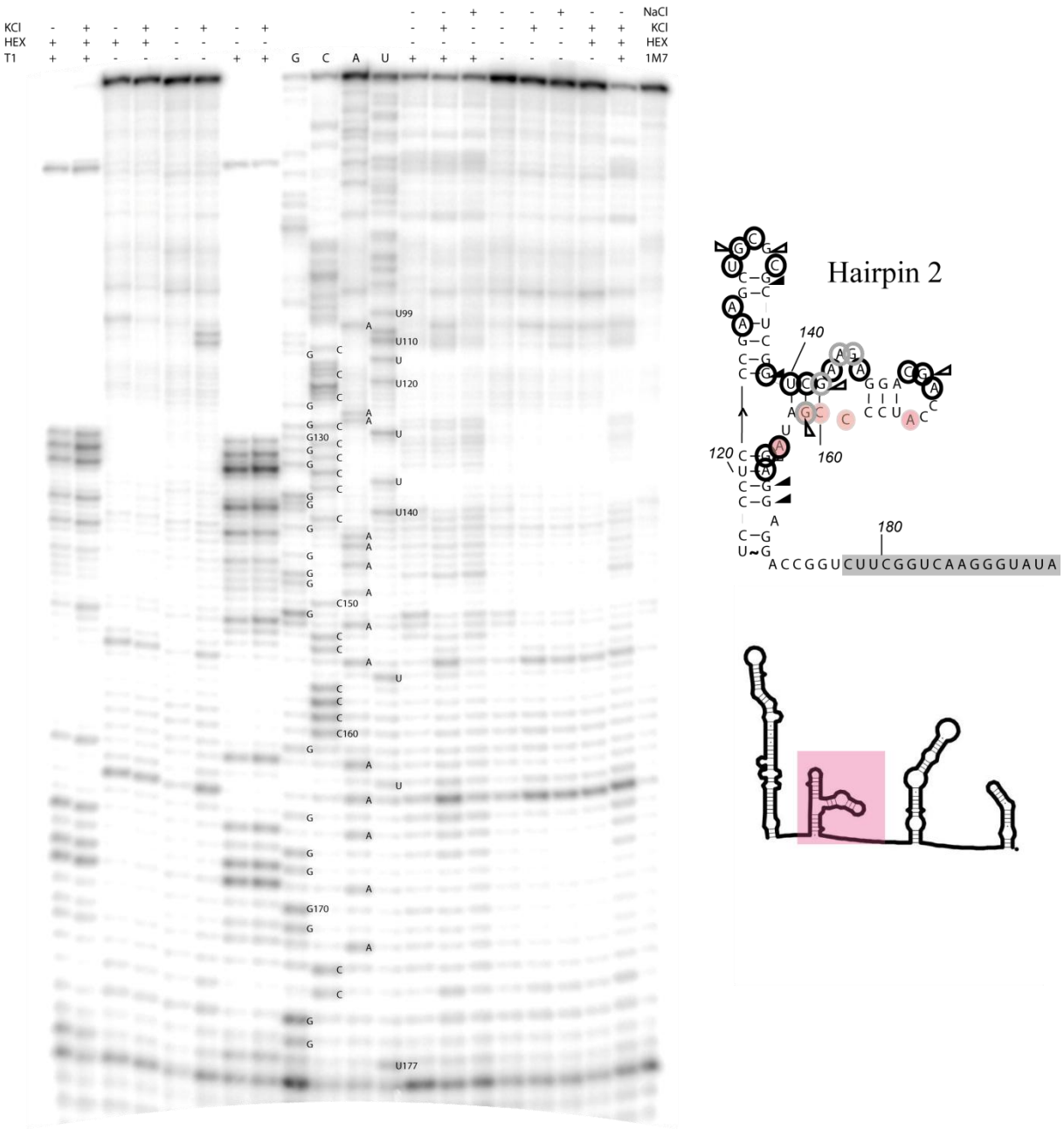



Figure VII.7. Analysis of primer B extension. Label code as in Figure VII.6.

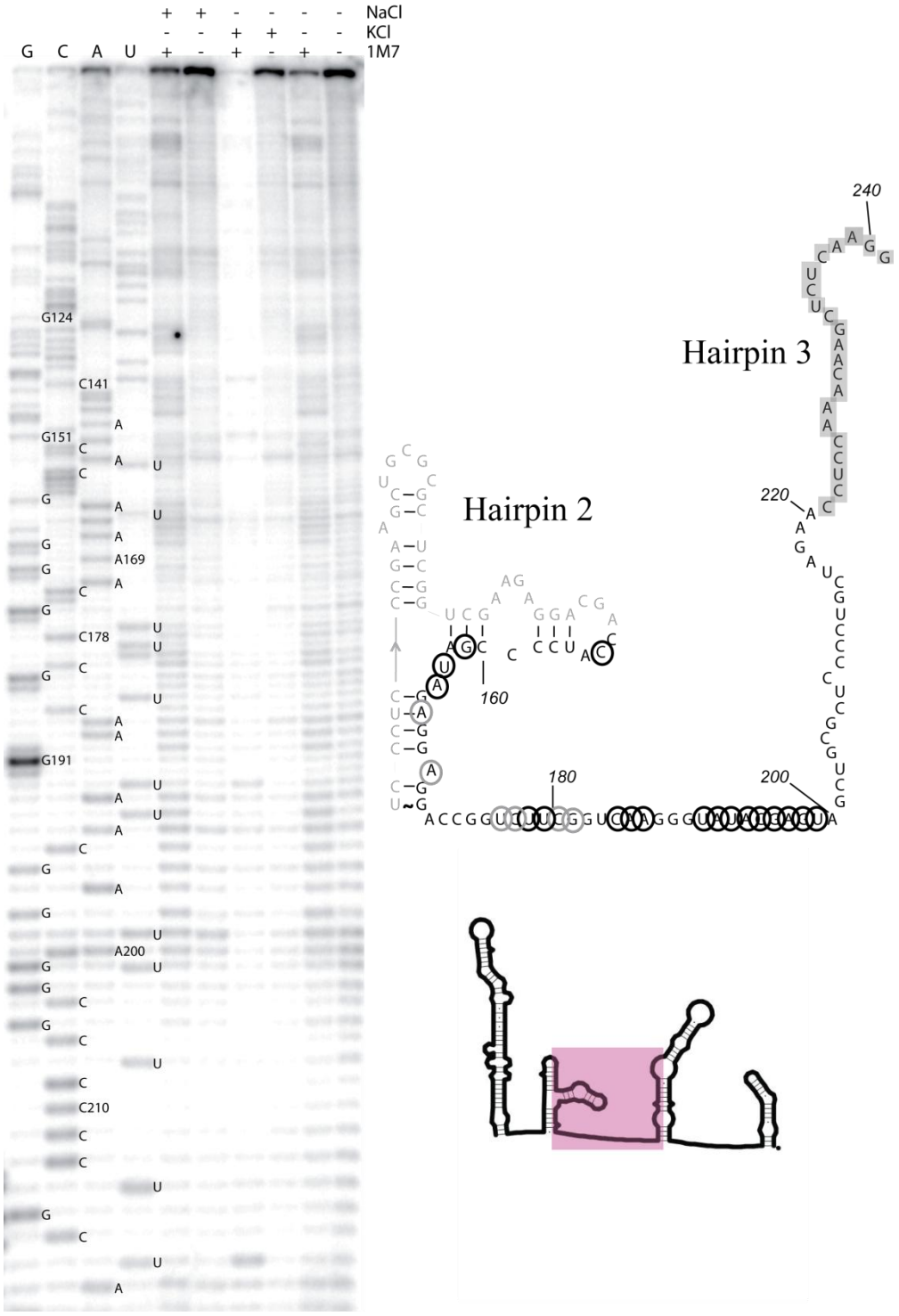



Figure VII.8. Analysis of primer G extension. Label code as in Figure VII.6.

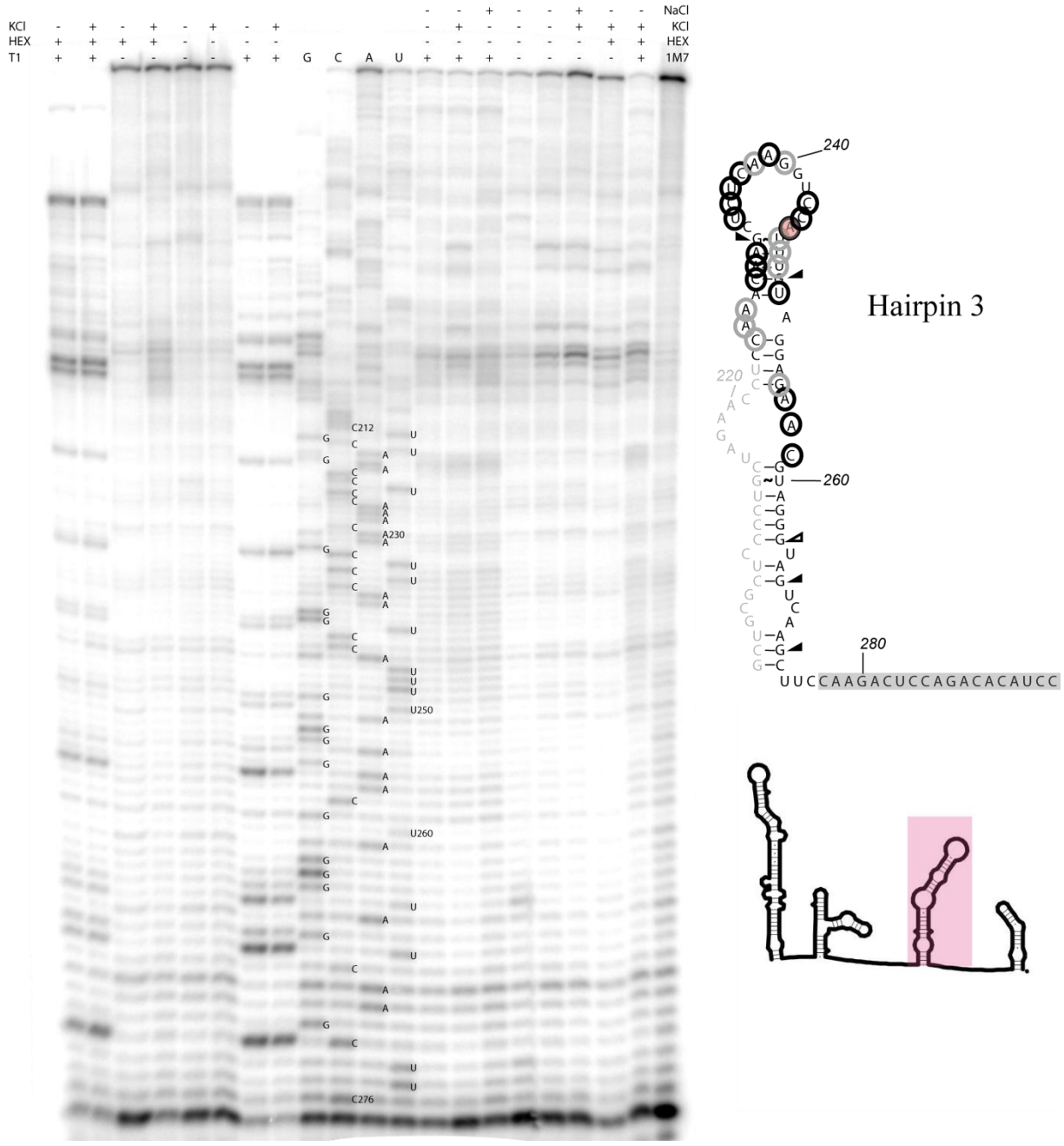



Figure VII.9. Analysis of primer E extension. Label code as in Figure VII.6.

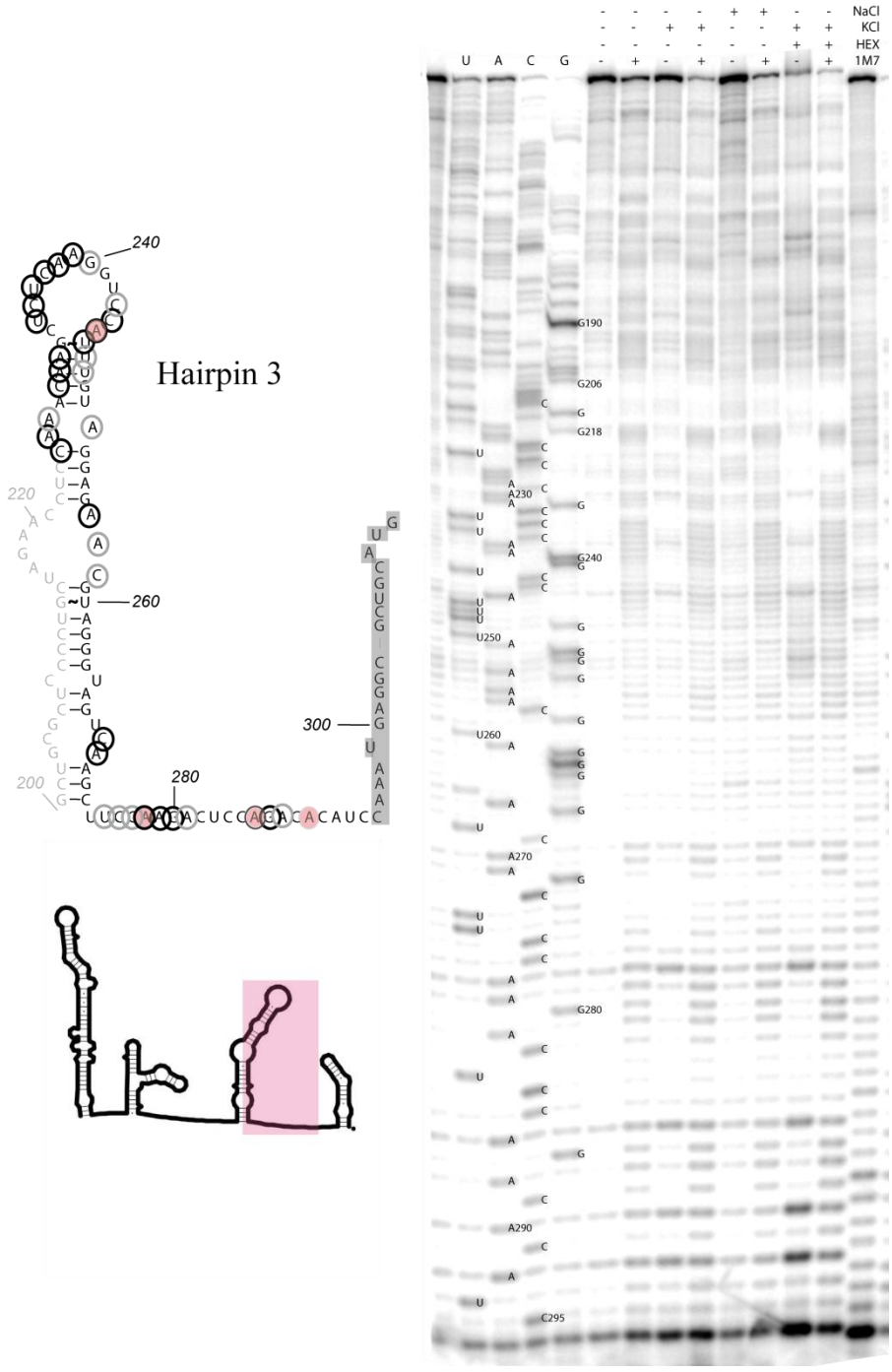



Figure VII.10. Analysis of primer D extension. Label code as in Figure VII.6.

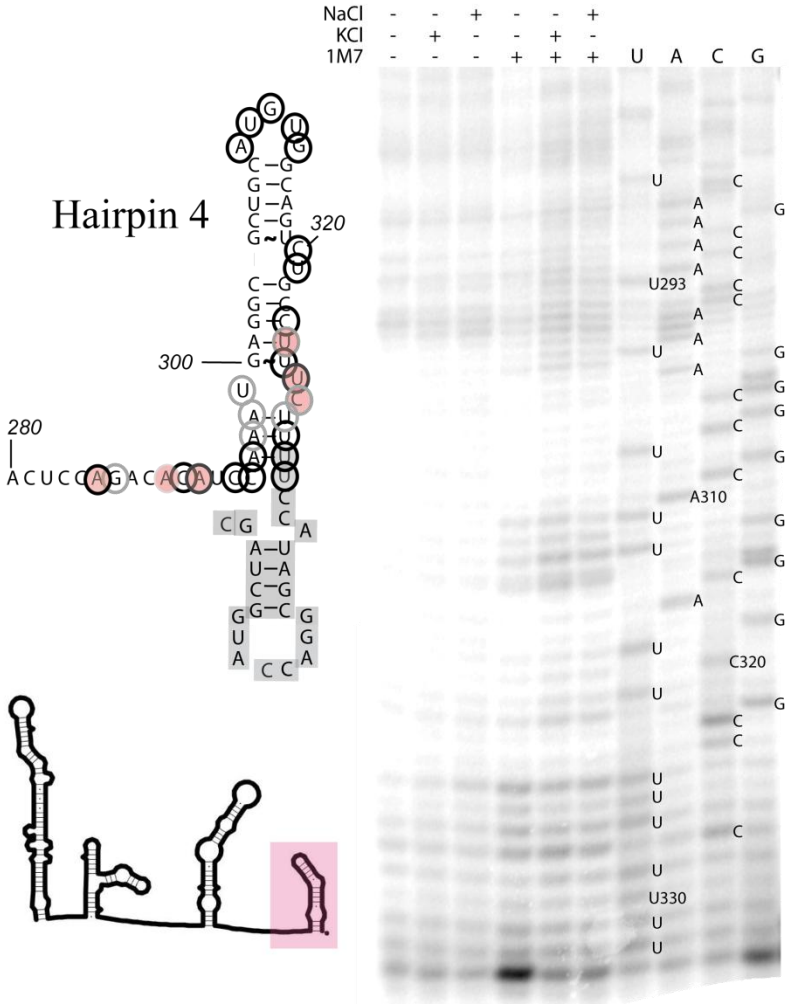



Figure VII.11. Analysis of primer RT extension. Label code as in Figure VII.6.

b. Further probing analysis and quantification

SHAPE information can be used to create highly accurate models for an RNA secondary structure (Deigan et al. 2009; Wilkinson et al. 2008). Currently, about 73% of known base pairs for a RNA are predicted by free energy minimization for sequences with fewer than 700 nucleotides (Mathews et al. 2004). The accuracy can be significantly improved by coupling with chemical modifications constraints when secondary structure is poorly predicted by free energy minimization alone. For example, only 26.3% of base pairs are correctly predicted for the *E. coli* 5S rRNA without constraints which improves to 86.8% when constraints are taken into account (Mathews et al. 2004).

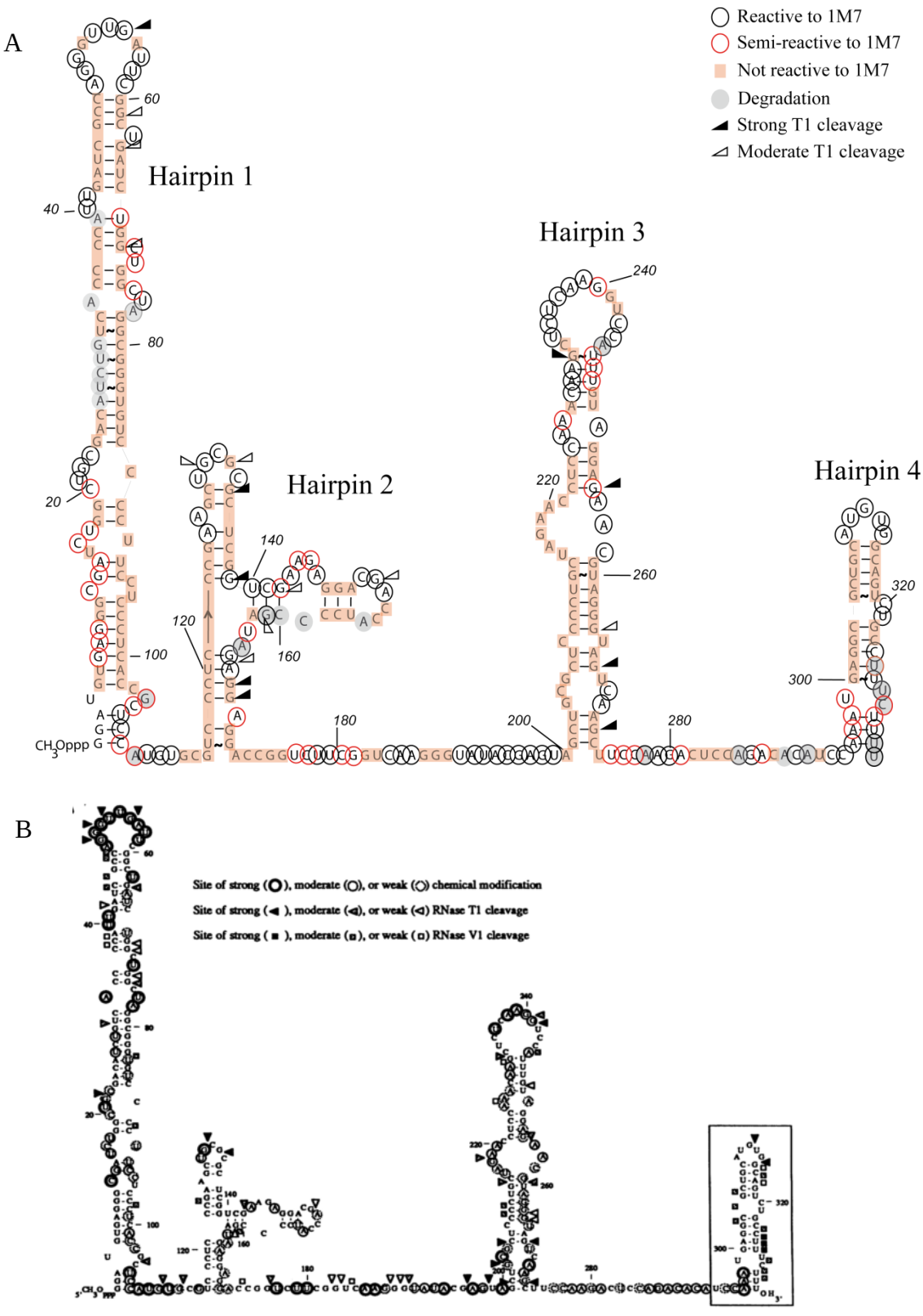



Figure VII.12. Comparison with Wasserman and Steitz results. A) The different probing results are shown in Wasserman and Steitz model. B) Wasserman and Steitz results (Wassarman et al. 1991).

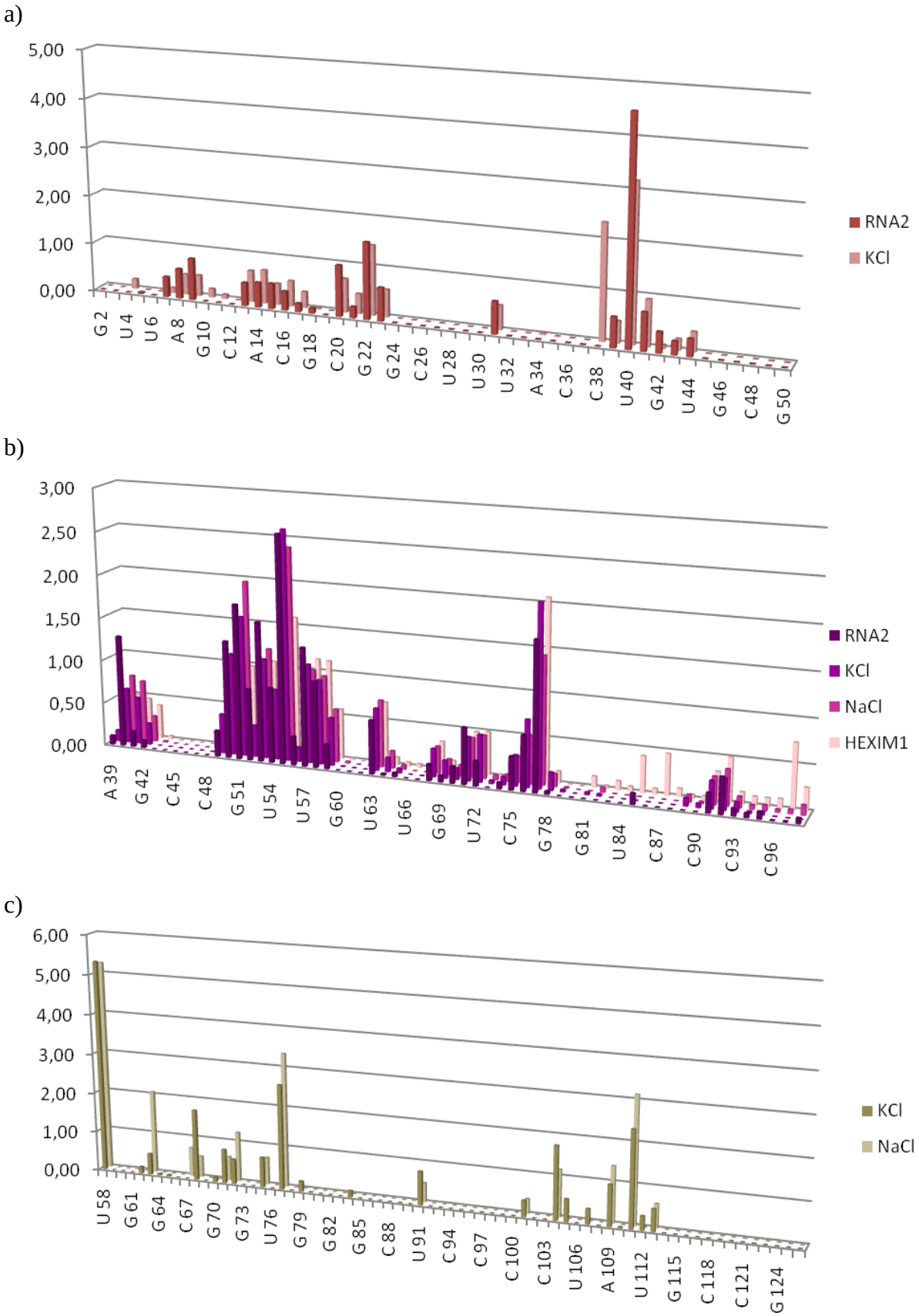
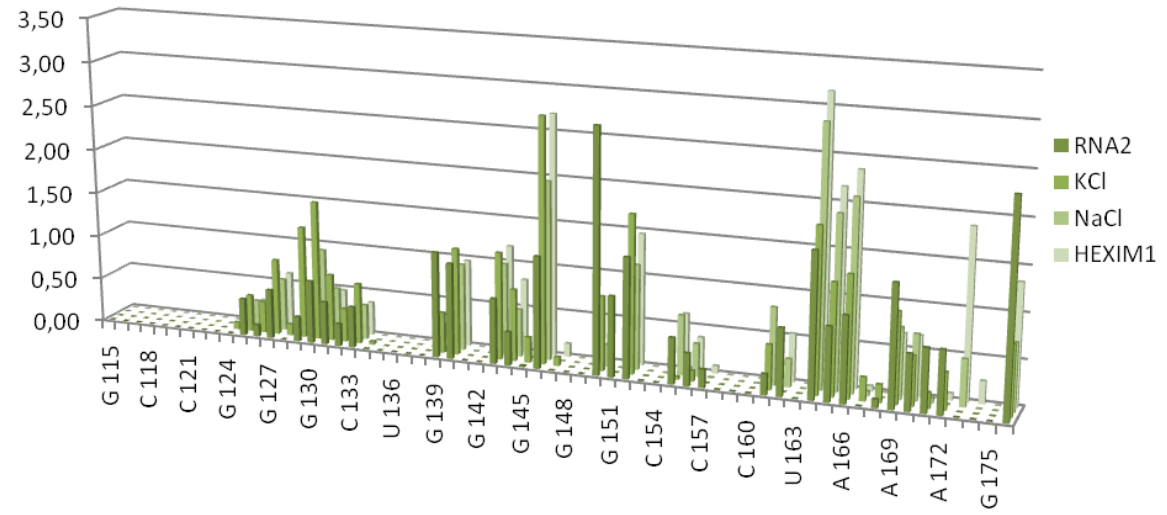
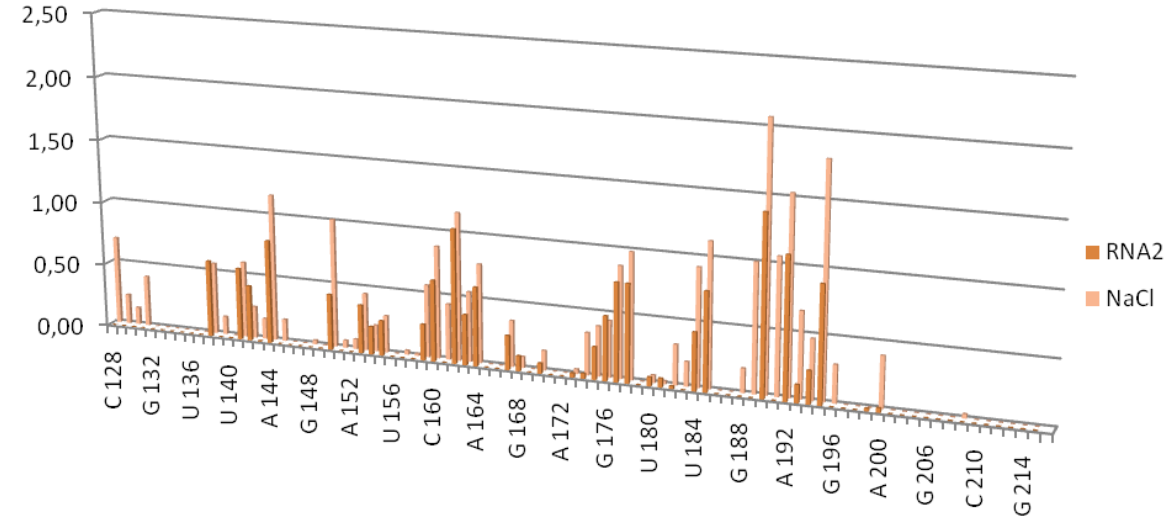



Figure VII.12. IM7 reactivities (continued).

d)




e)




f)

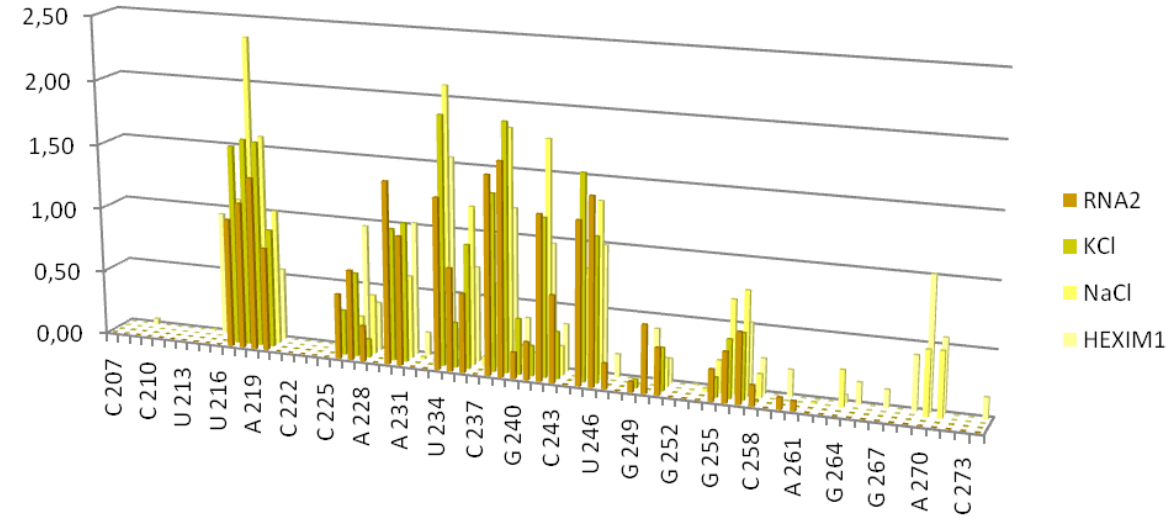
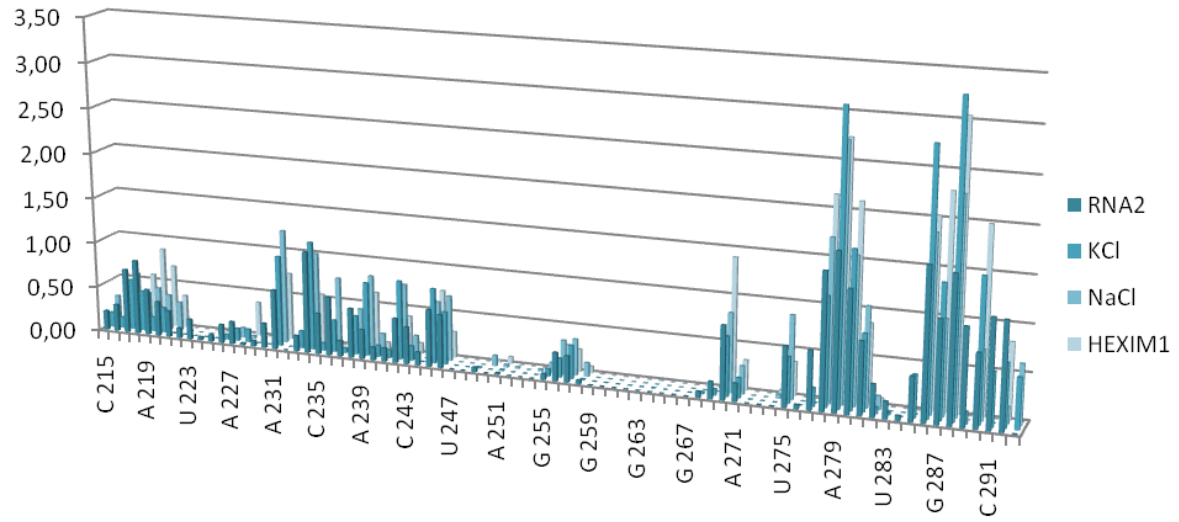



Figure VII.12. IM7 reactivities (continued).

g)




h)

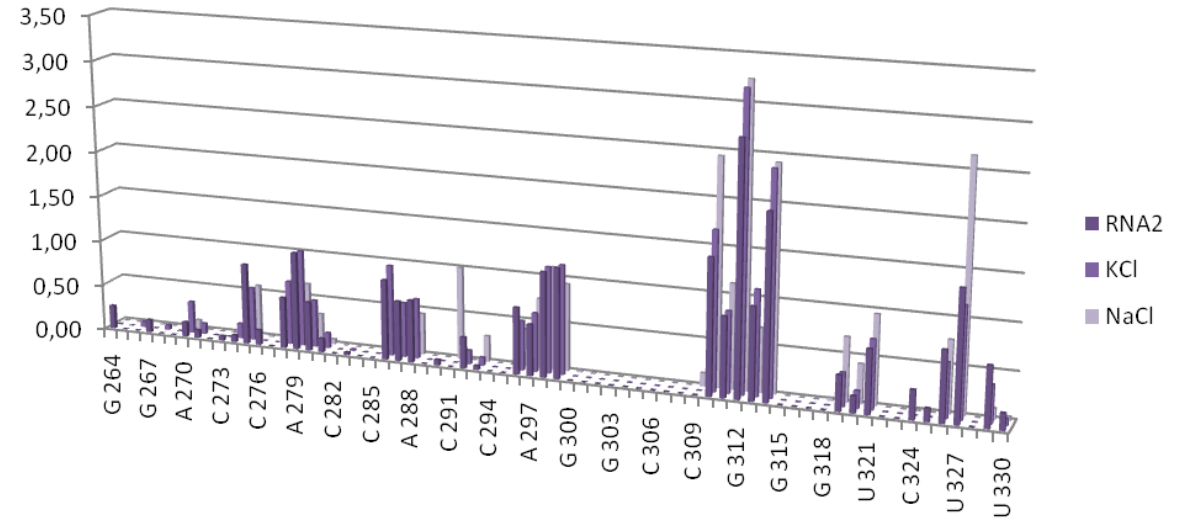



Figure VII.13. IM7 reactivities. The normalized bands intensities under different conditions obtained from primer H (a), primer C (b), primer I (c), primer B (d), primer G (e), primer E (f), primer D (g), or primer RT (h) extension are shown. The gels were rendered using SAFA (Laederach et al. 2008) and data normalized using noRNAlize (Vicens et al. 2007). Subtraction of background and graphs were performed using Microsoft Office Excel. 7SK sequence is indicated.

In this respect, SHAPE methodology has demonstrated to give trustworthy results. For instance, *E. coli* 16S rRNA is predicted with 97% of accuracy when quantitative, nucleotide-resolution information from SHAPE experiments is included (Deigan et al. 2009).

Thus, since SHAPE reactivity patterns typically yield quantitative information for nearly every nucleotide, SHAPE profile of a RNA can be helpful in ruling out incorrect or incomplete models.

Hence, to further exploit the SHAPE data, Semi-Automated Footprinting Analysis (SAFA) software was used to extract the information contained in the gels and yield the corresponding band intensities (Laederach et al. 2008). Briefly, from digitized gel image input, SAFA corrects geometric distortion of lanes and bands, and fits a peak model to accurately quantify the individual bands intensities.

The raw band intensities generated by SAFA were normalized to correct variations in the amount of sample loaded on the gel or in reverse transcription reactions, and to allow averaging several experiments. A recently developed noRNAlize program was used for this purpose (Vicens et al. 2007). Each gel was processed independently. The normalization allows the inclusion of control lanes (-1M7). Only after normalization -1M7 values were subtracted from +1M7 ones.

The Figure VII.13 shows the bands intensities corresponding to the SHAPE reactivities obtained for each primer after integration, normalization, and control subtraction.

At this level, buffer effect at each single nucleotide could be examined. The 1M7 reactivities in the three different buffer conditions were globally similar with only few nucleotides showing differences in flexibility. Most of these nucleotides presented higher 1M7 reactivity in presence of monvalent salts. For example, the positions 164 to 166 showed more flexibility in the presence of monovalent salts, as well as the positions 189 to 195, 279 to 280 and 286 to 288.

However, the SHAPE profile of 7SK was globally the same under the three conditions, suggesting that there is not considerable change of 7SK into a different conformation. SHAPE experiments in presence of HEXIM protein are discussed in the Chapter IV.

In order to merge the reactivities resulting from different experiments and from different primers, all intensities were normalized to the unity. Then all the intensities at each position were averaged (Figure VII.14).

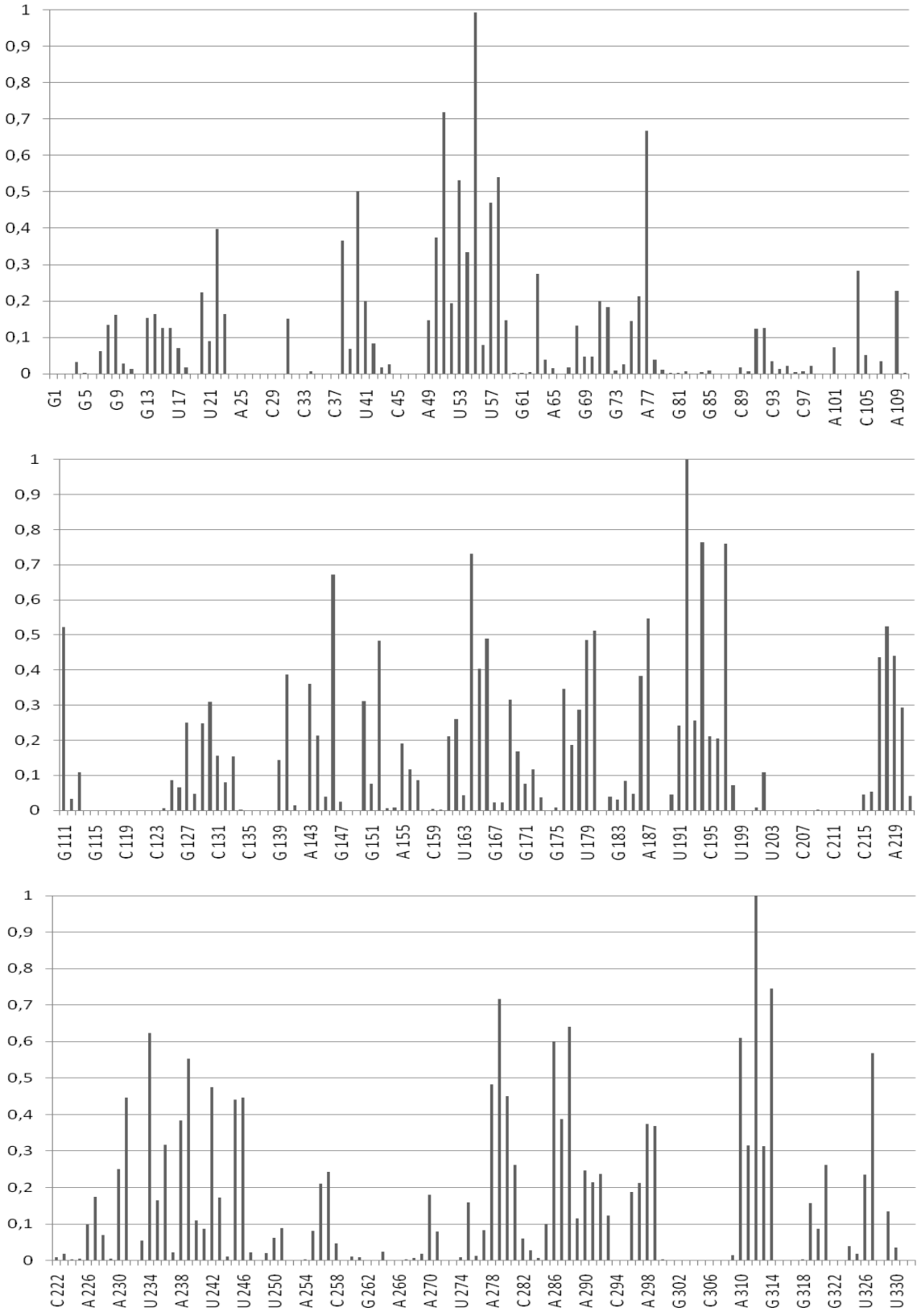



Figure VII.14. Averaged SHAPE reactivities. Normalized to unity and averaged IM7 reactivities as function of the 7SK sequence.

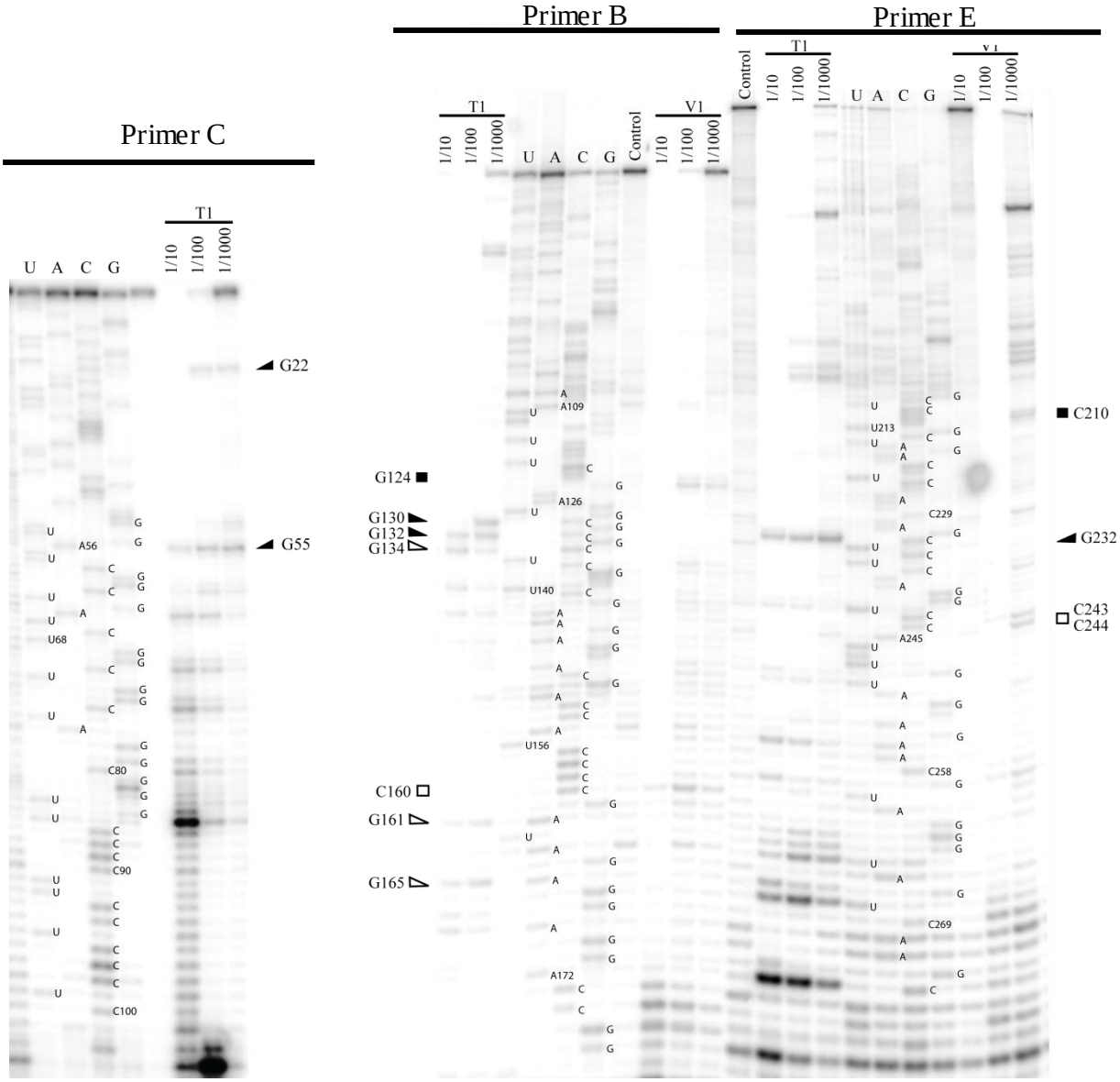



Figure VII.15. Enzymatic probing. The reverse transcription products from primer extension on the RNAse treated 7SK and fractionated in polyacrylamide gels are shown. Sequencing reactions were migrated in parallel to map the cleaved positions, and control lanes (non-treated 7SK) are shown to locate local degradation. RNAses dilutions are indicated. Only the bands corresponding in the lanes of the most diluted RNAses were considered as cleavage sites. The RNAse T1 and V1 cleavages sites are indicated: black triangles correspond to strong RNAse T1 cleavage, open triangles to moderated RNAse T1 cleavages, black squares to strong RNAse V1 cleavages, and open squares to moderated RNAse V1 cleavages.

Next, the secondary structure of 7SK was predicted using the program RNAstructure (Reuter et al. 2010). The algorithms implemented in RNAstructure use nearest neighbor parameters to predict the stability of secondary structure. The lowest free energy structure is found by using empirical thermodynamic parameters fitted against a large database of model structures with known stability (Turner et al. 2010). RNAstructure can also constrain the calculation with enzymatic, chemical mapping, SHAPE, or NMR data. This implementation makes RNAstructure a suitable tool for our prediction.

The RNAstructure energy function is modified by adding pseudo-free energy change terms derived from SHAPE reactivities (Low & Weeks, 2010). This approach is based on the observation that SHAPE reactivities correlate strongly with local nucleotides flexibility and therefore with the probability that a nucleotide is single stranded. These additional energetic terms provide a knowledge-based correction to nearest neighbor energy function.

The different steps for the SHAPE data analysis are summarized in the Table VII.2. SHAPE reactivities corresponded to the averaged intensities shown in Figure VII.14.

To add more information for the structure prediction, data obtained from enzymatic probing was also included in the RNAstructure prediction (see Figure VII.15). Hence, positions cleaved by RNAse T1 were indicated as sites of modification, while positions cleaved by the RNAse V1 were set as double-stranded nucleotides.

**Table VII.2 SHAPE data analysis**

| Step | Program used | Description | Input format | Output format | Reference |
|---|---|---|---|---|---|
| **1** | **ImageQuant** | General image analysis software used by Typhoon scanner | | .gel | |
| **2** | **SAFA** | Semi-Automated Footprinting Analysis developed for rapidly quantifying the bands intensities from probing gels at single nucleotide resolution | .gel<br>.seq | .txt | (Laederach et al. 2008) |
| **3** | **noRNAlize** | Normalizes the raw intensities rendered by SAFA and subtracts control intensities in an automated fashion. This program works under Matlab (Mathworks) environment | .txt from SAFA | .txt | (Vicens et al. 2007) |
| **4** | **Excel** | Intensities are normalized to unity. Results from different experiments are averaged and merged. A SHAPE file is prepared as a .txt file, it contains two columns: the numerical nucleotide position and SHAPE reactivity at that position. | .txt | .txt | |
| **5** | **RNAstructure** | Software package for RNA secondary structure prediction and analysis. Constraints from experimental or co-variation data can be included. | .txt<br>.seq | .ct<br>.txt<br>(helix) | (Reuter et al. 2010) |
| **6** | **XRNA** | Java based suite of tools for creation, annotation, edition and display of RNA secondary structure diagrams of publication quality. | .txt<br>(helix)<br>.seq | .xrna<br>.png | |

The nucleotides selected (and set as chemical modified in RNAstructure) were then assigned as in loops, helix ends, GU pairs, or adjacent to GU pairs. These nucleotides were: 22, 55, 130, 132, 134, 161, 165 and 232. The nucleotides selected as double-stranded were: 124, 160, 210, and 211. The enzymatic cleavage sites were considered only as a supporting criterion to refine the secondary structure since enzymatic probing was not performed exhaustively.

### c. Analysis of calculated models

Using constraints from SHAPE and enzymatic probing data, the RNAstructure program predicted 15 different structures (see Annexes 3). A survey of these structures showed that most of them share several structural features. These are HP4, HP1, and a consensus fold in the region encompassing the nucleotides 210 to 264. For the rest of the sequence some recurrent features could also be identified, although the models proposed different conformations.

**Table VII.3 Evaluation of the 7SK models**

| # structure | Free energy (kcal/mol) | HP1 | HP4 | Unreactive regions | Reactive regions | Consistent features |
|---|---|---|---|---|---|---|
| **1** | -246.4 | Yes | Yes | 4/4 | 3/3 | 4 |
| **2** | -241.8 | Yes | Yes | 4/4 | 3/3 | 4 |
| **3** | -238.3 | Yes | Yes | 4/4 | 2/3 | 3 |
| **4** | -238.1 | Yes | Yes | 3/4 | 2/3 | 2 |
| **5** | -235.2 | Yes | Yes | 3/4 | 3/3 | 3 |
| **6** | -235.2 | Yes | Yes | 3/4 | 2/3 | 2 |
| **7** | -234.9 | No | Yes | | | 1 |
| **8** | -230.1 | Yes | No | | | 1 |
| **9** | -230.0 | No | Yes | | | 1 |
| **10** | -229.6 | No | Yes | | | 1 |
| **11** | -226.7 | No | Yes | | | 1 |
| **12** | -225.0 | Yes | No | | | 1 |
| **13** | -224.8 | No | Yes | | | 1 |
| **14** | -223.2 | Yes | Yes | 3/4 | 3/3 | 3 |
| **15** | -222.6 | No | Yes | | | 1 |

However, seix of these 15 structures did not present HP1 and they were not considered as plausible structures (see Table VII.3). Indeed, the existence of HP1would be justified by

functional tests since it has been shown to be the HEXIM binding domain of 7SK (Egloff, Van Herreweghe, & Kiss, 2006; François Bélanger, Huricha Baigude, 2009; and see Chapters IV and V). Isolated HP1 can be produced and its fold is the same than in the 7SK-context (see chapter IV).

A structure presenting a rearranged HP4 was not considered either. It has been shown that LaRP7, a stable component of the 7SK snRNP (He et al. 2008; Krueger et al. 2008) and essential to maintain its integrity in cells, binds the 3' end poly-U tail of 7SK (He et al. 2008). Functional tests, performed in our tem, have also shown that LaRP7 has a higher affinity for the entire isolated HP4 (nucleotides 300 to 332) than by the poly-U tail alone, suggesting that HP4 contains some of the determinants for LaRP7 binding (personal communication of Emiko Uchikawa). The structure of the isolated HP4 (modified to form a stable stem-loop structure) has been solved by NMR (Durney et al. 2010) and agrees with the proposed HP4.

Table VII.3 summarizes all the features taken into account to restrain the choice of the 7SK secondary structure model. Interestingly, of the seven remaining secondary structures, six represented the most stables structures calculated by the RNAstructure program.

The structures were then manually classified according to their consistency with the experimental SHAPE data (see Figure VII.14). Four areas were considered as unreactive:

- 81-90,
- 115-124,
- 203-214
- 259-269

Three regions were considered as highly reactives:

- 51-57,
- 192-197
- 310-314

The two structures that agreed with most of the constraints are shown in Figure VII.16. Four stem-loop structures are strictly identical in both models:

- 24-87 which corresponds to HP1;
- a long stem-loop structure 88-190, which is nearly identical to the domain L2 of the Eilebrecth model (see above), including the small stem-loop structure 172-184, also present is Marz model (M6);
- 210-264, which is similar to HP3 (Wassarman et al. 1991) and to M7 (Marz et al. 2009);

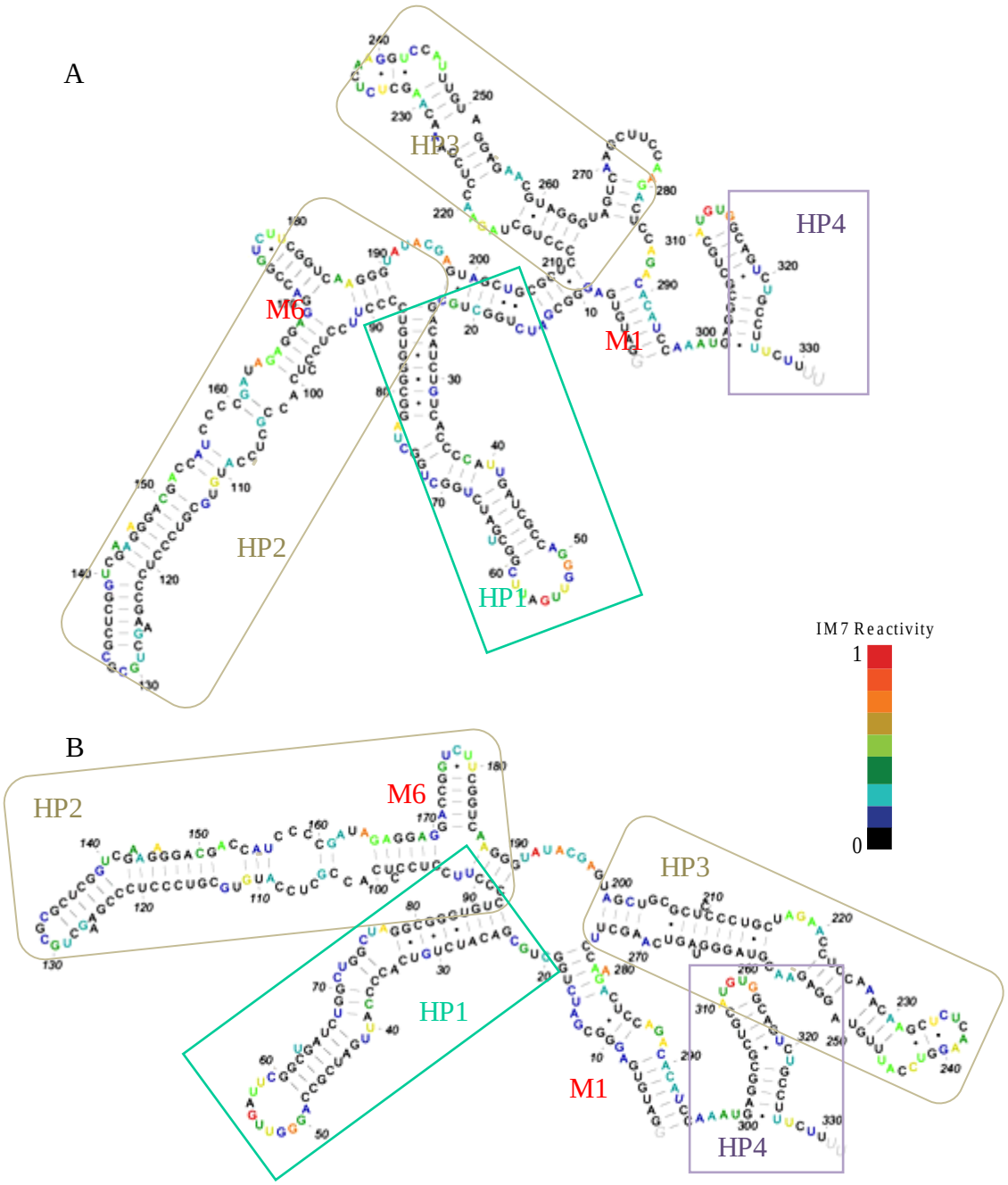



Figure VII.16. 7SK secondary structure models constructed from SHAPE data. 7SK secondary structure predictions were generated with RNAstructure software (Reuter et al. 2010) incorporating the normalized and averaged SHAPE reactivities into the energy function, and data from enzymatic probing. Images were generated using XRNA software (http://rna.ucsc.edu/rnacenter/xrna/xrna.html). A) structure #1 and B) structure #2. Nucleotides are colored according to their reactivity to 1M7 as shown in Figure VII.14. Nucleotides in gray were not explored. HP1 and HP4 are indicated, as well as M1 and M6.

- And 300-326, corresponding to HP4, is suggested to have a free poly-U tail as in Marz model.

Interestingly, a recurrent feature was the stem formed by nucleotides 1-7 base pairing nucleotides 289-295, initially hypothesized in the Marz model (stem M1).

### d. Additional information from sequence analysis

To refine and validate our secondary structure model of 7SK, a final criterion was considered: the evolution. Most functional RNAs exhibit a characteristic secondary structure that is highly conserved in evolution (Hofacker et al. 2002). Phylogenetic information contained in the sequence variation, and most importantly co-variation, can drastically improve the prediction of the secondary structure of RNAs. Namely, comparative sequence alignments can be used to identify nucleotides that exhibit compensatory mutations which are strong indicators of structural conservation.

Nevertheless, although the comparison sequence analysis is a powerful strategy for RNA secondary structure prediction (>95% of predicted pairs correct), one must be careful in its interpretation because multiple, homologous sequences, and high quality alignments are required.

These requirements may be hard to meet for RNAs highly conserved, but also for RNAs with high levels of variability (Mathews et al. 2010). Indeed, highly conserved RNAs have no covariance information, and thus highly diverged sequence carry most covariance information. However, highly diverged sequences are difficult to align correctly (Lindgreen et al. 2006). It has been also shown that the quality of the prediction depends on the proper selection of the sequences, on the number of sequences and on the representativeness in the phylogenetic tree (Yeang et al. 2007). However, sequences showing less than 60% identity are inaccurately aligned, which destroys secondary structure information (Gardner et al. 2005). Besides, an additional problem is to discriminate if co-variation reflects a secondary or a tertiary interaction.

7SK is highly conserved in vertebrates (Gürsoy et al. 2000). In general, the 5' and 3' end regions show the highest conservation (Egloff et al. 2006). Sequence conservation seems to decline rapidly outside jawed vertebrates (Gruber et al. 2008a). However, improved cloning strategies and computational homology searches have allowed detecting divergent 7SK RNAs

in lower vertebrates and even in invertebrates (Gruber et al. 2008a; Gruber et al. 2008b; Marz et al. 2009).

Analysis of local alignments of the best-conserved regions, and in particular of co-variation, resulted in structural information about 7SK (Gruber et al. 2008a). It was shown that the 5' and 3' end structures (HP1 and HP4 respectively) are common to all 7SKs. Besides, evidences for a short hairpin located next to the 5'end hairpin structure, as well as for a vertebrate-specific stem-loop (corresponding to HP3) were also provided. The new sequences identified by Marz and col. allowed then to perform a global multiple sequence alignment from which a consensus model was proposed (Marz et al. 2009; see above).

In order to identify co-variation in the 7SK sequence, we used the global alignment published by Marz and col. (Marz et al. 2009; the structural alignment is provided as supplementary data at www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/ 09-010). However, *Caenorhabditis* sequences were removed for prediction as they still remain controversial (personal communication with Olivier Bensaude).

We compared two approaches for the co-variation analysis. A first analysis was conducted using all 7SK sequences to ensure representativeness, however around half of the sequences have less than 60% identity. The second co-variation analysis was performed using only sequences with more than 60% identity to human 7SK to ensure a more accurate alignment. Only vertebrates (but not reptiles and amphibians) have more than 60% sequence identity (37 sequences). However most of these sequences correspond to mammalians with >95% sequence identity, and therefore contain low co-variation information.

For the secondary structure prediction we used RNAalifold (Bernhart et al. 2008) a tool contained in the Vienna RNA websuite (Gruber et al. 2008c). RNAalifold program predicts the consensus structure for a set of aligned sequences taking into account both thermodynamic stability and sequence co-variation (Hofacker et al. 2002).

Figure VII.17 shows both results. When all sequences were used for the analysis (Figure VII.17A), ten base-pairs including the 42GAUC-GAUC67, signature of 7SK, were strictly conserved in all sequences (highlighted in red in the figure). Nine co-variations were found, distributed almost homogeneously along the sequence. As reported before (Gruber et al. 2008a; Gruber et al. 2008b), several conserved base-pairs accumulated in HP1 and HP4, strongly supporting the existence of these structures, and also in the apical region of the HP2.

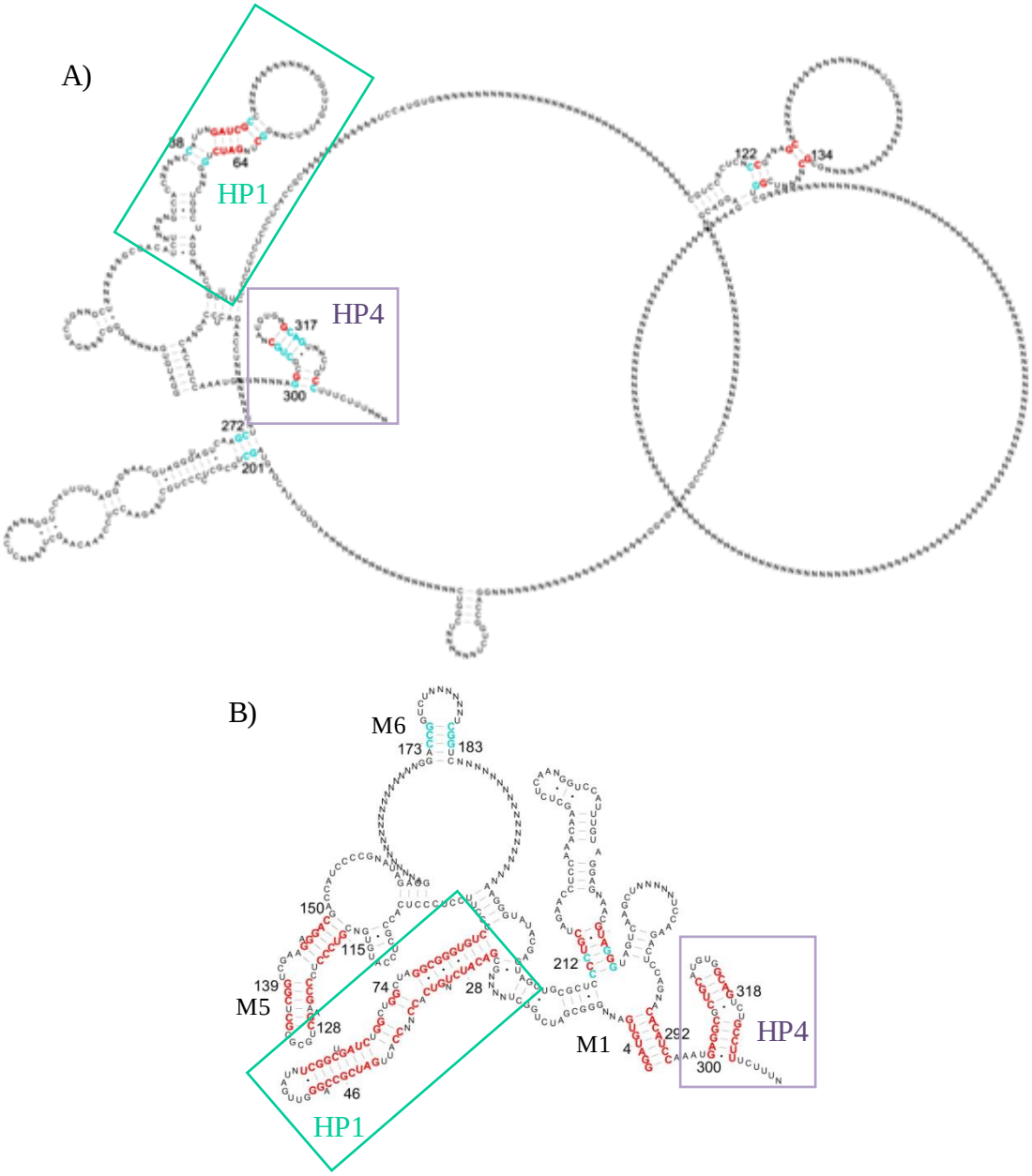
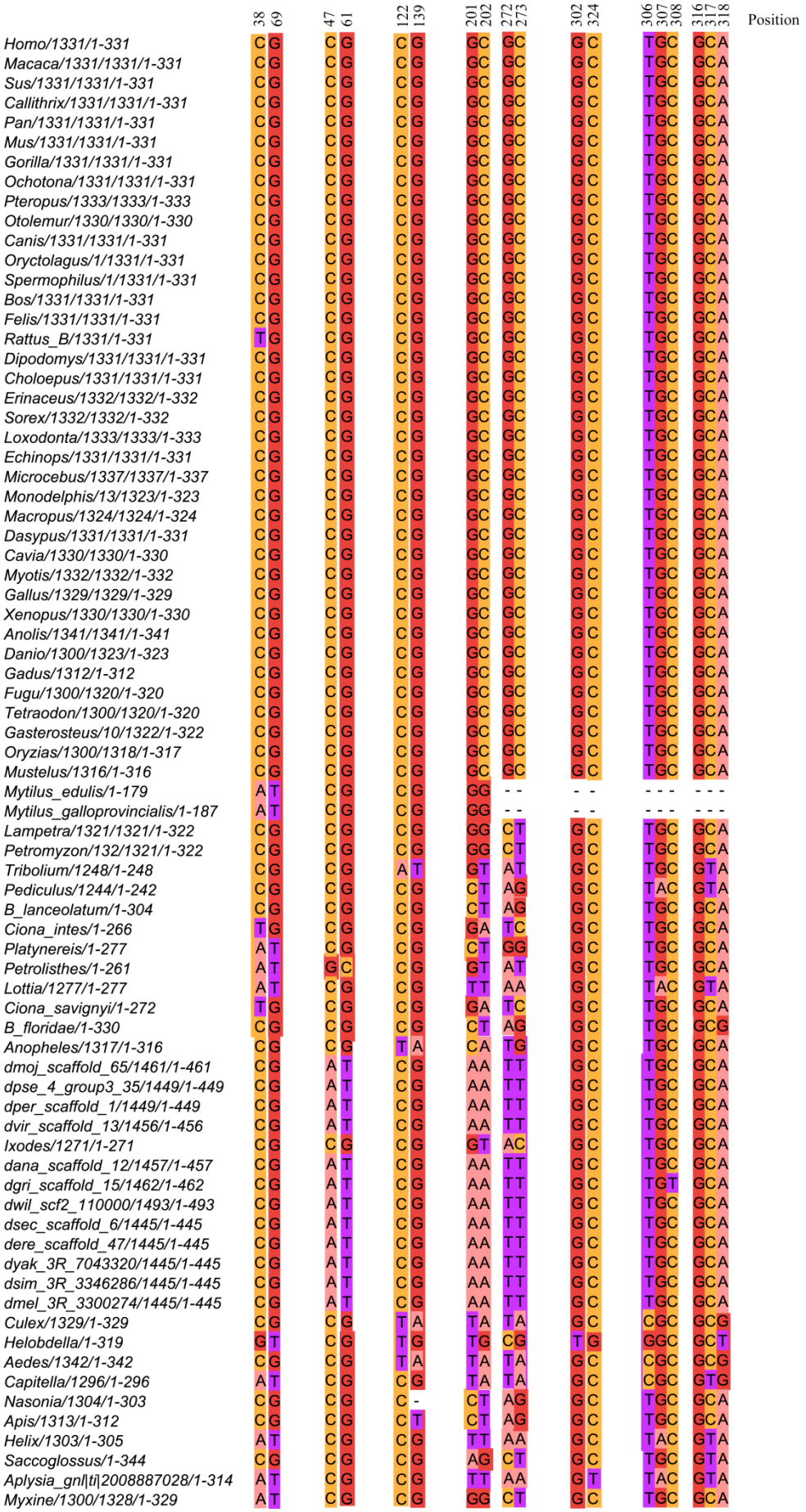



Figure VII.17. Consensus 7SK secondary structure from co-variation analysis. Predictions performed by RNAalifold (Bernhart et al. 2008) using all sequences (A) or only sequences with more than 60% identity (B). Base-pairs showing co-variation are colored in blue, while those conserved in red. The corresponding positions in the human 7SK sequence are noted and the structures corresponding to HP1 and HP4 indicated. Images generated using XRNA software.

A


38 69
47 61
122 139
201 202
272 273
302 324
306 307 308
316 317 318
Position
Homo/1331/1-331
Macaca/1331/1331/1-331
Sus/1331/1331/1-331
Callithrix/1331/1331/1-331
Pan/1331/1331/1-331
Mus/1331/1331/1-331
Gorilla/1331/1331/1-331
Ochotona/1331/1331/1-331
Pteropus/1333/1333/1-333
Otolemur/1330/1330/1-330
Canis/1331/1331/1-331
Oryctolagus/1/1331/1-331
Spermophilus/1/1331/1-331
Bos/1331/1331/1-331
Felis/1331/1331/1-331
Rattus_B/1331/1-331
Dipodomys/1331/1331/1-331
Choloepus/1331/1331/1-331
Erinaceus/1332/1332/1-332
Sorex/1332/1332/1-332
Loxodonta/1333/1333/1-333
Echinops/1331/1331/1-331
Microcebus/1337/1337/1-337
Monodelphis/13/1323/1-323
Macropus/1324/1324/1-324
Dasypus/1331/1331/1-331
Cavia/1330/1330/1-330
Myotis/1332/1332/1-332
Gallus/1329/1329/1-329
Xenopus/1330/1330/1-330
Anolis/1341/1341/1-341
Danio/1300/1323/1-323
Gadus/1312/1-312
Fugu/1300/1320/1-320
Tetraodon/1300/1320/1-320
Gasterosteus/10/1322/1-322
Oryzias/1300/1318/1-317
Mustelus/1316/1-316
Mytilus_edulis/1-179
Mytilus_galloprovincialis/1-187
Lampetra/1321/1321/1-322
Petromyzon/132/1321/1-322
Tribolium/1248/1-248
Pediculus/1244/1-242
B_lanceolatum/1-304
Ciona_intes/1-266
Platynereis/1-277
Petrolisthes/1-261
Lottia/1277/1-277
Ciona_savignyi/1-272
B_floridae/1-330
Anopheles/1317/1-316
dmoj_scaffold_65/1461/1-461
dpse_4_group3_35/1449/1-449
dper_scaffold_1/1449/1-449
dvir_scaffold_13/1456/1-456
Ixodes/1271/1-271
dana_scaffold_12/1457/1-457
dgri_scaffold_15/1462/1-462
dwil_scf2_110000/1493/1-493
dsec_scaffold_6/1445/1-445
dere_scaffold_47/1445/1-445
dyak_3R_7043320/1445/1-445
dsim_3R_3346286/1445/1-445
dmel_3R_3300274/1445/1-445
Culex/1329/1-329
Helobdella/1-319
Aedes/1342/1-342
Capitella/1296/1-296
Nasonia/1304/1-303
Apis/1313/1-312
Helix/1303/1-305
Saccoglossus/1-344
Aplysia_gnl|ti|2008887028/1-314
Myxine/1300/1328/1-329

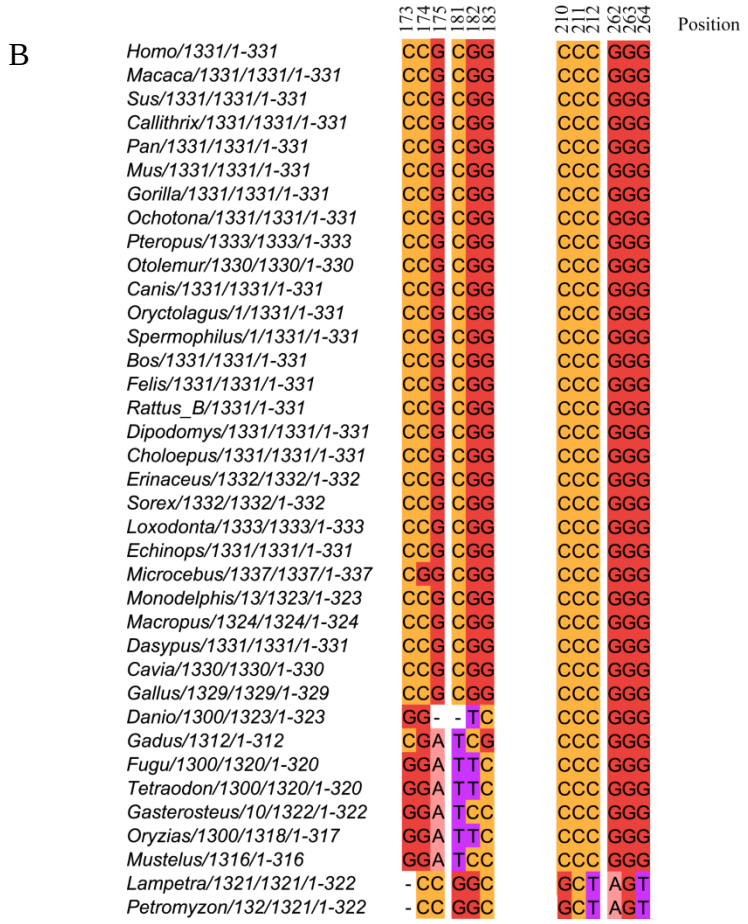



Figure VII.18. Co-variation in 7SK sequence alignment. Extract of 7SK sequence multialignment showing co-variation positions as indicated. Alignement was acquired and modified from (Marz et al. 2009) and computed by RNAalifold (Bernhart et al. 2008). A) All 7SK sequences (but *Caenorhabditis*) alignment analysis; B) Sequences with >60% identity alignment analysis. Figures were created in JalView.

When only the >60% identity sequences were used for the co-variation analysis, several patches of conservation and co-variation were observed. As might be expected given the high percentage of identity between sequences, many conserved base-pairs were predicted. Most conserved base-pairs were found once again in HP1 and HP4. Interestingly, now conserved base-pairs are observed in the apical region of HP2. Conspicuous co-

variations were found in the middle region of HP3 and in M6. Conserved base-pairs were also predicted in M1 stem and in the middle stem of HP3.

The position of co-variation having a reliability index higher than 90% and given into the human 7SK numbering are summarized in the Table VII.4. The complete list is given in Annexes 3. Alignments extracts illustrating the co-variations are shown in Figure VII.18.

**Table VII.4 Co-variation analysis**

| | Nucleotides | Likelihood |
|---|---|---|
| **All sequences (75)** | 38-69 | 99.9% |
| | 47-61 | 100% |
| | 122-139 | 99.4% |
| | 201-273 | 100% |
| | 202-272 | 100% |
| | 302-324 | 100% |
| | 306-318 | 100% |
| | 307-317 | 100% |
| | 308-316 | 100% |
| **Sequences with >60% identity (37)** | 173-183 | 100% |
| | 174-182 | 100% |
| | 175-181 | 99.7% |
| | 210-264 | 98.7% |
| | 212-262 | 99.5% |

e. Combining SHAPE and sequence information

Both candidate models constructed from SHAPE data show nearly all the conserved base-pairs found by both approaches. Only two base-pairs that lead to a slightly rearrangement in the apical region of HP1 were missed, G50-C59 and G51-U58, which were predicted when only sequences with more than 65% identity were used for the calculation. These base-pairs lead to a smaller apical loop with six instead of eleven nucleotides.

The structure #2 contained also all the co-variations. In contrast, structure #1 missed the co-variation at 201-273 and 202-272, in the proximal part of HP3. These two co-variations suggest that the structure #2 is the most plausible model.

## 4. Evaluation and Discussion

We propose a more compact and structured 7SK, in comparison to Wassarman and Marz models (Figure VII.19). Our model share some features with the previously published models, and some of its domains are clearly supported by experimental and co-variance data.

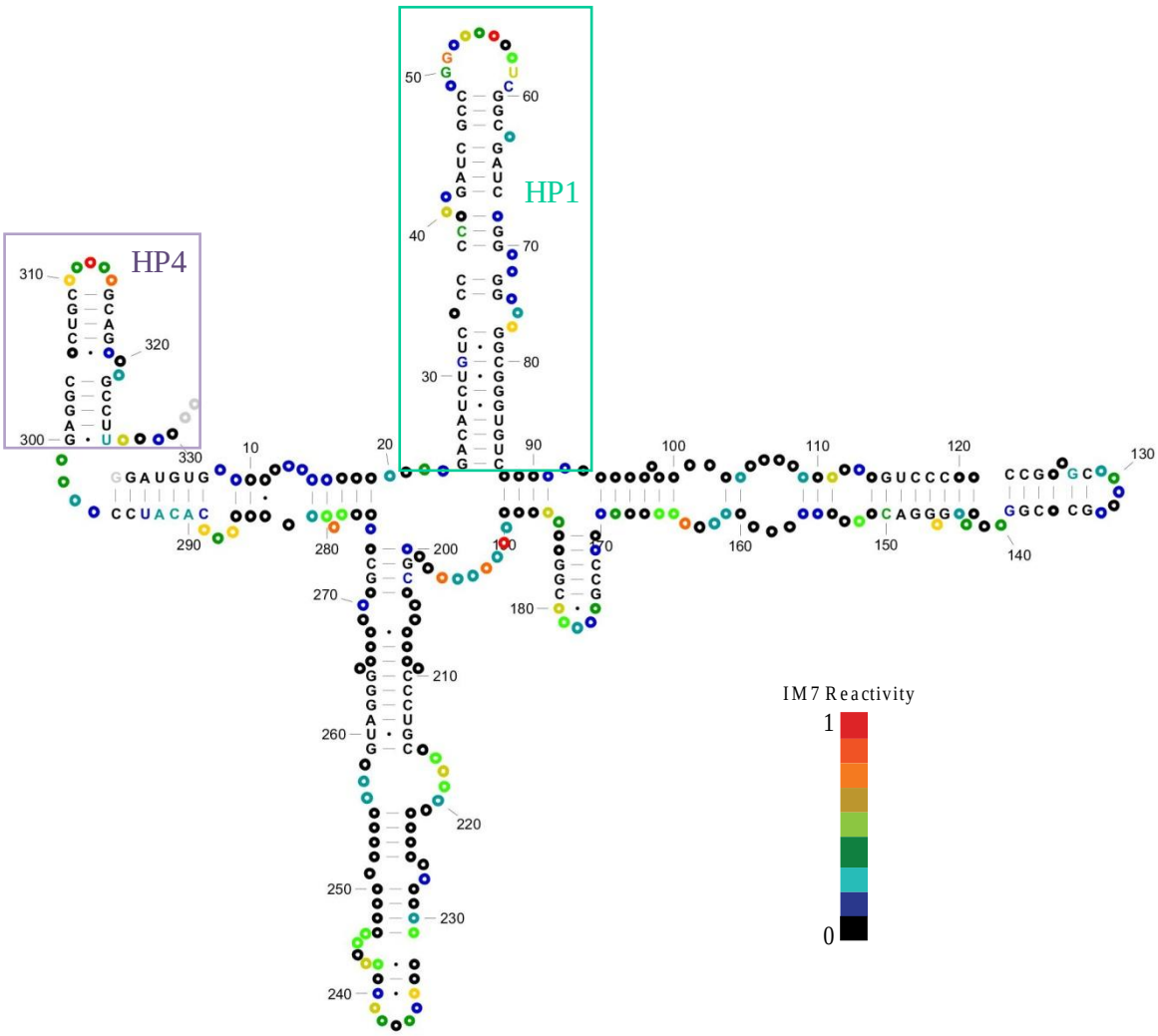



Figure VII.19. Our 7SK secondary structure model. The model is consistent with our SHAPE and enzymatic data, and with co-variation information. Nucleotides predicted in base-pairs by co-variation analysis are indicated. Nucleotides in gray were not explored. HP1 and HP4 are indicated.

In our model, different arms radiate from a central loop:

- HP1 (nucleotides 24 to 87) is proposed with a conformation like in Wassarman and Eilenbrecht models. Sequence analysis predicts two additional base-pairs G50-C59 and G51-U58. However, SHAPE data does not support the formation of these base-pairs due to their high 1M7 reactivity.
- HP2 is similar to Eilenbrecht L2, but without the short hairpin encompassing nucleotides 103-115 that was not supported by sequence analysis. Indeed, sequence analysis suggests that the region 115-118 base-pairs 147-150 which leads to a different arrangement of the region. In contrast, the apical stem-loop is constant in all models. Named stem M6 in Marz model, but also present in Eilenbrecht model, it was supported by our experimental and co-variation data. The existence of the apical hairpin in HP2 is also supported by functional tests since it has been shown that it is involved in the binding of HMGA1 (High Mobility Group Protein) a chromatin factor protein (Eilebrecht et al. 2010). The domain 2 forming a 3-way junction proposed by Wassarman and Steitz is not consistent with co-variation information.
- HP3 is globally similar to HP3 of Wassarman and M7 of Marz models. The main differences with both models were found in the distal region. We propose a HP3 with an apical tetraloop as in Marz model but with asymmetrical internal loops, similar to that proposed by Gruber and collaborators (Gruber et al. 2008a; Gruber et al. 2008b), which is more consistent with our SHAPE data.
- HP4 is the most constant structure in 7SK. Experimental and sequence analysis data suggest that the 3' end poly-U tail is free, which is also consistent with its function as LaRP7 binding site.
- Our model also agrees with the M1 stem of Marz model. However, we propose an extended stem structure containing two short internal loops.

Our model does not present the M2 stem from Marz model. Indeed, M2 is poorly supported by our experimental data, with moderate 1M7 reactivity. Wassarman and Steitz (1991) also showed chemical accessibilities in this region. When a co-variation analysis is performed with all sequences, the prediction suggests the existence of this stem; however, the plausibility for these base-pairs is less than 56%. Nevertheless, the conformation of the regions encompassing nucleotides 8-19 and 276-295 is not clear, since it shows a high 1M7 reactivity suggesting a more open structure than that presented by our model. The sequence analysis did not allow discriminating a plausible secondary structure for this region.

In this chapter, a new model of the secondary structure of 7SK has been proposed. It takes into account experimental and co-variation data, and was calculated with a free energy minimization program. The combination of these different strategies should increase our confidence in our 7SK model. However, further analyses are required to validate it. We plan to use a similar approach than that proposed by Kladwang et al. (2010), this is to mutate nucleotides showing low local flexibility (or accessibility) and therefore suggested as base-paired in our model. If these nucleotides participate actually in base pairing, their mutation should also result in a change of the SHAPE profile of their corresponding nucleotide pairs. In another hand, some insights about the secondary structure can also be provided by the characterization of the three-dimensional structure of 7SK at low resolution as we will see in the following chapter.

# CHAPTER VIII:
# STUDY OF THE SOLUTION STRUCTURE OF 7SK

Structural studies of flexible RNAs and natively disordered proteins represent a challenge. Due to their intrinsic flexibility, obtaining good quality, exploitable crystals for their structural analysis by X-ray crystallography is difficult or even impossible. In theory, for this kind of systems Small-Angle X-ray Scattering (SAXS) is a well suited technique since it allows the study of the structure and interactions of biological macromolecules in solution. However, in absence of crystalline order only low resolution information can be obtained from the scattering data. In fact, SAXS information can be only used to determine the overall molecular shape of proteins and RNAs. However, defining the shapes and conformational space of a biomolecule in solution marks a critical step toward understanding its functional roles.

## 1. WHAT IS SAXS

X-rays are electromagnetic radiations with wavelength in the range of 0.01 to 10 nm. There are two main interactions of X-rays with matter: absorption and scattering. When X-rays hit a material, a fraction will pass through the sample, a fraction will be absorbed and transformed into other forms of energy, and a fraction will be scattered into other direction of propagation. In scattering, the incident wave on a sample and the scattered wave can be described with their respective wavevectors k0 and k1. Scattering can occur with (incoherent scattering) or without (coherent or elastic scattering) loss of energy. Since SAXS considers only the elastic scattering, the incident and the scattered wavevectors both have the magnitude $2\pi/\lambda$. The difference between the two are usually referred to as the scattering vector and is defined by k1–k0 = q (Figure VIII.1). A Fourier transform take us from “real space” of coordinates represented as (r) to the “reciprocal space” of scattering vectors (q).

The length of the scattering vector is then

$$q = 2|k|\sin\theta = 4\pi/\lambda \sin\theta$$

where $2\theta$ is the scattering angle with respect to the incident ray. The units of q are the inverse of units used in the wavelength, typically $Å^{-1}$ or $nm^{-1}$.

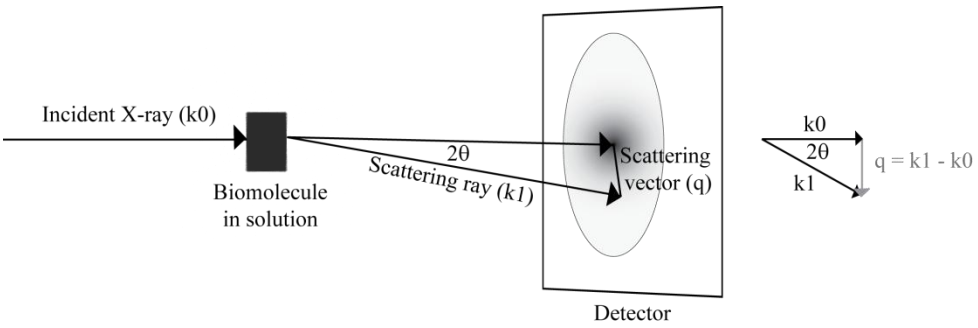



Figure VIII.1. Principle of basic X-ray scattering. Scheme representing a SAXS measurement where an incident X-ray is scattered by the molecules in solution and the scattering intensity is recorded as a function of the scattering angle $2\theta$.

In a SAXS experiment, the scattering intensities I are recorded as a function of the scattering angle $2\theta$. For mathematical convenience, the data are converted to an intensity function I(q) related to the scattering vector q (also called momentum transfer, Figure VIII.2). The scattering intensity I(q) is related to the scattering amplitude by

$$I(q) = A(q)\, A^*(q)$$

where the scattering amplitude A(q) is defined by the Fourier transform of the difference in the electron density, $\Delta\rho(r)$, of the particle of interest, $\rho(r)$, and the bulk solvent $\rho_s$ per unit volume (the contrast).

Since in a SAXS experiment the particles are in solution, they have random distributions, positions and orientations. In diluted solutions where particles do not interact with each other, the intensity from the entire ensemble will reflect the scattering from a single particle averaged over all orientations, what is called the "form factor". If the particles interact with each other, local correlations between the neighbouring particles must be taken into account; this inter-particle interference is called the "structure factor". Hence, the scattering function I(q) is a product of the form factor P(q) and the structure factor S(q):

$$I(q) = P(q)S(q)$$

I(q) can be related to the molecular mass and radius of gyration, Rg, of the molecule by the Guinier equation, which is graphically represented by the Guinier plot [ln I(q) vs $q^2$]. For monodisperse systems, the plot is linear and the molecular mass and Rg are provided by the I(0) intercept and the slope of the line, respectively (Figure VIII.3, red panel). A linear plot of log I(q) vs log q can be used to determine the form of a polypeptide sample, where a slope of approximately 2 indicates a Gaussian chain; 1.66 a chain with excluded volume; and 1 a rod shape.

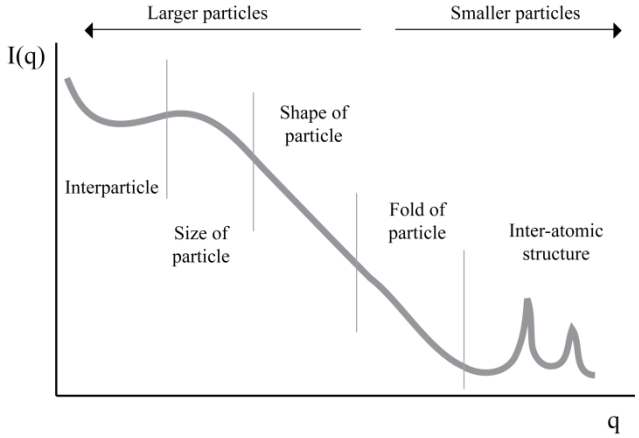



Figure VIII.2. Scheme of the intensity versus scattering vector of a biomolecule. The different structural information than can be theorically obtained from the plot is indicated.

SAXS can be also useful for identifying and characterizing biomolecules without folded domains. The Kratky plot [$q^2$I(q) as function of q], which can be calculated directly from the scattering curve, provides a tool for evaluation the folding of samples.

By means of Fourier transform, I(q) can be converted into the real space pair distance distribution function, P(r), which provides direct information about the distances between electrons in the scattering particles within a given volume. The maximum dimension particle ($D_{max}$) can be then estimated from P(r) by the value at which the function approaches zero. P(r) allows the graphical displaying of the features of the particle shape. For simple shapes, P(r) can provide a straightforward and intuitive representation of the data for visual inspection. Figure VIII.4 presents typical scattering patterns and pair distance distribution function of geometrical bodies with the same maximal size.

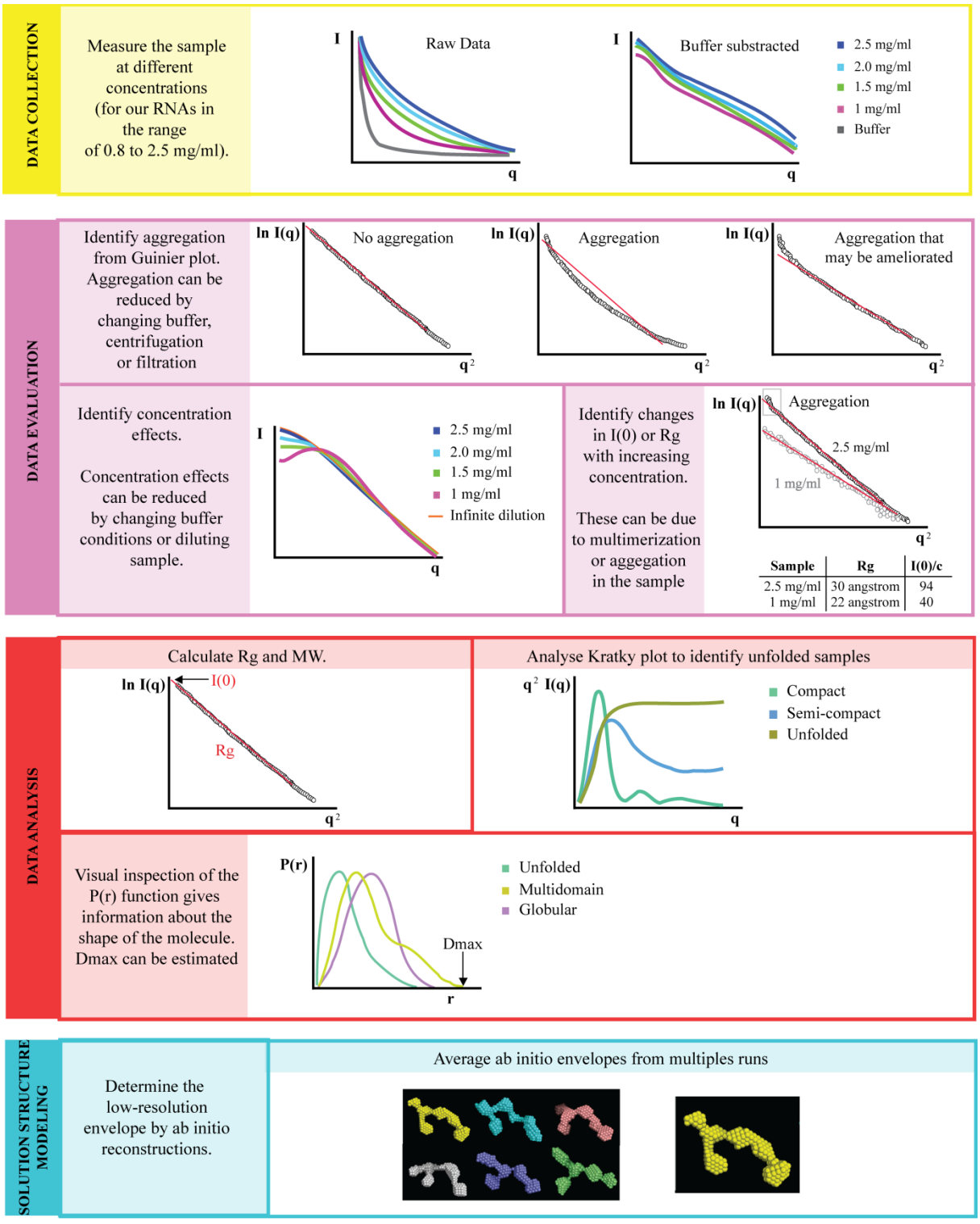



Figure VIII.3. Biological SAXS study strategy. Scheme representing the different steps during a SAXS experiment: data collection, evaluation and analysis, and shape model reconstruction [adapted from (Putnam et al. 2007)].

More complex shapes, often encountered in proteins or large RNAs, cannot easily be interpreted from visual inspection of P(r). Instead, computer based methods must be employed to identify models that can satisfy the scattering data. The general approach taken by these methods is to propose shapes, calculate scattering curves or P(r) functions and optimize the agreement to the experimental data.

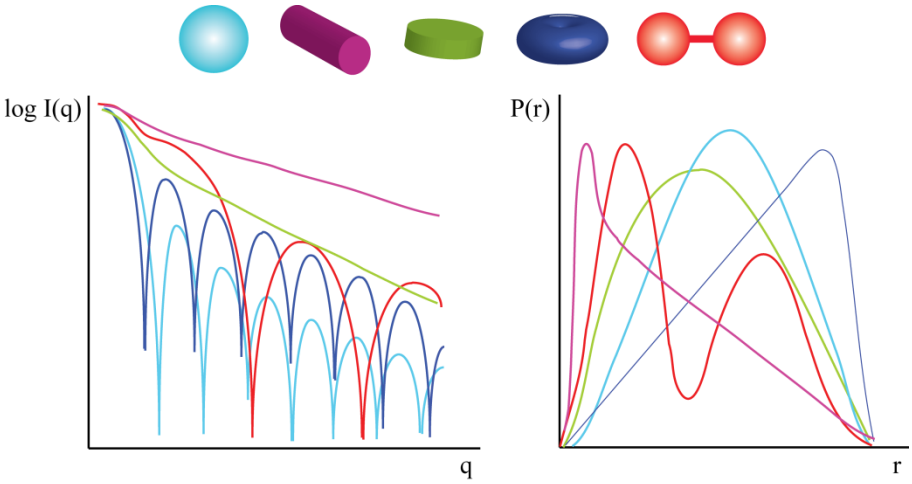



Figure VIII.4. Scattering intensities and pair distance distribution function of geometrical bodies. Globular particles (light blue) display bell-shaped P(r) functions with a maximum at about $D_{max}/2$. Elongated particles have skewed distributions with a clear maximum at small distances corresponding to the radius of the cross-section (pink). Flattened particles display a rather broad maximum (green), also shifted to distances smaller than $D_{max}/2$. A maximum shifted towards distances larger than $D_{max}/2$ is usually indicative of a hollow particle (blue). Particles consisting of well-separated subunits may display multiple maxima, the first corresponding to the intrasubunit distances, the others yielding separation between the subunits (red).

Many of the most commonly used relationships relevant in monodisperse, diluted solutions (without inter-particle interactions) are shown in Table VIII.1. The general strategy used for the study of biomolecule by SAXS is summarized in Figure VIII.3. Some reviews explaining more detailed theory and data analysis of SAXS are in (Putnam et al. 2007; Svergun et al. 2003; Mertens et al. 2010).

**Table VIII.1. Relationships commonly used biological SAXS analysis**

| Parameter | Formula |
|---|---|
| Radius of gyration (Rg): Guinier approximation | $\ln[I(q)] = \ln[I(0)] - \frac{q^2Rg^2}{3}$ |
| Radius of gyration (Rg): Debye approximation | $I(q) = \frac{2\,I(0)}{q^4Rg^4}(q^2Rg^2 - 1 + e^{-q^2Rg^2})$ |
| Radius of gyration (Rg): defined by P(r) | $Rg^2 = \int_0^{Dmax} r^2P(r)dr \Big/ \int_0^{Dmax} P(r)dr$ |
| Pair distance distribution function P(r) | $P(r) = \frac{r}{2\pi^2}\int_0^{\infty} I(q)qsin(qr)dq$ |
| Maximum dimension ($D_{max}$) | $D_{max}$ is the value of r at $P(r) = 0$ for large r |
| Particle volume (V) defined by Porod Invariant | $V = 2\pi^2 I_{exp}^2(0) \Big/ \left(\int_0^{\infty} I(q)q^2dq\right)$ |
| I(0): Intensity at q=0 which is also proportional to mass and volume | $I(0) = 4\pi\left(\int_0^{Dmax} P(r)dr\right)$ |
| Mass (M) | $M = \frac{I(0)\mu^2}{N_A\left(1 - (\rho_S/\rho_p)\right)^2}$ |

## 2. SAXS Study of the Three-Dimensional Envelope of 7SK and its Subdomains

Until now there is no information about the three-dimensional structure of 7SK. 7SK is a large snRNA and it has been suggested as a dynamic scaffold where conformational changes, for which an intrinsic flexibility is needed, would play an important functional role (Van Herreweghe et al. 2007; Marz et al. 2009; Krueger et al. 2010; Lebars et al. 2010). Moreover, SHAPE analysis of 7SK also shows signs of a flexible molecule (see Chapter VII). Previous secondary structure models based on experimental, sequence analysis and energy minimization, as well as our model from SHAPE and sequence analysis data, propose a modular organization of 7SK, with independent subdomains connected by flexible hinges. Indeed, evidences suggest that isolated 7SK hairpins are functional and able to recruit independently their protein partners (Bélanger et al. 2009; Lebars et al. 2010; Muniz et al. 2010; Eilebrecht et al. 2010; Durney et al. 2010). Thus, 7SK is pictured as a dynamic modular scaffold where each hairpin would be a platform for protein 7SK partners, and these

interactions would trigger 7SK conformational changes important for its regulation and function. On the whole, this makes 7SK an interesting system for its structural study in solution. A SAXS based *ab initio* model of 7SK should serve as a suitable structural framework for model building. Hence, we undertook a three-dimensional characterization of 7SK by SAXS. All SAXS measurement were carried out using the Beamline X33 at the DORIS storage ring of the European Molecular Biology Laboratory (EMBL) at Deutsches Elektronen Synchrotron (DESY), Hamburg, Germany, in collaboration with Michal Gadja and Dmitri I. Svergun.

## 2.1. Strategy: "divide and conquer"

If 7SK has a modular nature, it should be possible to dissect it into its subdomains and measure them separately, to give more interpretable SAXS data to reconstruct *ab initio* models. Then, by comparing these models with the full length construct, we should be able to build a solutions model of the entire 7SK. Our SHAPE data and the 7SK secondary structure models available (Wassarman et al. 1991; Marz et al. 2009; Eilebrecht et al. 2010), or our own proposition (Chapter VII), were a starting point for the design of the different construction. Reciprocally, SAXS data analysis should allow us to validate or improve our 7SK secondary structure model. It should be noted that the interpretation of the solution envelopes obtained from SAXS data strongly depends on secondary structure information.

Thus, we designed different constructions of 7SK where one subdomain was deleted, constructions of regions comprising two subdomains, and constructions consisting of individual subdomain isolated (Figure VIII.5).

Small constructions such as HP1, HP1u, L2, L3, HP3, and HP4 were systematically purified by Mono Q chromatography. Large constructions were refolded by thermal treatment as previously described, in a buffer at low ionic strength. Then, all RNAs were dialysed for > 16 h against a buffer containing monovalent salt (100 mM NaCl). The homogeneity of each sample was monitored by agarose gel (Figure VIII.6). In general, SAXS data were collected at RNA concentrations ranging from 0.8 to 3 mg/ml in Cacodylate pH 6.5, 100 mM NaCl, 6 mM $MgCl_2$, and 0.25 mM EDTA. Samples were centrifugated for 15 min at 14,000 g immediately before data collections.

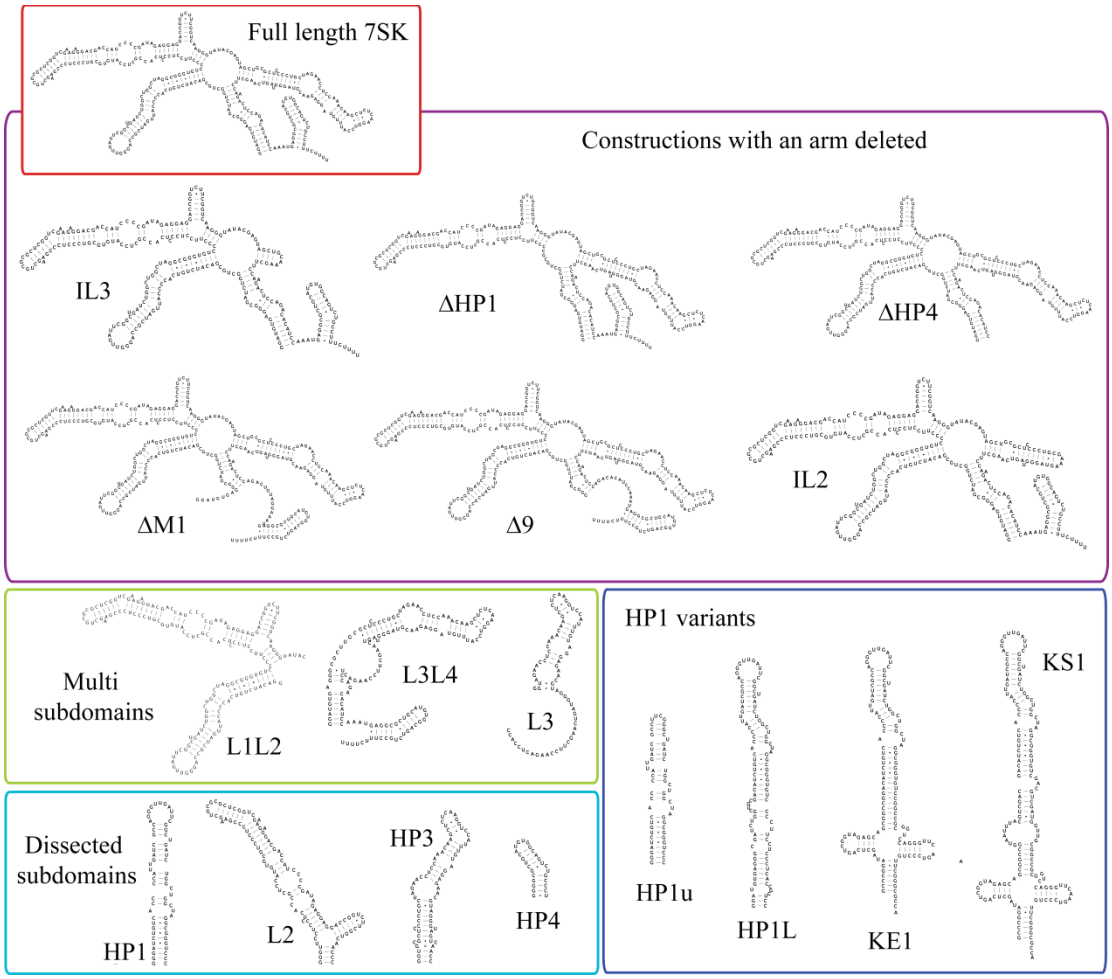



Figure VIII.5. 7SK constructions for SAXS analysis. Schematic representations of each construction (as indicated) are shown; for sequence details see Table III.1 in Chapther III "Molecules Preparation". These were drawn according to our proposed secondary structure model, but were designed with respect to Wassarman, Marz or Eilebrecht models.

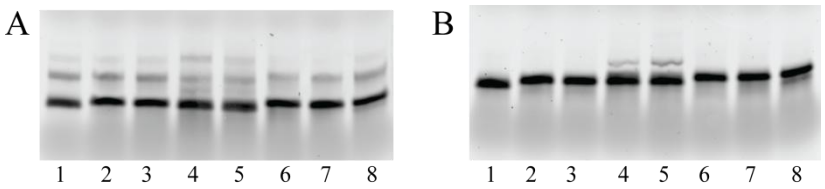



Figure VIII.6. RNAs samples for SAXS measurements. Agarose gel showing ΔHP1 (1), IL2 (2 and 3), ΔHP4 (4 and 5), Δ9 (6), M1 (7), and 7SK (8) before (A) and after (B) thermal treatment.

## 2.2. SAXS data evaluation

Concentration effects, indicating aggregation or interparticle interference effects, were observed when using a buffer without monovalent salt. A buffer with higher ionic strength (100 mM NaCl) was suitable to overcome the electrostatic repulsions between RNAs due to their highly negatively charged sugar-phosphate backbone (Figure VIII.7).

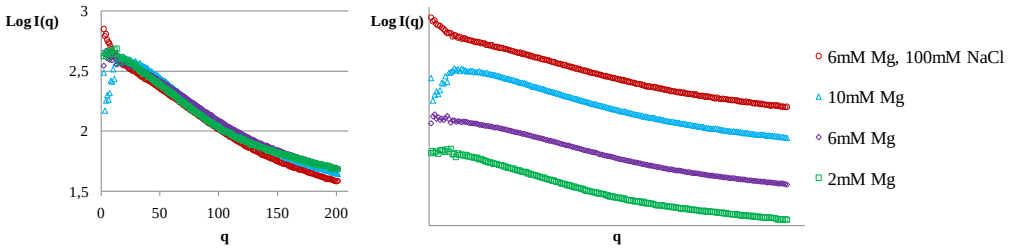



Figure VIII.7. 7SK SAXS profiles at different conditions. All measurements were performed at 1.2 mg/ml with the same 7SK batch in a buffer consisting of Cacodylate pH 6.5 and the indicated salt. Right, the curves are presented for clarity without superimposition. Decrease of intensity at very small q indicates the presence of repulsion forces (Putnam et al. 2007).

The biophysical parameters obtained by SAXS for most of the measured RNAs are summarized in Table VIII.2. As mentioned above, the I(0) determines the MW since it is related to the number of electrons in the scatterer (Putnam et al. 2007). Experimentally, the MW is estimated by comparison with a standard with a known MW, generally the Bovine Serum Albumin (BSA), since it is not possible to measure the absolute intensity of the scatterer directly. It might be noted that the BSA may not be a suitable standard for RNA MW determination since the ratio between the molecular weight and the number of electrons depends on the chemical composition of the molecule. Unexpectedly, we found that the ratio between the experimental and theorical MW was higher for our large constructs than for the individual hairpins. In general, hairpins showed a ratio around 4, similar to that for the $tRNA^{Thr}$, measured as reference. Large RNAs, however, showed a ratio around 6 or higher. This could indicate aggregation, contamination, or equilibrium between monomers and dimers. However our RNA preparation protocol did not show dimers after thermal treatment.

Indeed, equilibrium between monomers and dimers should be sensitive to the concentration and not be constantly.

**Table VIII.2. SAXS parameters obtained for some RNAs**

| | Rg | $D_{max}$ | Volume | $Vol/MW_{th}$ | $MW_{exp}$ | $MW_{exp}/MW_{th}$ |
|---|---|---|---|---|---|---|
| **7SK** | 9.2 | 29.7 | 485.6 | 4.8 | 602.2 | 6.0 |
| **M1** | 9.5 | 31.0 | 586.0 | 6.1 | 554.2 | 5.8 |
| **D9** | 10.2 | 34.2 | 749.0 | 7.6 | 587.2 | 5.9 |
| **DHP4** | 10.3 | 35.0 | 700.3 | 7.8 | 604.9 | 6.7 |
| **DHP1** | 9.3 | 33.0 | 507.0 | 6.2 | 531.6 | 6.5 |
| **IL2** | 9.8 | 34.4 | 630.1 | 7.1 | 605.1 | 6.8 |
| **HP1** | 2.5 | 8.7 | 37.4 | 1.9 | 77.1 | 4.0 |
| **HP1U** | 2.3 | 8.0 | 33.5 | 1.9 | 74.2 | 4.3 |
| **KE1** | 4.1 | 14.3 | 77.4 | 1.9 | 130.2 | 3.1 |
| **KS1** | 5.0 | 17.5 | 130.0 | 2.6 | 202.1 | 4.0 |
| **HP3** | 3.0 | 10.6 | 43.9 | 1.8 | 93.8 | 4.2 |
| **$tRNA^{Thr}$** | 2.2 | 7.5 | 37.0 | 1.7 | 93.6 | 3.7 |

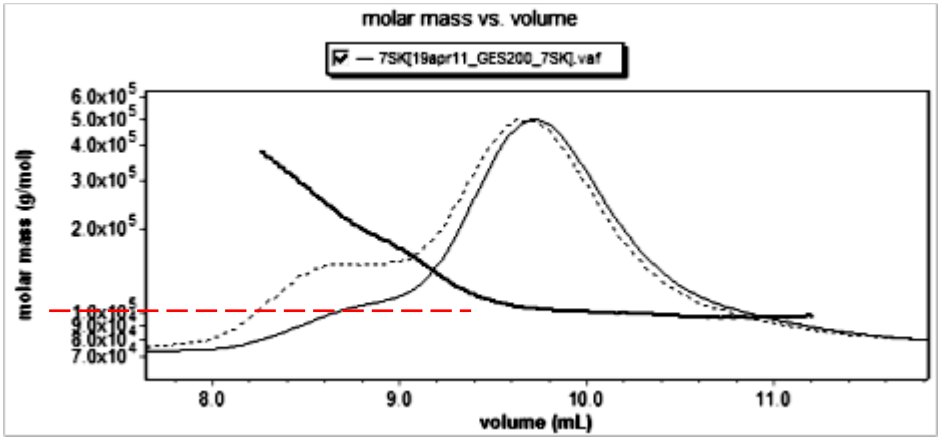



Figure VIII.8. MALS measurement of 7SK. A molar mass of ~100 kDa is calculated for the main peak of 7SK (red dashed line), consistent with a monomeric molecule.

To further monitor the homogeneity of our 7SK sample and asses its oligomeric state in solution we could, at the end of the project, use multiangle light scattering (MALS; Figure VIII.8). MALS measurement of 7SK was largely consistent with a monomer (MW of 110 kDa). However, there was indeed a small amount of aggregates that could not be removed by centrifugation. A recent study showed that most of the inconsistencies that we observed may

be explained by the heterogeneity of the folded RNA (Rambo et al. 2010). The authors obtained improved SAXS data by using SEC and concluded that a constant control of the samples by using MALS is required. In our experiments, the high quality samples for SAXS measurements were achieved by using Mono Q chromatography. However, only poor yields were obtained for large RNAs, so SEC may be a more suitable method to eliminate heterogeneity for these RNAs.

## 2.3. SAXS data analyses

### a. Larger constructions

Visual inspection of the different graphical representation of SAXS data provides some insights about the molecule compactness and shape. The Kratky plot is typically used to assess the "folded-ness" of a molecule (see Figure VIII.3 "Data Analysis" panel). It is well known that the folding stability of RNAs is highly sensitive to $Mg^{2+}$ ions concentrations (Leroy et al. 1977). Probing experiments suggested that after thermal treatment at 2 mM $MgCl_2$, incubation of 7SK at 6 mM $MgCl_2$ resulted in a more compact molecule. We measured 7SK at different Mg concentrations when setting up the SAXS buffer conditions (Figure VIII.9). 7SK showed a Kratky profile of a slightly more compact molecule at 6 mM Mg. As seen in Figure VIII.7, at this Mg concentration 7SK also showed less aggregation.

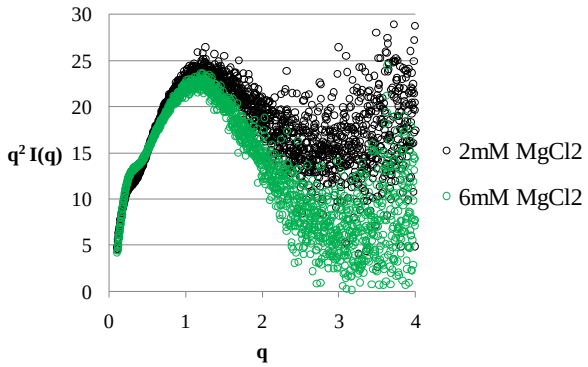



Figure VIII.9. Kratky plots of 7SK. Kratky representation of 7SK SAXS data at two different $MgCl_2$ concentrations. Plots were calculated with PRIMUS (Konarev et al. 2003).

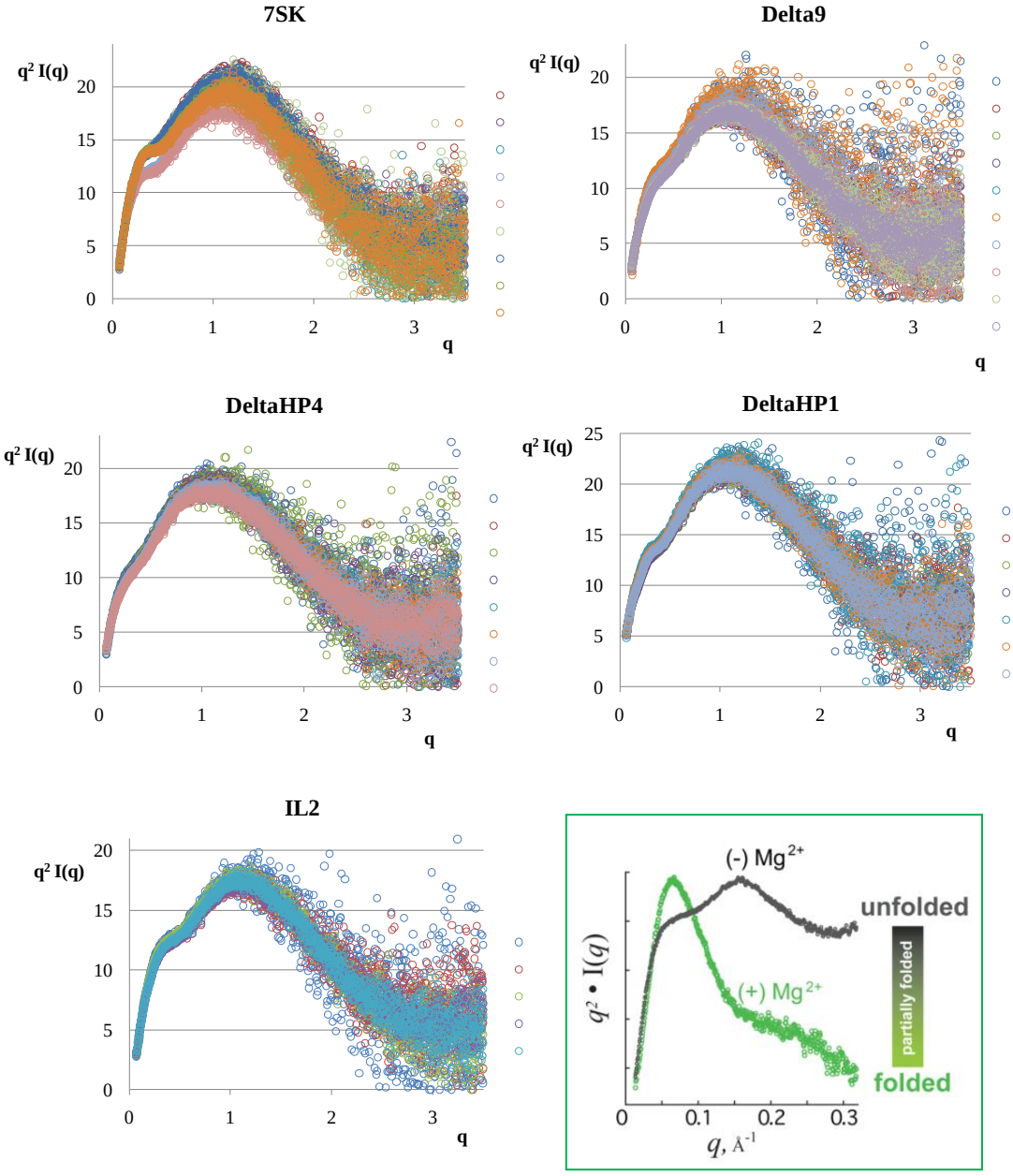



Figure VIII.10. Kratky plots of large 7SK constructions. Superposition of scattering profiles in Kratky representation for different samples (different batches and concentrations) of 7SK (9 samples), Δ9 (10 samples), ΔHP4 (8 samples), ΔHP1 (7 samples), and IL2 (5 samples). Plots were calculated with PRIMUS (Konarev et al. 2003). An example form the literature, the lysine riboswitch in the presence or not of Mg, is shown in the green box to illustrate the difference of a folded and an unfolded molecule (Rambo and Tainer 2011a).

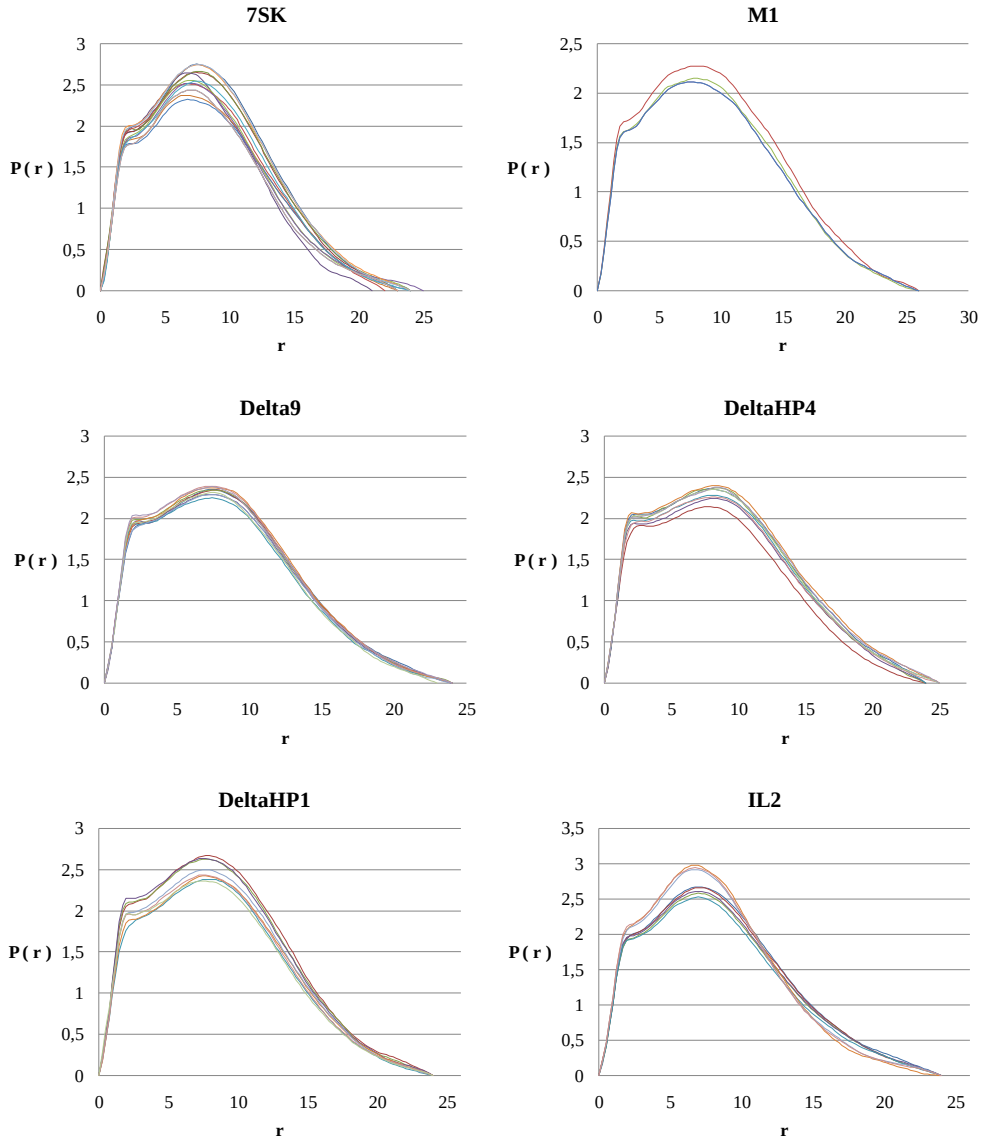



Figure VIII.11. P(r) function of large 7SK constructions. Superposition of P(r) functions obtained for different samples (different batches and concentrations) of 7SK (14 samples), M1 (4 samples), Δ9 (10 samples), ΔHP4 (9 samples), ΔHP1 (9 samples), and IL2 (8 samples) are shown. Functions were calculated with GNOM (Svergun, 1992).

The scattering profiles in Kratky representation and the pair distribution function, P(r), of all large RNA constructions (7SK, M1, Δ9, ΔHP4, ΔHP1, and IL2) are shown in Figures VIII.10 and VIII.11, respectively. The consistency of the measurements was verified by the superposition of the Kratky plots for different samples of large RNA constructions.

While most of the constructions presented reliable superpositions, 7SK showed more inconsistencies between the different samples. All RNAs showed a profile consistent with a semi-compact, flexible molecule, with a slow decreasing curve at large q values (Rambo and Tainer 2011a). Two peaks could be distinguished, in particular for 7SK and IL2, probably indicating the presence of two elements, one more structured and a second one more flexible.

Indeed, the 7SK SAXS profiles at different concentrations were superimposable when using the same batch of sample, but the reproducibility between different batches of samples (different beamtimes) was not satisfactory. This was also reflected by parameters like Rg and $D_{max}$ obtained from SAXS data (Table VIII.3).

**Table VIII.3. Some SAXS parameters obtained for 7SK**

| Run | mg/ml | Rg (nm) | $D_{max}$ (nm) | I(0) |
|---|---|---|---|---|
| **March 2010** | 2.2 | 9.346 +/- 0.01 | 32.710 | 551.8 |
| | 1.6 | 9.196 +/- 0.01 | 32.190 | 545.8 |
| | 1.1 | 9.327 +/- 0.03 | 31.950 | 567.4 |
| | 0.8 | 9.304 +/- 0.08 | 31.870 | 597.3 |
| **May 2010** | 1.9 | 8.361 +/- 0.001 | 28.640 | 461.4 |
| | 1.1 | 8.069 +/- 0.31 | 27.640 | 502.8 |
| | 1.0 | 8.622 +/- 0.16 | 30.180 | 553.2 |
| | 0.6 | 8.931 +/- 0.22 | 29.250 | 592.6 |

In another hand, P(r) can be acceptably superimposed, 7SK and IL2 showing the highest inconsistencies. P(r) functions presented two peaks, suggesting again a multidomain molecule. The main peak showed a maximum at smaller distances than D/2, which theorically corresponds to a flattened molecule. The second peak showed a maximum at very small distances, which is typical of elongated molecules. Hence, 7SK can be described as a flexible multidomain molecule, flat and elongated.

In order to emphasize the structural differences, the Kratky plot and the P(r) function of the different RNA constructions are superimposed in Figure VIII.12. The Kratky plot of IL3, which is deleted of the whole HP3, showed a partially unfolded molecule that may explain the low stability observed for this RNA, which degraded very easily. The rest of the constructions showed similar compactness, IL2 and ΔHP1 presenting a first peak easy to distinguish. Even if the P(r) functions of the different constructions are not completely superimposables, the general shape of the function is conserved, suggesting that the deletion of the corresponding subdomains does not cause significant disruption of the structure. Only

the deletion of the complete HP3 seemed to affect the core of 7SK structure. These results suggest that 7SK is a modular RNA with independent subdomains.

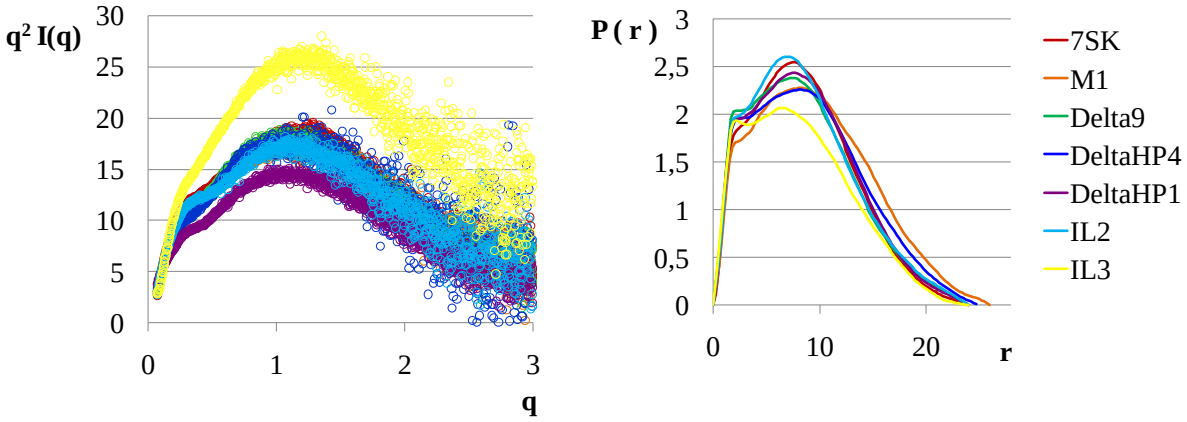



Figure VIII.12. Superposition of Kratky plots and P(r) functions of large 7SK constructions. The same color code is used for both graphs. Superposition of P(r) functions obtained for different samples (different batches and concentrations) of 7SK (14 samples), M1 (4 samples), Δ9 (10 samples), ΔHP4 (9 samples), ΔHP1 (9 samples), and IL2 (8 samples) are shown.

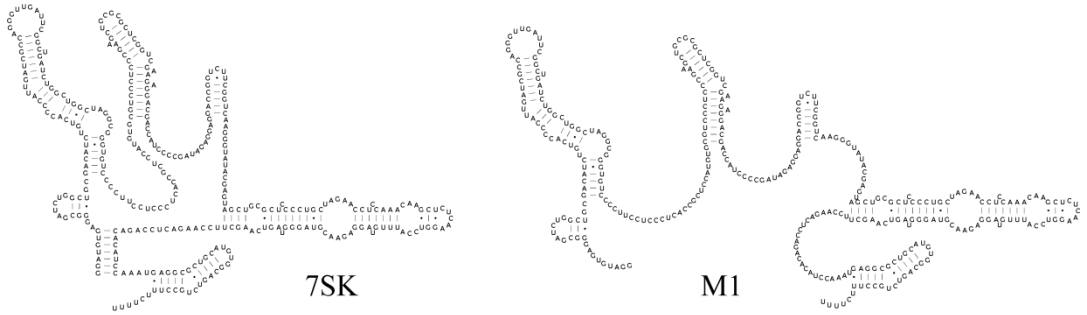



Figure VIII.13. M1 RNA according to Marz model. Schematic representation of 7SK (left) and M1 (right) RNAs, which has a more open structure according to Marz model.

Interestingly, the suppression of the M1 stem described by Marz et al. (2009), either by mutation (M1 RNA) or deletion (Δ9 RNA) of the nine 5'end nucleotides, did not result in

significantly more extended molecules, which would be reflected in the $D_{max}$ (the value at which the P(r) approaches zero). The Marz model describes 7SK as a molecule circularized by the base pairing of the first 5'end nucleotides to the region just before HP4 (see Figure VII.3 of Chapter VII). The resulting stem was called M1. Hence, according to the Marz model the suppression of M1 would release the 5' and 3'ends leading to an extended molecule (as schematized in Figure VIII.13). These observations were therefore more consistent with our secondary structure model, where a longer deletion is required to give an open molecule (Figure VIII.5).

b. Smaller constructions

Isolated subdomains showed a much more reproducible SAXS parameters (see Table VIII.2), such as I(0) and Rg, which were constant between the different measurements (Figure VIII.14).

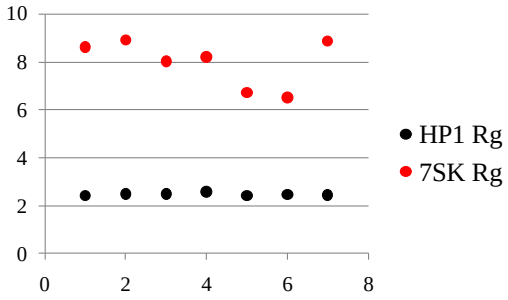



Figure VIII.14. Rg values for HP1 and 7SK. Rg values estimated from Guinier plot from SAXS data for different samples of HP1 (black circles) and 7SK (red circles).

In line with this, their Kratky plots and the P(r) functions were nicely superimposables, as shown in Figure VIII.15 for HP1 and HP3. Kratky plots suggested that HP1 and HP3 are well folded RNAs. The P(r) function showed a typical profile for an elongated, rod-like molecule as expected for hairpins structures.

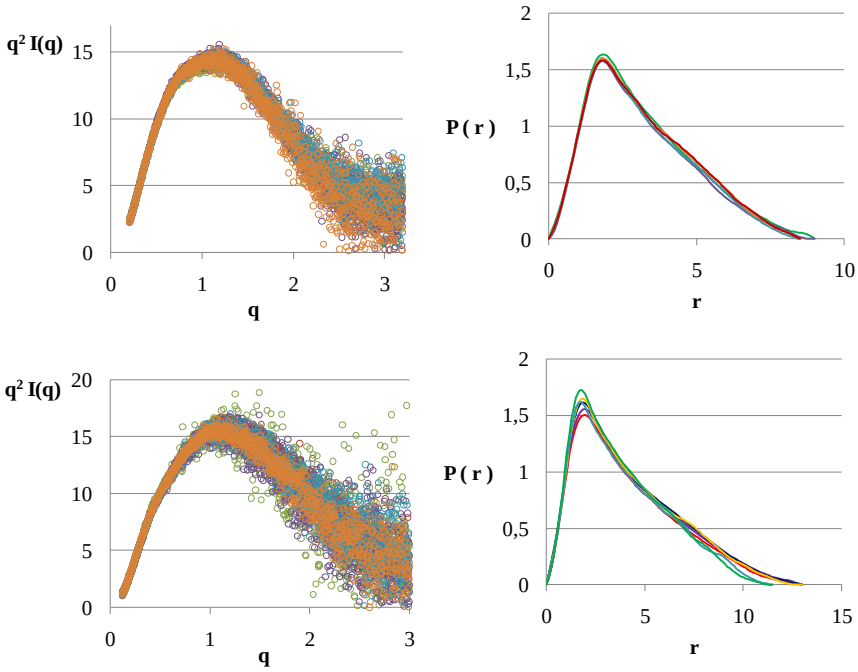



Figure VIII.15. Kratky plots and P(r) functions of HP1 and HP3. Upper panel, HP1 (6 samples); lower panel HP3 (6 samples).

The Kratky plot and the P(r) function of $tRNA^{Thr}$ and KE1 are shown in Figure VIII.16. The P(r) function emphasizes differences between these two RNAs.

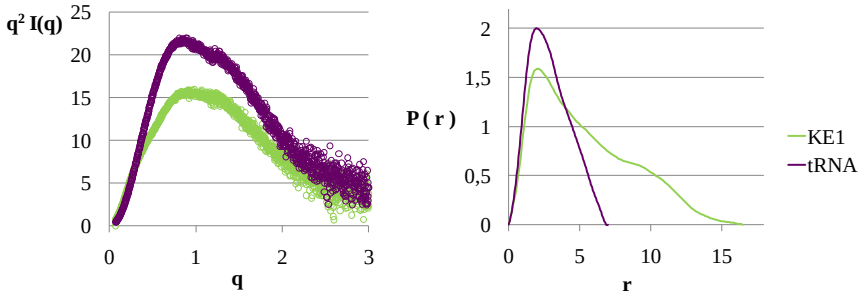



Figure VIII.16. Kratky plots and P(r) functions of KE1 and $tRNA^{Thr}$. The same color code is used for both graphs.

## c. Multi-subdomains RNAs

The isolated L2 or the constructions L1L2 showed $D_{max}$ of approximately 220 and 440 Å, respectively, which is much higher than expected (see Table VIII.2) and suggesting that these RNAs were unfolded. These results indicate either that L2 is a very flexible region whose

stability depends on the rest of the 7SK structure, either that the secondary structure of L2 is incorrect and therefore the designed construction resulted in a completely unfolded RNA. The design of L2 construction was based on the Eilenbrecht 7SK model presented in the Figure VII.4 of Chapter VII. Indeed, it was reported that L2 mediates the interaction between 7SK and HMGA1 (high mobility group protein), a chromatin factor and transcription regulatory hub (Eilebrecht et al. 2010). The authors showed that the apical portion of L2 substructure of 7SK (nucleotides 113 to 154), inserted in the viral EBER2 snRNA (which probably stabilized L2), was able to bind HMGA1 and to promote positive or negative regulatory activity. Our predictions of the secondary structure of 7SK largely agreed with this L2 model (see Chapter VII), and sequence analysis showed co-variation confirming its apical region and M6 conformation. However, it should be noted that the basal region of L2 showed some inconsistencies with SHAPE data. Thus, we propose to test the conformation of the basal region by coupling mutational analysis and SHAPE. This illustrates the necessity of well established RNA secondary structure model for carrying out SAXS studies.

In contrast, the L3L4 construction showed a $D_{max}$ around 180 Å and data quality that could be used for further analysis (see below). This also suggested that the L3L4 construction is rather structured RNA, probably as modeled in Figure VII.5.

## 2.4. *Ab initio* reconstructions

We attempted to construct envelopes of RNAs by *ab initio* modelling with DAMMIN (Dummy Atom Model Minimisation; Svergun, 1999). DAMMIN represents a molecule as a collection of densely packed beads inside a constrained (usually spherical) volume, with a maximum diameter defined by the experimentally determined $D_{max}$. Each bead is randomly assigned to the solvent or solute and the shape reconstruction is conducted starting from a random initial approximation by simulated annealing (SA). At each step in the SA procedure the assignment of a single bead is randomly changed leading to a new model, and refined against SAXS data. The solution is constrained by the penalty term, requiring that the beads must be connected and the model compact.

In agreement with the reproducible SAXS data for individual subdomain, ab initio reconstructions resulted in reproducible envelopes (Figure VIII.17). HP1, HP3 and HP4 showed elongated shape consistent with hairpin structures. HP1 presented an asymmetrical

shape, with a bulge and a bent next to the center of the molecule. HP3 presented a more symmetrical shape. HP4 showed an asymmetrical envelope with a thinner tail at one of its ends.

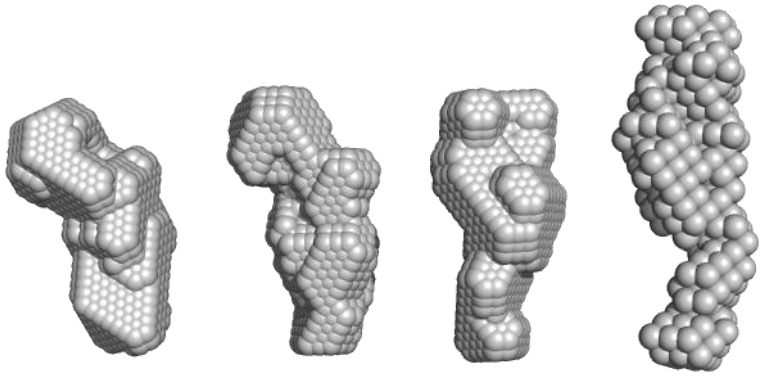


Figure VIII.17. Solution envelopes of 7SK subdomains. From left to right, the envelopes of HP1u, HP1, HP4 and HP3 are shown. *Ab initio* reconstructions were calculated by DAMMIN (Svergun, 1999), and images were created in PyMOL.

## a. Modelling hairpins: the problem of orientation

Given the low resolution of the envelopes, it was not obvious to assess the orientation of the molecules, and to locate the apical loop. This is illustrated by the experiment shown in Figure VIII.18. Three-dimensional model were built by MC-Sym (Macromolecular Conformation Symbolic programming; Parisien et al. 2008), which explores RNA structure database and assembles substructures taking into account base pairing and base stacking. MC-Sym calculates a huge number of all-atoms models for each secondary structure proposition. These models can be refined according to geometrical constraints, evaluated (various criteria are available) or clustered. We also generated all-atoms models using RFR (Michal Gadja, manuscript in preparation) that uses SAXS data to constrain the three-dimensional solution during calculation.

The fit between the theoretical scattering curves of the models and the SAXS experimental data was then evaluated by CRYSOL (Svergun et al. 1995). In Table VIII.4, the different models are representative from its corresponding cluster for HP3 (5000 models were clustered in 5 groups, showing rmsd up to ~8Å); the $\chi^2$ values typically obtained for the fit are shown.

| Table VIII.4. $\chi^2$ of the fit of theoretical and experimental curve | | |
|---|---|---|
| **HP3** | **SAXS data 1** | **SAXS data 2** |
| **MC-Sym model 1** | 1.027 | 1.086 |
| **MC-Sym model 2** | 1.078 | 1.129 |
| **MC-Sym model 3** | 1.028 | 1.083 |
| **MC-Sym model 4** | 1.063 | 1.118 |

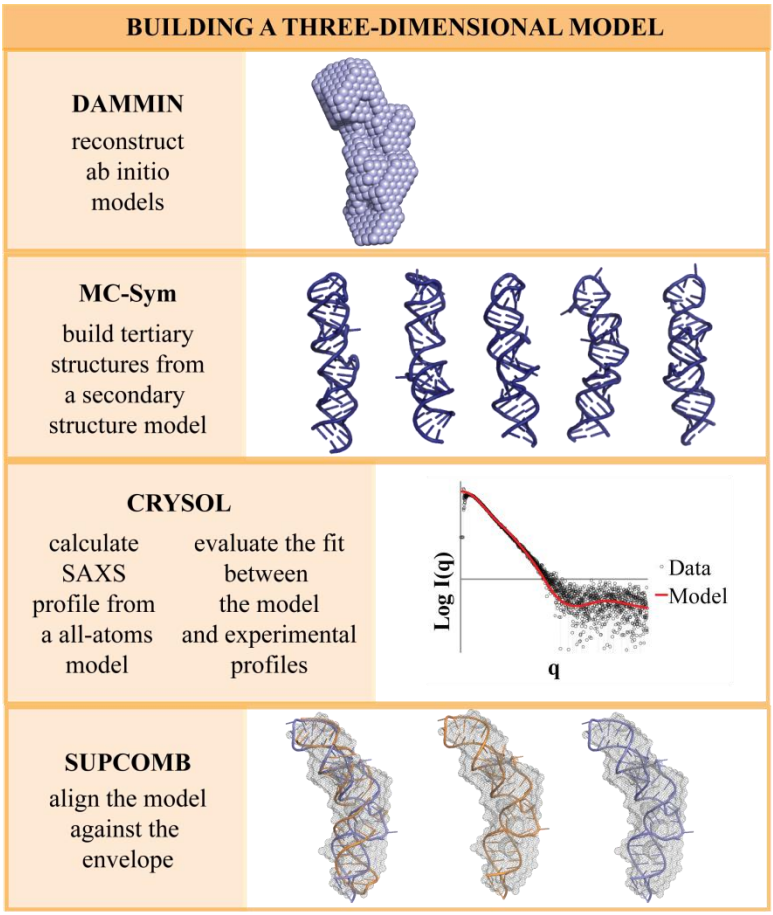



Figure VIII.18. Strategy used to construct 7SK subdomain models. The programs that we used during three-dimensional modelling of the 7SK subdomains are indicated. The references can be found in the text.

The models with a good fit were then aligned with the corresponding envelope using SUPCOMB (Kozin et al. 2001). Surprisingly, despite the good fit of the models, SUPCOMB failed to locate the apical loop and alignments with both orientations were obtained, as can be noted in Figure VIII.19. The envelopes of the hairpins were probably too symmetric to locate the apical loop.

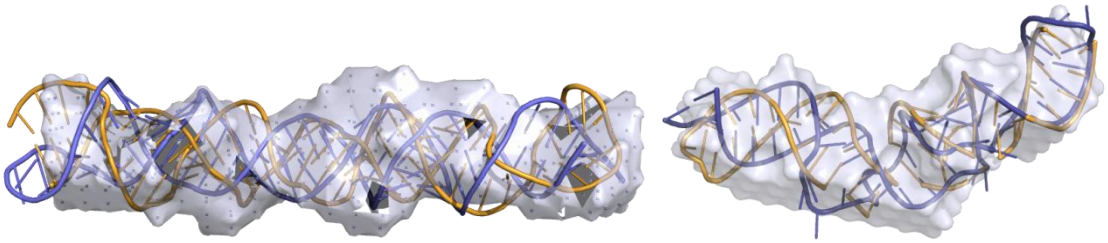


Figure VIII.19. SUPCOMB fitting. Two models were fitted in the envelope constructed from SAXS data (surface representation) of HP3 (left) and HP1 (right). Note that they are in opposing orientations.

Given these observations, we measured more asymmetrical constructions of the subdomains to attempt to orient the envelopes. For example, to orient HP1, we measured HP1u with a smaller apical loop, or KEI and KSI with a tRNA fused to its basal stem. The solution envelope of HP1u could be nicely superimposed on the HP1 one (Figure VIII.20). The overall HP1u envelope seemed slightly smaller than that of HP1. However, the size difference of the apical loop was insufficient to assign it with full confidence at any end.

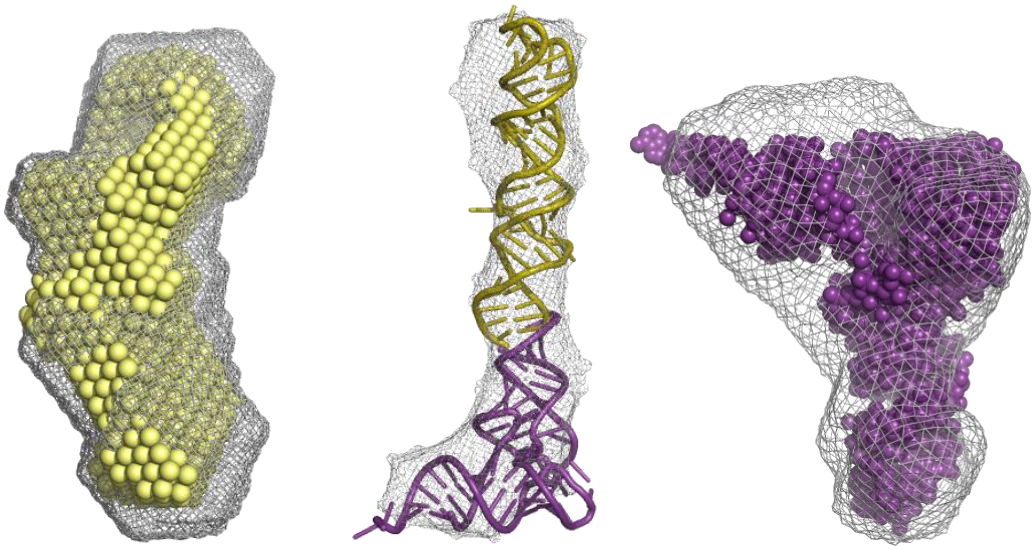


Figure VIII.20. Envelopes of HP1 variants. Left, HP1u envelope (yellow) was manually fitted in the HP1 one (mesh representation). Middle, an all-atom model of KE1 was manually fitted in its envelope (mesh representation). Right, the crystal structure of a tRNA (purple; PDB 3LOU) was manually fitted in the solution envelope of $tRNA^{Thr}$ (mesh representation).

To get an insight and a reference of the method, we also performed SAXS measurements and calculated the envelope of $tRNA^{Thr}$. The structure of the tRNA could be satisfactorily fitted. Then we analysed KE1 and KS1. The irregular envelope calculated for KE1 places considerable limitations on how the envelopes of HP1 and tRNA (as well as the crystal structure of the tRNA) can be fitted, particularly in view of the overall length and thickness (similar observations for KS1). Finally, we could propose an all-atoms model of KE1 fitted in the SAXS envelope (Figure VIII.20). Unfortunately, only a small number of SAXS dataset were collected with KE1. More should be measured to be able to attain the same precision than for HP1 and be able to visualize and locate the characteristic bent and bulges of HP1.

### b. 7SK *ab initio* reconstructions

*Ab initio* model reconstructions of 7SK failed to yield a unique interpretable scattering envelope (Figure VIII.21). This is not surprising given the inconsistencies found during data analysis. The intrinsic flexibility of 7SK may also account for this outcome (Kazantsev et al. 2011). Since all conformations in solution contribute to the overall scattering, each data set may result in a different reconstruction. However, some features seemed to be constant in 7SK envelopes. 7SK showed a twisted structure with most of the time two bulges at one of its ends (sometimes in both ends). Unfortunately, given the significant differences between 7SK envelopes, it was not possible to distinguish main differences between 7SK and constructions with a deleted subdomain.

To gain an insight into 7SK organization, we nevertheless performed several attempts to fit the envelopes of the isolated subdomain into the one of the 7SK calculated from one of the best SAXS data obtained. No satisfactory fit resulted (Figure VIII.22). As observed for KE1 and KS1, the length and thickness of 7SK envelope seemed generally small to accommodate all the isolated subdomains. Interestingly, no limb was long enough to fit HP3 or HP1L (the 5'end hairpin as presented in the Wassarman model), only HP1 or HP4. Moreover, the central body of 7SK appeared too thin to pack two helices parallely. A possible explanation may be that the conformations of the isolated subdomains are not the same as in the 7SK context.

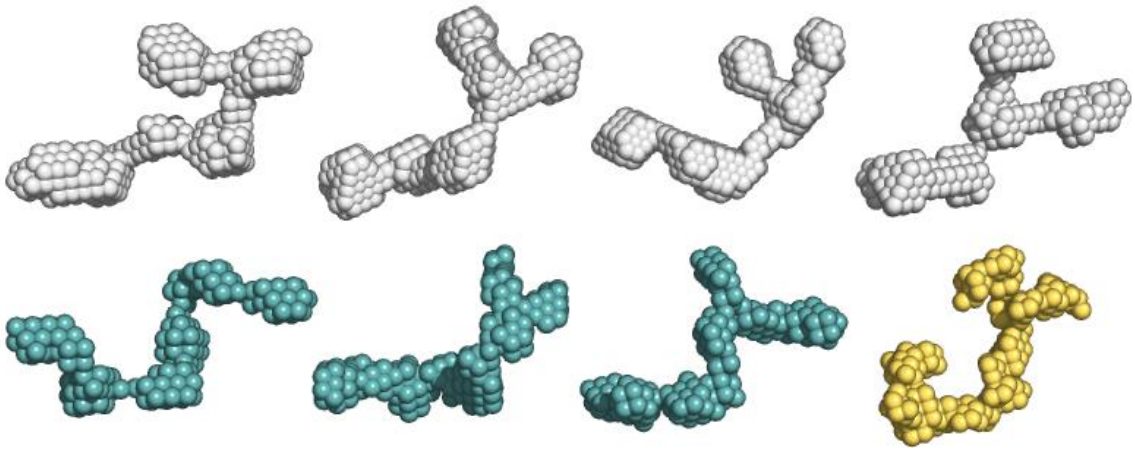


Figure VIII.21. Solution envelopes of 7SK. *Ab initio* reconstructions from different SAXS data calculated by DAMMIN (Svergun, 1999). The images were created in PyMOL. All data were measured in the same buffer. Models from different 7SK batches are coloured different.

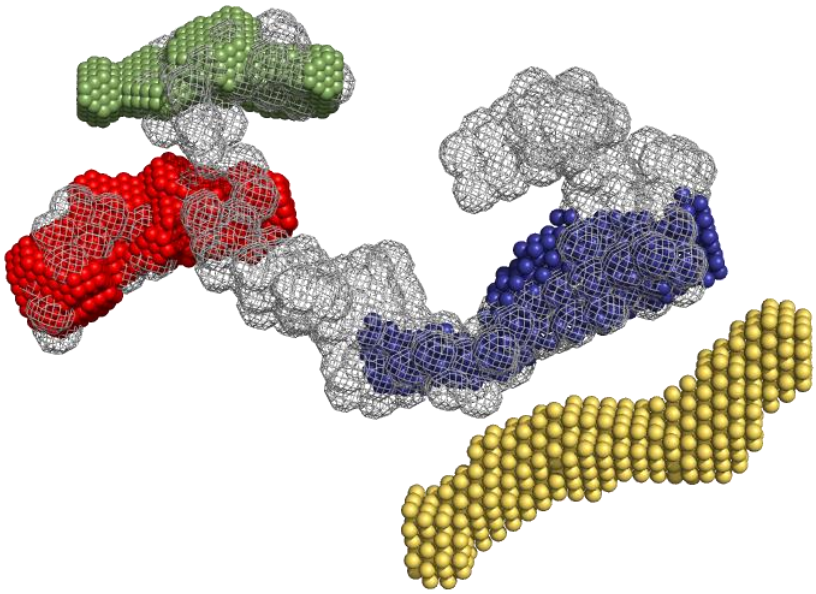


Figure VIII.22. Solution envelopes of 7SK and subdomains. HP1 (red), HP3 (blue) and HP4 (green) solution envelopes were placed into the solution envelope of 7SK (mesh representation). HP1L (orange) is also shown (apart). This representation highlights that the size of HP1 or HP4 could tally to the observed limbs (leaving aside the topological issue). But this does not hold for HP3, which requires a strong bending or shortening. HP1 L is clearly too long.

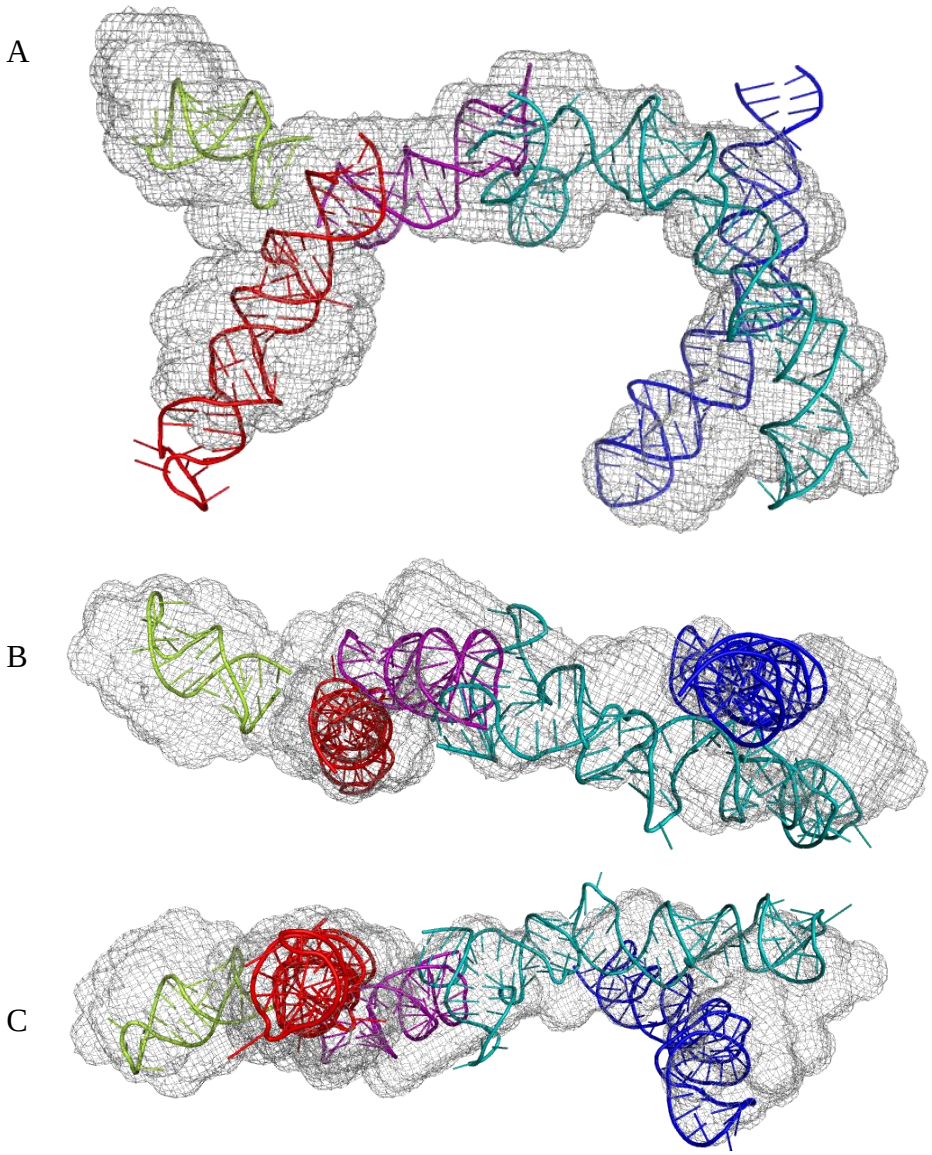



Figure VIII.24. Solution envelope of 7SK and all-atoms models of its subdomains. Side (A), top (B) and bottom (C) views of 7SK are shown. HP1 (red), HP2 (cyan) HP3 (blue), HP4 (green) and M1 (purple) all-atoms models were manually placed in an averaged solution envelope of 7SK (mesh representation).

To further discuss the strategy of "divide and conquer", we turned to L3L4, a construction designed according to Eilebrecht model. Good data was obtained for this construction, suggesting that L3L4 should be structured, and *ab initio* model was constructed

(Figure VIII.24). HP3 and HP4 can be accommodated in the envelope, even if they seemed slightly long.

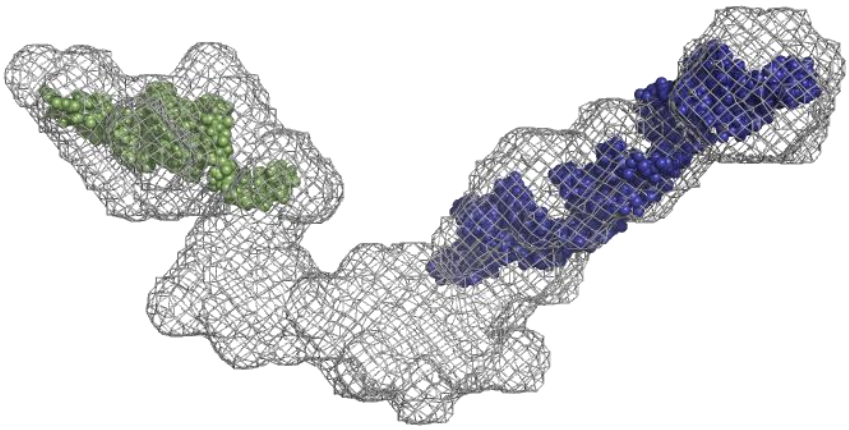


Figure X.24. Solution envelopes of L3L4. Left, HP4 (green) and HP3 (blue) all-atom models were placed in the solution envelope of L3L4 (mesh representation).

Finally, in order to illustrate the aim of the strategy used during this work, a 7SK SAXS envelope containing the all-atom models for all the subdomains generated by MC-Sym is shown in Figure VIII.23.

## 3. Discussion and Conclusions

During our SAXS study of 7SK, we grasped that different considerations must be taken in account for a successful insight into the three-dimensional shape of RNAs:

- The sample must be carefully prepared. This includes the establishment of a proper folding protocol and assay to asses sample homogeneity. Indeed, it was showed in a recent study that co-transcriptional folding (by adjusting the salt conditions during in vitro transcription) coupled with a non-denaturing protocol (using chromatography) for RNA purification yielded more compact and homogeneous molecule of an RNA (RNAse P) and interpretable data (Kazantsev et al. 2011). Our first biochemical experiments suggested that a more homogeneous (less degraded) 7SK is obtained when using denaturing methods, followed by a thermal treatment at low (2 mM) Mg concentration and a subsequent incubation at 6 mM $MgCl_2$. But, of course, it would be interesting to test the non-denaturing purification protocol for SAXS measurements of 7SK. It might be that criterion for good biochemical investigation such as SHAPE, which requires non-degraded RNA, is not the same than for SAXS in which a conformational homogeneity and strict avoidance of stable misfolded molecules is required. An alternative to improve homogeneity of samples (purified by denaturing gels) has been recently proposed (Rambo et al. 2010). The authors showed that HR (High Resolution) SEC is a suitable technique to achieve the required homogeneity for SAXS studies of riboswitch. Our experience showed that Mono Q chromatography can be useful for some RNAs, in our case for hairpins but these were unfortunately difficult to apply to large RNAs constructions. However, SEC proved to be applicable to RNAs of widely differing sizes, so it would be interesting to test this technique for 7SK and large constructions. Indeed, Figure VIII.25 shows that 7SK conformational state can be monitored with SEC. Besides, MALS is an important tool not only to monitor the homogeneity of the sample, but also to gain insights about the oligomerization state of the RNA in solution, which is essential for a correct SAXS data interpretation.
- The secondary structure of the RNA is an essential information for SAXS data interpretation. In our case it was also essential for a pertinent design of the different 7SK variants constructions. Ideally, the secondary structure of the different RNA constructions should be monitored, by probing experiments for instance. We controlled the secondary

structure of several constructions (HP1, HP1L, 7SKΔHP4, IL2, and IL3; the last one appeared to be more fragile than 7SK) using SHAPE, but this work revealed to be hard and tedious. This important contribution to SAXS approach should gain efficiency in the near future by the introduction of new techniques avoiding the use of sequencing gels and allowing more accurate quantification (Wilkinson et al. 2008). SHAPE data and functional information were useful for identify hinges regions in 7SK to design the constructions. Reciprocally, some constructions were designed to test the secondary structure itself, such as M1 and 7SKΔ9 that were conceived to test the Marz model of 7SK. Importantly, a good model of the secondary structure of the RNA is essential to interpret SAXS data, to evaluate and guide the modelling of the solution envelope of the RNA and, eventually, to generate all-atoms models. Our secondary structure model, presented in Chapter VII, was the result of a thorough analysis of SHAPE and sequences data, using programs only recently available. New constructions should now be designed accordingly.

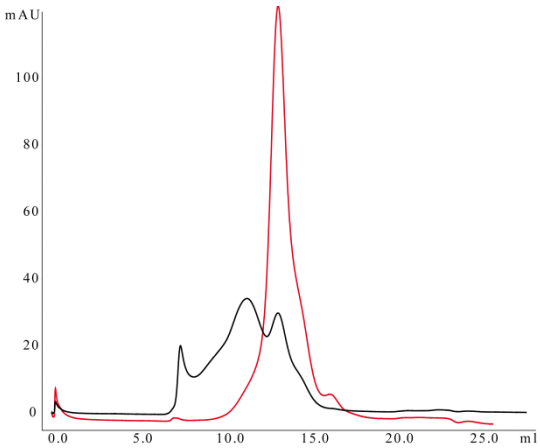



Figure VIII.25. 7SK SEC profiles. The 7SK SEC (Superose 6) profile before (dark line) and after (red line) thermal treatment are shown.

- Symmetrical molecules lead to difficulties for the orientation of the molecule during modelling. It should be noted that two drawbacks of the SAXS method that complicate modelling are: (1) the resolution is too low to locate the apical loop (*i.e.* orient each end of a rod); and (2) since the measured data are orientationally averaged, it is not possible to distinguish enantiomorphs (both enantiomers of the model should be used for fitting models). However, it should be noted that highly complex envelopes could be also difficult to correctly model (Volkov et al. 2003).

- SAXS provides structural information at low resolution. In line with this, only information about the size and shape of molecules in solution should be expected. Therefore, further assumptions require supplementary information from other methods.
- The available programs for analysis of SAXS data have been mainly conceived for proteins. *Ab initio* structure programs have been often designed for globular proteins and may be not appropriate for RNA, which usually have more extended, branch-like shapes. Also, these programs often do not correctly take into account the contribution of the hydration layer specific for RNAs, thus affecting the three-dimensional calculated envelope. In a recent investigation, a combination of coarse-grain normal mode analysis (NMA) and the Ensemble Optimization Method [EOM; (Bernadó et al. 2007)] was used to generate the solution envelope of the RNAse P (Kazantsev et al. 2011). This would be an interesting alternative for the treatment of our SAXS data. Likewise, a coarse-grained approach has been used to develop a method recently published [Fast-SAXS-RNA(S. Yang et al. 2010)] for calculating the SAXS profile from nucleic acids structures, which should be also suitable for the treatment of our SAXS data. Unfortunately, this program is not available for the moment. Apparently, there is nowadays a strong interest in SAXS combined with RNA modelling, which should open new possibilities in the near future.

Despite that our attempts to reconstruct a solution envelope of the complete 7SK at low resolution in order to build a three-dimensional structure model failed, some interesting information may be suggested by our SAXS measurements. 7SK may be described like a semi-compact, flexible molecule, with at least three, functional and structural, autonomous subdomains: HP1, HP3 and HP4. These subdomains consisted in hairpins, which in the 7SK context may involve tertiary interactions at their basal stem that constrain in some extent their conformation. Even if 7SK is semi-compact, our first attempts of *ab initio* reconstructions may indicate that 7SK does not consist of several subdomains tethered in a single strand as described previously by (Wassarman et al. 1991). Moreover, the envelope obtained for HP1L (the 5'end hairpin proposed in the Wassarman model) seemed difficult to fit in the 7SK envelope. Hence, 7SK may be circularized as proposed by (Marz et al. 2009). However, 7SK constructions where M1 was mutated or suppressed did not show a more extended conformation since no clear increase of $D_{max}$ was observed. This may indicate that the interactions determining a semi-compact conformation to 7SK are somewhere else, or that M1

stem is longer than expected (like in our model, for example). But further analyses are still required to finalize these preliminary results.

# CHAPTER IX:
# CRYSTALLIZATION TRIALS OF A FUNCTIONAL SUBDOMAIN OF 7SK

## 1. X-RAY CRYSTALLOGRAPHY

X-ray crystallography requires the generation of crystals. Crystals are ordered arrays of atoms (lattices) where the smallest repeating unit, called the unit cell, is repeated by translations in three dimensions. For three-dimensional crystals, the shape and size of the unit cell is defined by the length of three axes (a, b, and c) and angles between these axes ($\alpha$, $\beta$, and $\gamma$). The asymmetric unit is the smallest portion of structural information required to reconstruct the entire lattice through symmetry operations. Indeed, the unit cell is built by rotations and translations of the asymmetric unit, in special combinations called space groups.

When electromagnetic waves, X-ray strikes a crystal, the waves scattered by each electron in the crystal lattice interfere with each other either constructively or destructively, producing a diffraction pattern. Hence, according to Bragg's Law, when a crystal composed by parallel planes (hkl, where h,k and l, are the Miller indices) separated for a distance d is exposed to a beam of X-ray of wavelength $\lambda$ at an angle $\theta$, the maxima of the reflected rays occurs when $\sin \theta = n \lambda/2d$, where n is an integer.

The intensities of the diffracted X-rays are dictated by the atomic arrangements in the unit cell. Each diffraction spot corresponds to a point in the reciprocal lattice and represents a wave with an amplitude and a relative phase, which is described by the structure factor, F(h,k,l), for the lattice planes (h,k,l). Unfortunately, the data collection only allows measurements of the intensities, I(h,k,l), which are the square of the amplitude of the structure factor F(h,k,l), but not the relative phase information necessary to calculate electronic distribution in the unit cell.

The structure factors for each point on the reciprocal lattice correspond to the Fourier transform of the electron density distribution within the unit cell of the crystal. Reciprocally, the inverse Fourier transform of the structure factor is the electron density. Thus, if we can

can obtain phase estimates, their combination with the intensities recorded for each diffraction spot will give us access to the electron density of the asymmetric unit. This is the fundamental step in solving the structure. The next step is to construct a 3D model fitting the electron density, with satisfying geometry.

One of the experimental techniques to determine the phase of each reflection relies upon introducing atoms which modify the diffraction, but not the molecule nor the crystal. This can be done with heavy atoms or with anomalous scattering.

If a heavy atom can be attached to a unique location or locations on the macromolecule of interest without changing its structure or symmetry and without destroying the ability of crystal to diffract, the Patterson function can be used to solve the position of the heavy atom. This function is a Fourier transform of the set of squared but not phased amplitudes (h k l $F^2$). It does not produce an electron density map of the contents of the unit cell rather a density map of the vectors between scattering objects in the cell. Because the densities in the Patterson map go as squares of the numbers of electrons of the scattering atoms, the Patterson map of crystals that contain heavy atoms is dominated by the vectors between heavy atoms, and allow interpretion of the position of the heavy atom(s). This provides an initial estimate of phases.

The electronic differences needed to create difference maps from which the heavy atom positions and initial phases can be solved, can originate from any kind of differences in scattering behavior. The atomic scattering factor has three components: a normal scattering term that is dependent on the Bragg angle and two terms that are not dependent on the scattering angle but on wavelength. These latter two terms represent the anomalous scattering that occurs when X-ray energies are near electronic excitations, so these introduced atoms will absorb X-rays at a particular wavelength. These leads to the breakdown in Friedel's law.. Friedel's law tells that F(h,k,l) and F(-h,-k,-l) have the same magnitude and phases $\varphi$(h,k,l,) = $\varphi$(-h,-k,-l). However, anomalous behavior introduces a contribution such that the reflections F(h,k,l) and F(-h,-k,-l) have different intensities and their phases are no longer complementary, giving rise to anomalous differences that can be used to locate the anomalous scatterers. In contrast to the normal scattering factor $f^0$, the anomalous dispersion corrections F' and F'' depend only on the wavelength $\lambda$ of the X-rays used for the diffraction experiments and do not diminish with the diffraction angle. The interpretation of the Patterson difference map reveals the location of the anomalous scatterer in the unit cell. This allows both amplitude and phase of the atom to be determined.

A different approach to solve structures is by molecular replacement (MR). This method takes advantage of the facts that the basic backbone architecture of related proteins is similar. MR enables the solution of the crystallographic phase problem by providing initial estimates of the phases of the new structure from a previously known structure. All possible orientations and positions of the model in the unknown crystal are tried to find where the predicted diffraction best matches the observed diffraction. The phases for the reflections of the unknown crystal are then "borrowed" from the phases calculated from the model as if it were the model that had crystallized in the unknown crystal and an initial map is calculated with these borrowed phases and the experimental observed amplitudes. The crystallographer therefore relies on the measured amplitudes to supply the information for rebuilding of the model so that it more closely resembles the target structure. The MR method raises a number of issues:

(1) How to choose a suitable model and how to improve models.
(2) How to score each orientation and position so as to find when the models best fits the target structure: different target functions will have different degrees of discrimination between the solution and noise.
(3) How to search for solutions: strategies for exploring rotations and translations.

Each molecule needs six parameters to define orientation and position: three rotation angles and three translations. An exhaustive search in six dimensions can take a very long time, so two searches can be separated and the translational search only carried out for the best points found in the rotation search.

## 2. RNA AND CRYSTALLIZATION

The aim of this thesis was the structural characterization of the 7SK/HEXIM1 complex. A mechanistic understanding of RNA and protein function requires detailed knowledge of the three-dimensional configuration of the atoms involved in their function. One of the structural approaches that we used was X-ray crystallography because it allows determining the structure of biological macromolecules at atomic resolution.

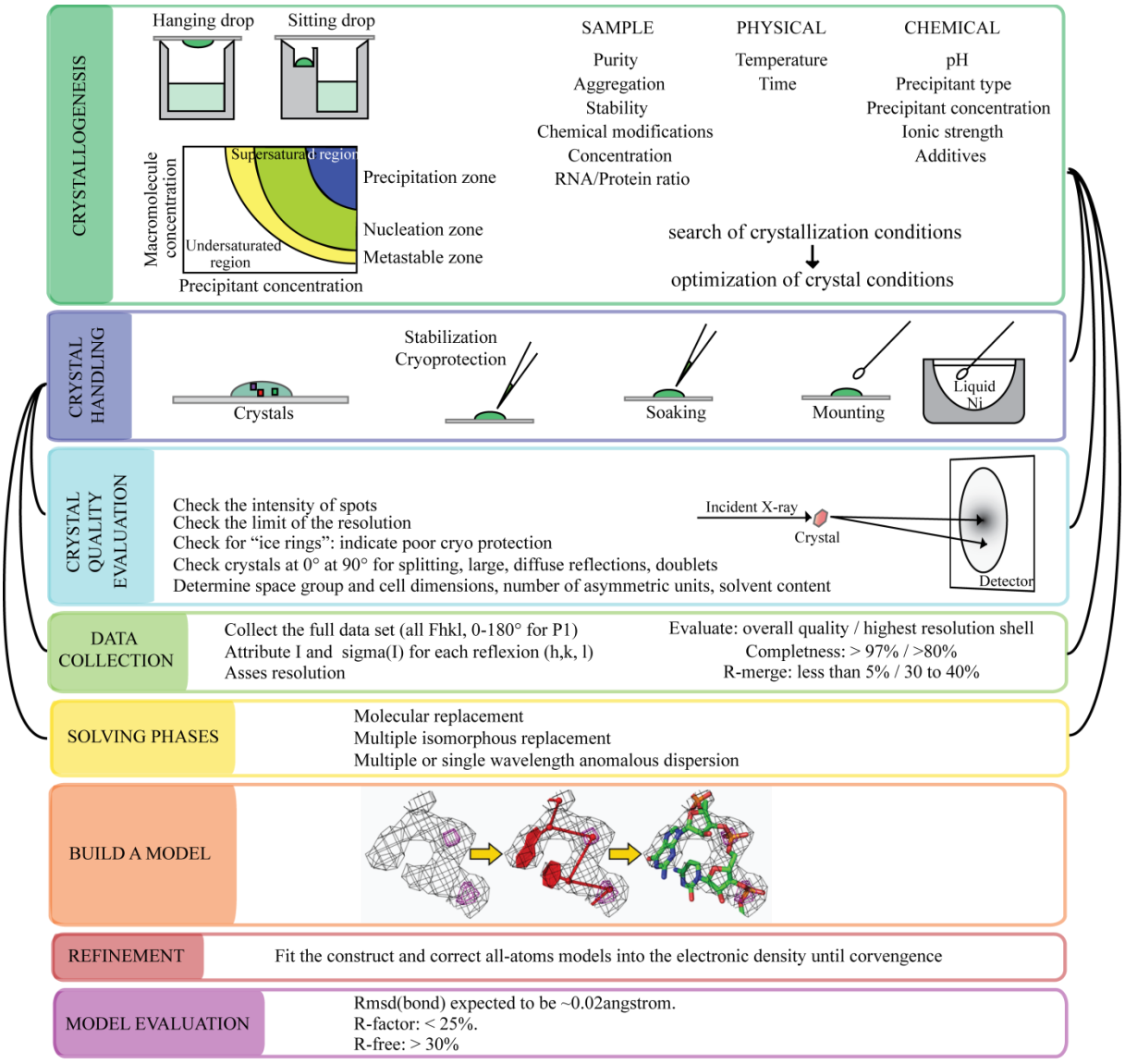



Figure IX.1. General strategy for X-ray crystallization. The basic steps for solving a crystal structure of a macromolecule are presented. The lines connect the processes that are often iterated.

However, even if X-ray crystallography has been largely developed for proteins, it remains a challenging task for RNAs. Indeed, before 2000, only three different types of biological RNA crystal structures were available: tRNAs (the first RNA crystal structures available), hammerhead ribozymes, and the P4-P6 fragment of the group I ribozyme. One of the reasons may be that obtaining well-ordered crystals seems more difficult for RNAs than for proteins, and frequently when crystals are obtained, they diffract X-rays to only low

resolution (Ke et al. 2004). This may be ascribed to different reasons. Dynamics of nucleic acids is very different compared to proteins, RNA being especially prone to kinking. A non-negligible effect of "breathing" is also expected from the ribose, as it is subjected to changes of "pucker". Crystal packing of helical objects is very peculiar. It has been observed that helices packing upon each other are a common feature of RNA crystals, which may lead to rare contacts (with respect to the surface of the molecules) and loose intermolecular interactions. In addition, RNAs are very sensitive to degradation by RNAses, but also to alkaline hydrolysis, restricting the useful pH range. Hydrolysis can also be catalyzed by metal ions or induced by magnesium ions. The solvatation of RNAs is also very different than for proteins.

Throughout my thesis project many crystallization trials were carried out. Several protein and RNA constructions were tested independently or in complex. A flowchart of the experimental strategy for crystallography studies is presented in the Figure IX.1.

## 3. HP1U CRYSTALLIZATION

### 3.1. Crystallogenesis and data collection

Several commercial crystallization screens are available in the Structural Biology and Genomics platform of our laboratory. For RNA/protein complexes and isolated protein constructions, we typically tested Index Classics, JCSG+, ProComplex, screens from Hampton Research, Nucleix (from Qiagen), and Wizard I and II (from Emerald Biosciences). We systematically tried two temperatures, 4°C and 17°C, but unfortunately all our attempts failed.

#### a. First crystals of HP1u

For RNAs, after some unsuccessful attempts with 7SK, we focused on HP1 because it has been recognized as the determinant subdomain of 7SK for HEXIM1 interaction (Bélanger et al. 2009; and our own investigations), and more recently for HIV-1 Tat interaction (Muniz et al. 2010). We used the crystallization screen for RNA of Sigma-Aldrich at 4°C and 17°C in

96-well sitting drop plates and tried HP1, HP1u, KE1, and HP1L. We obtained small, sea urchin-shape crystals only for our HP1u construction in two different conditions (named "hits" in Table IX.1) at both temperatures.

**Table X.1. HP1u crystallization conditions**

| | RNA | Precipitant | NaCl (mM) | $MgCl_2$ (mM) | Buffer | Additives | Remark |
|---|---|---|---|---|---|---|---|
| Hit (Sigma screen) | 5' Tri-P | PEG1000 25% or 30% | 50 or 100 | 50 or 100 | Tris pH 7.5 | | November, 2009 |
| 1st generation | 5' Tri-P | PEG1000 25% or 30% | 50 or 100 | 50 or 100 | Tris pH 7.5 | | Spherulites then urchin after > 2 months |
| 2nd generation | 5' Mono-P | PEG1000 30% | 50 or 75 | 100 | Tris pH 7.5 | DMSO | Spherulites then cubes |
| 3rd generation | 5' OH | PEG1000 25% (fresh) | 50 | 75 | Tris pH 7.5 | $BaCl_2$, $CaCl_2$, $CdCl_2$, $ZnCl_2$, $CoCl_2$, or $Co(NH_4)_6$ | |

We then created a 96-conditions optimization screen. In this screen with tested the buffer (Cacodylate or Tris), PEG type (1000 or 8000) and concentration (20 to 30%), NaCl concentration (20 to 400 mM) and $MgCl_2$ concentration (0 to 200 mM) at 4°C and 17°C. Only two conditions (see Table IX.1) at 4°C showed some small, sea urchin-shape crystals.

These conditions were further optimized with a 48-conditions "home-made" screen and larger (hanging) drops at 4°C. After three days, some spherulites appeared. These generally evolved into very thin sea urchin-shape crystals. However in one condition, and two months later the spherulites evolved into crystals of cubic shape. These were stabilized, mounted and the data was successfully collected at 3.2 Å resolution, by Alastair McEwen (IGBMC) using the home diffractometer. A second crystal obtained from the same drop was measured at the ESRF synchrotron (beamline BM30) gave a native dataset at 3.1 Å resolution (Table IX.2)

## b. Problems of reproducibility of HP1u crystals

In order to improve the resolution, but essentially to get enough crystals to search for heavy atoms derivatives for phasing, we then attempted to reproduce these crystals. This proved to be very difficult, a fact linked to the very long crystallization time observed. First, because each hypothesis tested, and each optimization step tried, needed a long time to give an answer. Second, we were worried about the state of the RNA.

We suspected that some degradation could be occurring in our RNA, leading to crystallize only fragments of HP1u. However, controls by denaturing gel electrophoresis did not show significant degradation of HP1u in the drops, even after several months (Figure X.2). This indicated that our solutions and cover-slips were safe from RNAse, but was not a definitive proof of the crystal contents. A direct check of the state of HP1u in the crystals could unfortunately not be performed, because the crystals were too small and too rare.

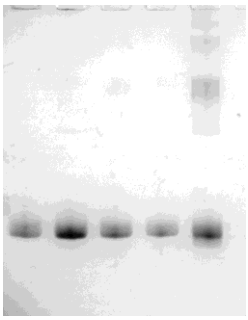


Figure IX.2. HP1u degradation control. From left to right, reference, nine, five, four and two months incubated HP1u.

The long crystallization time could however lead to smaller modifications, such as dephosphorylation at the 5'-triphosphate end. Our initial stocks of RNA transcribed with T7 Polymerase and a mixture of NTPs, produce RNAs with 5'-triphosphate, which is not very stable. Indeed, a mass spectrometry control of any such RNA gives often 3 masses corresponding to the presence of 3, 2 or 1 phosphates. The addition of GMP (in excess) was then tried for in vitro transcription, producing similar yield of product. In the drops set-up with that modified RNA, were observed first spherulites, then crystals three weeks later, in two conditions. These crystals were used for heavy atoms soak (see below).

Another attempt to modify the RNA preparation to speed up crystallization was to try to remove all phosphates by a phosphatase. We tried Calf Intestine Phosphatase (CIP), at low concentration but long incubation time. Crystals were obtained with that stock (and additives, see below), but they were of the urchin shape and their diffraction (after soaking with heavy atoms, see below) was of too bad quality to collect data.

In parallel with investigation of the RNA preparation, we performed classical optimization of crystallization by slight variations around the conditions, such as changing RNA, salt and PEG concentration, and Na/Mg ratio. Temperature was fixed at 4°C, were most of the crystals appeared. We tested also some additives. These were polyamines (spermine),

and cobalt hexamine, known to stabilize RNA, or chemicals known to modify the stability of RNAs (DMSO, alcools, glycerol). Finally, seeding was tried, but lead to obtain only very small crystals (these could however be used as seeds in future attempts) but most seeding tests failed.

Over all the trials, the reproducibility problem seemed to come, at least in some extent, from the PEG 1000 solution. The stock PEG 1000 in the laboratory was provided as powder, and the apparition of spherulites (which were indicative of crystallization) seemed to depend on the "freshness" of the solution (but not each new prepared solution yielded spherulites). We then used a brand new commercial PEG 1000 solution (Sigma-Adrich). Surprisingly, the behavior of the crystallization drop was completely different; usually precipitation was observed immediately at the preparation step, but with the new PEG 1000 no precipitation was observed. Moreover, spherulites were observed only when the fresh PEG was mixed with the old PEG solution (we tested several ratios).

All these observations led us to hypothesize that a contamination on the PEG powder could be implicated in crystallization, probably a metal. Hence, we prepared crystallization reservoirs with traces (25 µM) of different metals: Ca, Ba, Cd, Zn and Co and set-up drops with the HP1u stock after phosphatase treatment. About five weeks later, all conditions produced sea urchin-shape crystals bypassing the spherulite stage. Unfortunately, these crystals showed multiple lattices and data could not be collected.

Photos from some of the HP1u crystals obtained during this work are presented in Figure IX.3.

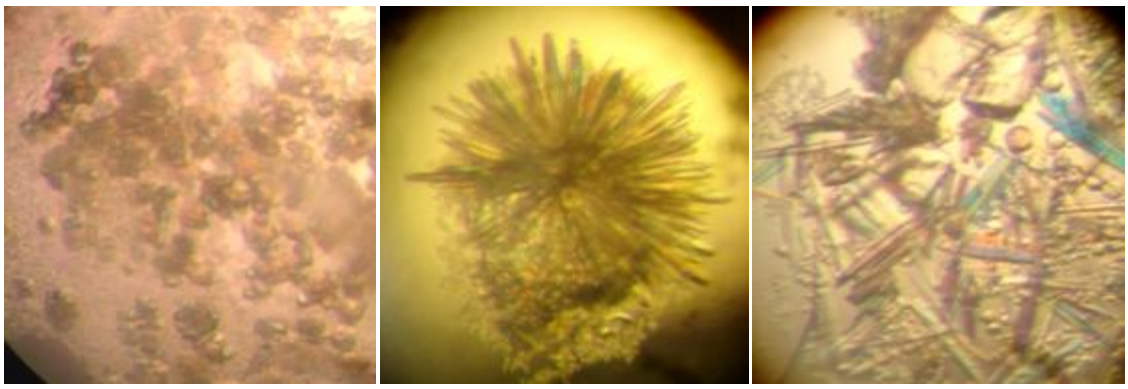


Figure IX.3. HP1u crystals. Left, crystals growing from spherulites. Middle, sea urchin-shape crystal. Rigth, the sea urchin-shape crystal was broken to mount individual crystal

c. Crystals handling and soaking

All RNA crystals are very fragile, and we chose, with the initial crystals to submit them to a "stabilization" treatment to try to improve the lattice order. Our first attempt was successful, and consisted in soaking for 20 min in a solution similar to the reservoir condition, but with an increased PEG concentration (0.5 µL of solution at 40% PEG was added to the drop giving a rough estimate of 30% PEG). The same 40% PEG solution was used for soaking with heavy atoms.

**Table IX.2. Data collected from HP1u crystals**

| | Initial Native | Native (Au) | Os |
|---|---|---|---|
| Beamline | ESRF BM30 | Soleil Proxima1 | Soleil Proxima1 |
| Wavelength (Å/keV) | 0.9797 / 12.6549 | 1.0401 / 11.9200 | 1.1396 / 10.8800 |
| Space Group | P21 | P21 | P21 |
| Unit Cell Parameters (Å) | A = 45.87, b = 47.96, c = 68.80, β = 98.994° | A = 47.44, b = 47.96, c = 69.39, β = 105.54° | A = 47.20, b = 47.84, c = 69.15, β = 105.12° |
| Resolution (Å) | 68.8 – 2.70 (2.75 – 2.70) | 20 – 2.74 (2.79 – 2.74) | 25 – 2.75 (2.80 – 2.75) |
| *Rmerge* (%) § | 0.073 (0.304) | 0.097 (0.520) | 0.106 (0.513) |
| $I/\sigma I$ | 20 (4.6) | 28.7 (3.82) | 13.9 (1.73) |
| Completeness (%) | 99.7 (100.0) | 98.7 (96.3) | 94.7 (97.9) |
| Multiplicity | 3.7 (3.7) | 6.1 (5.7) | 2.1 (2.1) |
| No. of unique reflections | 8322 | 8177 | 7945 |
| Wilson *B* (Å2) | 81.78 | 73.47 | 64.61 |
| NCS: Pseudo-translation | 0.167, 0.500, 0.963 | 0.167, 0.500, 0.981 | 0.167, 0.500, 0.981 |

§ $R\text{merge} = \Sigma h \Sigma i |\langle I h \rangle - I h,i| / \Sigma h \Sigma i I h,I$ where $I h,I$ is the i-th observed intensity of a measured reflection of Miller index h and $|\langle I h \rangle$ is the average intensity of this unique reflection

A first attempt to derive with gadolinium was done with the initial drop, without success, although the other crystals of the same drop were diffracting. We tried to maximize our panel of heavy atoms assays by transferring crystals (thin plates of the same generation) into crystallization solutions (at 40% PEG) containing various heavy atoms (Au, Pt, Zn or Mn

salts). The transferred crystals showed no diffraction, and we suspected then that they were too fragile to be handled in that way.

With the next 2 drops (cubic shapes), we added a solution containing either Au or osmium hexamine (kindly provided by Marat Yusupov group, IGBMC). The crystals were measured at SOLEIL synchrotron (beamline Proxima 1). The Au-soaked crystals were native, but the Os-soaked crystals showed some fluorescence signal of Os-binding. These crystals diffracted slightly better than the initial crystals (Table IX.2). Unfortunately (see below), the anomalous signal for the Os-soaked crystals was not strong enough to carry on with phasing. Interestingly, all crystals tested just after soaking (several hours for Os, 2 days for Au) diffracted correctly (but with a strong anisotropy).

At the next generation of crystals, obtained with trace amounts of metals, we soaked in either the same or other metal, for various time and concentration, but diffraction was of bad quality, and no data measured.

## 3.2. Data analyses and phase problem

For solving the phases by molecular replacement, we first tried using an HP1u model constructed by Frabrice Jossinet, IBMC) with his program ASSEMBL. We realized soon that the model is not straightforward to build. The 10 base-pairs probably helical region and the tetraloop can be built easily with ASSEMBL by analogy with structures in the PDB. However, we did not find obvious models for the internal loops. We tried to guide modeling by the SAXS envelope, but the symmetry of the envelope was a blockade, as discussed in Chapter VIII. We then tried a more systematic search by using MC-sym (Parisien et al. 2008), producing ~5000 models, that were filtered by their fit to the SAXS envelope. The ~500 best models were clustered in 5 groups according to their geometrical differences. At least one model of each group was tried for MR, using Phaser in the CCP4 package, but without success. In fact, most of the time, the program proposes a position for only one molecule, while there are 2 molecules in the asymmetric units.

We then tried the method described in (Robertson et al. 2008). Namely, the idea is to use fragments of model for the search in Phaser. When a fragment is placed, the residues which are not in density are eliminated, the others are fixed and another search launched for the missing parts. This iterative process was successful in the case of the ribozyme. In our

hand, the only piece of model that could be placed, from the beginning, and whatever the model we tried, was the helical part of the stem (about 8 base-pairs). This was unfortunately not enough to phase the rest of the molecule, and we decided to turn to phasing with anomalous signal.

### a. Phasing by anomalous signal

Phasing with anomalous signal, or heavy atom, requires to obtain data from a crystal were an anomalous scatterer (or heavy atom) has been bound in a fixed position. For protein, the production of such crystal can be obtained by crystallizing a protein containing selenium atoms (SeMet instead of Met). For RNA, signal for phasing can be obtained by introducing iodo-U or bromo-U at precise positions. This can be done by direct chemical synthesis of the RNA, which is quite expensive for large (>40 residues) RNAs, or by reconstituting the molecule, from a synthetic, labeled fragment, and a T7-product, which may poses some problems homogeneity of the mixture. In our case, we did not wish to try that, because we did not master the crystallization process.

We turned then to classical search for heavy/anomalous atoms. We faced great difficulties, due to the lack of crystals, problems with handling the crystals (which did not stand transfer in another drop), and could obtain only one drop (3 datasets) with an Os derivative, for which we had hopes of success. Unfortunately, all phasing programs tested failed to find the position of the Os atom(s). After discussion with crystallographers, the main explanation may be a too weak signal. This can be due to the short duration of soaking or dilute concentration of the Os hexamine (since high concentration of derivative can destroy crystals, we stayed on the safe side). It can also be due to weak (or diffuse) binding of the Os atom. Os hexamine have been shown to bind in the major groove of RNA helices, like cobalt hexamine. A very interesting paper describes the features required for a strong binding of $Co(NH4)_6$, opening the possibility to design such a site into the stem of HP1u (Keel et al. 2007). Like for MR, phasing with anomalous signal was strongly impaired by the presence of a non-crystallographic symmetry. This is further complicated by the fact that this symmetry is translational. Finally, anomalous signal is naturally weak, and as such, very sensitive to errors of measurements. This is directly linked to the quality of the crystals, which is very variable for our crystals showing anisotropic diffraction. The measurement quality is improved by

redundant measurements of the same Ihkl. We were limited, in our case, by the monoclinic space group, requiring data collection over 180° (an orthorhombic space group would require only 90°) for a full set of reflections, and the X-ray induced decay of crystals (dying after about 200°).

In summary, this crystallographic project we have has met several very annoying drawbacks: lack of reproducibility of crystals, unfavorable space group and symmetry, as well as fragility of crystals. We are however still hoping to increase our control on the crystallization process, by combining fresh PEG, mono-phosphate 5'-end and additives. This would open the way to crystals with chemically synthesized RNA bearing fixed anomalous scatterer, or with engineered Co hexamine binding site.

# GENERAL CONCLUSION

The eukaryotic transcription is a highly regulated process, and its alteration leads to the development of several diseases. Several ncRNAs, such as 7SK snRNA, have been recently highlighted as important regulators of transcription. 7SK snRNA, along with the HEXIM1 protein, inhibits P-TEFb, a transcription factor responsible for the productive elongation.

The purpose of the current study was the structural characterization of the 7SK snRNA and the identification of its structural determinants for its function and interaction with its protein partners, and in particular with HEXIM1.

Taken together, the results of the structural studies of 7SK by SHAPE and SAXS suggest that 7SK is a semi-compact, flexible and modular molecule. At least three structural, autonomous subdomains can be identified: HP1, HP3 and HP4. Our functional studies showed that HP1 (nucleotides 24 to 87) is also functionally autonomous.

One of the most significant finding to emerge from this study is the precise identification of the binding site of HEXIM1 in 7SK. NMR mapping showed that the ARM of HEXIM1 is able to specifically bind the conserved GAUC repeated motif stem in the apical region of HP1. The bulged Us encompassing this motif have an essential role in the recognition. Upon the binding, the GAUC motif stem opens and the base pair A39/G68 is formed. The Pro157, and particularly the Ser158, in the ARM are important for this effect. Using EMSA, we found that mutations of the GAUC motif, of the bulged Us or of the internal loop in the middle region of HP1, highly impair the binding to HEXIM1, whereas mutations in the basal stem region do not affect the interaction.

Our investigation by MS showed that the HEXIM1 binds preferentially HP1 as a dimer. Using a monomer HEXIM1 (lacking of the dimerization domain), we found that two monomers interact with HP1. These results support the existence of a second binding site in HP1, close to the GAUC motif. Further work needs to be done to establish the precise location of this second binding site.

Unfortunately, our numerous efforts to crystallize the 7SK/HEXIM1 complex using different variants of these molecules were unsuccessfully. Nevertheless, we obtained crystals of a HP1 with a modified apical loop. Although the reproducibility of the crystals and the determination of the structure have proved difficult, recent results have opened some insights for obtaining HP1u crystals, which would allow testing new strategies to solve the structure.

The present study confirms previous findings and contributes with additional information that enhances our understanding of the structural determinants for the interaction between 7SK snRNA and HEXIM1. This research should serve as a base for future structural studies of the 7SK snRNP which should reveal a mechanism of regulation of the kinase activity of P-TEFb.

# ANNEXES 1:
# MATERIAL AND METHODS

## 1. RNA PRODUCTION

### 1.1 Plasmids

A pBluescript plasmid containing the human 7SK sequence was obtained from Oliver Bensaude team. The 7SK sequence was then cloned in the laboratory into a pHDV vector, a kind gift from G. Conn (Walker et al. 2003) using standard protocols (Figure A.1). We labeled this construction pHDV_7SK.

*Escherichia coli* (DH5α) strain were transformed and selected on agar (ampicillin) plate. Overnight cultures of 1L LB were performed from one colony. Plasmid was recovered by maxipreparations, from usually ~350ml of overnight culture using NucleoBond® Xtra Maxi from Macherey-Nagel. Until 2 mg of plasmid were recovered by this method, however yield depended on construction.

### 1.2 Linearization

Run off transcription requires the plasmid to be linearized. The pHDV plasmid contains a 3' XbaI site for generating linear template (Figure A.1). Typically 1.5 mg of plasmid were linearized in 600µl reactions as below:

| | **Stock Conc** | **Final Conc** | **Volume** |
|---|---|---|---|
| pHDV-Insert | ≤ 3 mg/ml | ≤ 2.65 mg/ml | 530 µl |
| Tango buffer | 10 X | 1 X | 60 µl |
| XbaI | 10 U/µl | 100 U | 10 µl |

Reactions were incubated overnight at 37°C. A correct linearization of the plasmid was verified by agarose gel‡ (Figure A.2; ‡ **protocol or solution presented in Annexes 2**).

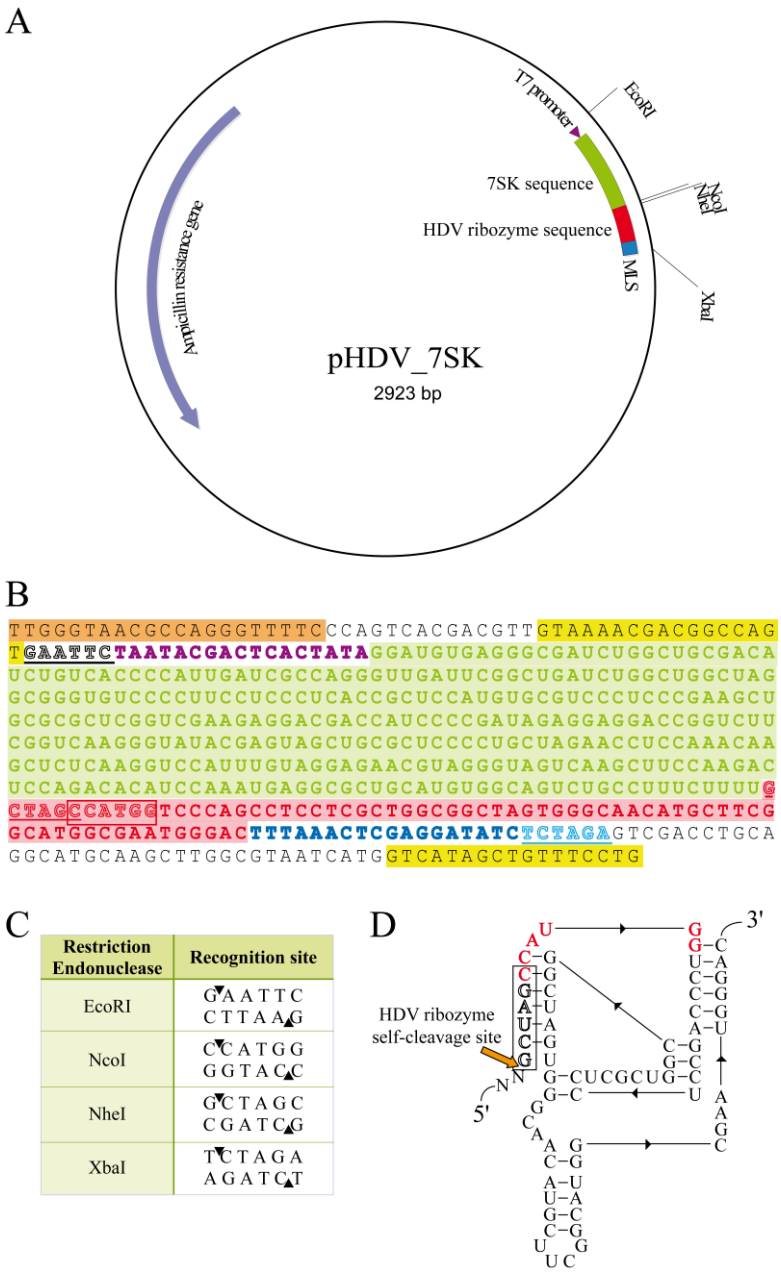



Figure A.1. pHDV vector. A) Circular map of the pHDV_7SK plasmid. B) Extract of the pHDV_7SK sequence (color code as in A); AFP287 primer (used for sequencing, orange), M13 forward and reverse primers hybridation sites (yellow), and EcoR I (underlined in black), Xba I (underlined in blue), NheI (underlined in red), and NcoI (red square) restriction sites are shown. C) Endonucleases recognition sites, cleavage sites are indicated by black arrows. D) HDV ribozyme model from pHDV vector [modified from (Walker et al. 2003)]. The self-cleavage site (arrow), and the restriction sites of NheI (black square) and NcoI (in red) are shown.

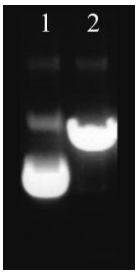



Figure A.2. Checking pHDV linearization. 200ng of pHDV_7SK plasmid (1) and digested pHDV_7SK (2) were loaded into a 1% agarose gel‡.

Phenol:chloroform extraction‡ and ethanol precipitation‡ were performed. Linear plasmid was suspended in water to achieve a concentration of 1mg/ml, suitable for carrying out the transcription reaction later. Typically more than 80% of linear plasmid was recovered.

## 1.3 Template generated by PCR

For some RNA constructs, particularly those used for functional tests and whenever possible, we used templates generated by PCR. Since all RNAs constructs correspond to 7SK substructures, pHDV_7SK was typically used as template DNA. The forward primers annealed the upstream region of T7 promoter, or contained a T7 promoter sequence upstream of the hybridization site (in this case, we added a G at the 5' end of sequence to ensure a proper in vitro T7 transcription). The reverse primers were designed according to the target, and could or not include the HDV ribozyme sequence.

For 1 ml of transcription mix, 100µl of PCR reaction were used. The PCR reaction was performed as follow

| | **Stock Conc** | **Final Conc** | **Volume** |
|---|---|---|---|
| Milli Q Water | | | 61 µl |
| pHDV-Insert | 25 ng/µl | 0.5 ng/µl | 4 µl |
| Buffer 5X | 5X | 1X | 20 µl |
| dNTP | 5 mM | 0.5 mM | 10 µl |
| T7 Primer | 100 µM | 2 µM | 2 µl |
| RT* Primer | 100 µM | 2 µM | 2 µl |
| Phusion | 2U/µl | 0.02 U/µl | 1 µl |

PCR thermocycle

| Cycle | | Temperature | Duration |
|---|---|---|---|
| Denaturation | | 98°C | 30 seconds |
| Amplification (50X) | Denaturation | 98°C | 7 seconds |
| | Annealing | 55°C | 20 seconds |
| | Elongation | 72°C | 7 seconds |

The PCR product was used without further purification.

## 1.4 Template from synthetic oligonucleotides

Since only the T7 promoter region (from nucleotide -17 to -1) need to be double stranded (Milligan et al., 1987), another strategy was to anneal a T7 promoter DNA oligonucleotide to a single-stranded template. Furthermore, since the coding region (from +1) can be single stranded, the T7 promoter oligonucleotide required to form the double strand could be used with multiples templates. This technique was used as a faster alternative to produce RNAs, but was limited to 120 nucleotides. When RNA were used currently (for instance, for crystallization experiments) the cloning into pHDV was preferred.

For 1 ml of transcription mix, 1µM template was needed which corresponds to 100µl of annealing reaction. The annealing reaction was performed as follow

| | Stock Conc | Final Conc | Volume |
|---|---|---|---|
| MilliQ Water | | | 70 µl |
| Tris pH 7.6 | 1M | 5mM | 5 µl |
| $MgCl_2$ | 1M | 5mM | 5 µl |
| T7 Primer | 100 µM | 10µM | 10 µl |
| Template Oligo | 100 µM | 10µM | 10 µl |

The reaction was carried out in a thermocycler.

| Cycle | | Temperature | Duration |
|---|---|---|---|
| 1 | | 98°C | 5 minutes |
| 2 | Steps (80X) | From 98 to 18 °C | 1°C/minute |
| 3 | | 10°C | ∞ |

## 1.5 T7 in vitro transcription

Generally 5ml of transcription were carried on, but distributed in 1ml reactions.

| | Initial Concentration | Final Concentration |
|---|---|---|
| Linear plasmid | 1 mg/ml | 500µg |
| 5X Transcription Buffer‡ | 5 X | 1 X |
| $MgCl_2$ | 1M | 10mM |
| DTT | 1M | 5mM |
| rNTPs | 20mM | 4mM |
| GMP (used in RNA for crystallization) | 125mM | 20mM |
| T7 Polymerase | 1mg/ml | 0.1mg/ml |

A final concentration of 0.5 U/ml of Pyrophosphatase was added to avoid pyrophosphate accumulation. Reactions were incubated at 37°C for at least 4 h. Longer incubations (even overnight) were occasionally performed, but no significant increase of yield was observed. When Ribozyme cleavage was required, $MgCl_2$ was added to achieve a concentration of 40mM, taking into account the $MgCl_2$ (16mM) included in the transcription mix, and a folding treatment was performed. Reactions were incubated at 65°C for 10 minutes, and then at 37°C for 20 minutes to allow self-cleavage to proceed. Reactions were stopped by adding EDTA to a final concentration of 25mM. The transcripts were monitored by polyacrylamide gel electrophoresis‡ (Figure A.3).

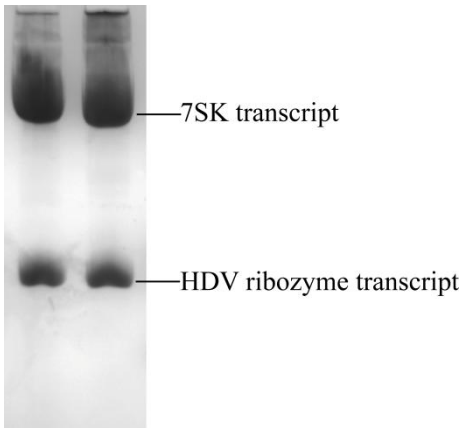



Figure A.3. Checking in vitro transcription. 5µl of transcription mix were loaded into an analytical polyacrylamide gel‡. Bands fro 7SK and the HDV ribozyme transcripts are shown.

## 1.6 In vivo production of RNAs

### a. Cloning

An alternative strategy was to produce recombinant RNA in *Escherichia coli* using a pBSTNAV vector initially used for tRNA production (Meinnel et al. 1988) and modified for in vivo production of RNAs (Ponchon et al. 2007). In this thesis the vector will be designated pKEa or pKSa (where K stands for tRNA Lys, S for sephadex aptamer, and E indicates that the sephadex aptamer is absent, see Figure A.4).

This thechnique is only applicable to hairpins. HP1, HP1u, HP1a and HP3 sequences were cloned into the pBSTNav vector using standard protocols to give pKS_HP1, pKS_HP1u, pKSA_HP1a and pKS_HP3, respectively.

### b. Culture and extraction

Competent cells of *Escherichia coli* strain JM101 were transformed with the recombinant pBSTNav vector that includes an ampicillin resistance gene. The tranformants were selected on LB (ampicillin)-agar plates. 4ml of sterile LB (ampicillin) medium was inoculated with a colony picked from agar plates and growth in an incubator shaker at 37°C for no more than 6 hours. The suitable volume of preculture was added to inoculate 0.5 L of sterile LB (ampicillin) medium at $1\times10^{-4}$ OD. The bacteria were growth in an incubator shaker at 37°C overnight (<15 hours). The culture was stopped before reaching 2.5 O.D. One liter culture at 0.1 O.D. was started by inoculation with the suitable volume of the overnight culture. The O.D. was monitored, and culture were stopped and harvested when O.D. is just < 2.5 (Figure A.5). This procedure ensures that the culture never reach the stationary (saturation) phase.

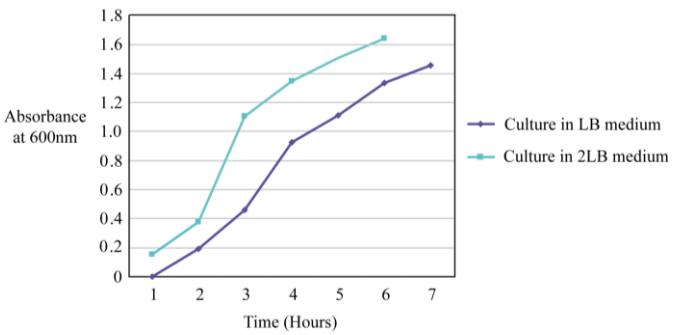



Figure A.5. Growth curve of DH5α *E. coli* cells transformed with pKEa vector. Growth in two different media is shown. Note that cultures were stopped before the stationary phase.

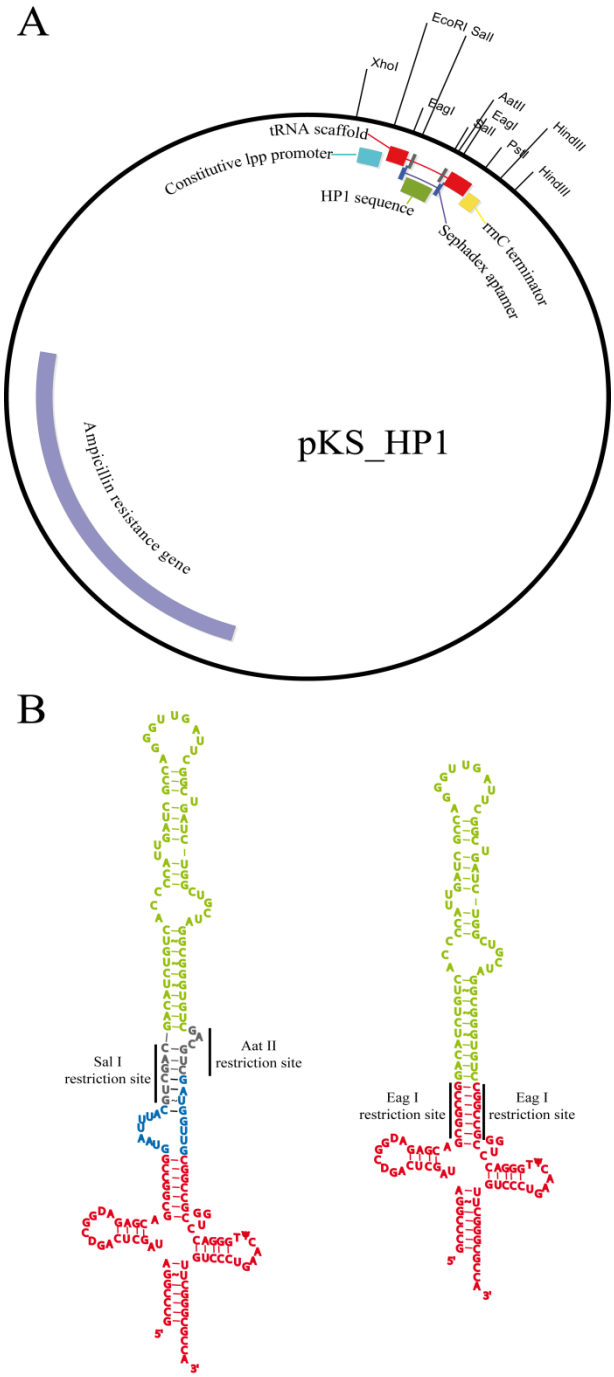



Figure A.4. pKS_HP1. A) Circular map of pKS_HP1 showing its main features: tRNA scaffold (red), inserted HP1 (green), sephadex aptamer (dark blue). B) Left, pKS_HP1; Right pKE_HP1. The color code as in A, and endonucleases restrictions sites are indicated.

## 1.7 Purification of RNAs

RNA was then purified with gel or chromatography. Cells were suspended in 10 ml of RNA2 buffer‡. A phenol:chloroform extraction‡ and ethanol precipitation‡ were then performed. RNA was suspended in Urea Loading Buffer‡ and then loaded in a preparative 10% polyacrylamide gel‡.

### a. Gel Purification

A phenol:chloroform extraction‡ and ethanol precipitation‡ was performed after the T7 in vitro transcription. RNA was suspended in Urea Loading Buffer‡ and then loaded in a preparative polyacrylamide gel‡. Different polyacrylamide percentages were used according to the RNA (see Table A.1). For 7SK substructures comprising single hairpins we used 10% to 12% polyacrylamide gels, while for full length 7SK and longer substructures we used 7.5% polyacrylamide gels. The electrophoresis was performed in TBE buffer‡ at 35mA, and monitored using the markers dyers.

**Table A.1**

| % Acrylamide | **RNA range size** | **Xylene cyanol** | **Bromophenol** |
|---|---|---|---|
| 7.5 | 60-450 nucleotides | 80 nucleotides | 20 nucleotides |
| 10 | 45-300 nucleotides | 55 nucleotides | 12 nucleotides |
| 12 | 35-250 nucleotides | 45 nucleotides | 10 nucleotides |
| 15 | 20-150 nucleotides | 28 nucleotides | 8 nucleotides |

RNA was visualized by UV shadowing (Figure A.6) and the RNA band was excised. RNA was eluted in Elution Buffer‡ overnight at 4°C with agitation. The eluted RNA was filtered through home made glass wool filter (adapted onto a syringe), then with a 0.2 µm poresize Minisart filter. The filtered solution was precipitated in ethanol‡. RNA was suspended in RNA2 Buffer‡. When further purification was needed, the RNA was diluted to achieve a concentration ≤ 1.5 mg/ml.

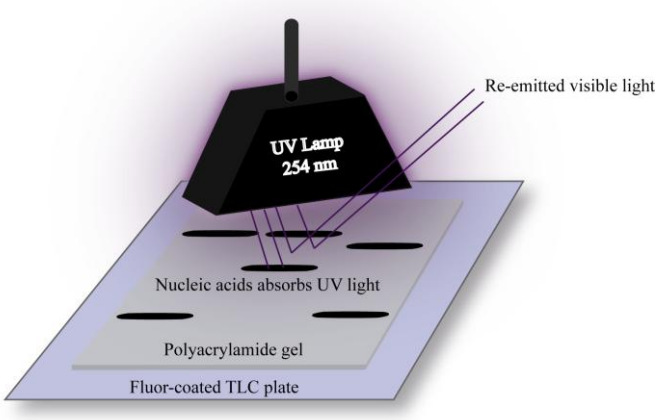



Figure A.6. UV shadowing. The RNA band was detected in the polyacrylamide gel without staining, using the UV shadowing method. The gel is placed over a Fluor-coated TLC plate and under a UV source with a wavelength of 254nm. Dark areas are observed where RNA in gel absorbs the UV light.

### b. MonoQ chromatography

MonoQ is based on a 10µm beaded hydrophilic polystyrene/divinyl benzene resin which has been substituted with quaternary amine groups to yield the strong anion exchanger (Ad 2002b), hence MonoQ is a suitable column for RNAs purification. When high purity of RNA was required for experiment, as for crystallization or SAXS, RNAs were further purified by anion exchange chromatography using a MonoQ column with linear gradient of 0.4 to 0.8 M NaCl in buffer containing 20mM Bis-Tris pH 7.0 and 0.25mM EDTA. For 7SK a thermal treatment (heating at 85°C for 1 minute, then ice for 5 minutes, see below) was needed to achieve a single conformation before MonoQ column (see Figure A.7). Fractions enriched with the RNA of interest were pooled and ethanol precipitated. RNA was recovered in RNA2 buffer‡.

c. Final conditioning of RNAs

*Thermal Treatment*

7SK and several of its substructures showed different conformations. These conformations were detected by agarose gel, gel filtration, and by MonoQ chromatography (Figure A.7). The thermal treatment allowed obtaining a unique conformation. This thermal treatment was required after each ethanol precipitation or after freezing. The conditions for a suitable thermal treatment were determined experimentally. Hence, thermal treatment was systematically performed in RNA2 buffer‡.

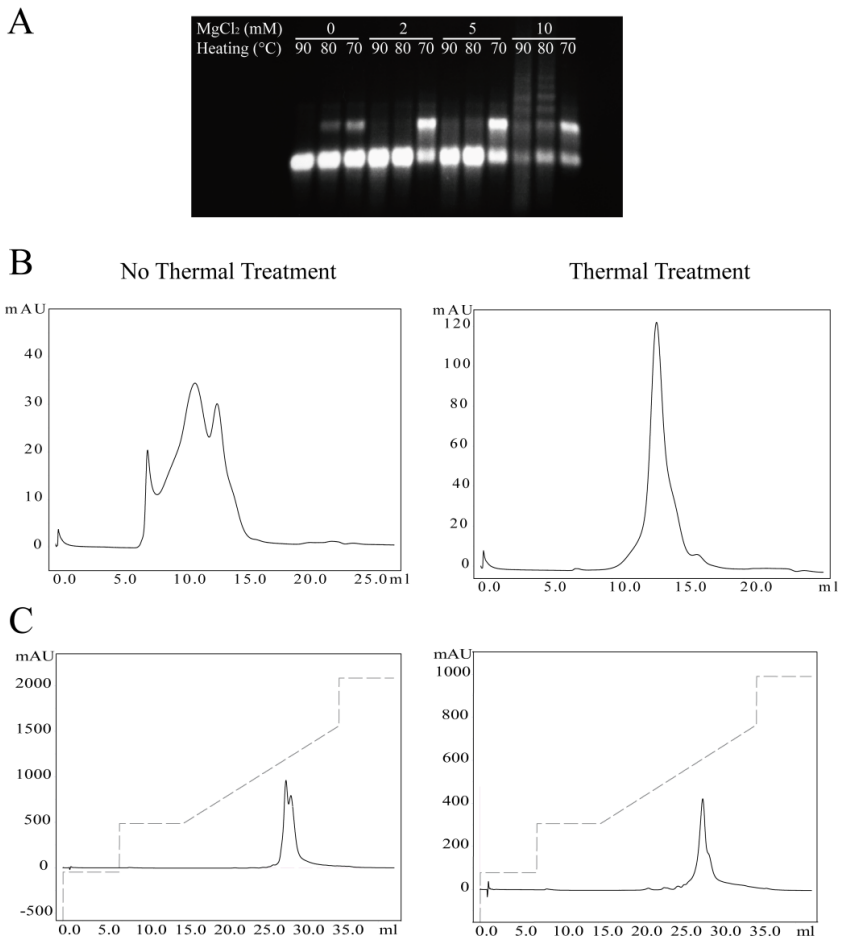



Figure A.7. 7SK thermal treatment. A) Agarose gel. Several thermal treatments conditions were tested and different 7SK conformations were visualized by analytical agarose gel‡. B) GF chromatography. Without thermal treatment different peaks of 7SK were observed in Superose6. After thermal treatment, only one peak was observed. C) MonoQ chromatography. No thermal treated 7SK showed two peaks in a Mono Q column, whereas only one peak was present after thermal treatment.

*Dialysis*

For crystallization and SAXS experiments a dialysis of RNA was required to eliminate salt excess or buffer exchange. Dialysis against RNA2 buffer‡ or SAXS buffer was performed using a GeBaFlex device (Gene-Bio Applications), overnight at 4°C. Once again, a thermal treatment is performed just before dialysis.

# 2. PROTEIN PRODUCTION

## 2.1 Protein expression

Competent cells of *Escherichia coli* expression strain BL21(DE3) were transformed with the recombinant plasmid that includes an antibiotic resistance gene. The transformants were selected on LB-agar plates containing the antibiotic. In general, 4ml of sterile LB medium with the corresponding antibiotics was inoculated with a colony picked from agar plates and growth in an incubator shaker at 37°C for ~6 hours. These pre-cultures served to inoculate 1L of sterile auto-inductive medium‡. The bacteria were grown in an incubator shaker at 25°C for 18 hours. After harvesting and centrifugation at 4000 rpm ad 4°C for 20 minutes to eliminate the medium, cells were suspended in the lysis buffer in presence of one tablet of COmplete Protease Inhibitor Cocktail (Roche) per 1L culture and sonicated. The lysate was clarified by centrifugation at 20,000 rpm and 4°C for 1 hour.

## 2.2 Affinity Chromatography

Most of the proteins were fused to a tag to facilitate their purification and/or crystallization (see Table III.3). Hence, the soluble extract was added to a batch of prepared affinity resin previously equilibrated in binding buffer and incubated for >2 hours at 4°C with agitation. Successive washes at low and high salt concentration were perfromed before elution to remove the RNAs non-specifically bound to proteins. In the case of $(His)_6$ tag fused proteins purification, an extra wash at 20mM imidazole was needed to eliminate the contaminants non-specifically bound to the resin. Then the protein was eluted or the tag was cleaved with the

corresponding protease. For crystallization, tag removal by enzymatic cleavage was tested as an optimization strategy. The protein was incubated at 4°C overnight in presence of the P3C protease while bound to the resin. Then the protein was recovered by washing the resin. Otherwise, the protein was eluted with an elution buffer containing 20mM Glutathione or 200mM Imidazole for proteins purified by a GST or $(His)_6$ tag, respectively.

## 2.3 Ion exchange chromatography

The eluted protein was then loaded into an ion exchange column adapted into an AKTA (GE Healthcare) chromatography system. For some proteins (see Table III.3), the cationic exchange HiTrap SP HP column (GE Healthcare) was used for RNAses removal. For other proteins, bacterial RNAs unspecifically bound to the protein were completely removed using the anionic exchange HiTrap Q FF column (GE Healthcare). In both cases, a linear ascendant gradient of NaCl was performed to elute the proteins.

## 2.4 Hydrophobic interaction chromatography

For the particular case of HEXIM136-273 fused to MBP, some proteolysis was observed even after ion exchange chromatography. To completely remove the proteases, a hydrophobic interaction chromatography was included before the ion exchange column. The ionic strength of the protein was increased by adding 4M ammonium sulphate to reach a concentration of 0.5M. The sample was then loaded onto a Phenyl Toyopearl column adapted in an AKTA chromatography system. The protein was eluted with a decreasing linear gradient of ammonium sulphate. This step greatly increased the quality of the purified protein.

## 2.5 Gel filtration chromatography

We usually used a gel filtration chromatography as final polishing step and removal of aggregates. In this way the protein is conditioned in the suitable buffer for analysis, storage or crystallization. Different types of gel filtration columns were used according to the size and

amount of the protein available: HiLoad 16/60 Superdex 200, Superdex 200 10/300 GL and Superdex 75 10/300 GL. Finally the protein was concentrated in an Amicon (Millipore) device.

| | **HiLoad 16/60 Superdex 200** | **Superdex 200 10/300 GL** | **Superdex 75 10/300 GL** |
|---|---|---|---|
| Exclusion limit ($M_r$)* | 1 300 000 | 1 300 000 | 100 000 |
| Separation range ($M_r$)* | 10 000 - 600 000 | 10 000 - 600 000 | 3 000 - 70 000 |
| Recommended sample volume | ≤ 5 ml | 25 – 250 µl | 25 – 250 µl |
| Bed volume | 120 ml | 24 ml | 24 ml |

*For globular proteins

Each purification step was monitored by SDS-PAGE‡, and yield was determined by measuring the absorbance at 280nm in a ND-1000 NanoDrop spectrophotometer or by Bradford chromogenic method‡.

The purification protocols of proteins prepared during my project and not included in the Table III.3 are presented in the next table:

| **Prot** | **Tag** | **Purification 1** | | **Purification 2** | | **Purification 3** | | **Yield** |
|---|---|---|---|---|---|---|---|---|
| | | **Step** | **Buffer** | **Step** | **Buffer** | **Step** | **Buffer** | |
| CTD | GST | GSH | 20mM Tris pH 7.6 250mM NaCl | Dialysis | 20mM Tris pH 7.6 250mM NaCl | | | ~2 mg |
| P-TEFb | Strep | Strep Resin | 20mM Tris pH 7.6 250mM NaCl | | | | | NE |
| LaRP7 | $(His)_6$ | NI-NTA | 100mM K Phosphate pH 8.0 500mM KCl 5mM CHAPS | Dialysis SP | 20mM KHepes pH 7.2 0.2-1M KCl 5mM CHAPS | S200 | 20mM KHepes pH 7.2 200mM KCl 5mM CHAPS 7mM βM | <1 mg |
| T7 RNA POL | $(His)_6$ | Ni-NTA | 100mM Na Phosphate pH 8.0 500mM NaCl 1.4mM βM | S200 | 50mM NaMES pH 6.5 100mM KCl | SP | 50mM NaMES pH 6.5 0.1-1M KCl | ~7 mg |
| P3C | $(His)_6$ | NI-NTA | 100mM Na Phosphate pH 8.0 250mM NaCl | Dialysis | 50mM Tris pH 8.0 150mM NaCl 0.5mM EDTA 2mM DTT | | | ~5 mg |

Yield in mg/L of culture; βM: β–Mercaptoethanol

# 3. EMSA

## 3.1 RNA labeling by in vitro T7 transcription

In EMSA experiment, RNA was radioactively labeled with P32. Two different types of labelling were carried out: 5' labeling or co-transcriptional labeling. The co-transcriptional labeling was widely preferred because more $P^{32}$ is incorporated into RNA with higher specific activity. The 5' labeling of RNA is described on "Probing methods". The labeling of RNA was performed by in vitro transcription in presence of alpha-$P^{32}$-CTP in a total volume of 20µl, as follow:

| | **Stock Conc** | **Final Conc** | **Volume** |
|---|---|---|---|
| Milli Q Water | | | 1.5 µl |
| Transcription Buffer | 5 X | 1 X | 4 µl |
| $MgCl_2$ | 100 mM | 10 mM | 2 µl |
| DTT | 100 mM | 5 mM | 2 µl |
| (AUG)TP | 10 mM | 1 mM | 2 µl |
| CTP | 1 mM | 0.1 mM | 2 µl |
| Template | 1 µg/µl | 2 µg | 2 µl |
| RNAsine | 40 U/µl | 1 U/µl | 0.5 µl |
| Alpha-$P^{32}$-CTP | | | 2 µl |
| T7 RNA Polymerase | 1 mg/ml | 0.1 mg/ml | 2 µl |

Reaction was incubated at 37°C for at least 2 hours. For the Ribozyme cleavage, $MgCl_2$ was added to achieve a concentration of 40mM. Reactions were incubated at 65°C for 10 minutes, and then incubated at 37°C for 20 minutes to allow self-cleavage to proceed. Reactions were stopped by adding 20 µl of Urea Denaturing Loading Buffer‡. After denaturation (2 minutes at 90°C) the labeled RNA was separated on a 10% polyacrylamide denaturing gel at 700 volts, 30mA and 15W for 90 minutes. Labeled RNA was localized by autoradiography (5 minutes exposition); the band was cut out and eluted overnight at room temperature. The eluted RNA was precipitated in ethanol‡. RNA was suspended in 10µl RNA2 Buffer‡. The c.p.m. were measured in a Beckman LS 6000SC scintillation counter by dissolving 1µl of radioactive material in 5 ml of scintillation liquid. Using this method, the RNA radioactivity typically ranged from 500 000 to 1 000 000 cpm.

## 3.2 Electrophoretic Mobility Shift Assay

The protocol is illustrated in Figure A.8. EMSA's reactions were done in 10µl of total volume with 50,000 c.p.m RNA. Labeled RNAs was folded by the thermal treatment as previously described. Then, 1µl of labeled RNA (50,000 c.p.m) diluted if needed in RNA2 Buffer, was mixed with 5µl of 2X EMSA Buffer‡, 3µl of increasing concentration solutions of protein and 1µl of MilliQ water. A control reaction was performed in absence of protein. The protein used was freshly purified, diluted in the Protein Dilution Buffer‡, and range of protein concentration was usually from 0.1 to 1µM.

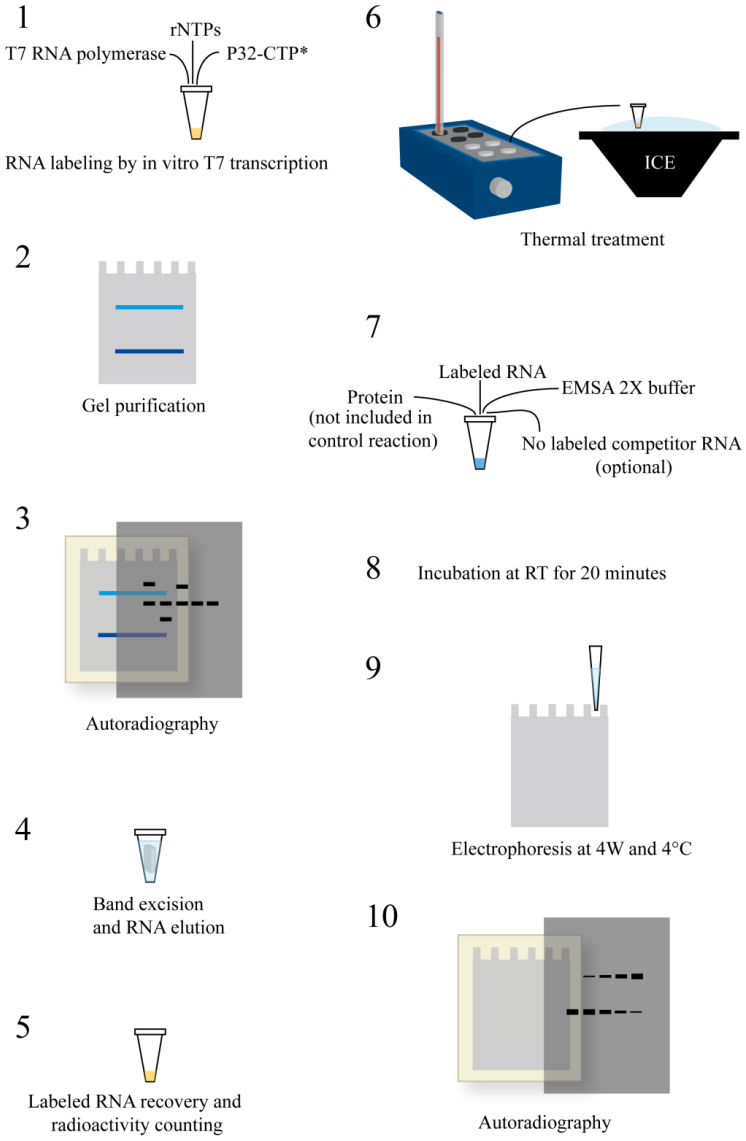



Figure A.8. EMSA method. The different steps of EMSA, from RNA labeling to autoradiography of results, are shown.

All reactions were incubated at room temperature for 30 minutes. The tRNA included in the 2X buffer, in large excess compared to the tested RNA, acts as a competitor to prevent non-specific RNA-protein interactions.

For competition experiments, a 5 minutes pre-incubation of 2µl of increasing concentration solutions of protein, 2µl of non-labeled competitor (refolded by thermal treatment) and 5µl of 2X EMSA Buffer was performed before adding the labelled RNA.

After incubation, samples were loaded in a 4% non-denaturing polyacrylamide gel‡ in 0.5X TBE‡ or TG buffer‡. Electrophoresis was performed at 4W and 4°C. Gel was pre-run for 30 minutes at least. At the end of the electrophoresis, the gel was transferred to a Whatman sheet paper and dried at 80°C under vacuum for 1 hour. Bands were visualized by overnight autoradiography at -80°C using an intensifying screen, or by 5 hours exposure to a phosphorimaging plate at room temperature.

# 4. RNA PROBING

## 4.1 RNA 5'-end labeling

End labeling of RNA with $P^{32}$ was a condition for direct secondary structure analysis. Free 5'-OH group was required for labeling. 5' phosphate was removed using Calf Intestine Alkaline Phosphatase (CIP) prior the labeling reaction. A reaction of 20µl was performed as follow:

| | **Stock Conc** | **Final Conc** |
|---|---|---|
| RNA | | 10 µg |
| CIP Buffer | 10 X | 1 X |
| CIP | 20 U/µl | 4 U |

Reaction was incubated at 37°C for 1 hour.

The transfer of the gamma $P^{32}$ from ATP to the 5'OH group was catalyzed using the Polynucleotide Kinase (PNK). For this radioactive labeling, a 20µl reaction was carried out as below:

| | Stock Conc | Final Conc | Volume |
|---|---|---|---|
| Milli Q Water | | | 2 µl |
| 5'OH RNA | 0.5 µg/µl | 0.25 µg/µl | 5 µl |
| PNK Buffer | 10 X | 1 X | 1 µl |
| ATPγP$^{32}$ | | | 1 µl |
| PNK | 10 U/µl | 1 U | 1 µl |

Reaction was incubated at 37°C for 1 hour. Then it was stopped by adding 10 µl of denaturing loading buffer. After denaturation (2 minutes at 90°C) the labeled RNA was separated on a 10% polyacrylamide denaturing gel‡ at 700 volts for 90 minutes. Labeled RNA was localized by autoradiography (5 minutes exposition); the band was cut out and eluted overnight at room temperature. The eluted RNA was ethanol precipitated‡. To increase recovery of labeled RNA during precipitation, 50µg of glycogen was added as co-precipitant. RNA was suspended in 10µl RNA2 Buffer‡. The counts per minute (c.p.m.) were measured in a Beckman LS 6000SC scintillation counter by dissolving 1µl of radioactive material in 5 ml of scintillation liquid.

The labeled RNA was refolded by a thermal treatment at 90°C for 1 minute and then fast cooling on ice for 5 minutes.

## 4.2 Enzymatic Probing analyzed by direct labeled RNA detection

For enzymatic probing, 250,000 c.p.m. per reaction was used. Probing reactions were performed in 200mM KCl, 2mM MgCl2, and 1 µg of non-labeled RNA as carrier RNA. Three different dilutions of each RNAse were prepared in MilliQ water and 1µl of such dilution was added to the reaction. Tubes were incubated at room temperature for 5 minutes. Additionally, to identify the cleavage position in the RNA sequence, the same reactions were performed in denaturing conditions using the Sequence Buffer provided by Ambion and incubation at 50°C for 5 minutes. Incubation controls in the absence of nucleases were performed in order to detect non-specific cleavage in RNA. At least two different alkaline reactions were also done using the Alkalyne Buffer by Ambion and incubating at 90°C for 4 and 8 minutes. All reactions were stopped on ice and by adding 10µl of Urea Denaturing Loading Buffer‡ and then electrophoresed in a 15% polyacrylamide denaturing gel‡ at 1.7 KV, 200mA and 35W for 150 minutes. Gel was pre-run for at least 30 minutes. Bands were visualized by overnight autoradiography at -80°C using an intensifying screen.

## 4.3 Enzymatic probing and selective 2´ -hydroxyl acylation analyzed by primer extension

### a. Enzymatic probing

RNA was refolded by the thermal treatment as mentioned above. Digestion reactions of 20µl volume were performed at a final concentration of 3µM RNA in 200mM KCl and 2mM $MgCl_2$. Appropriate dilutions of enzyme were done in MilliQ water, and 1µl of such dilutions was added to the reaction. Tubes were incubated at room temperature for 5 minutes. Incubation controls in the absence of enzyme were performed to detect non-specific cleavage in RNA. Reactions were stopped by adding 40µl of 300mM sodium acetate and by ethanol precipitation‡. Digested RNA was suspended in 20µl of MilliQ water.

### b. Selective 2´ -Hydroxyl Acylation

The following protocol was used to analyze the RNA sequence from each primer (see below) as described in (Wilkinson et al. 2006):

1. 10µM RNA was refolded in RNA2 Buffer by heating at 90°C and snap cooling on ice for 5 minutes. The final concentration of the RNA in the SHAPE reaction was 1µM.
2. 2µl of refolded RNA were distributed in three sterile tubes
3. For each tube, 16µl of one of the 1.25 X SHAPE Buffers‡ were added. Each SHAPE experiment was systematically performed under these three different buffer conditions:
    a. 1.25 X SHAPE Buffer A‡
    b. 1.25 X SHAPE Buffer B‡
    c. 1.25 X SHAPE Buffer W‡
4. Reaction were incubate at room temperature for 30 minutes
5. 9µl reaction were distributed in two sterile tubes:
    a. Tube (-) 1M7: 1µl of neat DMSO is added. This tube is a control in absence of 1M7 in order to detect non-specific cleavage in RNA.

b. Tube (+) 1M7: 1 µl of 80mM 1M7 (dissolved in DMSO) was added. This concentration was determined experimentally as suitable for 7SK. The recommended concentration of this solution varies with the RNA length(Wilkinson et al. 2006) (Wilkinson et al. 2006). Since 1M7 is hydrolysable, it was stored in a dry atmosphere as well as DMSO.

6. Reactions were incubated at 37°C for 90 seconds.
7. Reactions were terminated by adding 100 µl of SHAPE Stop Solution‡.
8. 350 µl of 100% ethanol were added to the reaction for ethanol precipitation‡.
9. Modified RNA was suspended in 10 µl of MilliQ water.

## c. Primer Extension Analysis

For the primer extension method, different DNA primers were designed to anneal different regions of 7SK so that all sequence was covered (Figure A.9).

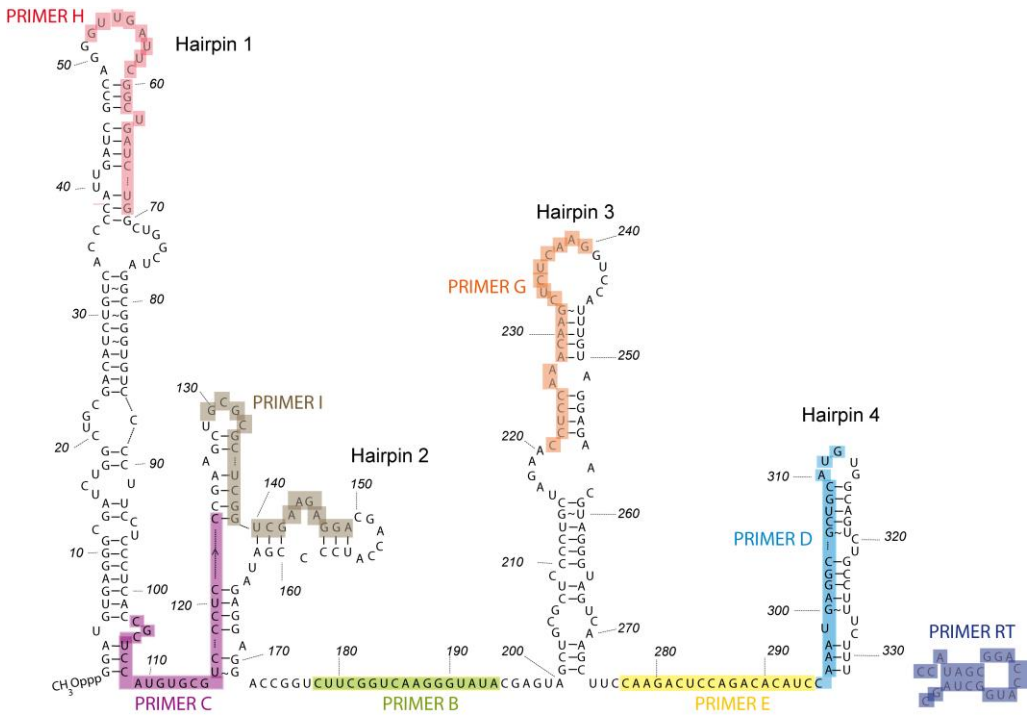



Figure A.9. Primers. The hybridation sequences for the different primers used for primer extension method are coloured on the Wassarman and Steitz 7SK model.

*Radioactive labelling of primers*

The primers were labeled with $P^{32}$ at their 5'-end using the PNK as indicated bellow. Since the DNA primers are produced by chemical synthesis, not prior 5'-end dephosphorylation was required.

| | Stock Conc | Final Conc | Volume |
|---|---|---|---|
| Milli Q Water | | | 6 µl |
| DNA primer | 10 µM | 1 µM | 1 µl |
| PNK Buffer | 10 X | 1 X | 1 µl |
| ATPγP32 | | | 1 µl |
| PNK | 10 U/µl | 1 U | 1 µl |

Reaction was incubated at 37°C for 1 hour. Then it was stopped by adding 10 µl of denaturing loading buffer and purified on a 10% polyacrylamide denaturing gel‡ at 700 volts for 90 minutes. Labeled DNA primer was visualized by autoradiography; the band was cut out and eluted overnight at room temperature. The eluted DNA primer was ethanol precipitated‡, in presence of 50µg of glycogen as co-precipitant. Labeled primer was suspended in 10µl MilliQ water. The c.p.m. were measured in a Beckman LS 6000SC scintillation counter by dissolving 1µl of radioactive material in 5 ml of scintillation liquid.

*Primer extension reaction*

Primer extension reactions were done in 20µl of total volume and using 60,000 c.p.m. each. Hence, 3µl of labeled primer at 20,000 c.p.m. were added to 10µl of SHAPE modified RNA or 3µl of enzymatic cleaved RNA (to which were added 7µl of MilliQ water to make up the volume). In parallel, a reverse transcription control and sequencing reaction were prepared by mixing 15µl of labeled primer at 20,000 c.p.m., 5µl of untreated RNA at 6µM and 30µl of MilliQ water; 11µl of this mix were distributed in five sterile tubes. All reaction tubes were then heated at 90°C for 10 minutes and then incubated at room temperature for 20 minutes to allow the annealing of the primer. Six microliters of Reverse Transcription Solution consisting of 4 volumes of SuperScript II Buffer (included with the commercial enzyme), 1 volume of 100mM DTT and 1 volume of 2.5mM dNTPs were added to all tubes. Finally, reactions were started by adding 1 µl of the SuperScript II Reverse Transcriptase (Invitrogen) at 20 U/µl, and incubated at 42°C for 45 minutes.

All reactions were stopped with 1µl of 4M NaOH and heating at 90°C for 10 minutes which completely degraded the RNA. Samples were precipitated with 60µl of 100% ethanol and 1 µl of 1M sodium acetate and incubated overnight at -20°C. Samples were recovered and suspended in 10µl of Acid Loading Buffer‡. The unbuffered acid Tris included in this buffer was essential for the proper bands migration on gel. DNA fragments were denatured at 90°C for 2 minutes and electrophoresed on 15% polyacryalmide denaturing gel‡ at 1.7kV, 200mA and 35W for approximately 150 minutes. Gel was pre-run for 30 minutes at least. Bands were examined either by overnight autoradiography at -80°C using an intensifying screen, or by 6 hours exposure to a phosphorimaging plate at room temperature.

## 5. Size Exclusion Chromatography

We used a Superose 6 10/300 GL (GE Healthcare) consisting of highly cross-linked porous agarose particles with a fractionation range from $5\times10^3$ to $5\times10^6$ for globular proteins. This column was adapted to an AKTA Purifier System (GE Pharmacia) at 4°C room and the UV absorbance at 280 nm monitored. The column was equilibrated with 20mM K Hepes pH 7.2, 200mM KCl, 2mM $MgCl_2$ and 7mM β-mercaptoethanol at a flow rate of 0.5 ml/min. The column was calibrated using molecular weight standards from BioRad. The concentrations of RNAs and proteins typically used were from 5 to 20 µM. All complexes were incubated 30 minutes at 20°C prior to the injection of 100 µl of complex.

## 6. Mass Spectrometry

Mass spectrometry is an analytical technique that measures mass-to-charge ratio of ions in vacuum. Mass spectrometers consist into three fundamental parts: ion source, analyzer and detector. Two main strategies for biomolecules ionizations exist: matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI). We used ESI since this method allows maintaining non-covalent interactions in complexes during their transferring from solution to gas phase. The ions were separated in a TOF (time-of-flight) analyzer.

Because the proper ionization of RNAs is limiting for MS and it depends on the length of the RNA, we used HP1 and HP1-L for MS analysis which have been reported as the HEXIM1 interaction elements of 7SK and showed a similar interaction than 7SK by EMSA. The buffer of the samples, proteins and RNAs, was exchanged against 250 mM $NH_4Ac$ pH 7.5 using MicroBioSpin (BioRad) columns or GeBAflex dialysis devices, since $NH_4$ ions are more volatile than the counterions contained in the samples buffer. The experiments were performed in at concentration of 2.5 to 20 µM of RNAs and proteins. Measurements were carried out in a LCT (Waters) instrument.

## 7. CRYSTALLOGENESIS

For RNA/protein complexes and isolated protein constructions, we typically tested Index (from Hampton Research), Classics, JCSG+, ProComplex, Nucleix (from Qiagen), and Wizard I and II (from Emerald Biosciences) at 4°C and 17°C. For RNAs we used the crystallization screen of Sigma-Aldrich at 4°C, 17°C and 24°C. The initial trials of crystallization were performed in 96-well sitting drop plates (Innovaplate or Swissci), and using the Honeybee system (robot) for protein crystallization. For RNA/protein complexes, drops were usually were prepared by mixing 200 or 100 nl of complex with 100 nl of reservoir. We used RNA:protein ratios of 1:1 or 1:2 according to the protein construction. The complexes were previously incubated at 20°C for 20 min. For RNAs trials, drops were prepared mixing 100 nl of complex with 100 nl of reservoir.

For HP1u crystallization, a first 96-conditions optimization screen was prepared using a Tecan Miniprep pippeting station and using “home-made” solutions. Drops were prepared as described above. These conditions were further optimized with a 48-conditions “home-made” screen. For this optimization, we used a 48-well hanging drop plate (Hampton Research). The cover-slips were previously washed with ethanol and dried. Drops were prepared by mixing 1 µl of HP1u and 1 µl of reservoir at 20°C. HP1u trials were typically performed at 10 mg/ml and at 4°C.

# ANNEXES 2:
# BASIC PROTOCOLS AND SOLUTIONS

## Phenol:Chloroform extraction

Solutions were mixed with 1.0 volumes phenol:chloroform solution, vortex 1 minute, and centrifuged 1 minute at 12 000 g. Aqueous upper phase was transferred to a new tube. To increase the yield of recovery, 0.1 volumes of water or RNA2 buffer was added. The mix is vortex 30 seconds, and centrifuged 1 minute at 12 000 g. Aqueous upper phase was pooled with the precedent one.

## Ethanol precipitation

Ethanol precipitation was performed by mixing the solution containing DNA or RNA with 2.0 volumes of 100% ethanol, vortexing and incubating overnight at -20°C. Then DNA or RNA was recovered by centrifuging at 12 000 g and 4°C for 30 minutes. The pellet was washed with 70% ethanol and centrifuged once more at 12 000 g and 4°C for 15 minutes. DNA or RNA was dried and suspended in the suitable buffer.

## Preparative denaturing polyacrylamide gel for RNA

The dimensions of the preparative gel were 15cm × 15cm × 1.5mm. To attain the desired polyacrylamide concentration, an appropriate volume of a concentrated polyacrylamide stock solution containing 15% Polyacrylamide:Bisacrylamide 19:1, 8M Urea, and 1X TBE was diluted with a solution of 8M Urea, and 1X TBE to reach a final volume of 40ml. 0.01 and 0.001 volumes of 10% ammonium persulfate and TEMED, respectively, were added. The polymerization was allowed to procced for 30 minutes at least. Then the gel was pre-run in 1X TBE for 30 minutes before loading the sample.

## Analytical denaturing polyacrylamide gel for RNA

Gels were prepared using the MiniProtean® sytem from BioRad. The dimensions were 8.6cm × 6.8cm × 1mm. 6ml of polyacrylamide solution were prepared as for preparative gels previously described.

## Denaturing polyacrylamide gel for primer extension

The dimensions of this gel were 40cm × 28cm × 0.3mm.

| | |
|---|---|
| 1X | TBE |
| 8M | Urea |
| 15% | Acrylamide:Bisacrylamide (19:1) |

The polymerization was allowed to procced for 30 minutes at least. Then the gel was pre-run in 1X TBE for 45 minutes before loading the samples.

## Non-denaturing polyacrylamide gel for EMSA

The dimensions of native gel were 15cm × 15cm × 1.5mm.

| | |
|---|---|
| 0.5X | TBE or TG |
| 4% | Acrylamide:Bisacrylamide (19:1) |

The polymerization was allowed to procced for 30 minutes at least. Then the gel was pre-run in 0.5X TBE for 30 minutes before loading the samples.

## Protein Quantification by Bradford Assay

5X Bradford Reagent (BioRad) was diluted in MilliQ water.

1 ml of Bradford reagent was disposed in a cuvette.

A volume into the range 5 to 50 µl of protein solution was added.

Mix was incubated for 5 minutes.

The blank was read and sample absorbance at 594 nm was measured in a spectrophotometer.

When absorbance was into a range of 0.1 and 0.9, the protein concentration could be calculated from the next equation:

$$\text{mg/ml protein} = \frac{O.D._{595nm} \times 17.5}{\mu l \text{ of protein solution added}}$$

## SDS-PAGE

Gels were prepared using the MiniProtean® sytem from BioRad

### 12% Separating Gel

| | |
|---|---|
| 375mM | Tris-HCl pH 8.8 |
| 0.1% | SDS |
| 12% | Acrylamide:Bisacrylamide (40:1) |

### Stacking gel

125mM Tris pH 6.8
0.1% SDS
5% Acrylamide:Bisacrylamide (40:1)

### Electrophoresis buffer

25mM Tris
250mM Glycine
0.1% SDS

### Coomassie dye solution

10% Acetic acid
30% Ethanol
0.05% Coomassie Brilliant Blue R250

### Destaining solution

10% Acetic acid
30% Ethanol

### Phenol:Chloroform solution

50% Water saturated phenol
48% Chloroform
2% Isoamyl alcohol

### RNA2 Buffer

10mM NaCacodylate pH 6.5
2mM $MgCl_2$
0.25mM EDTA

### 5X Transcription Buffer

150mM Na Hepes pH 8.0
30mM $MgCl_2$

| | |
|---|---|
| 10mM | Spermidine |
| 0.05% | Triton |

### 2X Urea Denaturing Loading Buffer

| | |
|---|---|
| 7M | Urea |
| 0.002% | Xylen cyanol blue |
| 0.002% | Bromophenol blue |

### Elution Buffer

| | |
|---|---|
| 0.5M | Ammonium acetate |
| 2mM | $MgCl_2$ |
| 0.1% | SDS |
| 1mM | EDTA |

### 1X TBE

| | |
|---|---|
| 90mM | Tris |
| 90mM | Boric acid |
| 2mM | EDTA |

### Autoinductive medium

For 1L of autoinductive medium:

| | |
|---|---|
| 930 ml | ZY medium |
| 20ml | 50X 5052 solution |
| 50ml | 20X NPS solution |
| 600µl | 1M Ammonium sulfate |

### Sterile ZY medium

| | |
|---|---|
| 10 g | Tryptone |
| 5 g | Yeast extract |
| 930 ml of | Demineralised water |

## 50X 5052 solution

| | |
|---|---|
| 22% | Glycerol |
| 139mM | Glucose |
| 292mM | Lactose |

Filter

## 20X NPS

| | |
|---|---|
| 0.5M | $(NH_4)_2SO_4$ |
| 1M | $KH_2PO_4$ |
| 1M | $Na_2HPO_4$ |

Filter

## 2X EMSA buffer

| | |
|---|---|
| 100mM | K Hepes pH 7.2 |
| 400mM | KCl |
| 4mM | $MgCl_2$ |
| 0.2mg/ml | BSA |
| 0.6mg/ml | tRNA |
| 2mM | TCEP |
| 1mM | EDTA |
| 30% | Glycerol |
| 0.02% | NP40 |

## Protein Dilution Buffer

| | |
|---|---|
| 50mM | KHepes pH 7.2 |
| 500mM | KCl or NaCl |
| 5mM | $MgCl_2$ |
| 0.1mg/ml | BSA |
| 1mM | TCEP |

## TG buffer

| | |
|---|---|
| 25mM | Tris |

250mM Glycine

### 1.25X SHAPE Buffer A

62.5mM NaHepes pH 7.6
2.5mM $MgCl_2$
0.31mM EDTA

### 1.25X SHAPE Buffer B

62.5mM NaHepes pH 7.6
250mM KCl
12.5mM $MgCl_2$
0.31mM EDTA

### 1.25X SHAPE Buffer W

125mM NaHepes pH 8.0
125mM NaCl
7.5mM $MgCl_2$

### SHAPE Stop Solution

200mM NaCl
0.2mg/ml Glycogen
2mM EDTA

### Acid Loading Buffer

25 volumes Urea Loading Buffer
4 volumes Unbuffered acid Tris

### Elution Buffer

0.5M Ammonium acetate
0.1% SDS
2mM $MgCl_2$
1mM EDTA

## Analytical agarose gel for RNA

| | |
|---|---|
| 1.6% | Agarose |
| 0.5X | TBE |
| 0.004% | GelRed (Interchim) |

## 6X Native Loading Buffer for RNA or DNA

| | |
|---|---|
| 60% | Glycerol |
| 10mM | Tris pH 8.0 |
| 1mM | EDTA |
| 0.03% | Xylene cyanol |
| 0.03% | Bromophenol blue |

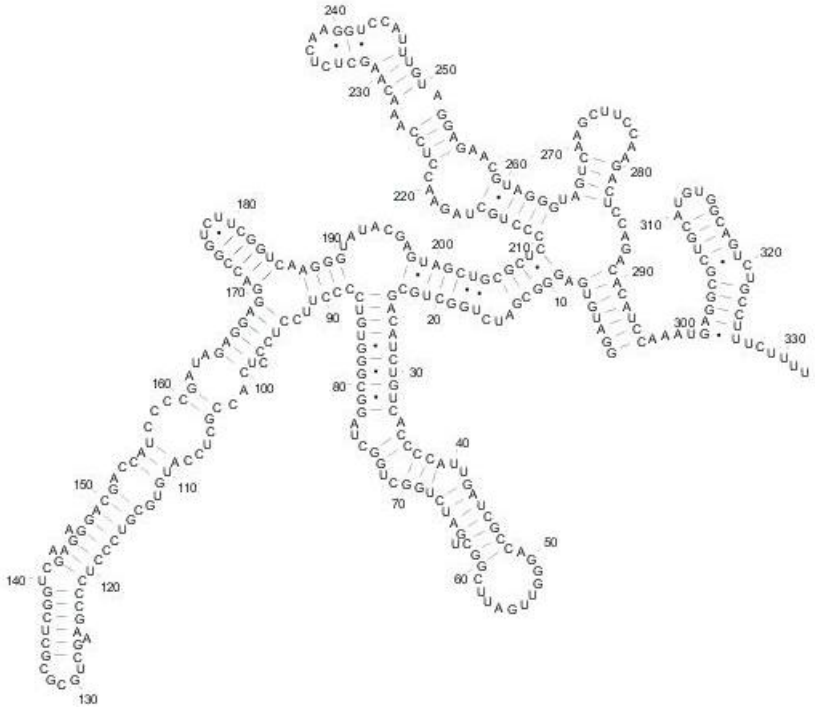
## Agarose gel for DNA

| | |
|---|---|
| 1 to 1.5% | Agarose |
| 1X | TAE |
| 0.004% | Ethidium Bromide |

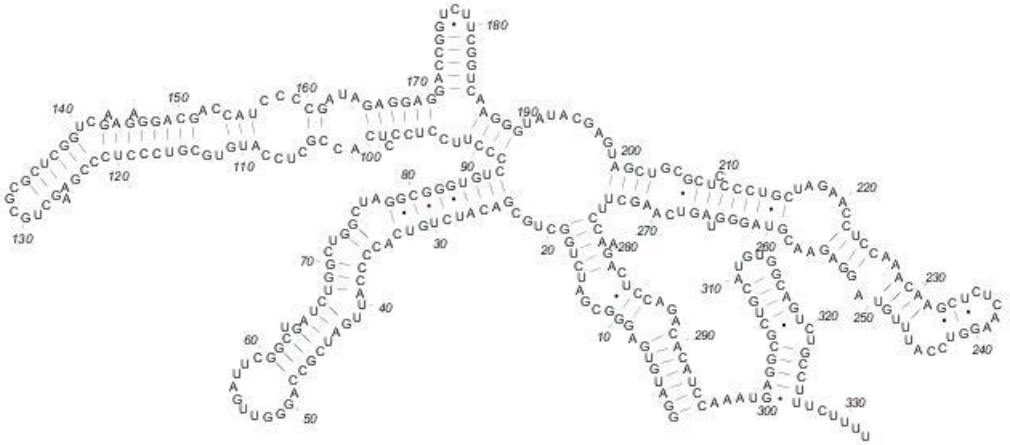
## 1X TAE Buffer

| | |
|---|---|
| 40mM | Tris |
| 20mM | Acetic acid |
| 1mM | EDTA |

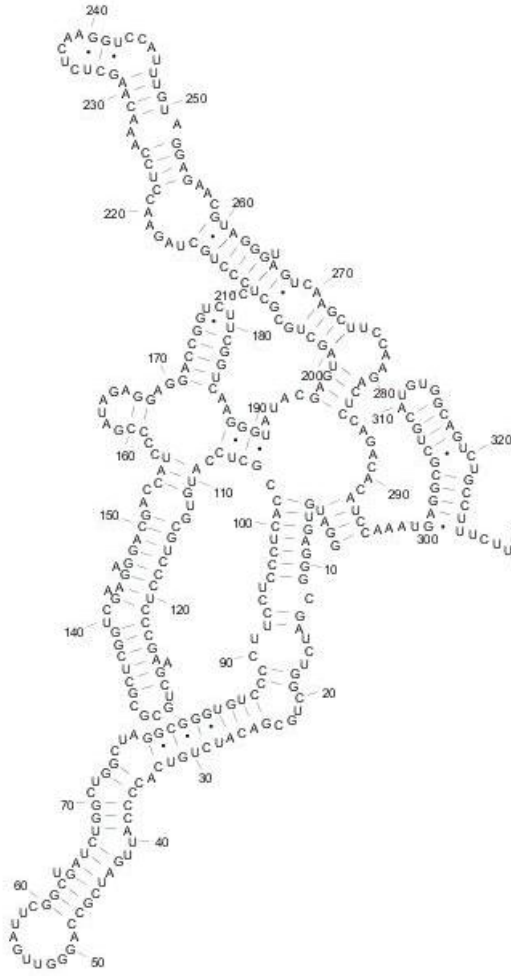
# ANNEXES 3:
# 7SK MODELS CALCULATED USING SHAPE DATA



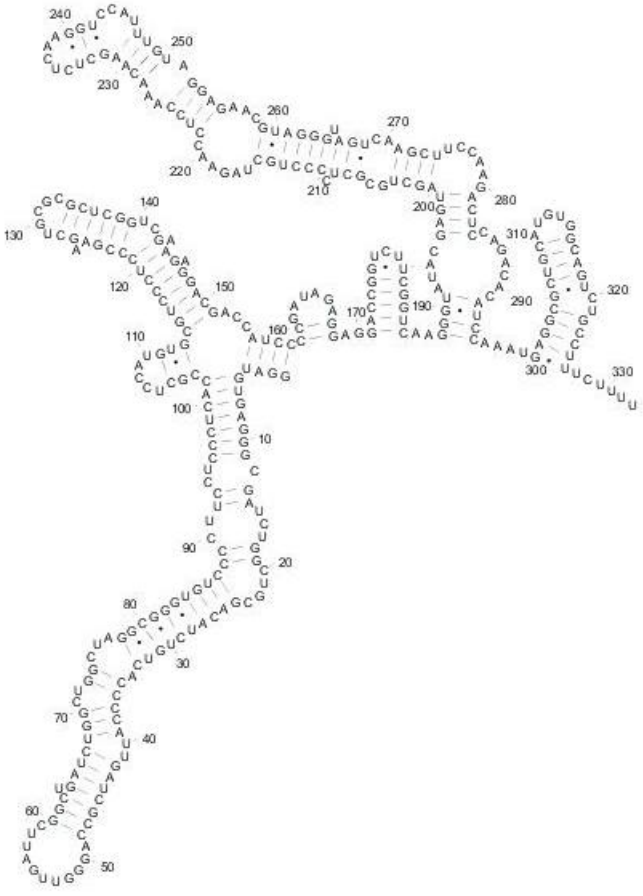
1. ΔG = -246.4 kcal/mol



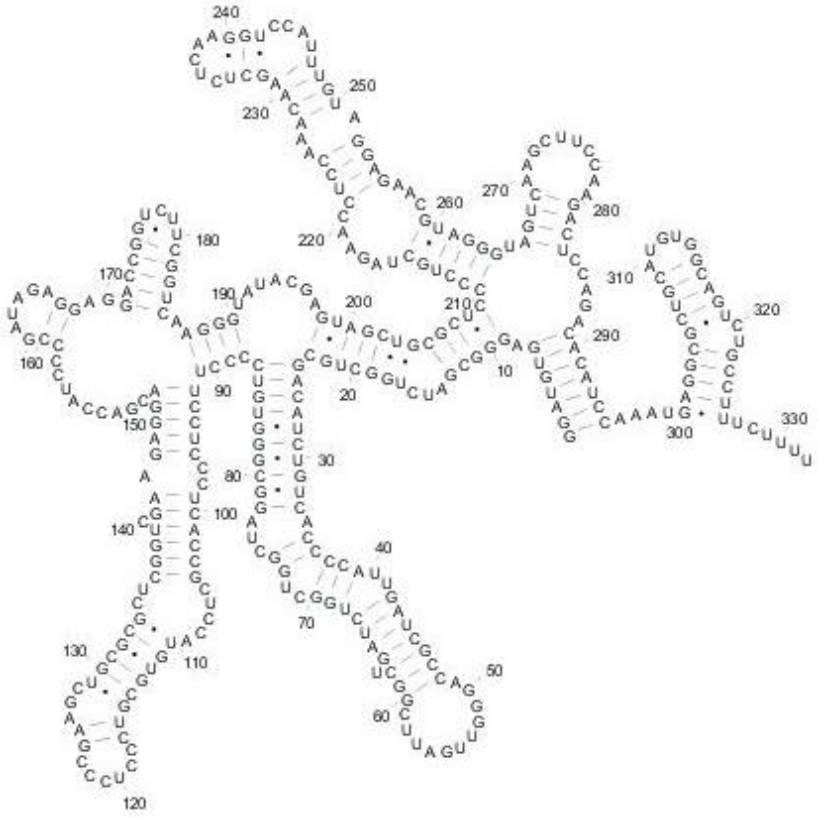
2. ΔG = -241.8 kcal/mol



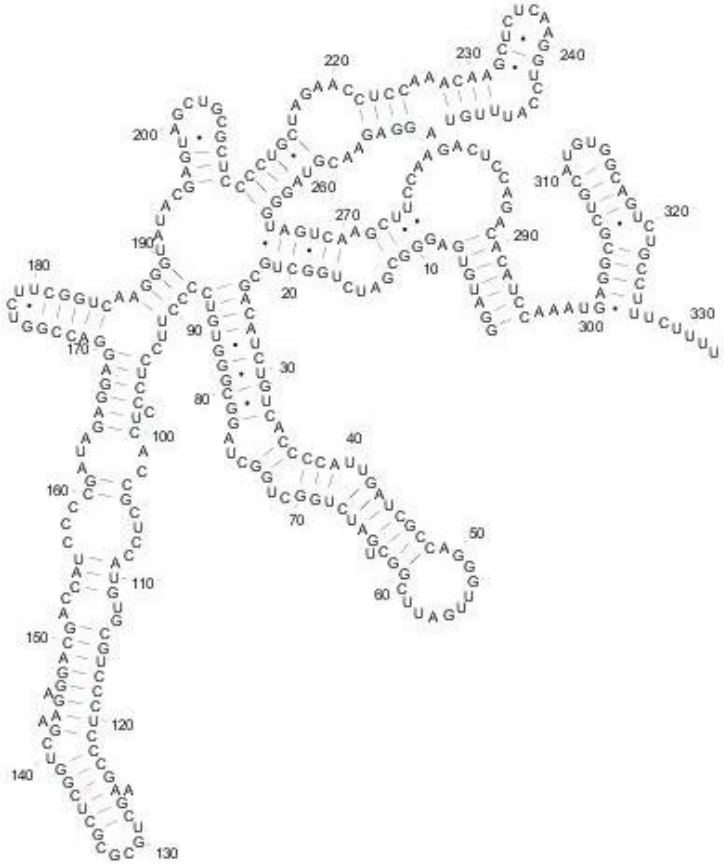
3. ΔG = -238.3 kcal/mol



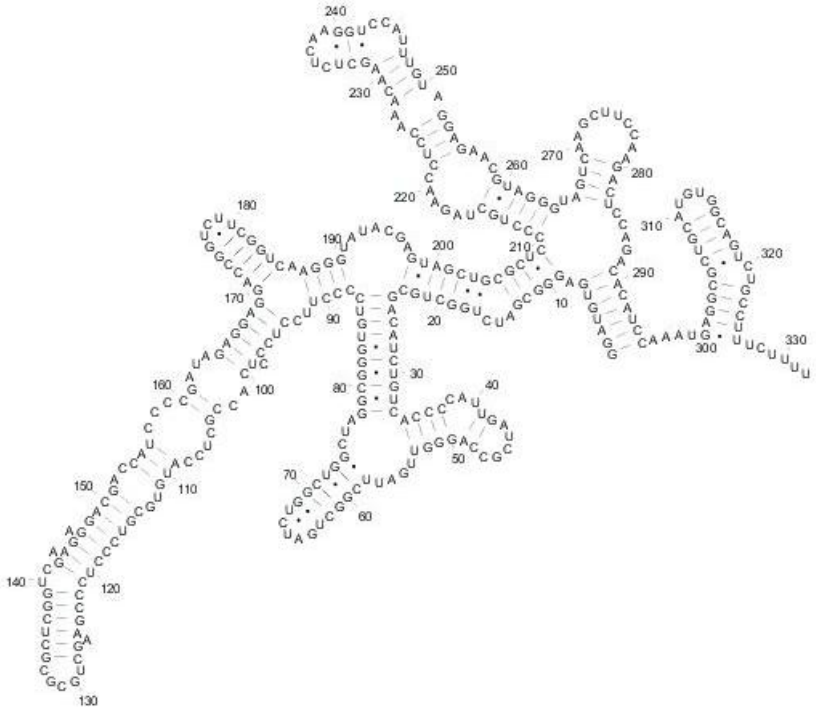
4. ΔG = -238.1 kcal/mol



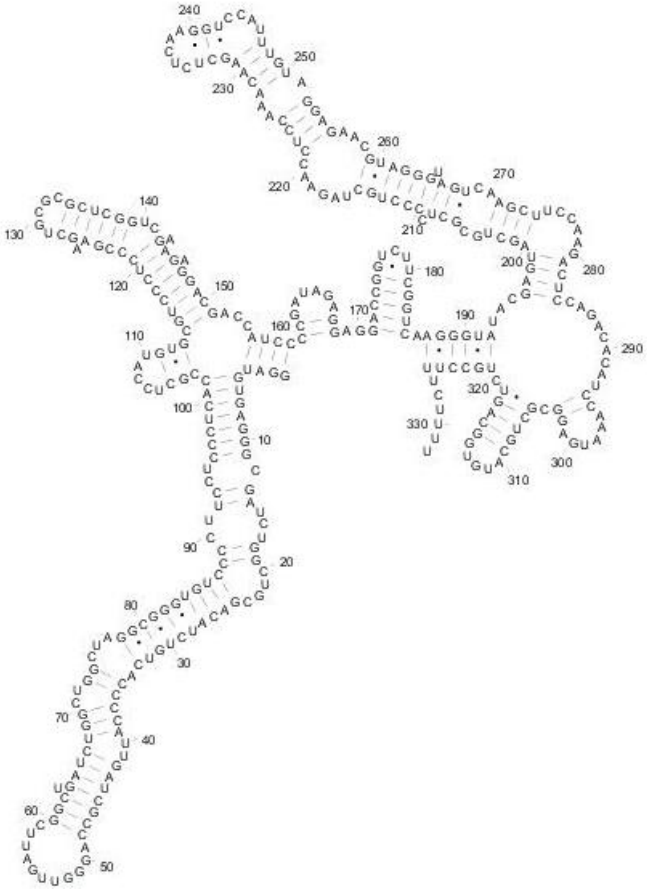
1. ΔG = -235.2 kcal/mol



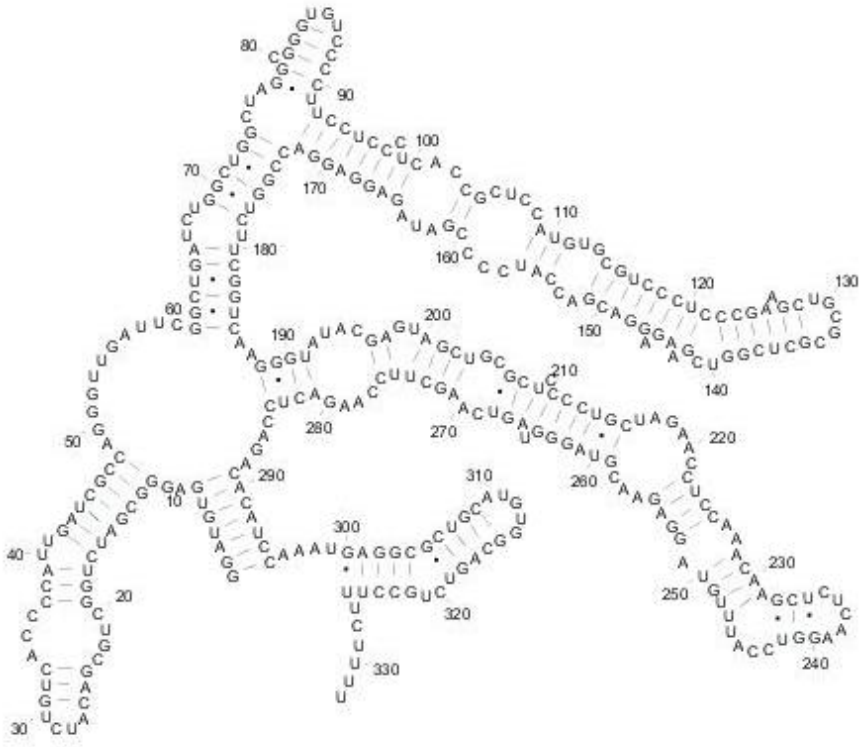
6. ΔG = -235.2 kcal/mol



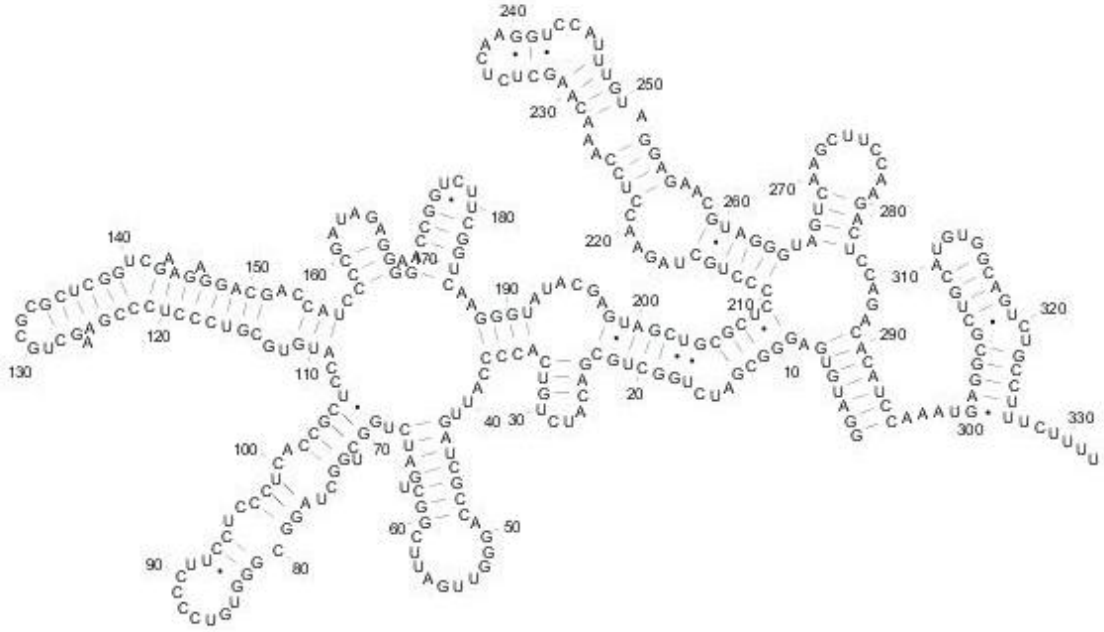
7. ΔG = -234.9 kcal/mol



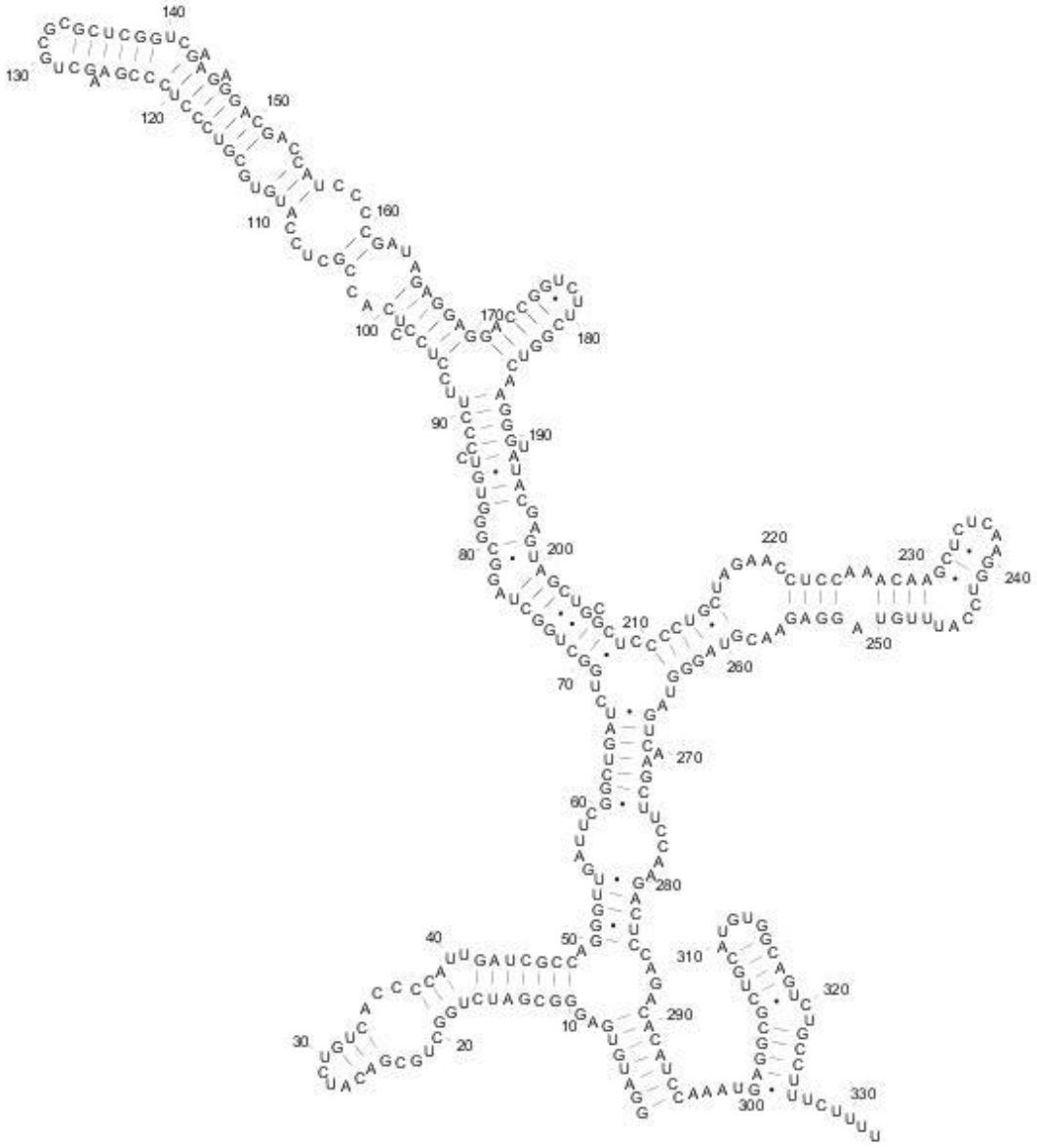
8. ΔG = -230.1 kcal/mol



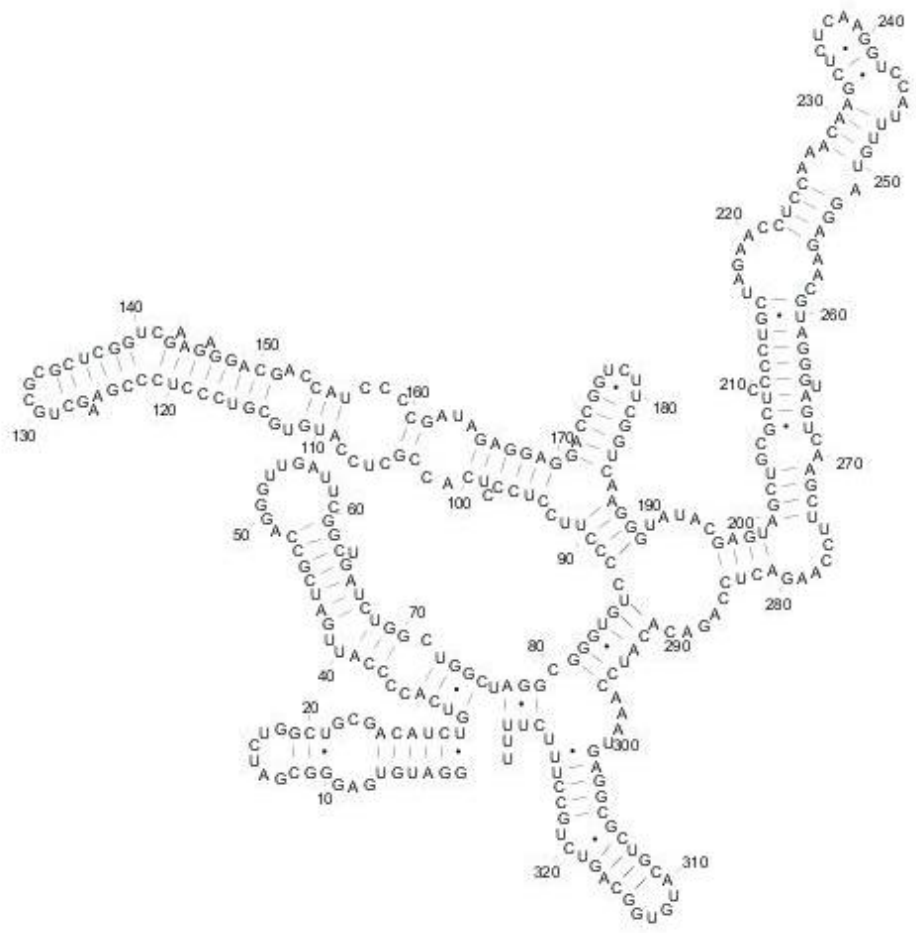
9. ΔG = -230.0 kcal/mol



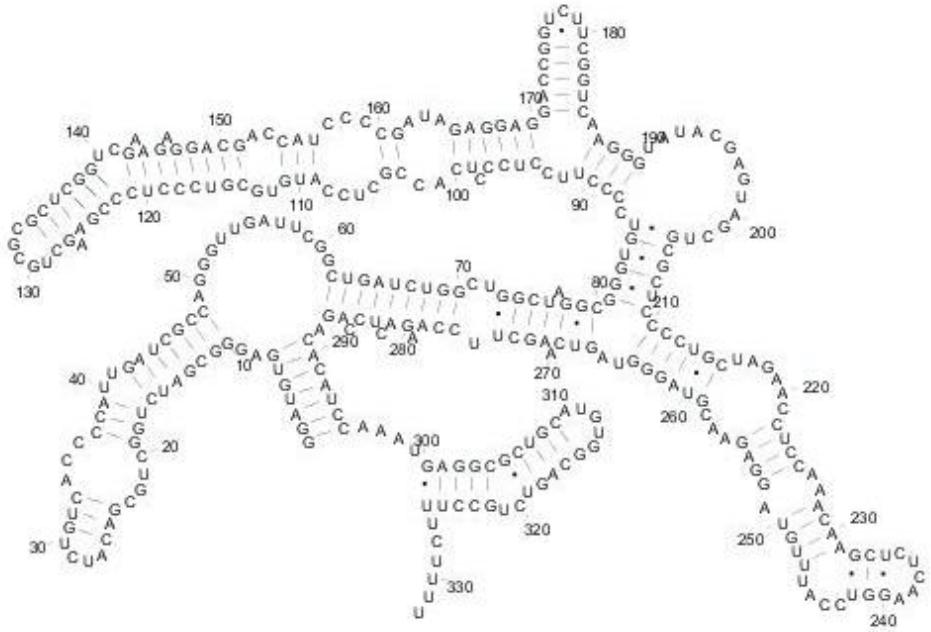
10. ΔG = -229.6 kcal/mol



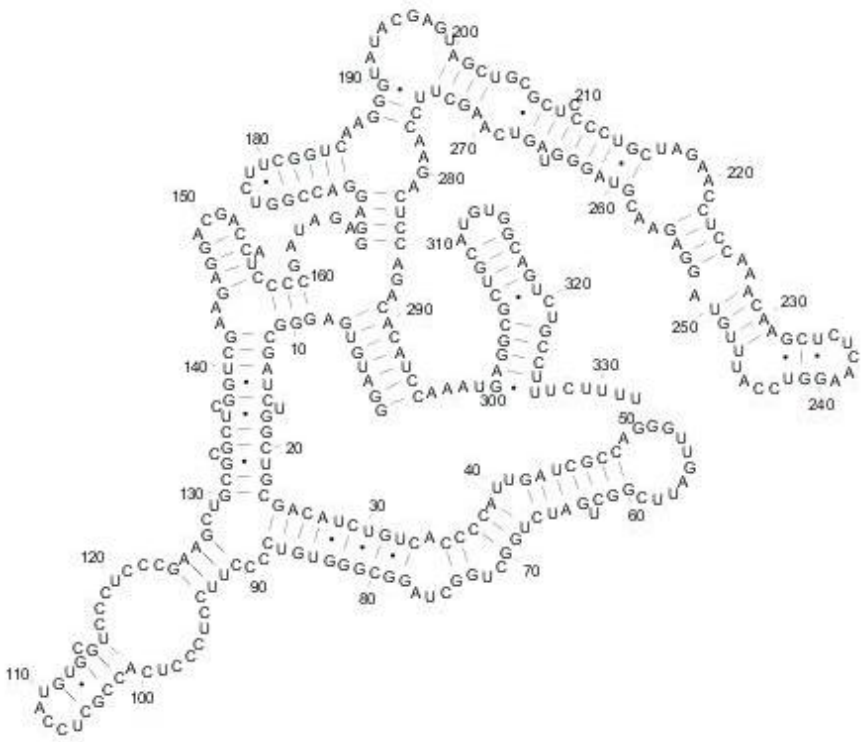
11. ΔG = -226.7 kcal/mol



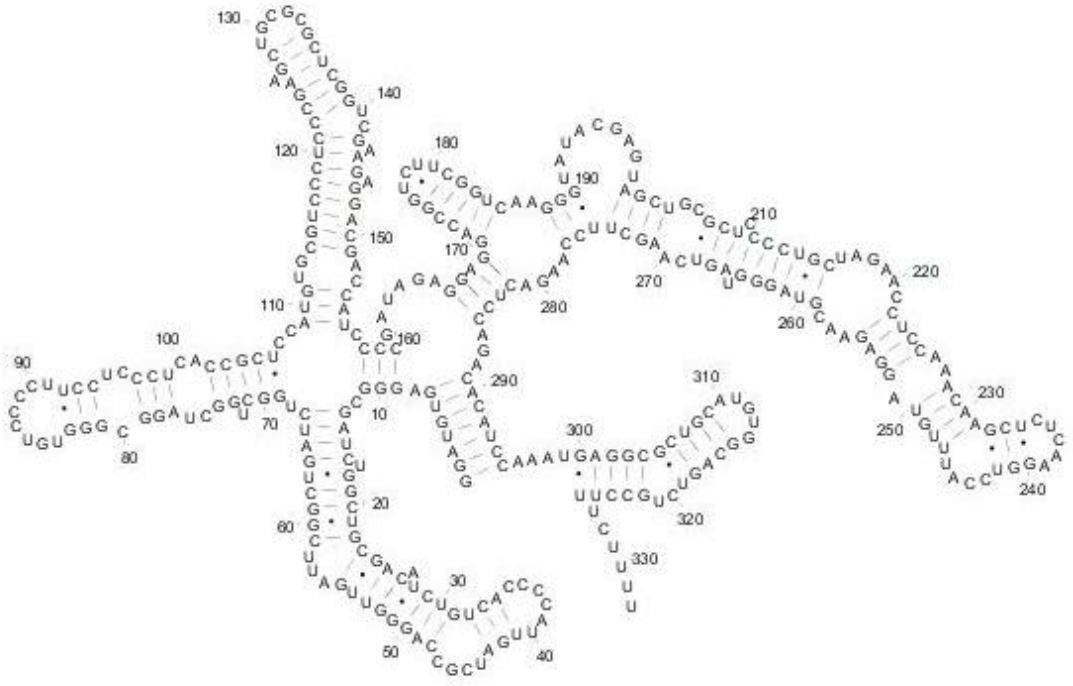
12. ΔG = -225.0 kcal/mol



13. ΔG = -224.8 kcal/mol



14. ΔG = -223.2 kcal/mol



15. ΔG = -222.6 kcal/mol

| Sequences with >60% identity (37) | | | |
|---|---|---|---|
| **Nucleotides** | **Co-variation** | **Conservation** | **Likelihood** |
| 1-295 | | ✓ | 93.9% |
| 2-294 | | ✓ | 93.9% |
| 3-293 | | ✓ | 93.7% |
| 4-292 | | ✓ | 92.8% |
| 5-291 | | ✓ | 90.9% |
| 6-290 | | ✓ | 90.8% |
| 7-289 | | ✓ | 90.5% |
| 24-87 | | ✓ | 98.5% |
| 25-86 | | ✓ | 98.5% |
| 26-85 | | ✓ | 98.5% |
| 27-84 | | ✓ | 98.5% |
| 28-83 | | ✓ | 98.4% |
| 28-82 | | ✓ | 98.5% |
| 30-81 | | ✓ | 98.4% |
| 31-80 | | ✓ | 98.5% |
| 32-79 | ✓ | | 98.4% |
| 33-78 | | ✓ | 98.4% |
| 35-74 | | ✓ | 97.5% |
| 36-73 | | ✓ | 97.5% |
| 46-70 | | ✓ | 97.6% |
| 47-69 | | ✓ | 98.7% |
| 42-67 | | ✓ | 99.7% |
| 43-66 | | ✓ | 99.7% |
| 44-65 | | ✓ | 99.7% |
| 45-64 | | ✓ | 99.7% |
| 46-62 | | ✓ | 99.3% |
| 47-61 | | ✓ | 99.7% |
| 48-60 | | ✓ | 99.7% |
| 50-59 | | ✓ | 93.4% |
| 51-58 | | ✓ | 90.7% |
| 115-150 | | ✓ | 99.8% |
| 116-149 | | ✓ | 99.8% |
| 117-148 | | ✓ | 99.3% |
| 118-147 | | ✓ | 98.8% |
| 119-145 | | ✓ | 98.2% |
| 122-139 | | ✓ | 99.8% |
| 123-138 | | ✓ | 99.8% |
| 124-137 | | ✓ | 99.8% |
| 127-133 | | ✓ | 95.1% |
| 128-132 | | ✓ | 95.0% |
| 173-183 | ✓ | | 100% |
| 174-182 | ✓ | | 100% |
| 175-181 | ✓ | | 99.7% |

| | | | |
|---|---|---|---|
| 210-264 | ✓ | | 98.7% |
| 211-263 | | ✓ | 99.1% |
| 212-262 | ✓ | | 99.5% |
| 213-261 | | ✓ | 99.4% |
| 214-260 | | ✓ | 99.4% |
| 215-259 | | ✓ | 99.4% |
| 300-326 | | ✓ | 90.6% |
| 301-325 | | ✓ | 95.6% |
| 302-324 | | ✓ | 100% |
| 303-323 | | ✓ | 100% |
| 304-322 | | ✓ | 100% |
| 306-318 | | ✓ | 100% |
| 307-317 | | ✓ | 100% |
| 308-316 | | ✓ | 100% |
| 309-315 | | ✓ | 100% |

| **All sequences (75)** | | | |
|---|---|---|---|
| **Nucleotides** | **Co-variation** | **Conservation** | **Likelihood** |
| 38-69 | ✓ | | 99.9% |
| 42-67 | | ✓ | 100% |
| 43-66 | | ✓ | 100% |
| 44-65 | | ✓ | 100% |
| 45-64 | | ✓ | 100% |
| 46-62 | | ✓ | 100% |
| 47-61 | ✓ | | 100% |
| 122-139 | ✓ | | 99.4% |
| 123-138 | | ✓ | 99.5% |
| 124-137 | | ✓ | 98.1% |
| 127-135 | | ✓ | 95.2% |
| 128-134 | | ✓ | 95.3% |
| 201-273 | ✓ | | 100% |
| 202-272 | ✓ | | 100% |
| 302-324 | ✓ | | 100% |
| 303-323 | | ✓ | 100% |
| 306-318 | ✓ | | 100% |
| 307-317 | ✓ | | 100% |
| 308-314 | ✓ | | 100% |
| 309-315 | | ✓ | 99.9% |

# ANNEXES 4:
# PUBLICATION

Lebars, I., Martinez-Zapien, D., Durand, A., Coutant, J., Kieffer, B., and Dock-Bregeon, A.-C. (2010). HEXIM1 targets a repeated GAUC motif in the riboregulator of transcription 7SK and promotes base pair rearrangements. Nucleic acids research 38, 7749-63

# BIBLIOGRAPHY

Ad, E. (2002a). Affinity Chromatography: Principles and Methods E. Ad, ed. (Handbook from Amersham Biosciencies).

Ad, E. ed. (2002b). Ion Exchange Chromatography (Handbook from Amersham Biosciencies).

Anand, K., Schulte, A., Vogel-Bachmayr, K., Scheffzek, K., and Geyer, M. (2008). Structural insights into the cyclin T1-Tat-TAR RNA transcription activation complex from EIAV. Nature structural & molecular biology 15, 1287-92.

Armache, K.-J., Kettenberger, H., and Cramer, Patrick (2005). The dynamic machinery of mRNA elongation. Current opinion in structural biology 15, 197-203.

Barberis, A., and Petrascheck, M. (2003). Transcription Activation in Eukaryotic Cells. Life Sciences.

Barboric, M., Kohoutek, Jirí, Price, J. P., Blazek, D., Price, David H, and Peterlin, B. M. (2005). Interplay between 7SK snRNA and oppositely charged regions in HEXIM1 direct the inhibition of P-TEFb. The EMBO journal 24, 4291-303.

Barboric, M., Yik, J. H. N., Czudnochowski, N., Yang, Zhiyuan, Chen, R., Contreras, X., Geyer, M., Matija Peterlin, B., and Zhou, Q. (2007). Tat competes with HEXIM1 to increase the active pool of P-TEFb for HIV-1 transcription. Nucleic acids research 35, 2003-12.

Barrandon, C., Bonnet, F., Nguyen, Van Trung, Labas, V., and Bensaude, Olivier (2007). The transcription-dependent dissociation of P-TEFb-HEXIM1-7SK RNA relies upon formation of hnRNP-7SK RNA complexes. Molecular and cellular biology 27, 6996-7006.

Barrick, J. E., Sudarsan, N., Weinberg, Z., Ruzzo, W. L., and Breaker, R. R. (2005). 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. RNA (New York, N.Y.) 11, 774-84.

Batey, R., Rambo, R., and Doudna, J. (1999). Tertiary Motifs in RNA Structure and Folding. Angewandte Chemie (International ed. in English) 38, 2326-2343.

Baumli, S., Lolli, G., Lowe, E. D., Troiani, S., Rusconi, L., Bullock, A. N., Debreczeni, J. E., Knapp, S., and Johnson, L. N. (2008). The structure of P-TEFb (CDK9/cyclin T1), its complex with flavopiridol and regulation by phosphorylation. The EMBO journal 27, 1907-18.

Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M., and Svergun, Dmitri I (2007). Structural characterization of flexible proteins using small-angle X-ray scattering. Journal of the American Chemical Society 129, 5656-64.

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, 474.

Blazek, D., Barboric, M., Kohoutek, Jiri, Oven, I., and Peterlin, B. M. (2005). Oligomerization of HEXIM1 via 7SK snRNA and coiled-coil region directs the inhibition of P-TEFb. Nucleic acids research 33, 7000-10.

Bomsztyk, K., Denisenko, O., and Ostrowski, J. (2004). hnRNP K: one protein multiple processes. BioEssays : news and reviews in molecular, cellular and developmental biology 26, 629-38.

Buratowski, S. (2009). Progression through the RNA polymerase II CTD cycle. Molecular cell 36, 541-6.

Byers, S. a, Price, J. P., Cooper, J. J., Li, Qintong, and Price, David H (2005). HEXIM2, a HEXIM1-related protein, regulates positive transcription elongation factor b through association with 7SK. The Journal of biological chemistry 280, 16360-7.

Bélanger, F., Baigude, H., and Rana, T. M. (2009). U30 of 7SK RNA forms a specific photo-crosslink with Hexim1 in the context of both a minimal RNA-binding site and a full reconstituted 7SK/Hexim1/P-TEFb ribonucleoprotein complex. Journal of molecular biology 386, 1094-1107.

Cairns, B. R. (2009). The logic of chromatin architecture and remodelling at promoters. Nature 461, 193-8.

Cheetham, G. M., and Steitz, T. a (2000). Insights into transcription: structure and function of single-subunit DNA-dependent RNA polymerases. Current opinion in structural biology 10, 117-23.

Chelm, B. K., and Geiduschek, E. P. (1979). Gel electrophoretic separation of transcription complexes: an assay for RNA polymerase selectivity and a method for promoter mapping. Nucleic Acids Research 7, 1851-1867.

Chen, R. et al. (2008). PP2B and PP1alpha cooperatively disrupt 7SK snRNP to release P-TEFb for transcription in response to Ca2+ signaling. Genes & development 22, 1356-68.

Chen, R., Yang, Zhiyuan, and Zhou, Q. (2004). Phosphorylated positive transcription elongation factor b (P-TEFb) is tagged for inhibition through association with 7SK snRNA. The Journal of biological chemistry 279, 4153-60.

Cho, S., Schroeder, S., Kaehlcke, K., Kwon, H.-S., Pedal, A., Herker, E., Schnoelzer, M., and Ott, M. (2009). Acetylation of cyclin T1 regulates the equilibrium between active and inactive P-TEFb in cells. The EMBO journal 28, 1407-17.

Chowrira, B. M., Pavco, P. a, and McSwiggen, J. a (1994). In vitro and in vivo comparison of hammerhead, hairpin, and hepatitis delta virus self-processing ribozyme cassettes. The Journal of biological chemistry 269, 25856-64.

Clarkson, J., and Campbell, I. D. (2003). Studies of protein-ligand interactions by NMR. Biochemical Society transactions 31, 1006-9.

Cramer, P. (2000). Architecture of RNA Polymerase II and Implications for the Transcription Mechanism. Science 288, 640-649.

Cramer, P, Bushnell, D. a, and Kornberg, R. D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science (New York, N.Y.) 292, 1863-76.

Cuatrecasas, P., Wilchek, M., and Anfinsen, C. B. (1968). Selective enzyme purification by affinity chromatography. Proceedings of the National Academy of Sciences of the United States of America 61, 636-43.

Czudnochowski, N., Vollmuth, F., Baumann, S., Vogel-Bachmayr, K., and Geyer, M. (2010). Specificity of Hexim1 and Hexim2 complex formation with cyclin T1/T2, importin alpha and 7SK snRNA. Journal of molecular biology 395, 28-41.

Dames, S. a, Schönichen, A., Schulte, A., Barboric, M., Peterlin, B. M., Grzesiek, S., and Geyer, M. (2007). Structure of the Cyclin T binding domain of Hexim1 and molecular basis for its recognition of P-TEFb. Proceedings of the National Academy of Sciences of the United States of America 104, 14312-7.

Davanloo, P., Rosenberg, a H., Dunn, J. J., and Studier, F. W. (1984). Cloning and expression of the gene for bacteriophage T7 RNA polymerase. Proceedings of the National Academy of Sciences of the United States of America 81, 2035-9.

Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, Kevin M (2009). Accurate SHAPE-directed RNA structure determination. Proceedings of the National Academy of Sciences of the United States of America 106, 97-102.

Diebold, M.-L., Fribourg, S., Koch, M., Metzger, T., and Romier, C. (2011). Deciphering correct strategies for multiprotein complex assembly by co-expression: Application to complexes as large as the histone octamer. Journal of structural biology, 1-11.

Do, C. B., Woods, D. a, and Batzoglou, S. (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics (Oxford, England) 22, e90-8.

Document_not_found Document not found ((Karen M Wassarman et al. 2006)).

Doetsch, M., Fürtig, B., Gstrein, T., Stampfl, S., and Schroeder, R. (2011). The RNA annealing mechanism of the HIV-1 Tat peptide: conversion of the RNA into an annealing-competent conformation. Nucleic acids research 39, 4405-18.

Dreyfuss, G., Kim, V. N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nature reviews. Molecular cell biology 3, 195-205.

Dulac, C., Michels, A. a, Fraldi, A., Bonnet, F., Nguyen, Van Trung, Napolitano, G., Lania, L., and Bensaude, Olivier (2005). Transcription-dependent association of multiple positive transcription elongation factor units to a HEXIM multimer. The Journal of biological chemistry 280, 30619-29.

Durney, M. a, and D'Souza, V. M. (2010). Preformed protein-binding motifs in 7SK snRNA: structural and thermodynamic comparisons with retroviral TAR. Journal of molecular biology 404, 555-567.

D'Orso, I., and Frankel, A. D. (2010). RNA-mediated displacement of an inhibitory snRNP complex activates transcription elongation. Nature structural & molecular biology 17, 815-21.

Egloff, S., Herreweghe, E. Van, and Kiss, Tamás (2006). Regulation of polymerase II transcription by 7SK snRNA: two distinct RNA elements direct P-TEFb and HEXIM1 binding. Molecular and cellular biology 26, 630-42.

Egloff, S., and Murphy, S. (2008). Cracking the RNA polymerase II CTD code. Trends in genetics□ : TIG 24, 280-8.

Ehresmann, C., Baudin, F., Mougel, M., Romby, P., Ebel, J.-P., and Ehresmann, B. (1987). Probing the structure of RNAs in solution. Nucleic acids research 15, 9109-9128.

Eilebrecht, S., Brysbaert, G., Wegert, T., Urlaub, H., Benecke, B.-J., and Benecke, A. (2010). 7SK small nuclear RNA directly affects HMGA1 function in transcription regulation. Nucleic acids research 39, 2057-2072.

Espinosa, J. M. (2010). The meaning of pausing. Molecular cell 40, 507-8.

Espinoza, C. a, Allen, T. a, Hieb, A. R., Kugel, J. F., and Goodrich, J. a (2004). B2 RNA binds directly to RNA polymerase II to repress transcript synthesis. Nature structural & molecular biology 11, 822-9.

Falco, G. D., and Giordano, A. (2002). CDK9□ : From Basal Transcription to Cancer and AIDS. Cancer Biology & Therapy 1, 342-347.

Fong, B. a, Wu, W.-Y., and Wood, D. W. (2010). The potential role of self-cleaving purification tags in commercial-scale processes. Trends in biotechnology 28, 272-9.

Fraldi, A., Varrone, F., Napolitano, G., Michels, A. a, Majello, B., Bensaude, Olivier, and Lania, L. (2005). Inhibition of Tat activity by the HEXIM1 protein. Retrovirology 2, 42.

Fried, M., and Crothers, Donald M. (1981). Equlibria an kinetics of lac repressor-operator interaction by polyacrylamide gel electrophoresis. Nucleic Acids Research 9, 6505-6525.

Fu, J., Yoon, H.-G., Qin, J., and Wong, J. (2007). Regulation of P-TEFb elongation complex activity by CDK9 acetylation. Molecular and cellular biology 27, 4641-51.

Fuda, N. J., Ardehali, M. B., and Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature 461, 186-192.

Fürtig, B., Richter, C., Wöhnert, J., and Schwalbe, H. (2003). NMR spectroscopy of RNA. Chembiochem□ : a European journal of chemical biology 4, 936-62.

Galperin, M. Y., and Cochrane, G. R. (2009). Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. Nucleic acids research 37, D1-4.

Garber, M. E., Mayall, T. P., Suess, E. M., Meisenhelder, J., Thompson, N. E., and Jones, K. a (2000). CDK9 autophosphorylation regulates high-affinity binding of the human immunodeficiency virus type 1 tat-P-TEFb complex to TAR RNA. Molecular and cellular biology 20, 6958-69.

Gardner, Paul P, Wilm, A., and Washietl, S. (2005). A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic acids research 33, 2433-9.

Gilchrist, D. a, Santos, G. Dos, Fargo, D. C., Xie, B., Gao, Y., Li, L., and Adelman, K. (2010). Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation. Cell 143, 540-551.

Gilmour, D. S., and Lis, J. T. (1986). RNA Polymerase II Interacts with the Promoter Region of the Noninduced hsp7O Gene in Drosophila melanogaster Cells. Molecular and cellular biologybiology 6, 461-1659.

Goodrich, J. A., and Kugel, J. F. (2010). Dampening DNA binding. Rna Biology 7, 305-309.

Goodrich, J. A., and Kugel, J. F. (2006). Non-coding-RNA regulators of RNA polymerase II transcription. Nature reviews. Molecular cell biology 7, 612-616.

Gruber, A. R., Kilgus, C., Mosig, A., Hofacker, I. L., Hennig, W., and Stadler, P. F. (2008). Arthropod 7SK RNA. Molecular biology and evolution 25, 1923-30.

Gruber, A. R., Koper-Emde, D., Marz, M., et al. (2008). Invertebrate 7SK snRNAs. Journal of molecular evolution 66, 107-15.

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., and Hofacker, I. L. (2008). The Vienna RNA websuite. Nucleic acids research 36, W70-4.

Gurumurthy, M. et al. (2008). Nucleophosmin interacts with HEXIM1 and regulates RNA polymerase II transcription. Journal of molecular biology 378, 302-17.

Gürsoy, H.-C., Koper, D., and Benecke, B.-J. (2000). The Vertebrate 7S K RNA Separates Hagfish ( Myxine glutinosa ) and Lamprey ( Lampetra fluviatilis ). Journal of molecular evolution 50, 456-464.

Haaland, R. E., Herrmann, C. H., and Rice, A. P. (2005). siRNA depletion of 7SK snRNA induces apoptosis but does not affect expression of the HIV-1 LTR or P-TEFb-dependent cellular genes. Journal of cellular physiology 205, 463-70.

Hammarström, M., Hellgren, N., Berg, S. V. A. N. D. E. N., Berglund, H., and Härd, T. (2002). Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli. 313-321.

He, N., Jahchan, N. S., Hong, E., Li, Qiang, Bayfield, M. a, Maraia, R. J., Luo, Kunxin, and Zhou, Q. (2008). A La-related protein modulates 7SK snRNP integrity to suppress P-TEFb-dependent transcriptional elongation and tumorigenesis. Molecular cell 29, 588-99.

Hendrix, D. K., Brenner, S. E., and Holbrook, S. R. (2005). RNA structural motifs: building blocks of a modular biomolecule. Quarterly reviews of biophysics 38, 221-43.

Herreweghe, E. Van, Egloff, S., Goiffon, I., Jády, B. E., Froment, C., Monsarrat, B., and Kiss, Tamás (2007). Dynamic remodelling of human 7SK snRNP controls the nuclear level of active P-TEFb. The EMBO journal 26, 3570-80.

Hofacker, I. L., Fekete, M., and Stadler, P. F. (2002). Secondary structure prediction for aligned RNA sequences. Journal of molecular biology 319, 1059-66.

Hogg, J. R., and Collins, K. (2007). RNA-based affinity purification reveals 7SK RNPs with distinct composition and regulation. RNA (New York, N.Y.) 13, 868-880.

Holbrook, S. R. (2005). RNA structure: the long and the short of it. Current opinion in structural biology 15, 302-8.

Jang, M. K., Mochizuki, K., Zhou, M., Jeong, H.-S., Brady, J. N., and Ozato, K. (2005). The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. Molecular cell 19, 523-34.

Jeronimo, C. et al. (2007). Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. Molecular cell 27, 262-74.

Johnson, L. N. (2009). Protein kinase inhibitors: contributions from structure to clinical compounds. Quarterly reviews of biophysics 42, 1-40.

Karn, J. (1999). Tackling Tat. Journal of molecular biology 293, 235-54.

Kazantsev, a V., Rambo, R. P., Karimpour, S., Santalucia, J., Tainer, J. a, and Pace, N. R. (2011). Solution structure of RNase P RNA. RNA, 1159-1171.

Keel, A. Y. Rambo, R. P., Batey, R. T., and Kieft, J. S. (2007). A general strategy to solve the phase problem in RNA crystallography. Structure 15, 761-772.

Kettenberger, H., Eisenführ, A., Brueckner, F., Theis, M., Famulok, M., and Cramer, Patrick (2006). Structure of an RNA polymerase II-RNA inhibitor complex elucidates transcription regulation by noncoding RNAs. Nature structural & molecular biology 13, 44-8

Kladwang W., and Das, R. (2010). A mutate-and-map strategy for inferring base pairs in structured nucleic acids: proof of concept on a DNA/RNA helix. Biochemistry 49, 7414-7416.

Klocko, A. D., and Wassarman, Karen M. (2009). 6S RNA binding to Eσ70 requires a positively charged surface of σ70 region 4.2. Molecular microbiology 73, 152-164.

Kochetkov, S. N., Rusakova, E. E., and Tunitskaya, V. L. (1998). Recent studies of T7 RNA polymerase mechanism. FEBS letters 440, 264-7.

Kohoutek, Jiri (2009). P-TEFb- the final frontier. Cell division 4, 19.

Kohoutek, Jiri, Blazek, D., and Peterlin, B. M. (2006). Hexim1 sequesters positive transcription elongation factor b from the class II transactivator on MHC class II promoters. Proceedings of the National Academy of Sciences of the United States of America 103, 17349-54.

Kozin, M. B., and Svergun, D I (2001). Automated matching of high- and low-resolution structural models research papers Automated matching of high- and low-resolution structural models. October, 33-41.

Krueger, B. J. et al. (2008). LARP7 is a stable component of the 7SK snRNP while P-TEFb, HEXIM1 and hnRNP A1 are reversibly associated. Nucleic acids research 36, 2219-29.

Krueger, B. J., Varzavand, K., Cooper, J. J., and Price, David H (2010). The mechanism of release of P-TEFb and HEXIM1 from the 7SK snRNP by viral and cellular activators includes a conformational change in 7SK. PloS one 5, e12335.

Laederach, A., Das, R., Vicens, Q., Pearlman, S. M., Herschlag, D., and Altman, R. B. (2008). Semi-automated and rapid quantification of nucleic acid footprinting and structure mapping experiments. Nature protocols 3, 1395-1401.

Lebars, I., Martinez-Zapien, D., Durand, A., Coutant, J., Kieffer, B., and Dock-Bregeon, A.-C. (2010). HEXIM1 targets a repeated GAUC motif in the riboregulator of transcription 7SK and promotes base pair rearrangements. Nucleic acids research 38, 7749-63.

Leontis, N. B., and Westhof, E. (2003). Analysis of RNA motifs. Current Opinion in Structural Biology 13, 300-308.

Leroy, J. L., Guéron, M., Thomas, G., and Favre, a (1977). Role of divalent ions in folding of tRNA. European journal of biochemistry / FEBS 74, 567-74.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics (Oxford, England) 23, 127-8.

Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic acids research 39, 475-478.

Levine, M. (2011). Paused RNA polymerase II as a developmental checkpoint. Cell 145, 502-11.

Li, B., Carey, M., and Workman, J. L. (2007). The role of chromatin during transcription. Cell 128, 707-19.

Li, Qintong, Cooper, J. J., Altwerger, G. H., Feldkamp, M. D., Shea, M. a, and Price, David H (2007). HEXIM1 is a promiscuous double-stranded RNA-binding protein and interacts with RNAs in addition to 7SK in cultured cells. Nucleic acids research 35, 2503-12.

Li, Qintong, Price, J. P., Byers, S. a, Cheng, D., Peng, J., and Price, David H (2005). Analysis of the large inactive P-TEFb complex indicates that it contains one 7SK molecule, a dimer of HEXIM1 or HEXIM2, and two P-TEFb molecules containing Cdk9 phosphorylated at threonine 186. The Journal of biological chemistry 280, 28819-26.

Lindgreen, S, Gardner, P P, and Krogh, a (2006). Measuring covariation in RNA alignments: physical realism improves information measures. Bioinformatics (Oxford, England) 22, 2988-95.

Lindgreen, Stinus, Gardner, Paul P, and Krogh, A. (2007). MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. Bioinformatics (Oxford, England) 23, 3304-11.

Luo, Y., Kurz, J., MacAfee, N., and Krause, M. O. (1997). C-myc deregulation during transformation induction: involvement of 7SK RNA. Journal of cellular biochemistry 64, 313-27.

Mariner, P. D., Walters, R. D., Espinoza, C. a, Drullinger, L. F., Wagner, S. D., Kugel, J. F., and Goodrich, J. a (2008). Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. Molecular cell 29, 499-509.

Markert, A., Grimm, M., Martinez, J., Wiesner, J., Meyerhans, A., Meyuhas, O., Sickmann, A., and Fischer, U. (2008). The La-related protein LARP7 is a component of the 7SK ribonucleoprotein and affects transcription of cellular and viral polymerase II genes. EMBO reports 9, 569-75.

Martinez-Rucobo, F. W., Sainsbury, S., Cheung, A. C. M., and Cramer, Patrick (2011). Architecture of the RNA polymerase-Spt4/5 complex and basis of universal transcription processivity. The EMBO journal 30, 1302-10.

Marz, M., Donath, A., Verstraete, N., Nguyen, Van Trung, Stadler, P. F., and Bensaude, Olivier (2009). Evolution of 7SK RNA and its protein partners in metazoa. Molecular biology and evolution 26, 2821-30.

Mason, J. M., and Arndt, K. M. (2004). Coiled coil domains: stability, specificity, and biological implications. Chembiochem□ : a European journal of chemical biology 5, 170-6.

Matera, A. G., Terns, R. M., and Terns, M. P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nature reviews. Molecular cell biology 8, 209-20.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, Michael, 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, 7287-92.

Mathews, D. H., Moss, W. N., and Turner, D. H. (2010). Folding and finding RNA secondary structure. Cold Spring Harbor perspectives in biology 2, a003665.

Mathews, D. H., and Turner, D. H. (2006). Prediction of RNA secondary structure by free energy minimization. Current opinion in structural biology 16, 270-8.

Meinhart, A., Kamenski, T., Hoeppner, S., Baumli, S., and Cramer, Patrick (2005). A structural perspective of CTD function. Genes & development 19, 1401-15.

Meinnel, T., Mechulam, Y., and Fayat, G. (1988). Fast puification of a functional elongator tRNAmet expressed from a synthetic gene in vivo. Nucleic acids research 16, 8095-8096.

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. Nature reviews. Genetics 10, 155-160.

Merino, E. J., Wilkinson, K. a, Coughlan, J. L., and Weeks, Kevin 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, 4223-31.

Mertens, H. D. T., and Svergun, Dmitri I (2010). Structural characterization of proteins and complexes using small-angle X-ray solution scattering. Journal of structural biology 172, 128-41.

Michels, A. A. et al. (2004). Binding of the 7SK snRNA turns the HEXIM1 protein into a P-TEFb (CDK9/cyclin T) inhibitor. The EMBO journal 23, 2608-19.

Michels, A., Nguyen, Van Trung, Fraldi, A., Edwards, M., Lania, L., and Bensaude, Olivier (2003). MAQ1 and 7SK RNA Interact with CDK9 / Cyclin T Complexes in a Transcription-Dependent Manner. Molecular and cellular biology 23, 4859-4869.

Moon, A. F., Mueller, G. a, Zhong, X., and Pedersen, L. C. (2010). A synergistic approach to protein crystallization: combination of a fixed-arm carrier with surface entropy reduction. Protein science□ : a publication of the Protein Society 19, 901-13.

Moore, P. B. (1999). STRUCTURAL MOTIFS IN RNA. Annual review of biochemistry 68, 287-300.

Mortimer, S. a, and Weeks, Kevin M (2008). Time-resolved RNA SHAPE chemistry. Journal of the American Chemical Society 130, 16178-80.

Mortimer, S. a, and Weeks, Kevin M (2007). A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. Journal of the American Chemical Society 129, 4144-5.

Muniz, L., Egloff, S., Ughy, B., Jády, B. E., and Kiss, Tamás (2010). Controlling cellular P-TEFb activity by the HIV-1 transcriptional transactivator Tat. PLoS pathogens 6, e1001152.

Nasalean, L., Stombaugh, J., Zirbel, C. L., and Leontis, N. B. (2009). RNA 3D Structural Motifs: Definition, Identification, Annotation, and Database Searching. In Non-Protein Coding RNAs N. G. Walter, S. A. Woodson, and R. T. Batey, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 1-26.

Nechaev, S., Fargo, D. C., Santos, G. dos, Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science (New York, N.Y.) 327, 335-8.

Nechaev, S., and Adelman, K. (2010). Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. Biochimica et biophysica acta 1809, 34-45.

Nechaev, S., and Adelman, K. (2008). Promoter-proximal PolII. Cell Cycle 7, 1539-1544.

Nguyen, V T, Kiss, T, Michels, a a, and Bensaude, O (2001). 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. Nature 414, 322-5.

Ogba, N., Chaplin, L. J., Doughman, Y. Q., Fujinaga, K., and Montano, M. M. (2008). HEXIM1 regulates 17beta-estradiol/estrogen receptor-alpha-mediated expression of cyclin D1 in mammary cells via modulation of P-TEFb. Cancer research 68, 7015-24.

Orphanides, G., Lagrange, T., and Reinberg, D. (1996). The general transcription factors of RNA polymerase II. Genes & Development 10, 2657-2683.

Ouchida, R., Kusuhara, M., Shimizu, N., Hisada, T., Makino, Y., Morimoto, C., Handa, H., Ohsuzu, F., and Tanaka, H. (2003). Suppression of NF-kappaB-dependent gene expression by a hexamethylene bisacetamide-inducible protein HEXIM1 in human vascular smooth muscle cells. Genes to cells□ : devoted to molecular & cellular mechanisms 8, 95-107.

O'Gorman, W., Thomas, B., Kwek, K. Y., Furger, A., and Akoulitchev, A. (2005). Analysis of U1 small nuclear RNA interaction with cyclin H. The Journal of biological chemistry 280, 36920-5.

Pantano, S., Marcello, A., Sabò, A., Ferrari, A., Pellegrini, V., Beltram, F., Giacca, M., and Carloni, P. (2005). A Model of N-Terminal Cyclin T1 Based on FRET Experiments. Journal of Theoretical Medicine 6, 73-79.

Parisien, M., and Major, F. (2008). The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature 452, 51-5.

Patel, D. J. (1999). Adaptive recognition in RNA complexes with peptides and protein modules. Current Opinion in Structural Biology 9, 74-87.

Peterlin, B. M., and Price, David H (2006). Controlling the elongation phase of transcription with P-TEFb. Molecular cell 23, 297-305.

Phatnani, H. P., and Greenleaf, A. L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. Genes & development 20, 2922-36.

Ponchon, L., and Dardel, F. (2007). Recombinant RNA technology□ : the tRNA scaffold. Nature Methods 4, 571-576.

Price, D. H. (2000). P-TEFb, a Cyclin-Dependent Kinase Controlling Elongation by RNA Polymerase II. Molecular and Cellular Biology 20, 2629-2634.

Putnam, C. D., Hammel, M., Hura, G. L., and Tainer, J. a (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Quarterly reviews of biophysics 40, 191-285.

Rambo, Robert P., and Tainer, J. A. (2011a). Bridging the solution divide: comprehensive structural analyses of dynamic RNA, DNA, and protein assemblies by small angle X- ray scattering. Current opinion in structural biology 20, 128-137.

Rambo, Robert P, and Tainer, J. A. (2011b). Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. Biopolymers 95, 559-71.

Rambo, Robert P, and Tainer, J. A. (2010). Improving small-angle X-ray scattering data for structural analyses of the RNA world. RNA 16, 638-646.

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

Robertson, M. P., and Scott, W. G. (2008). A general method for phasing novel complex RNA crystal structures without heavy-atom derivatives. Acta crystallographica. Section D, Biological crystallography D64, 738-44.

Ryder, S. P., Recht, M. I., and Williamson, J. R. (2008). Quantitative analysis of RNA-Protein Interactions by gel mobility shift. Methods in Mollecular Biology 488, 99-115.

Saunders, A., Core, L. J., and Lis, J. T. (2006). Breaking barriers to transcription elongation. Nature reviews. Molecular cell biology 7, 557-67.

Schramm, L., and Hernandez, N. (2002). Recruitment of RNA polymerase III to its target promoters. Genes & development 16, 2593-620.

Schulte, A., Czudnochowski, N., Barboric, M., Schönichen, A., Blazek, D., Peterlin, B. M., and Geyer, M. (2005). Identification of a cyclin T-binding domain in Hexim1 and biochemical analysis of its binding competition with HIV-1 Tat. The Journal of biological chemistry 280, 24968-77.

Schönichen, A., Bigalke, J. M., Urbanke, C., Grzesiek, S., Dames, S. a, and Geyer, M. (2010). A flexible bipartite coiled coil structure is required for the interaction of Hexim1 with the P-TEFB subunit cyclin T1. Biochemistry 49, 3083-91.

Sedore, S. C., Byers, S. a, Biglione, S., Price, J. P., Maury, W. J., and Price, David H (2007). Manipulation of P-TEFb control machinery by HIV: recruitment of P-TEFb from the large form by Tat and binding of HEXIM1 to TAR. Nucleic acids research 35, 4347-58.

Seemann, S. E., Gorodkin, J., and Backofen, R. (2008). Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. Nucleic acids research 36, 6355-62.

Shimizu, N. et al. (2005). HEXIM1 forms a transcriptionally abortive complex with glucocorticoid receptor without involving 7SK RNA and positive transcription elongation factor b. Proceedings of the National Academy of Sciences of the United States of America 102, 8555-60.

Sikorski, T. W., and Buratowski, S. (2009). The Basal Initiation Machinery: Beyond the General Transcription Factors. Current opinion in cell biology 21, 344-351.

Smith, C. A., Calabro, V., and Frankel, A. D. (2000). An RNA-Binding Chameleon. Molecular cell 6, 1067-1076.

Smith, C., Heyne, S., Richter, A. S., Will, S., and Backofen, R. (2010). Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA. Nucleic acids research 38, W373-7.

Smyth, D. R., Mrozkiewicz, M. K., Mcgrath, W. J., Listwan, P., and Kobe, B. (2003). Crystal structures of fusion proteins with large-affinity tags. Protein Science, 1313-1322.

Sobhian, B., Laguette, N., Yatim, A., Nakamura, M., Levy, Y., Kiernan, R., and Benkirane, M. (2010). HIV-1 Tat assembles a multifunctional transcription elongation complex and stably associates with the 7SK snRNP. Molecular cell 38, 439-51.

Studier, F. (2005). Protein production by auto-induction in high-density shaking cultures. Protein Expression and Purification 41, 207-234.

Svergun, Dmitri I, and Koch, Michel H J (2003). Small-angle scattering studies of biological macromolecules in solution. Reports on Progress in Physics 66, 1735-1782.

Svergun, D., Barberato, C., and Koch, M H J (1995). CRYSOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates D . Svergun , C . Barberato and M . H . J . Koch. 768-773.

Tahirov, T. H., Babayeva, N. D., Varzavand, K., Cooper, J. J., C, S., and Price, David H (2010). NIH Public Access. 465, 747-751.

Tahirov, T. H., Babayeva, N. D., Varzavand, K., Cooper, J. J., Sedore, S. C., and Price, David H (2010). Crystal structure of HIV-1 Tat complexed with human P-TEFb. Nature 465, 747-51.

Terpe, K. (2003). Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. Applied microbiology and biotechnology 60, 523-33.

Thomas, M. C., and Chiang, C.-M. (2006). The general transcription machinery and general cofactors. Critical reviews in biochemistry and molecular biology 41, 105-78.

Trotochaud, A. E., and Wassarman, Karen M (2004). 6S RNA Function Enhances Long-Term Cell Survival. Journal of Bacteriology 186, 4978-4985.

Trotochaud, A. E., and Wassarman, Karen M (2006). 6S RNA regulation of pspF transcription leads to altered cell survival at high pH. Journal of bacteriology 188, 3936-43.

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, D280-2.

Vicens, Q., Gooding, A. R., Laederach, A., and Cech, T. R. (2007). Local RNA structural changes induced by crystallization are revealed by SHAPE. RNA Journal 13, 536-548.

Vitali, J., Ding, J., Jiang, J., Zhang, Y., Krainer, A. R., and Xu, R.-M. (2002). Correlated alternative side chain conformations in the RNA-recognition motif of heterogeneous nuclear ribonucleoprotein A1. Nucleic acids research 30, 1531-8.

Volkov, V. V., and Svergun, Dmitri I (2003). Uniqueness of ab initio shape determination in small-angle scattering. Journal of Applied Crystallography 36, 860-864.

Walker, S. C., Avis, J. M., and Conn, G. L. (2003). General plasmids for producing RNA in vitro transcripts with homogeneous ends. Nucleic acids research 31, 1-6.

Wang, Q., Young, T. M., Mathews, M. B., and Pe'ery, T. (2007). Developmental regulators containing the I-mfa domain interact with T cyclins and Tat and modulate transcription. Journal of molecular biology 367, 630-46.

Wassarman, D. a, and Steitz, J. a (1991). Structural analyses of the 7SK ribonucleoprotein (RNP), the most abundant human small RNP of unknown function. Molecular and cellular biology 11, 3432-45.

Wassarman, K M, and Storz, G. (2000). 6S RNA regulates E. coli RNA polymerase activity. Cell 101, 613-23.

Wassarman, Karen M (2007). 6S RNA: a regulator of transcription. Molecular microbiology 65, 1425-31.

Weeks, K M, and Crothers, D M (1993). Major groove accessibility of RNA. Science (New York, N.Y.) 261, 1574-7.

Weeks, Kevin M (2010). Advances in RNA structure analysis by chemical probing. Current opinion in structural biology 20, 295-304.

Werner, M., Thuriaux, P., and Soutourina, J. (2009). Structure-function analysis of RNA polymerases I and III. Current opinion in structural biology 19, 740-5.

Westhof, E., and Romby, P. (2010). The RNA structurome: high-throughput probing. Nature Methods 7, 965-967.

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

Wilkinson, K. a, Merino, E. J., and Weeks, Kevin M (2006). Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nature protocols 1, 1610-6.

Wittmann, B. M., Fujinaga, K., Deng, H., Ogba, N., and Montano, M. M. (2005). The breast cell growth inhibitor, estrogen down regulated gene 1, modulates a novel functional interaction between estrogen receptor alpha and transcriptional elongation factor cyclin T1. Oncogene 24, 5576-88.

Wittmann, B. M., Wang, N., and Montano, M. M. (2003). Identification of a Novel Inhibitor of Breast Cell Growth That Is Down-Regulated by Estrogens and Decreased in Breast Tumors. Cancer Research 63, 5151-5158.

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, W605-10.

Xu, R. M., Jokhan, L., Cheng, X., Mayeda, a, and Krainer, a R. (1997). Crystal structure of human UP1, the domain of hnRNP A1 that contains two RNA-recognition motifs. Structure (London, England □ : 1993) 5, 559-70.

Xue, Y., Yang, Zhiyuan, Chen, R., and Zhou, Q. (2010). A capping-independent function of MePCE in stabilizing 7SK snRNA and facilitating the assembly of 7SK snRNP. Nucleic acids research 38, 360-9.

Yakovchuk, P., Goodrich, J. a, and Kugel, J. F. (2009). B2 RNA and Alu RNA repress transcription by disrupting contacts between RNA polymerase II and promoter DNA within assembled complexes. Proceedings of the National Academy of Sciences of the United States of America 106, 5569-74.

Yang, S., Parisien, M., Major, F., and Roux, B. (2010). RNA structure determination using SAXS data. The journal of physical chemistry. B 114, 10039-48.

Yang, Z, Zhu, Q., Luo, K, and Zhou, Q. (2001). The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. Nature 414, 317-22.

Yang, Zhiyuan, Yik, J. H. N., Chen, R., He, N., Jang, M. K., Ozato, K., and Zhou, Q. (2005). Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. Molecular cell 19, 535-45.

Yeang, C.-H., Darot, J. F. J., Noller, H. F., and Haussler, D. (2007). Detecting the coevolution of biosequences--an example of RNA interaction prediction. Molecular biology and evolution 24, 2119-31.

Yik, J. H. N., Chen, R., Nishimura, R., Jennings, J. L., Link, A. J., and Zhou, Q. (2003). Inhibition of P-TEFb (CDK9/Cyclin T) kinase and RNA polymerase II transcription by the coordinated actions of HEXIM1 and 7SK snRNA. Molecular cell 12, 971-82.

Yik, J. H. N., Chen, R., Pezda, A. C., Samford, C. S., and Zhou, Q. (2004). A Human Immunodeficiency Virus Type 1 Tat-Like Arginine-Rich RNA-Binding Domain Is Essential for HEXIM1 To Inhibit RNA Polymerase II Transcription through 7SK snRNA-Mediated Inactivation of P-TEFb. Molecular and cellular biology 24, 5094-5105.

Yik, J. H. N., Chen, R., Pezda, A. C., and Zhou, Q. (2005). Compensatory contributions of HEXIM1 and HEXIM2 in maintaining the balance of active and inactive positive transcription elongation factor b complexes for control of transcription. The Journal of biological chemistry 280, 16368-76.

Yoshikawa, N., Shimizu, N., Sano, M., Ohnuma, K., Iwata, S., Hosono, O., Fukuda, K., Morimoto, C., and Tanaka, H. (2008). Role of the hinge region of glucocorticoid receptor for HEXIM1-mediated transcriptional repression. Biochemical and biophysical research communications 371, 44-9.

Young, T. M., Tsai, M., Tian, B., Mathews, M. B., and Pe'ery, T. (2007). Cellular mRNA activates transcription elongation by displacing 7SK RNA. PloS one 2, e1010.

Zhou, Q., and Yik, J. H. N. (2006). The Yin and Yang of P-TEFb regulation: implications for human immunodeficiency virus gene expression and global control of cell growth and differentiation. Microbiology and molecular biology reviews: MMBR 70, 646-59.

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