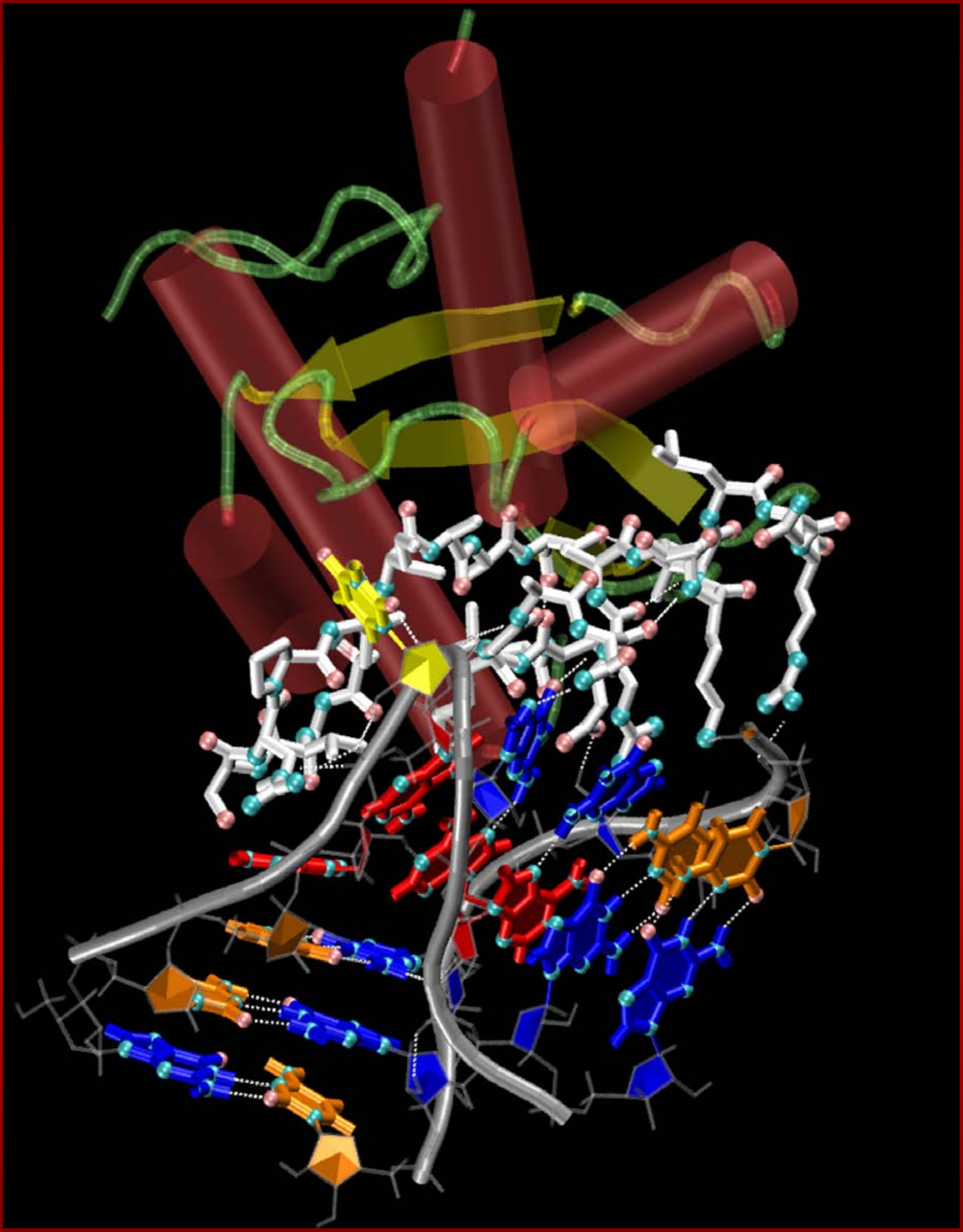
Molecular motions at the 5' stem-loop of U4 snRNA: Implications for U4/U6 snRNP assembly



Ph.D. thesis Vlad Cojocaru

Molecular motions at the 5' stem-loop of U4 snRNA: Implications for U4/U6 snRNP assembly

PhD Thesis

in partial fulfillment of the requirements
for the degree "Doctor of Philosophy (PhD)"
in the Molecular Biology Program
at the Georg August University Göttingen,
Faculty of Biology

2005

submitted by

Vlad Cojocaru

born in

Arad, Romania

"The intuitive mind is a sacred
gift and the rational mind
is a faithful servant.
We have created a society that
honors the servant and
has forgotten the gift"
Albert Einstein

"He who climbs upon the highest
mountains laughs at all
tragedies, real or
imaginary"
Friedrich Nietzsche

Although life has brought me to distant lands,
has showed me different cultures, has taught
me how to treasure different people and
has guided me into a great unknown,
I will always carry, deep in my soul,
the beauty of the Carpathians with
their enchanted valleys, the sanctity
of Bucovina's monasteries, the
treasures hiding in caves
untouched by humans
or the sound of the
traditional "Aşa-i românul".

This PhD thesis is dedicated to
my mom, Rodica, my dad Nicu,
my sister Ioana and
my love, Kerstin.

Această dizertaţie este dedicată
mamei mele Rodica,
tatălui meu Nicu,
sorei mele Ioana
si iubitei mele Kerstin

The present PhD thesis is based on research that I performed at:

Max Planck Institute for Biophysical Chemistry,
Department of Molecular Biology

Am Fassberg 11,
37077 Göttingen, Germany

Guided by:

Dr. Thomas M. Jovin
Reinhard Klement

Doctoral committee members:

Dr. Thomas M. Jovin
Prof. Dr. Reinhard Lührmann
Prof. Dr. Hans-Joachim Fritz

Table of contents

List of publications .... - 9 -

Acknowledgments .... - 10 -

Abstract .... - 11 -

1. Introduction .... - 12 -

1.1. Spliceosome assembly and function .... - 12 -

1.2. The U4/U6·U5 tri-snRNP .... - 15 -

1.3. The 15.5K protein .... - 18 -

1.4. The RNA K-turn motif .... - 23 -

1.5. Protein-assisted RNA folding .... - 26 -

1.6. Computer simulations: benefits and challenges .... - 28 -

1.7. Conformational parameters of nucleic acids .... - 29 -

1.8. Fluorescence of 2-aminopurine .... - 30 -

2. Theoretical background .... - 30 -

2.1. Introduction to molecular modeling .... - 30 -

2.2. Empirical force fields: molecular mechanics .... - 31 -

2.3. Force field parameter development .... - 33 -

2.3.1. Atom types, bond, angle and dihedral parameters .... - 34 -

2.3.2. Van der Waals (VDW) parameters .... - 36 -

2.3.3. Derivation of partial atomic charges .... - 37 -

2.4. Energy minimization .... - 39 -

2.5. Molecular dynamics simulations .... - 42 -

2.5.1. Choosing the time step .... - 45 -

2.5.2. The SHAKE / RATTLE algorithm .... - 45 -

2.5.3. Periodic boundary conditions .... - 47 -

2.5.4. The Particle Mesh Ewald method .... - 48 -

2.5.5. Constant temperature dynamics .... - 52 -

2.5.6. Constant pressure dynamics .... - 53 -

2.6. Locally enhanced sampling .... - 54 -

2.7. Electrostatic potential surfaces .... - 57 -

3. Materials and methods - 59 -
3.1. RNA constructs - 59 -
3.2. Molecular dynamics simulations - 59 -
3.3. Locally enhanced sampling - 60 -
3.4. Essential dynamics - 61 -
3.5. Generating electrostatic potential surfaces - 61 -
3.6. Steady-state fluorescence spectroscopy - 62 -
4. Aims, motivation and relevance - 63 -
5. Results - 64 -
5.1. Protein-assisted folding of KtU4 - 64 -
5.1.1. General analysis of the simulations - 65 -
5.1.2. Flexibility in the purine-rich internal loop - 67 -
5.1.3. Flexibility in the NC-stem - 71 -
5.2. Dynamics of the free KtU4 - 73 -
5.2.1. Opening-closing of KtU4 - 73 -
5.2.2. Opening of G-A base pairs - 75 -
5.2.3. Modified stacking patterns in the internal loop - 78 -
5.2.4. Opening of KtU4 is correlated with backbone flexibility in the stems - 80 -
5.2.5. Essential dynamics of the LES4 trajectory - 82 -
5.3. The role of the external UUUAU pentaloop - 84 -
5.3.1. Trajectories of the external loop in the bound and unbound RNAs - 84 -
5.3.2. Trajectories of the UGAA and UUAAUU loops in the bound RNAs - 86 -
5.3.3. Trajectories of the bound and unbound Kt15 - 88 -
5.4. Flexibility of the 15.5K protein - 91 -
5.4.1. Flexible regions in the 15.5K protein - 91 -
5.4.2. The protein-RNA interface - 94 -
5.4.3. PRO62 induces a hairpin-like structure - 96 -
5.4.4. The flexibility of α-helix 5 - 98 -
5.5. Steady-state fluorescence spectroscopy - 99 -
6. Discussion - 101 -
6.1. Protein-assisted RNA folding - 101 -
6.2. The role of G-A base pair formation in the folding of KtU4 - 103 -

6.3. The role of backbone flexibility in the folding of KtU4 .... - 106 -
6.4. RNA folding and U4/U6 snRNP assembly................................. - 107 -
6.5. Computer simulations of protein-assisted RNA folding .... - 108 -
6.6. Perspectives for simulating protein-assisted RNA folding...... - 110 -
6.7. Dynamics of 15.5K protein and U4/U6 snRNP assembly .... - 110 -
6.8. Structural changes of KtU4 monitored by 2AP fluorescence .. - 112 -
7. Summary and conclusions.. - 114 -
8. References .... - 116 -
List of figures.. - 130 -
List of tables .... - 131 -
Curriculum vitae .... - 132 -

List of publications

Scientific articles:

1. Vlad Cojocaru, Reinhard Klement, and Thomas M. Jovin. 2005. Loss of G-A base pairs is insufficient for achieving a large opening of U4 snRNA K-turn motif. *Nucleic Acids Research.* **33**:3435-3446.

2. Vlad Cojocaru, Stephanie Nottrott, Reinhard Klement, and Thomas M. Jovin. 2005. The snRNP 15.5K protein folds its cognate K-turn RNA: A combined theoretical and biochemical study. *RNA* **11**:197-209.

3. Jurg Stebler, Derek Spieler, Krasimir Slanchev, Klathleen A. Moyneaux, Ulrike Richter, Vlad Cojocaru, Victor Tarabykin, Chris Wylie, Michael Kessel, and Erez Raz 2004. Primordial germ cell migration in the chick and mouse embryo: the role of the chemokine SDF-1/CXCL12. *Dev Biol.* **272**:351-61.

Poster abstracts:

Vlad Cojocaru, Stephanie Nottrott, Reinhard Klement, and Thomas M. Jovin. 2004. The snRNP 15.5K protein folds its cognate K-turn RNA: A combined theoretical and biochemical study – presented at: (i) ‘Structures in Biology’ conference in Heidelberg, Germany, November 2004 (abstract book), and (ii) ‘RNA Structure and Function’ meeting in Edinburgh, Scotland, December 2004 (abstract book).

Oral presentations:

Vlad Cojocaru. The snRNP 15.5K protein folds its cognate K-turn RNA: A combined theoretical and biochemical study – presented at the ‘Computer Simulation & Theory of Macromolecules’ workshop; Hünfeld, Germany, April 22-24, 2005 (abstract book)

Acknowledgments

I am deeply grateful to Tom Jovin for guiding me throughout my Ph.D studies, to Reinhard Klement for introducing me into the world of Molecular Modeling and Simulation, and to the entire Molecular Biology department at Max Planck Institute for Biophysical Chemistry for great atmosphere and collegiality.

I would like to thank Reinhard Lührmann for constant support and many valuable advices, Stephanie Nottrott for a fruitful collaboration reflected in our RNA paper, Eric Westhof and Pascal Auffinger for the wonderful week spent in Strasbourg and for discussion regarding the simulations, Anne-Marie Schultz, Ping Li and Teresa Calomagno for sharing their views and experimental data and Bernd Rieger for the matlab crash course.

My participation in the International PhD Program Molecular Biology - International Max Planck Research School at the Georg August University has been a great experience and enriched my scientific background. Therefore, I express my gratitude to Stefffen Burkhard and the coordination office, to Thomas Pieler for having confidence in me regardless of my lack of laboratory experience and poor English during the interview, to Hans-Joachim Fritz for being a member of my thesis committee, and to Gabi, Olga, Ralf, Patrick, Tabrez, Ajay and Roland for a successful “Horizons Team”.

Special thanks to Tudor Oprea (currently at New Mexico University) for his lectures, advices and chats during my last year at the West University Timişoara where he guided me and his other students into the world of scientific research.

I would like to thank all my friends, apologize for not giving their names and assure them of my sincere friendship.

Last but not least, this thesis would have never been written if my mom, Rodica, my dad, Nicu and my sister, Ioana haven’t been by my side in all the tough decisions that I have taken including leaving Romania in the autumn of 2000. I am very lucky to have Kerstin fulfilling my life. Her love gave me strength and confidence and I thank her for that.

Abstract

The human 15.5K protein binds to the 5' stem-loop of U4 snRNA (KtU4), promotes the assembly of the spliceosomal U4/U6 snRNP, and is required for the recruitment of the 61K protein and the 20/60/90K protein complex to the U4 snRNA. In the crystallographic structure of the 15.5K-U4 snRNA complex, the RNA fold belongs to the family of kink turn (K-turn) motifs. This motif has a kink in the phosphodiester backbone that causes a sharp turn in the RNA helix. Two stems are connected by a purine-rich internal asymmetric loop, containing a flipped out uridine and two tandem sheared G-A base pairs. The shorter stem is attached to an external pentaloop. Using molecular dynamics simulations, I showed that the folding of KtU4 is assisted by protein binding. Conformational transitions such as the inter-conversion between alternative purine stacking schemes, the loss of G-A base pairs, and the opening of the K-turn (k-e motion) occurred only in the free RNA. The simulations provided the first atomic details of K-turn dynamics and were in excellent agreement with experimental data obtained by chemically probing the RNA structure and from single molecule FRET studies. In the free RNA, the k-e motion was triggered both by loss of G-A base pairs in the internal loop and backbone flexibility in the stems. However, the loss of G-A base pairs alone was insufficient for achieving a large opening of the free RNA. Essential dynamics showed that the loss of G-A base pairs is correlated along the first mode but anti-correlated along the third mode with the k-e motion. Based on these findings, I conclude that G-A base pair formation occurs upon binding to the 15.5K protein, thereby stabilizing a selective orientation of the stems.

The external loop was not revealed in the crystallographic structure of the 15.5K-KtU4 complex. In the simulations, it adopted a specific orientation which did not persist in the unbound RNA and did not form when the natively occurring external loop was replaced by different loops or by an extended helix. I propose that the lack of stacking interactions between the last base pair of the stem and the adjacent nucleotide in the external loop are important for the correct folding of the RNA and might play a role in the subsequent binding of the 61K protein to the U4 snRNA.

1. Introduction

1.1. Spliceosome assembly and function

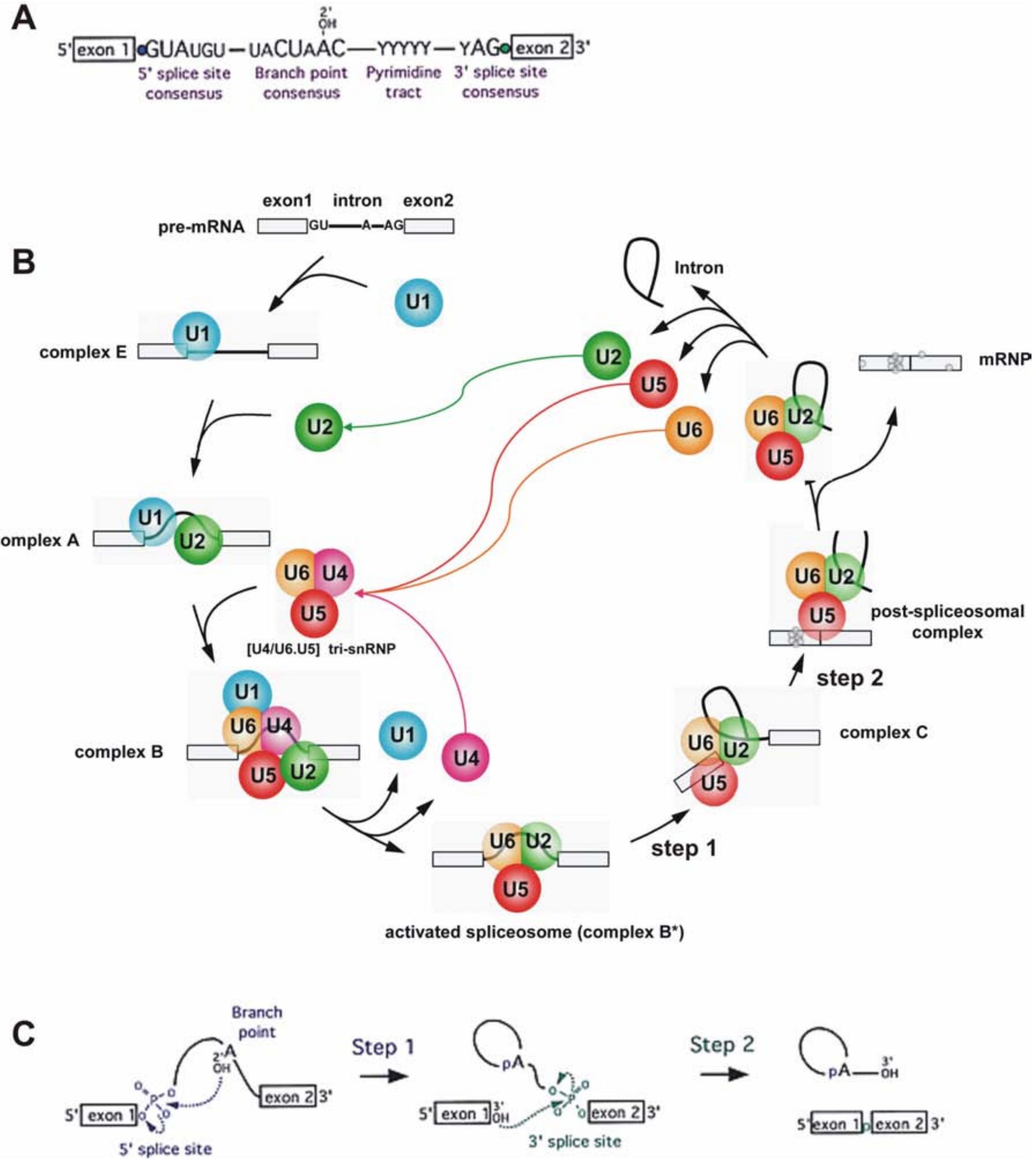


Figure 1: Overview of the splicing process.
(A) Consensus sequences within the mRNA introns [taken from Collins and Guthrie (5)]. (B) A schematic view of the assembly pathway highlighting all the intermediate steps; U1, U2, U4, U5 and U6 are the major spliceosome components (snRNPs) (courtesy to Reinhard Lührmann). (C) The mechanism of splicing reactions [taken from Collins and Guthrie (5)].

The spliceosome is the complex macromolecular machinery that catalytically removes the non-coding sequences (introns) from the newly transcribed messenger RNA precursors (pre-mRNA). Most pre-mRNA introns are removed by the U2-dependent (major) spliceosome, which is composed of the small nuclear ribonucleoprotein particles (snRNPs) U1, U2, U4/U6 and U5 and numerous non-snRNP protein factors. A rare class of pre-mRNA introns is removed by the U12-dependent (minor) spliceosome, which contains a different set of snRNPs, namely U11, U12 and U4atac/U6atac, but shares U5 snRNP with the major spliceosome [reviewed by Burge et al. (1)]. Each snRNP consists of specific and non-specific proteins that wrap around the RNA core. Different snRNPs recognize the introns via several consensus sequences (Figure 1A): (i) the 5' splice site, (ii) the branch point adenosine; (iii) the poly-pyrimidine track located downstream from the branch point, and (iv) the 3' splice-site. U1 joins the intron at the 5' splice-site and U2 recognizes the adenosine at the branch point via different splicing factors. U4, U5 and U6 are pre-assembled into a tri-snRNP before they join the spliceosome. During the activation step, U1 and U4 are released, and the catalytically active spliceosome acquires the capability to perform the splicing reactions. A detailed scheme of the major spliceosome assembly pathway showing all the intermediate complexes formed is presented in Figure 1B. The chemistry of the splicing process consists of two transesterification steps (Figure 1C): (i) the 2' OH group of the adenosine branch point attacks the phosphate group of the 5' splice site, and (ii) the terminal OH group of exon 1 formed as a leaving group in the first step, attacks the phosphate group at the 3' splice site, releasing the intron as a lariat [for detailed reviews see (2-4)].

Dramatic RNA structural rearrangements occur during spliceosome activation (Figure 2). Before joining the spliceosome, U4 base-pairs with U6 snRNA forming a RNA duplex while U2 is associated with the adenosine branch point of the pre-mRNA. Upon activation, U4 is released from the spliceosome and U6 base-pairs with U2 snRNA forming the catalytically active RNA fold. It has been long debated whether the spliceosome is actually a ribozyme. A network of critical RNA interactions, between the snRNAs with one another and with the pre-mRNA substrate, that could perform all the functions required for catalysis has been identified. U6 and U2 snRNAs interact with the intron near two sites of chemistry – the 5' splice site and the branchpoint via the 5' splice site helix and BP helix respectively (Figure 2, upper panel).

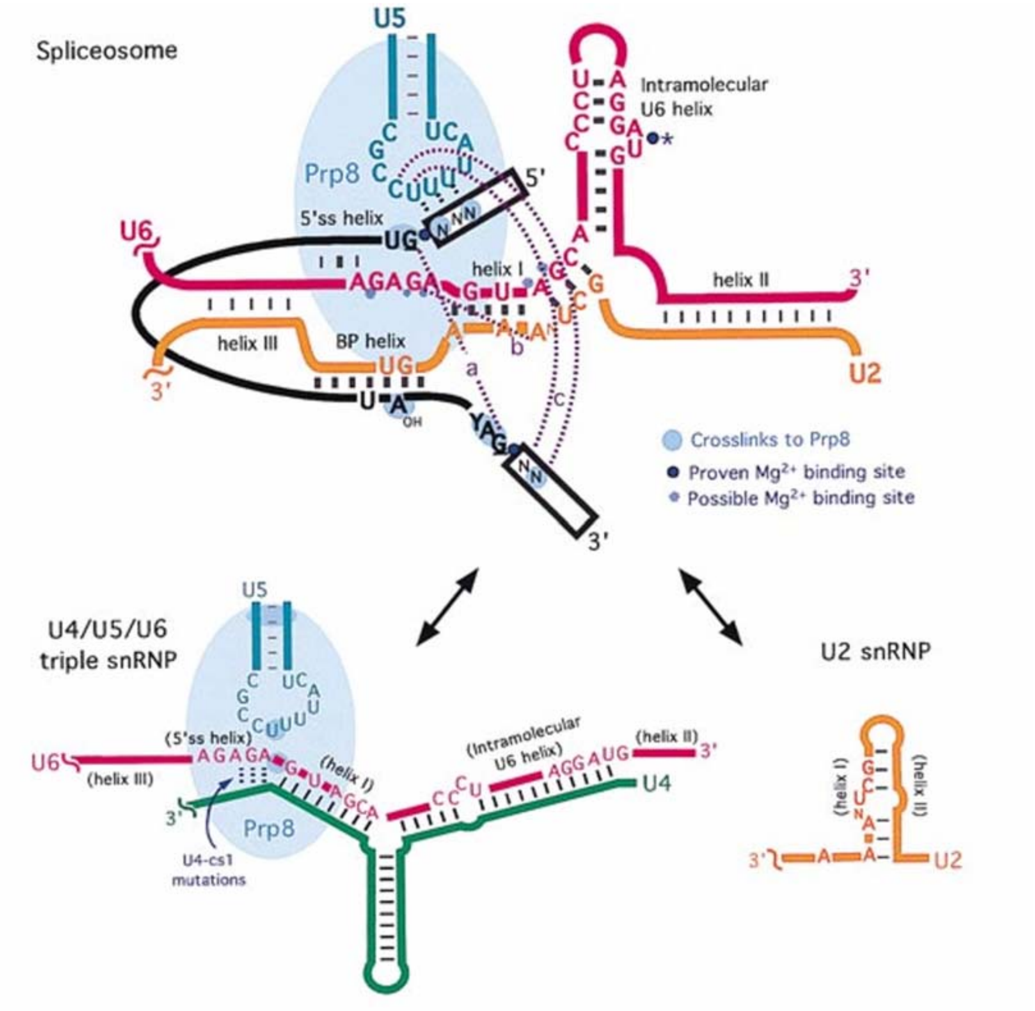


Figure 2: RNA rearrangements during spliceosome activation.
Large letters denote RNA sequences that are absolutely conserved in major and minor spliceosomes from mammals, worms, plants, yeast and trypanosomes. The exception (underlined), terminal intron Gs are replaced by covariant 5' C and 3' A in some introns. Black lines denote Watson-Crick base pairing interactions (the thinner lines denote interactions that are not absolutely conserved in all systems). Exons are denoted by rectangles, and the intron substrate is in black. Purple dotted lines indicate tertiary interactions a, b, and c. [taken from Collins and Guthrie (5)].

The adjoining U2/U6 helices I, II, and III could help juxtapose these two reactive sites for the first phosphoryl-transfer step. The highly conserved loop of U5 snRNA has been implicated in positioning the exons for ligation during the second phosphoryl-transfer step. Additional long range interactions have been identified (a, b, and c shown as dotted lines in Figure 2), suggesting the existence of a more compact RNA structure. These include an interaction between the first and last guanines of the intron (interaction a), and an invariant U6 RNA residue near the 5’ splice site helix with a budged U2 RNA residue in helix I [interaction b; for detail reviews see (5,6)]. Similarities between the spliceosome and the group II introns, a class of ribozymes that catalyze their own removal by a similar splicing reaction have been documented, providing evidence that the splicing reaction is catalyzed by RNA and the spliceosome evolved from RNA enzymes (7-10).

It was proposed that Prp8, a highly conserved spliceosomal protein, acts as a protein cofactor to the RNA enzyme (11). Prp8 crosslinks extensively with the pre-mRNA substrate near the sites of chemistry for both steps of the splicing reaction. Prp8 also forms crosslinks within snRNPs to U5 and U6 snRNAs (12,13). Prp8 is thought to stabilize interactions between variant exon sequences and the highly conserved loop of U5 snRNA, which are not likely to be strong enough on their own to hold onto the cleaved 5' exon between the first and second chemical steps (14,15). Protein cofactors are required by many ribozymes, including RNase P and most group I and group II introns, for catalysis under physiological conditions. However, protein-independent catalysis can occur in conditions that stabilize RNA structure, such as high concentrations of divalent metal ions, or in the presence of a crosslink that stabilizes a long range tertiary interaction (16,17).

RNA and protein might function intimately together in catalysis through a specific protein-RNA structure that cannot be achieved by the RNA alone. Many proteins that bind specifically to RNA have been observed to induce dramatic conformational changes in the RNA or to stabilize a unique RNA conformation (16,18,19).

Understanding the complexity of these processes requires data at atomic level on the active and inactive spliceosomes and their subcomponents. In the absence of such experimental data, the dynamics of small RNA motifs found in different snRNAs could provide valuable insights into the folding and dynamics of large RNAs, ultimately contributing to the elucidation of spliceosome assembly and function.

1.2. The U4/U6·U5 tri-snRNP

The association of the U4/U6·U5 tri-snRNP completes the assembly of the inactive spliceosome (complex B in Figure 1B). Within the U4/U6 di-snRNP, the U4 and U6 snRNAs form a phylogenetically highly conserved Y-shaped U4/U6 interaction domain, consisting of two intermolecular helices (stem I and II), which are separated by the 5' stem-loop of U4 (Figure 3). Both intermolecular helices are disrupted as the spliceosome become activated for catalysis; the region of U6 constituting stem II folds back on itself to form a new intramolecular stem-loop, and the region of U6 residing in stem I base-pairs with U2 snRNA to form part of the catalytic center. Concomitantly, U6 snRNA base-pairs with the 5’ end of the intron, while U1 snRNP dissociates from the 5’ splice site. In contrast to U6, U4 snRNA is released from the spliceosome, or remains only loosely attached to it [for detailed description of U4/U6 snRNP structure and function see (1,20-24)].

The mechanism by which the spliceosome is transformed into a catalytically active machine is only poorly understood at present. There is evidence that the yeast DExH/D-box RNA helicase Brr2p (human U5-200K) is one of the driving forces behind the disruption of the U4/U6 snRNA helices (25,26), but its precise mechanism of action is still unknown. Brr2p/U5-200K could unwind the U4/U6 duplex directly. However, the *in vitro* RNA helicase activity of purified U5-200K is a generic one, exhibiting no specificity for the naked U4/U6 duplexes (26). Alternatively, U4/U6 snRNA-binding proteins could play a role in regulating the stabilization/destabilization of the U4/U6 duplex. In this respect, information about interactions between U4/U6 snRNP proteins and the snRNAs is of particular importance.

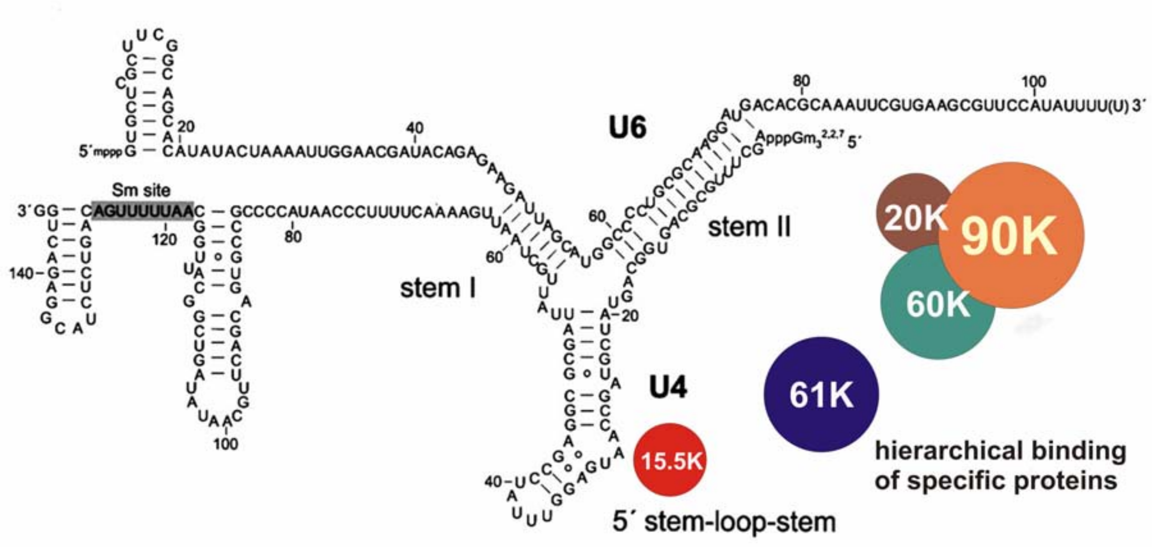


Figure 3: Secondary structure of U4/U6 snRNA.
The black lines indicate Watson-Crick base-pairs and the circles indicate non-Watson-Crick interactions. The U4/U6 snRNP-specific proteins, 15.5K, 61K and 20/60/90K are shown as color filled circles of size proportional to the molecular weight [taken from Nottrott et al. (37)].

Figure 3 shows a two dimensional view of the U4/U6 snRNA duplex. In addition to the seven Sm proteins that bind the Sm site of U4 snRNA, and the seven LSm (“like-Sm”) proteins (LSm 2-8) that are associated with the 3’ end of U6 snRNA, five proteins have been found by biochemical means to be associated with the human 13S U4/U6 snRNP [reviewed by (4)]. These include three proteins with molecular weights of 20, 60 and 90 kDa that form a stable, heteromeric complex [hereafter termed as 20/60/90K complex; (27,28)]. U4/U6 snRNP also contains a 15.5 kDa protein (15.5K) that binds directly to the 5’ stem-loop of U4 snRNA (29) and it is also present in box C/D snoRNPs, providing a link between the pre-mRNA and pre-rRNA processing machineries (30). Finally, a 61 kDa protein (61K) was identified and shown to be required for U4/U6·U5 tri-snRNP formation (31). It shares a homologous central domain with the proteins Nop56 and Nop58, which (like the 15.5K protein) are integral constituents of the box C/D snoRNPs (31,32). These U4/U6 snRNP-specific proteins are also associated with the HeLa U4atac/U6atac snRNP (33). Except for the 20K protein, the U4/U6 snRNP-specific proteins are evolutionary conserved, and orthologous proteins termed Snu13p (15.5K), Prp4p (60K), Prp3p (90K) and Prp31p (61K) are also associated with the *Saccharomycetes cerevisiae* U4/U6 snRNP particle (34-36).

Nottrott et al. showed that (37): (i) the 61K protein and the 20/60/90K complex bind U4 and U4/U6 snRNA, respectively, only in the presence of the 15.5K protein, (ii) the association of the 61K protein and the 20/60/90K complex with the U4/U6 snRNA can occur independently of each-other, (iii) the 20/60/90K complex binds to a minimal U4/U6 snRNA duplex comprising stem II and the 5’ stem-loop of U4 snRNA, (iv) the 61K protein requires the 5’ stem-loop and the 5’ end of U4 snRNA for binding, (v) the 61K protein can be cross-linked to two distinct sites on U4 snRNA in native tri-snRNP particles (positions 37-39, 28 and 19-20 – see Figure 3).

It was proposed that U4 snRNA acts as a chaperone that delivers U6 to the spliceosome, sequestering a catalytically active domain of U6 snRNA until the dissociation of the U4/U6 snRNA duplex activates this domain for its function in splicing (21). Upon spliceosome activation, U6 snRNA forms a duplex with U2 snRNA and U4 snRNA is released, becoming available for a new round of splicing. The mechanism of U4 snRNA dissociation is still under investigation. Nottrott et al. proposed that the 20/60/90K complex might stabilize stem II such as to permit the dissociation of stem I first (37). Consistent with the idea of sequential dissociation, an intermediate in the catalytic core of the U12-dependent spliceosome has been detected that has an intact U4atac/U6atac stem II, but in which parts of stem I are already base-paired with U12 snRNA (38). The role of U4/U6 snRNP-specific proteins in the spliceosome assembly and activation is not yet completely understood. Therefore, further analysis of the function of each protein is required.

1.3. The 15.5K protein

Mutational analysis of U4 snRNA in *Xenopus* oocytes and in the HeLa in vitro splicing system demonstrated that the 5’ stem-loop is essential for pre-mRNA splicing and spliceosome assembly (39,40). Since the U4 snRNA 5’ stem-loop is dispensable for U4/U6 base pairing in vitro, it has been suggested that it functions in spliceosome assembly at a stage subsequent to U4/U6 snRNP formation (40). Prior to the characterization of U4 snRNP-specific proteins, it was not clear whether the function of U4 snRNA 5’ stem-loop is to interact with other RNAs in the spliceosome or to recruit snRNP proteins into the spliceosome.

The 15.5K was identified and characterized by Nottrott et al. and it was shown to bind specifically to the 5’ stem-loop of U4 snRNA. Ortologues of 15.5K protein have been identified in *C. elegans*, *S. cerevisiae*, *S. pompe*, and *Arabidopsis thaliana* sharing 71-77% identity and 83-89% similarity with the human protein. Furthermore, the 15.5K protein shares significant homology with several other proteins in the database which are clearly not 15.5K orthologues. These include several ribosomal proteins such as L7AE from different organisms or the yeast ribosomal protein L30 (29).

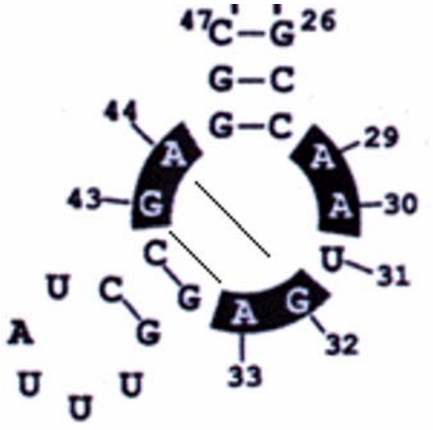


Figure 4: Minimal binding site for 15.5K protein.
RNA construct used for the crystallization of 15.5K-KtU4 complex [taken from Vidovic et al. (44)].

Figure 4 shows a detailed view of the minimal RNA binding site of the 15.5K protein. It consists of an internal asymmetric loop which is closed by the long stem of the 5’ stem-loop and a short stem. Close inspection of the 5’ stem-loop sequences of U4 snRNAs revealed that nucleotides U31, G32, A33, G43 and A44 are 100% conserved in all U4 snRNAs, except for U31 in *Trypanosoma brucei*. Furthermore, positions A29 and A30 are always purines, exception being A29 in *T. brucei* and *Physarum polycephalum*. The seven mentioned nucleotides proved to be crucial for the binding of 15.5K protein to the U4 snRNA. At positions 29 and 30, the requirement for a particular nucleotide is less stringent, in that the adenosines can be replaced individually by guanosines without any loss of 15.5K protein-binding activity. If both positions contain guanosines, protein binding is reduced only slightly. In contrast, if the two adenosines are replaced simultaneously by cytidines, protein-binding activity is lost. Furthermore, deletion of one adenosine residue has a deleterious effect on the capability of the resulting mutant to interact with the 15.5K protein. Thus, there is a preference for purines at positions 29 and 30 of the U4 snRNA stem-loop. In good agreement with their phylogenetic conservation (see above), the identity of nucleotides U31, G32, A33, G43 and A44 is crucial for protein-binding activity *in vitro* [for a detailed description of the binding of the 15.5K protein to the U4 snRNA see (29)].

Based on the sequence comparisons at the time of its characterization, the 15.5K protein did not exhibit obvious structural similarities with members of well-established families of RNA-binding proteins. However, it shared the homologous central region of 56 amino acids with several proteins from a variety of species which had one feature in common: they were all associated with RNP particles. Thus, the 15.5K was characterized as a novel RNA-binding protein.

The 15.5K protein has been found to perform a similar function in the minor spliceosome (29), the RNA binding site adopting a similar architecture as that observed in the major spliceosome (Figures 5A and 5B). Interestingly, in a subsequent study Watkins et al. have shown that the 15.5K protein is also a component of the box C/D snoRNPs and the U3 specific box B/C motif (30).

These RNPs are involved in the processing and maturation of the precursor ribosomal RNA [rRNA; (41-43)]. The binding sites of the 15.5K adopt a similar fold stem-loop-stem (Figure 5C and 5D) as that observed in the U4 snRNA. The 15.5K protein was the first protein shown to be a common component of the spliceosome and rRNA processing machinery. These findings raised the interesting possibility that the U4/U6 snRNP and the box C/D RNPs share a common ancestral snRNP. It is likely that this complex functioned as a chaperone in guiding RNA folding and perhaps, in some cases, developed the ability to methylate the substrate RNA. The 5' stem-loop complex of U4 snRNP may have evolved from a primitive box C/D snoRNP. Archaea possesses box C/D snoRNAs but do not contain pre-mRNA spliceosomal components, suggesting that the box C/D snoRNPs existed before the U4 snRNP [for detailed description see (30)]

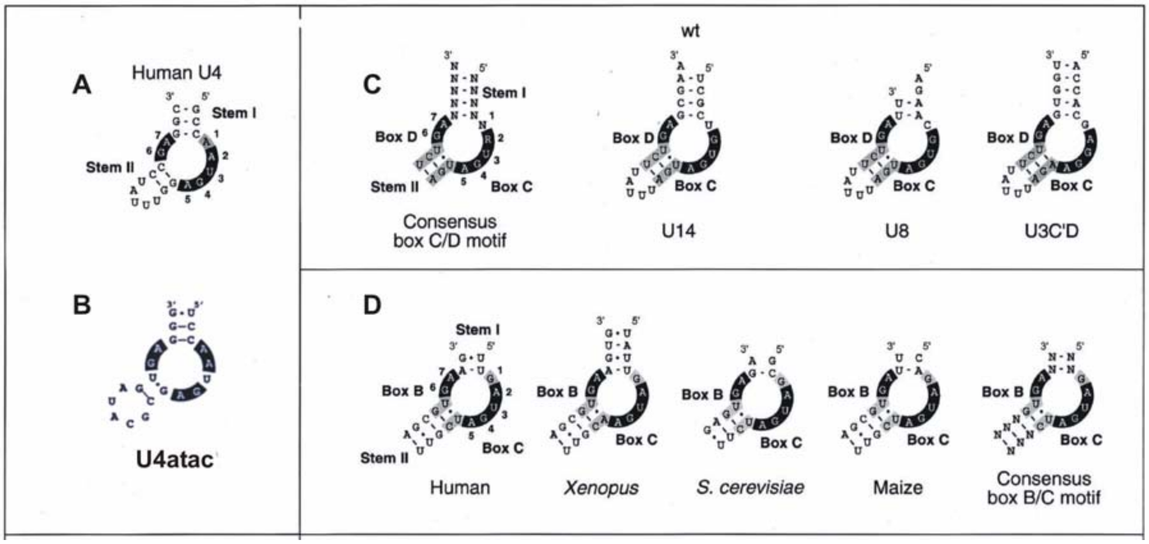


Figure 5: RNA motifs capable of binding the 15.5K protein.
The highly conserved nucleotides required for 15.5K binding are indicated in white on a black background. Gray boxes indicate the remaining conserved nucleotides. (A) U4 snRNA; (B) U4atac; (C) box C/D motifs; the consensus U14, U8, and U3 box C/D motifs are draw to demonstrate the structural similarity with U4 snRNA (D) proposed structure of the U3-specific box B/C motifs [taken from Watkins et al. (30)].

The crystal structure of the 15.5K-U4 snRNA was solved to a resolution of 2.9 Å and is shown in Figure 6. As described by Vidovic et al. (44), the RNA molecule folds into a compact structure that consists of two double-helical stems (stems 1 and 2) bridged by the (5+2) internal loop, which is asymmetric and highly structured. The two stems exhibit a slightly distorted A-form RNA double helix. A striking feature of the RNA structure is the strong bend at the internal loop, which adopts a complex fold: of its seven nucleotides, four form tandem G-A base pairs extending stem 2, while the remaining three are unpaired. Of these, one (U31) is flipped out: it protrudes away from the rest of the oligonucleotide chain and toward the protein. The sugar–phosphate backbone forms a sharp hairpin-like turn at this point. The other two unpaired bases, A29 and A30, are stacked onto the ends of stem 1 and stem 2: A29 is stacked onto the base pair G45-C28 and thus caps stem 1, while A30, which has a syn and a 2′-endo conformation, is stacked onto A44 of the opposite strand and thus caps stem 2. The two G-A base pairs are of the sheared type characterized by hydrogen bonds from the 2-amino group of guanine to N7 of adenine and from the 6-amino group of adenine to N3 of guanine. Between these two tandem-sheared G-A base pairs, the helix is strongly overwound, with a twist angle of 81 °. This extreme helix twist leads to cross-strand stacking of the two adenines. Both the high twist and a cross-strand stacking of the adenines and guanines are common in tandem sheared G-A base pairs (45,46).

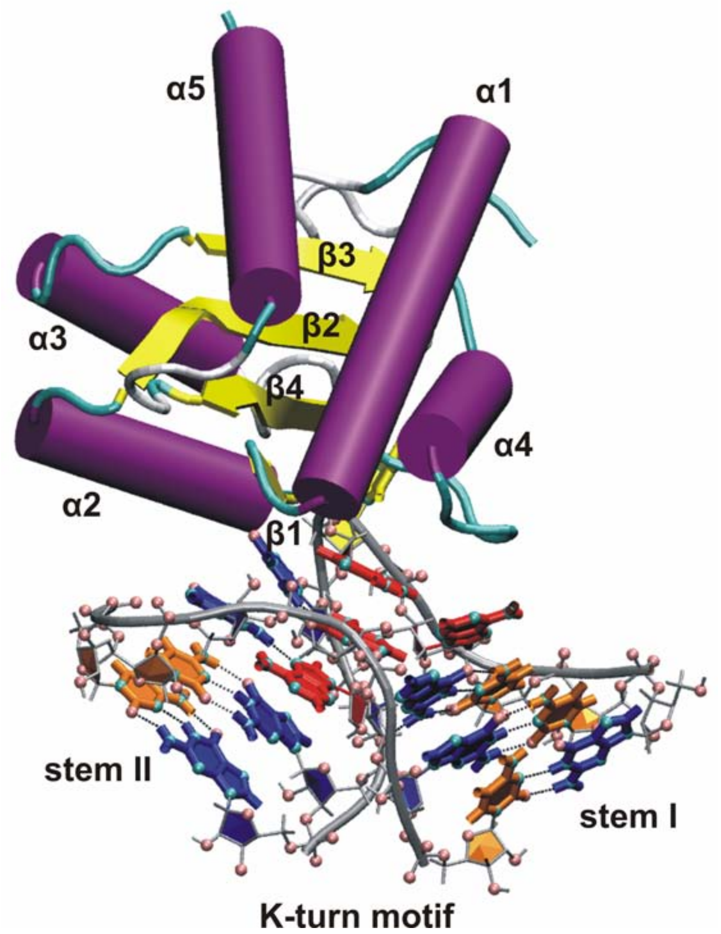


Figure 6: Crystal structure of 15.5K protein bound to U4 snRNA.
In the RNA, guanines are blue, adenines red, cytosines orange and uracils yellow; nitrogen atoms are indicated in cyan and oxygens in pink; hydrogen bonds are denoted as dotted black lines. The protein is drawn in cartoon representation: α-helices are purple, β-sheets yellow, coils white and turns cyan. Unless indicated otherwise, the coloring of this figure is transferred to all figures.

In the present RNP complex, only adenines A33 and A44 exhibit perfect cross-strand stacking, while the corresponding guanines G43 and G32 are displaced. Interestingly, the cross-strand stacking of A33 and A44 is continued by a third adenine, A30, which is one of the unpaired internal loop nucleotides; thus, the structure exhibits a three-adenine cross-strand stack, with A44 from the one strand sandwiched between A33 and A30 from the other (‘3+1’ stacking scheme). Besides the base pairing and base stacking interactions, a network of hydrogen bonds involving several ribose 2′-OH groups further stabilizes the fold of the internal loop. The 2′-OH group of A44, which has a 2′-endo conformation, is within hydrogen-bonding distance of the N6 of A30, while the N1 of A44 forms a hydrogen bond with the 2′-OH group of A29. Furthermore, the 2′-OH group of A33 forms hydrogen bonds to N3 of G45, and the 2′-OH of the flipped-out U31, which has also a 2′-endo conformation, forms a hydrogen bond with the phosphate group of A30. The 2′-OH groups of both guanosine nucleotides of the tandem G-A base pairs contribute to the hydrogen bond network as well. The G32 exhibits 2′-endo conformation and its 2′-OH group forms a hydrogen bond with N2 of G43. The conformation of the G43 ribose is 2′-endo, and its 2′-OH group forms a hydrogen bond with the phosphate of A44.A striking feature of the RNA structure is the sharp bend between the two double-stranded stems. An angle between the helical axes of about 65° was calculated (Figure 7A).

The 128 amino acid residues of the 15.5K protein fold into a single, compact globular domain of alternating α-helices and β-strands, forming an α-β-α sandwich structure. The central β-sheet consists of three antiparallel and one parallel β-strand positioned in the order β1, β4, β2, β3. Helices α1, α4, and α5 pack against one side of the β-sheet, while helices α2 and α3 are located on the other side.

In the 15.5K protein, the RNA binding surface consists of amino acid residues located in two α-helices (α2 and α4), one β-strand (β1), and three different loops (β1-α2, β2-α3, and α4-β4). These residues interact predominantly with the nucleotides of the (5+2) internal loop, and there are also contacts with the sugar–phosphate backbone of stem 2.

U31 is tightly bound in a pocket of the protein formed by Glu61 and Ile65, Lys86, and Ile100. The O4 of U31 forms two hydrogen bonds, with the amino group of Lys-86 and the main chain amide of Glu61, respectively, and the 3-imino group of U31 forms a hydrogen bond with the main chain oxygen of Glu61. Furthermore, hydrogen bonds are present between the phosphate group of U31 and the main chain amides of Ala39 and Ile100. In addition to these hydrogen bonds, the base of U31 is in van der Waals contact with the hydrophobic side chains of Ile65 and Ile100 and the hydrophobic part of the Lys86 side chain.

The amino acid residues that contact the tandem-sheared G-A base pairs are Asn40, Glu41, and Lys44 (in loop β1-α2 and helix α2), which bind to the G32 Watson-

Crick edge and the G43 Hoogsteen edge. The carboxylate group and the main chain amide of Glu41 are within hydrogen bonding distance of N1, N2, and O6 of G32; the ε-amino group of Lys44 is in hydrogen bond distance to N7 and O6 of G43; and the ND2 of Asn40, which forms a hydrogen bond with the N7 of G32 and the main chain amide of Asn40, is in hydrogen bond distance with the O6 of G32. In the crystal structure, the orientation of ASN40 proved to be incorrect (see 5.3.1)

The base of the unpaired nucleotide A29, which stacks on the base pair G45-C28 of stem 1, packs with its other side against the hydrophobic part of the side chain of Arg-97 in loop α4-β4. Likewise, the unpaired nucleotide A30, which extends the purine stacking of stem 2, packs on its opposite side against a hydrophobic protein surface provided by Lys37 and Val95 in loops β1-α2 and α 4-β4, respectively.

The negative charge of the RNA phosphates is neutralized by several basic protein residues. Lys44 and Arg97 are located within hydrogen bonding distance of the phosphates of C42 and A29, respectively. Additionally, Arg36, Lys37, and Arg48 are within 7–8 Å of the RNA and contribute significantly to the overall electrostatic picture.

The 5' stem-loop of U4 snRNA also contains an external pentaloop (U36-U37-U38-A39-U40), the conformation of which was not revealed in the crystallographic structure although it was present in the crystallization construct [for complete description of the crystal structure see (44)].

1.4. The RNA K-turn motif

The crystal structure of the 15.5K-U4 snRNA complex revealed a novel RNA structural motif (for description see 1.3) which was then observed in different large RNA structures. The motif, named the kink turn motif (K-turn) is a two-stranded, helix–internal loop–helix motif comprising ~15 nucleotides. The first helical stem, the 'canonical stem' or 'C-stem', ends at the internal loop with two Watson–Crick base pairs, typically C–Gs, while the second helical stem, the 'non-canonical stem' or 'NC-stem', which follows the internal loop, starts with two non-Watson–Crick base pairs and is extended into the internal loop by sheared G–A base pairs (Figure 7B). The internal loop between the helical stems is always asymmetrical, and usually has three unpaired nucleotides on one strand and none on the other. The 5'-most nucleotide in the long strand of the loop stacks on the C-stem, the second extends to stack on the NC-stem, and the third flips out from the compact RNA structure into protein pockets or into solution. The K-turn occurs six times in *H.marismortui* 23S rRNA, and twice in *T.thermophilus* 16S rRNA. Each one is designated 'Kt-#', with Kt standing for kink-turn and the number indicating the helix of rRNA in which it is found. Although these eight K-turns vary somewhat in sequence, each has essentially the same distinctive three-dimensional form, and a consensus sequence can be derived [Figure 7C; for detailed description of the K-turn motif see (47)]. K-turns were also identified in the structure of the box C/D snoRNA bound to archaeal L7AE protein (48-50) and in the yeast L30E-mRNA complex (51).

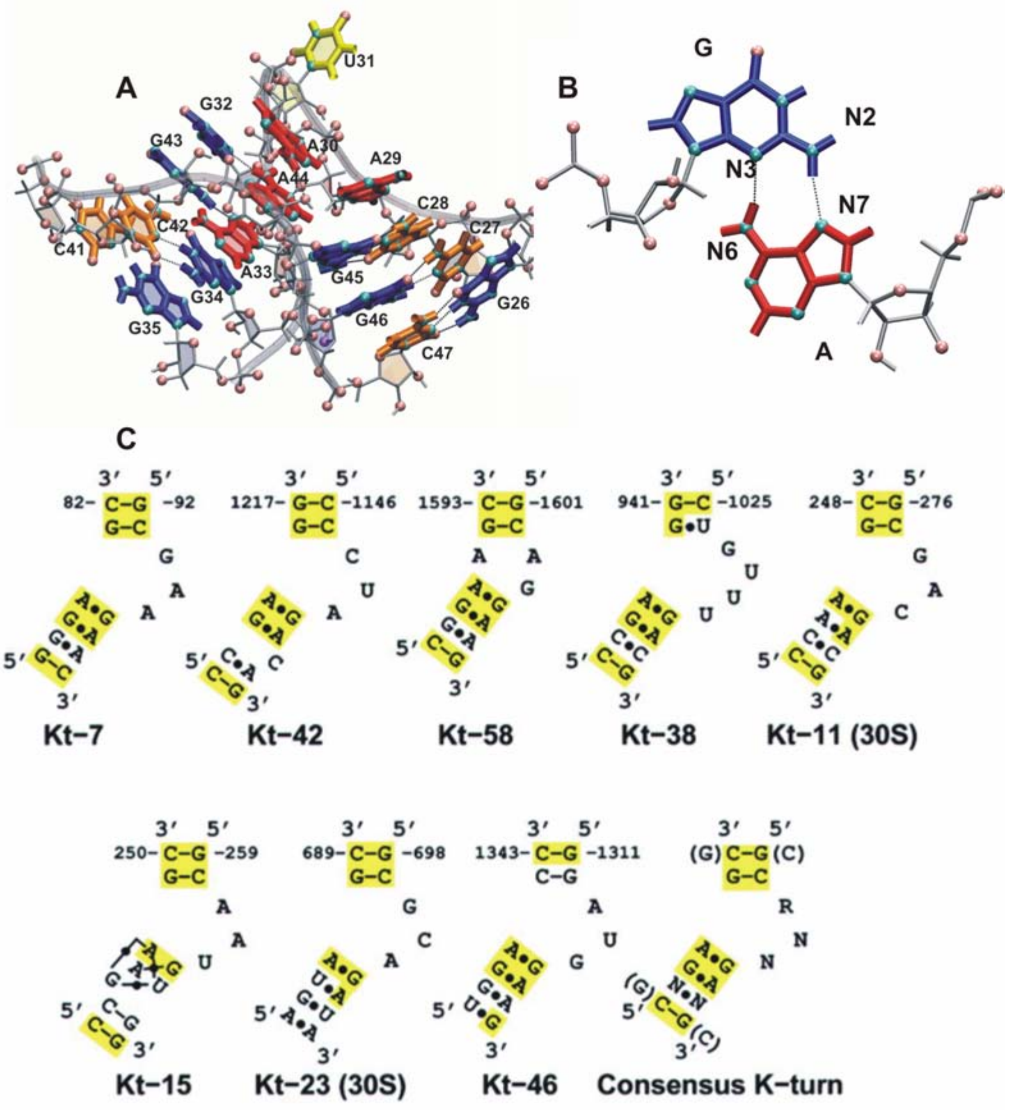


Figure 7: Structure of the K-turn motif.

(A) KtU4 as found in the crystal structure. φ is the angle between the P atoms of C47, U31, and G35. (B) Structural view of a sheared G-A base pair (for coloring see Figure 6). (C) Secondary structure diagrams of the eight K-turns found in the *H.marismortui* 50S and *T.thermophilus* 30S subunit structures and a derived consensus sequence. The names indicate in which rRNA helix each example of the motif is found. Solid lines represent Watson–Crick pairings between bases, and black dots represent mismatched base pairings. Yellow shading indicates nucleotides that conform to the derived consensus sequence [taken from Klein et al. (47)].

The K-turn is an important RNA recognition motif for the ribosomal proteins in the 50S subunit: five of the six K-turns in *H.marismortui* 23S rRNA make significant interactions with at least one ribosomal protein, and nine of the 28 observed proteins interact with K-turns. One of these, Kt-46, also interacts extensively with two distant regions of the rRNA, demonstrating that K-turns also function to stabilize RNA tertiary structure.

There is considerable variation in the way that K-turns interact with proteins in the ribosome. Four principal surface features are recognized: (i) the widened major groove of the C-stem; (ii) the flattened minor groove of the NC-stem; (iii) the sharply kinked sugar–phosphate backbone and the protruded nucleotide; and (iv) the exposed base planes. Recognition of these features involves complementary surfaces on proteins that allow the burial of significant hydrophobic surface area. These features enable a single K-turn motif to participate in many intermolecular interactions simultaneously, making it well suited to serve as a nucleation site around which large ribonucleoprotein assemblies can be built. Although these nine ribosomal proteins do not share a common structural domain that recognizes K-turns, there is at least one homologous family of RNA-binding domains that is specific for it. *Holoarcuta marismortui* L7AE, yeast L30E and the human 15.5K contain identical domain structures that bind K-turn RNA elements in the same fashion. It seems likely, therefore, that other proteins containing this RNA-binding motif will be found to bind to K-turns the same way (47).

The K-turn is a member of the larger family of RNA motifs that are defined as directed and ordered stacked arrays of non-Watson–Crick base pairs forming distinctive folds of the phosphodiester backbones of the interacting RNA strands. RNA motifs mediate the specific interactions that induce the compact folding of complex RNAs. RNA motifs also constitute specific protein or ligand binding sites. A given motif is characterized by all the sequences that fold into essentially identical three-dimensional structures with the same ordered array of isosteric non-Watson–Crick base pairs [for detail review of RNA motifs see (52)].

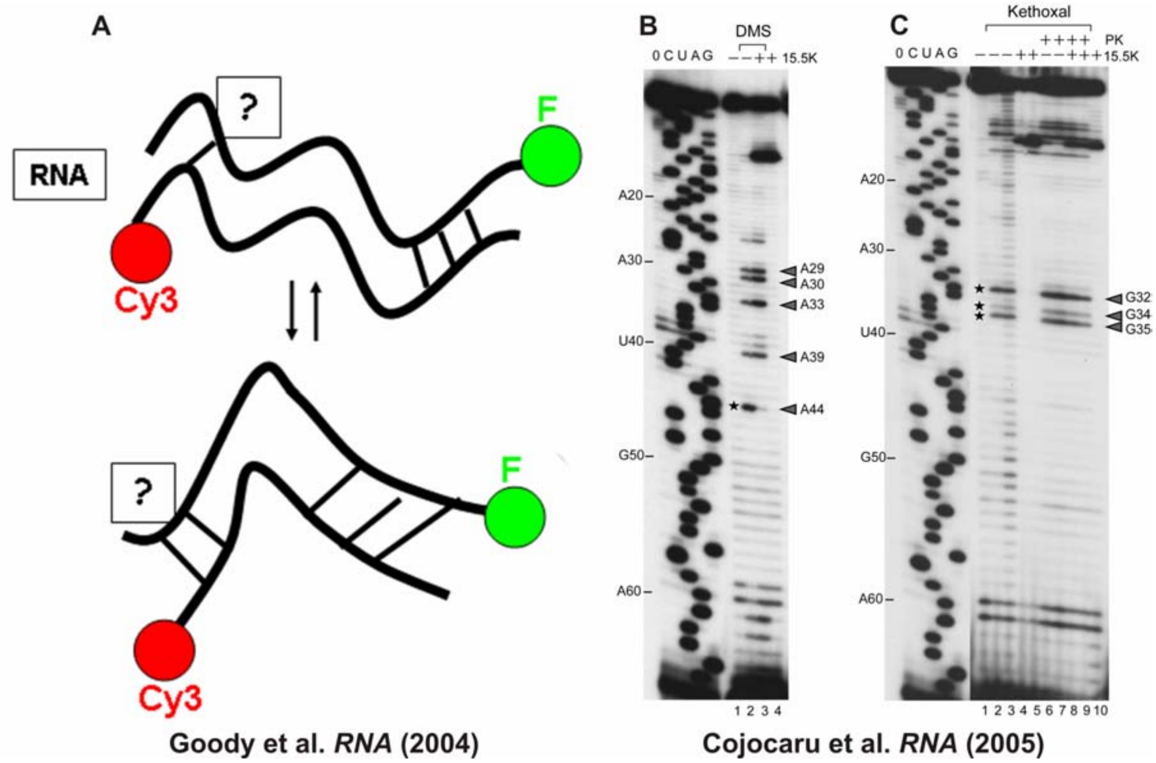
Out of the ten K-turns mentioned above, only one (Kt-38) has been found not to be associated with proteins, indicating that the motif could be a candidate for protein-assisted folding. From here on I will refer to the K-turn motif formed by the 5’ stem-loop of U4 snRNA as ‘KtU4’.

1.5. Protein-assisted RNA folding

Formation of a wide variety of protein–RNA complexes involves conformational changes in the protein, RNA, or both. In several cases, the folding of the RNA is assisted by protein binding. The terms ‘induced fit’ and ‘conformational capture’ were introduced for designating alternative pathways of conformational change upon complex formation. In the induced fit, the RNA undergoes a transition between two different well-defined conformations, whereas conformational capture refers to the stabilization by the protein of one specific conformation from a pool of conformations reflecting the inherent flexibility of the RNA (16,19,53). One example of protein-assisted RNA folding according to the induced fit mechanism is the binding of the 3' UTR of U1A pre-mRNA to the U1A protein (54-57). The conformational capture mechanism is harder to be identified experimentally because the free RNA is very flexible and therefore, not suited for crystallography or NMR studies.

Previous studies proposed a protein-assisted RNA folding for the K-turn motif by showing that the K-turn is a rather flexible entity in the unbound form (58,59). Single molecule fluorescence resonance energy transfer (FRET) studies performed by Goody et al. (58) provided evidence of large amplitude conformational transitions in the K-turn Kt7 that contains most of the motif’s consensus sequence. They observed that the K-turn is dimorphic, undergoing a transition between a closed (kinked) and an open (extended) conformation (‘k-e motion’). A schematic representation of the transition is shown in Figure 8A. The k-e motion depends on the ionic strength but a significant population of the extended structure was observed even at high concentrations of divalent cations. However, our understanding of the atomic details of such transitions remains inadequate.

Prior to the work presented in this doctoral thesis, Stephanie Nottrott has performed chemical RNA modification studies showing that the free KtU4 lacks several secondary and tertiary structure interactions that are present in the complex. Using dimethylsulfate (DMS) she observed that the N1 position of A44 is clearly accessible, permitting chemical modification in the absence, but not in the presence of 15.5K protein (Figure 8B, cf. lanes 2 and 3). These data suggests that the inter-stem contact between the N1 position of A44 and the 2’ OH group of A28 is established only upon protein binding. RNA structural probing with Kethoxal showed that the nucleotides G32, G34, and G35 are clearly accessible for modification in the absence, but not in the presence of 15.5K protein (Figure 8C, cf. lanes 2,3 and 6,7 with lanes 4,5). In addition, G32, G34, and G35 are also accessible for modification with Kethoxal after digestion of the bound 15.5K with Proteinase K (Figure 8C, lanes 8,9). The N2 atom of G32 is involved in the base-pairing interaction of G32 with A44, while the N1 and N2 positions of G34 and G35 form hydrogen bonds with C42 and C41 (60).



Goody et al. *RNA* (2004)

Cojocaru et al. *RNA* (2005)

Figure 8: Experimental data on the K-turn motif.
(A) Schematic view of the single molecule FRET studies on Kt7 by Goody et al. (58). In the RNA used for labeling with fluorescein (green) and Cy3 (red) the stems were extended with extra Watson-Crick base pairs. The question mark indicates the lack of atomic detail information about the structures of the two states represented. (B) Primer extension analysis of U4 snRNA after DMS treatment of KtU4 RNA either in the absence or presence of recombinant 15.5K protein (lanes 2,3). Lanes 1,4 are control lanes (no DMS modification). (C) Primer extension analysis of U4 snRNA after Kethoxal treatment either in the absence (lanes 2,3,6,7) or presence (lanes 4,5,8,9) of 15.5K protein. Lanes 6–9 show RNA modification after Proteinase K digestion. Lanes 1,10 are control lanes (no Kethoxal treatment). Modified nucleotides are indicated by an arrowhead; nucleotides that are clearly protected from chemical modification in the presence of bound 15.5K protein are marked by asterisks. The presence or absence of the 15.5K protein is indicated by "+" or "-," respectively. C, U, A, and G refer to dideoxysequencing reactions and correspond to the sequence of human U4 snRNA; 0 indicates a control primer extension with unmodified U4 snRNA where no ddNTPs were added to the reaction, and the position of every tenth nucleotide of the U4 snRNA is indicated on the left in panels A and B, respectively.

However, all these experiments did not provide any atomic details about the transitions that the free KtU4 undergoes upon binding to 15.5K protein. Furthermore, no structural data at atomic resolution is available to date for the unbound K-turns.

1.6. Computer simulations: benefits and challenges

Computer simulations have become very powerful tools for studying biological processes, largely due to the rapid increase in computer power and improved accuracy. Advances such as the explicit modelling of solvent, further refinement of force fields, and the advent of the Particle Mesh Ewald (PME) method for treating long range electrostatic interactions have led to increasingly fruitful simulations of biological systems and processes (61-64). Molecular dynamics (MD) simulations have been applied in studies of RNA structure (65-69), RNA-metal ion binding (65,70-73) or RNA-protein interfaces (74-78). However, limitations remain; for example, MD trajectories are restricted to tens of nanoseconds time scale, reducing the range of processes that can be studied to those occurring in this time range. Thus, conformational sampling is still poor in standard MD simulations and large conformational transitions are inaccessible. Simulating a protein-assisted RNA folding event poses several challenges: (i) the lack of structural data on the free RNA due to its flexibility; (ii) the time scale of the transitions relevant for the folding is often significantly larger than that accessible by standard MD protocols; (iii) the evolution of the system during MD simulations is largely dependent on the initial structure; (iv) the multitude of factors influencing the process occurring in the cell.

Several methods have been developed to increase the conformational sampling during MD simulations. Among them, Locally Enhanced Sampling (LES), a mean field based theory has been previously applied in several studies investigating conformational diversity of small regions in proteins or nucleic acids (79-84). Coupled with PME, LES constitutes a powerful tool for locating experimental structures when starting from different conformations (84). The application of LES leads to a smoother potential energy surface allowing conformational transitions that are otherwise inaccessible to standard MD simulations (85).

It was also shown that the application of LES triggers a large conformational transition in the lateral and diagonal thymine loops of DNA G-quartets (86). However, the structures to which the simulations converged were very different from the experimental structures. Since free energy calculations confirmed that the new structures were more stable, it was proposed that the inconsistencies arose from force field inaccuracies rather than artifacts introduced by the application of LES methodology.

1.7. Conformational parameters of nucleic acids

Throughout the present dissertation, I will refer to several conformational parameters that describe the RNA structure. A summary of all dihedral angles describing the conformation of a typical nucleotide is shown in Figure 9. The most important parameters are: (i) the sugar pucker, (ii) the χ angle, and (iii) the γ angle.

The five-membered furanose ring is generally non-planar. It can be puckered in a twist (T) form with two adjacent atoms (C2' and C3') displaced on opposite sites of a plane through the other three atoms (C1', O4' and C4'). There are two most abundant conformations: (i) C2'-endo if the C2' atom is above the plane and, (ii) C3'-endo if the C3' atom is above the plane (Figure 9, upper right panel). C2'-endo pucker is specific for B-type helices (B-DNA) while C3'-endo pucker is specific for A-type helices (A-RNA, A-DNA). In more general terms, the sugar pucker is described by a pseudorotation angle that is calculated using all the five dihedrals in the sugar (Figure 9).

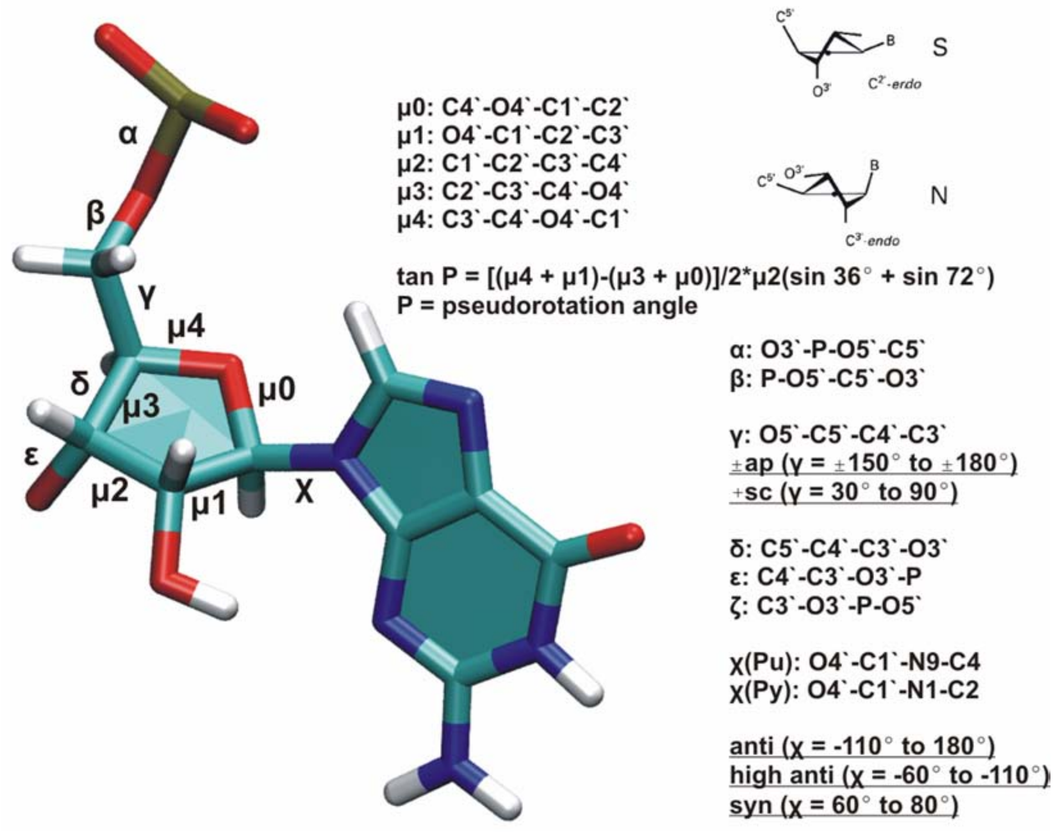


Figure 9: Conformational parameters in the RNA demonstrated on guanine. Carbons are shown in cyan, nitrogens in blue, oxygens in red and hydrogens in white. The two most abundant sugar puckers (C2'-endo and C3'-endo) are shown in the upper right panel.

The χ angle describes the orientation about the glycosyl bond. Relative to the sugar moiety, the base can adopt three main orientations: ‘anti’, ‘high-anti’ and ‘syn’ (for the corresponding values of χ see Figure 9). Nucleotides in A-type helices are generally in anti configuration, in B-type helices in high-anti, while the syn configuration is adopted only by purines (every second G in the left handed Z-DNA or A30 in the K-turn of U4 snRNA).

Rotation about the exocyclic C4’-C5’ bond (described by the γ angle) allows O5’ to assume different positions relative to the furanose (Figure 9).

A detailed description of all the parameters describing the conformational diversity of nucleic acids is available in ‘Principles of Nucleic Acid Structure’ by Wolfram Saenger (87)

1.8. Fluorescence of 2-aminopurine

The fluorescent adenine isomer, 2-aminopurine (2AP), is widely employed as a reporter of the structure and dynamics of nucleic acids because its ability to form Watson-Crick base pairs with thymine (uracil) and its fluorescence quantum yield is very sensitive to structural context. 2AP fluorescence is strongly quenched both is single- and double-stranded DNA (88), and this property has been exploited to probe the formation of R-DNA triplexes (89), the dynamics of melting (90-92), abasic sites (93,94), mismatched base pairs (95), and metal ion binding (94,96) as well as thermodynamics and kinetics of protein-induced DNA conformational transitions (97-104). Quenching of 2AP fluorescence in DNA has been mainly attributed to base stacking and hydrogen bonding (105). The mechanism of quenching has been investigated theoretically by examining the electronic structure of 2AP in different environments (106-109).

Although it is mostly documented in fluorescence studies of DNA, the use of 2AP was extended to RNA for studying Mg^{2+} - dependent conformational changes in the hammerhead ribozyme (96) or the dynamics of GNRA tetraloops (110).

2. Theoretical background

2.1. Introduction to molecular modeling

The rapid increase in the number of structures at atomic level of biologically relevant macromolecules provides a scaffold to explore macromolecular structural dynamics. Structures can only provide static views of the macromolecules, whereas the cellular processes are highly dynamic and based on temporary interactions between large numbers of macromolecules. Conformational transitions occurring during biological processes are studied experimentally using techniques such as nuclear magnetic resonance (NMR) or fluorescence resonance energy transfer (FRET) and theoretically using a variety of computational methods included in the general field of ‘molecular modeling and simulation’. In her book ‘Molecular modeling and simulation: An interdisciplinary guide’, Tamar Schlick gives the following definition: “Molecular modeling is the science and art of studying molecular structure and function through model building and computation. The model building can be as simple as plastic templates or metal rods, or as sophisticated as interactive, animated color stereographics and laser-made wooden sculptures. The computations encompass *ab initio* and semi-empirical quantum mechanics, empirical (molecular) mechanics, molecular dynamics, Monte Carlo, free energy and solvation methods, structure/activity relationships (SAR), chemical/biochemical information and databases, and many other established procedures. The refinement of experimental data, such as that obtained from NMR or x-ray crystallography, is also a component of biomolecular modeling” (111).

The most accurate theoretical description of a molecular system is achieved by quantum mechanical methods (*ab initio*) which account for the interactions between all the particles in the molecular system including the electronic interactions. However, the size of even small biomolecules is large enough to make quantum mechanical calculations not feasible. Therefore, approximations are required in order to investigate macromolecules, leading to the development of empirical force field models (molecular mechanics).

For the rest of this chapter, I will provide a detailed introduction into the theoretical background of the diverse methods that I applied. For further details, please refer to the references cited or to the book ‘Molecular Modeling: Principles and Applications’ by Andrew R. Leach (112).

2.2. Empirical force fields: molecular mechanics

Force field methods (also known as molecular mechanics) ignore the electronic motions and calculate the energy of the system as a function of nuclear positions only. Molecular mechanics is thus invariably used to perform calculations on systems containing significant numbers of atoms. In some cases force fields can provide answers that are as accurate as even the highest-level quantum mechanical calculations, in a fraction of the computer time. However, molecular mechanics cannot of course provide properties that depend upon the electronic distribution in a molecule.

That molecular mechanics works at all is due to the validity of several assumptions. The first of these is the Born-Oppenheimer approximation (the separation of electronic and nuclear motions), without which it would be impossible to contemplate writing the energy as a function of nuclear coordinates at all. Molecular mechanics is based upon a rather simple model of the interactions within a system with contributions from processes such as the stretching of bonds, the opening and closing of angles and the rotation about single bonds. Even when simple functions are used to describe these contributions, the force field can perform quite acceptably. Transferability is a key attribute of a force field, for it enables a set of parameters developed and tested on a relatively small number of cases to be applied to a much wider range of problems. Moreover, parameters developed from data on small molecules can be used to study much larger molecules such as polymers.

Many of the molecular modeling force fields in use today for molecular systems can be interpreted in terms of a relatively simple four-component picture of the intra- and inter-molecular forces within the system. Energetic penalties are associated with the deviation of bonds and angles away from their equilibrium values, there is a function that describes how the energy changes as bonds are rotated, and finally the force field contains interactions between non-bonded parts of the system. More sophisticated force fields may have additional terms, such as hydrogen-bonding term, but they invariably contain these four components. An attractive feature of this representation is that the various terms can be ascribed to changes in specific internal coordinates such as bond lengths, angles, rotation of bonds or movements of atoms relative to each other. This makes it easier to understand how changes in the force field parameters affect its performance, and also helps in the parameterization process. One functional form for such a force field is given in Equation (1).

$$\begin{aligned} V(r^N) = {} & \sum_{bonds} K_r (r_i - r_{eq})^2 + \sum_{angles} K_\theta (\theta_i - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}(1+\cos(n\varphi - \gamma)) \\ & + \sum_{i<j} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}} \right) \end{aligned} \qquad (1)$$

$V(r^N)$ denotes the potential energy, which is a function of the positons (r) of N particles (usually atoms). The first term in equation (1) models the interaction between pairs of bonded atoms, modeled here by a harmonic potential that gives the increase in energy as the bond length r_i deviates from the reference value r_{eq}. The second term is a summation over all valence angles in the molecule, again modeled using a harmonic potential (a valence angle is the angle formed between three atoms A-B-C in which A and C are both bonded to B). The third term in equation (1) is a torsional potential that models how the energy changes as a bond rotates. The forth contribution is the non-bonded term. This is calculated between all pairs of atoms (i and j) that are in the same molecule or in different molecules. In a simple force field the non-bonded term is usually modeled using a Coulomb potential term for the electrostatic interactions and a Lennard-Jones potential for van der Waals interactions.

Unlike the van der Waals potential, the Coulomb interactions decay slowly with distance. In fact, electrostatic interactions are important for stabilizing biomolecular conformations in solvent and associating distant residues in the primary sequence in a compact (folded) structure. The charge distribution in a molecule can be represented in a number of ways, one common approach being an arrangement of fractional point charges throughout the molecule. If the charges are assigned to the nuclear centers they are often referred to as ‘partial atomic charges’ or ‘net atomic charges’.

Besides the terms described above, the force field contains also a so called ‘out-of-plane’ bending term. This term is required to keep structures containing sp^2 atoms such as the nucleic acids bases planar. There are several ways in which out-of-plane bending terms can be incorporated into a force field. One approach is to treat the atoms as an ‘improper’ torsion angle (a torsion angle in which the four atoms are not bonded in the sequence 1-2-3-4). A torsional potential of the form in equation (2) is then used to maintain the improper torsion angle at 0° or 180°.

$$v(\omega) = k(1 - \cos 2\omega) \qquad (2)$$

Although there are other ways to incorporate out-of-plane bendings terms into a force field, the improper torsion definition is most widely used as it can then be easily included with the ‘proper’ torsional terms in the force field.

2.3. Force field parameter development

For all the simulations described in this dissertation I used the ‘AMBER99’ force field (113), which is an improved version of the Cornell et al. 94 force field [equation (1); (114)]. These force fields were developed by the Kollman group and are distributed together with the AMBER (Assisted Model Building with Energy Refinement) software package. The force field contains parameters for all the standard nucleic acids and protein residues. Parameter development is thus required only for new chemical compounds or unusual protein and nucleic acid residues. In this subchapter I will briefly describe how the force field parameters were developed by the Kollman group, emphasizing the strategy to be used to develop parameters for novel chemical entities.

2.3.1. Atom types, bond, angle and dihedral parameters

The first force field that was extensively used to simulate proteins and nucleic acids was developed by Weiner et al. (115,116). The Cornell et al. force field (AMBER94) is a second generation force field that was developed to adjust to the significant increase of computer power that allowed simulations using explicit solvent. Thus, the major changes that were introduced by Cornell et al. comparing to Weiner et al. force field were related to the non-bonded interactions.

The atom types are assigned to each atom in function of the hybridization state and environment. For example, there are nine different types assigned for sp^2 atoms depending of their substituents. An atom type is intrinsic to an array of distance, angle, and dihedral parameters involving the types of the neighboring atoms, as well as having its own van de Waals (VDW) parameters, and atomic mass (charge is not fixed per atom type). Therefore, if a new atom type is required, the first step is to attempt to reason by analogy and clone as many of the pre-existing parameters as possible to account for the environment of the new atom. It is instructive to consider the variability of the existing parameters, which tend to be duplicated over various combinations of atoms. This step may also be required if old atom types are used in a new topological relation.

Equilibrium bond lengths and angles may be obtained from tabulations of experimental data in the literature. Initial bond and angle force constants may be chosen based upon analogy to similar parameters in the force field or using the method of Hopfinger and Pearstein (117). For example, in complex fragments such as the nucleic acids bases, the r_{eq} and θ_{eq} values have been taken from X-ray structural data, the K_r values determined by linear interpolation between pure single and double bond values using the observed bond distances and the K_θ values taken from vibrational analysis of a simple sp^2 atom containing fragments such as benzene or N-methyl acetamide. That this approach was reasonably successful is supported by the reasonable agreement found in nucleic acid base vibrational analysis and suggested by the critical analysis of Halgren of the diagonal force constants used in different force fields (118).

During the development of the AMBER force field, the dihedral parameters were optimized on the simplest molecule possible and then applied to larger and more complex molecules. This approach is in contrast to one employed by other force field developers where the parameters are optimized to best reproduce the conformational energies of a large number of molecules. An advantage of the approach used by Cornell et al. is the lack of dependence of the resulting parameters on the particular molecules chosen for the test set.

The dihedral parameters, in conjunction with the atomic charges and van der Waals parameters are the primary determinants of the relative conformational energies of a molecule. The AMBER parameters 'IDIVF', 'PK', 'PN', and 'PHASE' are used to define the torsional potential energy function. Each bonded series of atoms I-J-K-L must have at least one set of these dihedral parameters in the force field (just as every bonded pair I-J or triplet I-J-K must have bond or angle parameters, except that for dihedrals multiple terms may be used). The torsional energy function formula (from the *AMBER* source code) is:

$$E_{tors} = \left(PK/IDIVF\right) \cdot \left[1 + \cos\left(PN \cdot \varphi - PHASE\right)\right] \qquad (3)$$

If atoms J and K are sp3 carbons (type CT) as in the molecule ethane (H_3C-CH_3), then the intrinsic barrier to rotation about the J-K bond is ~3 kcal/mol. This potential may be generic for torsions about CT-CT bonds (X-CT-CT-X), or explicit restricted to HC substituents (HC-CT-CT-HC). This choice determines IDIVF, which is the total number of torsions about a single bond that the potential applies to. If all atoms are explicit, then IDIVF=1 and the total potential for the bond (3.0 in this case) is divided by the number of torsions involved; since each substituent 'sees' the opposite 3 substituents, there are 3x3=9 torsions around the bond, as would be the case whenever the central bond is between two sp3 atoms. If the generic representation is chosen, then the entire potential is used and IDIVF=9. PK is equal to one-half of the barrier magnitude and would therefore be equal to 3.0 / 2.0 = 1.5 kcal/mol for the generic case, or 3.0 / 9 / 2.0 = 0.1667 for the specific case. The topology about the dihedral of interest has a three-fold periodicity (PN); that is, there are three potential barriers as the C-C bond is rotated -180 to 180 degrees. These barriers occur when the methyl hydrogens eclipse each other: at 0, -120, and 120 degrees. Since the dihedral formula is a Fourier series truncated to a single cosine term, no phase shift would be needed to reproduce the potential energy barriers and PHASE = 0 degrees. (PHASE = 0 degrees if an energy *maximum* is at 0 degrees; PHASE = 180 degrees if an energy minimum is at 0 degrees).

These same torsional parameters can be used for n-butane, and the results are in good agreement with experiment and higher-level calculations for the relative energy of ‘trans’ and ‘gauche’ minima and ‘cis’ and ‘skew’ energy barriers.

In the case of ethylene ($H_2C{=}CH_2$), the lowest-energy conformation of this molecule is planar with a two-fold (PN = 2), 60 kcal/mol barrier to rotation about the C=C bond. The barriers are found at dihedral angles of -90 and 90 degrees (energy minimum at 0 degrees), and can be reproduced by the truncated Fourier series only if a phase shift of 180 degrees (PHASE = 180.0 degrees) is used.

I used these two examples, as they are described in the AMBER user manual to illustrate how the dihedral parameters are derived for different atom types. A similar approach should be considered in case new dihedral parameters are needed and no analogy with existing parameters is possible.

2.3.2. Van der Waals (VDW) parameters

The shape of the VDW potential for a given atom type is specified in terms of the distance between two atoms of the same type at the minimum energy point. Half the interatomic distance at that point is treated as the basic radius, or ‘R*’, parameter for that type. The form for the radial potential for two atoms is the sum of the R* values of their types. The potential well depth (‘e’) of the minimum energy point between two atoms of the same type is combined with the potential of another atom type by taking the root of the product. (Other parametric forms can be used which tend to have different type-type ‘combining rules’). The simplest approach to deriving VDW parameters is to match a relevant experimental determination of the size of the atom in question. One source of such measurements can be diffraction data. Another variety of experimental data that can contribute to parameterization is the free energy of solvation in water or another relevant solvent. However, it is still not clear whether the combination of experimental size and solvation free energy is sufficient to determine unique R* and e parameters for an atom in relation to an existing type. A further complication arises because an atom type may come into contact with more than one other type, and nothing in principle guarantees that VDW parameters for a group of types can be fitted to yield uniformly correct pairwise potentials.

2.3.3. Derivation of partial atomic charges

The derivation of accurate and transferable partial charges on atomic centers is a key step in the calculation of the electrostatic interactions using the force field models. It was shown that the use of quantum mechanical electrostatic potential is an essential element in the derivation of charges that accurately represent the molecular multipole moments (115,116,119,120). Thus, the partial charge models most often involve a least-square fit between the model and the electrostatic potential. The charges derived might have been dependent on the *ab initio* basis set or semiempirical methodology, but a reasonable model of choice was the use of *ab initio*-derived charges using a 6-31G* basis set (121), which uniformly overestimates molecular polarity. This overestimate made such models relatively balanced with empirical solvent models such as TIP3P (119) or SPC (122) that are extensively used to explicitly represent the solvent during calculations. The water models were empirically calibrated to reproduce the density and enthalpy of vaporization of the liquid.

The electrostatic potential-derived charges suffered from two main disadvantages: (i) they were not transferable, chemically similar atoms often had variable charges, and (ii) the derivation of charges for large polymers became impractical, even with powerful computers. Moreover, spurious fluctuations were observed for the charges obtained in the least-squares fitting procedure for the statistically poorly determined (buried) centers. These problems were solved by the development and implementation of multiple-conformation fitting (121,123,124) and the RESP (restrained electrostatic potential) derived charges (123,124).

The detailed procedure of charge derivation for nucleic acid and protein residue was described by Cieplak et al. (125). Electrostatic potentials (ESP) at grids of points (120) around appropriate components of the nucleic acids and proteins were calculated with the Gaussian 90 program using the Hartree-Fock method with the 6-31G* basis set (new versions of Gaussian have been released since then or other quantum mechanical software may be used). Initially, calculations were performed for dimethylphosphate (DMP) in its gauche-gauche (g^+, g^+) conformation optimized at the 6-31G* level, as well as A and B standard forms of deoxyribonucleosides, standard forms of ribonucleosides, and all amino acids with appropriate CH_3-CO- and –NH-CH_3 blocking groups (dipeptides) optimized using molecular mechanics with the Weiner et al. force field. For each amino acid, ESPs were calculated for two sidechain conformers (or four, in the case of proline). However, with the increase of computer power, quantum mechanical calculations such as ESP calculations are now feasible for larger compounds such as the nucleosides or entire amino acids and should be applied if atomic charges are to be derived for new chemical entities.

The next step in the charge derivation procedure is the RESP fitting methodology. The restrained ESP charge fitting is achieved using Equation (4).

$$f\left(q_1,\ldots,q_{natom}\right)=\chi^2_{esp}+\chi^2_{hyp.restr}+\lambda_1 g_1+\ldots+\lambda_w g_w \qquad (4)$$

where:

$$\chi^2_{esp}=\sum_{i=1}^{ESPpo\text{int}s}\left(V_i-\sum_{j=1}^{natoms}\frac{q_j}{r_{ij}}\right)^2 \qquad (5)$$

and

$$\chi^2_{hyp.restr}=a\sum_{j=1}^{natoms}\left(\left(q_j^2+b^2\right)^{\frac{1}{2}}-b\right) \qquad (6)$$

In Equations (4), (5) and (6), V_i is the quantum mechanically calculated electrostatic potential (ESP) at point i, q_j are the resultant charges, a is a scale factor defining the asymptotic limits of the strength of the hyperbolic restraint according to Equation (6), and b defines the tightness of the hyperbola around the minimum. A value of 0.1 was found to be appropriate for b (123). The g_i are additional constraints imposed on resultant charges, and λ_i are Lagrange multipliers. The minimum of the $f(q_1,\ldots,q_{natoms})$ function is sought by requiring that

$$\frac{\partial f}{\partial q_k}=0,\frac{\partial f}{\partial \lambda_l}=0\text{, foreach } k, l \qquad (7)$$

which leads to the matrix equation of the type

$$Aq=B \qquad (8)$$

which must be solved for q. This is done iteratively when using nonzero hyperbolic restraints, since the left-hand side of equation (8) (matrix A) depends on charges q.

The RESP fitting involves a two-stage procedure with hyperbolic restraints. In the first stage, a weak hyperbolic restraint (a = 0.0005) to a target value of 0.0 is applied to all heavy atoms. Hydrogen atoms are not restrained because they are never buried within a molecule and are always well defined by the ESP points. In the second stage, charges on all atoms were kept frozen to their values obtained in the first stage, except those in methyl and methylene groups. CH_3 and CH_2 are refitted with the hydrogens within a given group constrained to have equivalent charges. The hyperbolic restraint applied during the second stage is twice as strong as the one is stage one (a = 0.001). The two-stage restrained ESP charges exhibit less conformational dependence compared to the standard ESP charges, result in excellent conformational energies, and give good results for hydrogen bonding energies and free energies of salvation (123,124).

The necessity of a two-stage fit arises from the need to constrain atoms which are not symmetrically equivalent within the static conformation of the molecule used for the calculation but which become equivalent under dynamical conditions when rotation can occur. One example of this would be the three methyl hydrogens in methanol. If these nonequivalent atoms are forced to have the same charge during a one-stage fit, the charge on the oxygen is reduced to a value which does not yield good free energies of solvation or interaction energies. The two-stage fit then allows for the 'best' charges to be fit on the heteroatoms during the first stage, with the maximum number of degrees of freedom available to the molecule. Then methyl or methylene hydrogens are constrained to be equivalent in the second stage of the fit. For a detailed description of how the charges were derived for the nucleic acid and protein residues see (125).

However, the RESP charges derived using the procedure presented above are dependent on the orientation of the optimized structure provided by the quantum mechanical software (different optimizations lead to different orientations of the same molecule) and on the conformation of the molecule. In order to overcome these limitations, a new program (R.E.D: www.u-picardie.fr/labo/lbpd/RED/index.php) was recently developed. It employs a reorientation algorithm to perform multiorientation and multiconformation fit to derive RESP charges. R:E.D. provides a platform for automated derivation of reproducible RESP charges.

2.4. Energy minimization

The minimization of the potential energy function (i.e., geometry optimization) involves a search for the minimum of a function, and, to be efficient, requires calculations of derivatives of a function (in this case, the potential energy) versus independent variables (in our case, coordinates). Most programs use cartesian coordinates as independent variables, however, in some cases, internal coordinates may be used. The derivatives of potential energy are denoted as:

$$g_i = \frac{\partial V\left(r^N\right)}{\partial r_i}, H_{ij} = \frac{\partial^2 V\left(r^N\right)}{\partial r_i \partial r_j} \qquad (9)$$

where g_i is the gradient (i.e., first derivative) of the potential energy $V(r^N)$ with respect to a cartesian coordinate r_i of an atom, and H_{ij} is the second derivative of the energy with respect to the cartesian coordinates. In most modern programs these derivatives are calculated analytically, i.e., the appropriate mathematical formulae for corresponding terms are incorporated into the program. Some older codes compute derivatives numerically by approximating the slope of an energy function (or its gradient in the case of second derivatives) from finite differences. The table of all possible second derivatives versus cartesian coordinates of atoms has 3N rows and 3N columns and is called a 'Hessian matrix'. The derivatives are used not only in function minimization but also yield forces acting on atoms (from energy gradients) and normal modes of vibration (from the Hessian matrix). There are three major approaches to finding a minimum of a function of many variables: (i) 'search methods' - utilize only values of the function itself. They are usually slow and inefficient, but are very simple to program, since deriving cumbersome formulas for derivatives is not necessary. In spite of their inefficiency, the search algorithms are infallible and always find a minimum. For this reason, they are often used as an initial step, when the starting point in optimization is far from the minimum. Another disadvantage of search techniques is that they are very inefficient for a large number of optimized variables and converge very slowly when the number of variables is more then 10, (ii) 'gradient methods' - utilize values of a function and its gradients. These are currently the most popular methods in molecular mechanics. They offer a much better convergence rate than search methods and do not require a lot of computer memory (only 3N first derivatives are needed). However, in some situations they fail to converge to a minimum. The 'conjugated gradient' algorithm is considered the most robust in this class, and (iii) 'Newton methods' - are the most rapidly converging algorithms which require values of function, and its first and second derivatives. The memory required for storing the Hessian matrix is proportional to N^2 (i.e., prohibitive for large macromolecules). The 'BFGS' algorithm is considered the most refined one.

In general, the minimization methods are iterative. They require on input some initial estimate for the position of the minimum, and provide a better estimate for the minimum as a result. This corrected estimate is used as an input into the next cycle (i.e., iteration) and the process is continued until there is no significant improvement in the position of the minimum. Most search methods and minimization methods using derivatives are the 'descent series methods', i.e., each iteration results in a solution which corresponds to a lower (or equal) value for the energy function:

$$V\left(x^{start}\right) \geq V\left(x^{1}\right) \geq V\left(x^{2}\right) \ldots \geq V\left(x^{\min}\right) \qquad (10)$$

As a consequence, these methods can only find the minimum closest to the starting estimate and will never cross to a minimum (however deep) if it is separated from the starting estimate by a maximum (however small). This situation is schematically illustrated in Figure 10 where arrows indicate the direction of geometry optimization depending upon the starting point.

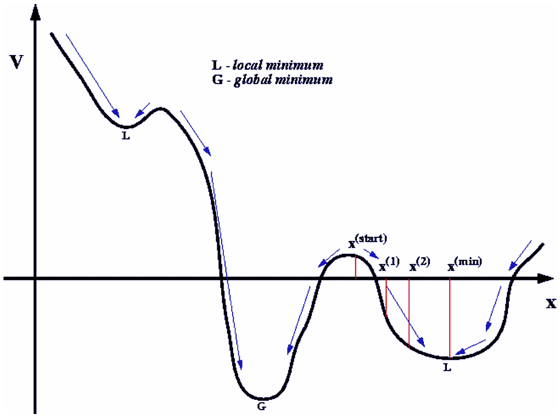


Figure 10: simplified view of the energy landscape.
The blue arrows indicate the direction of the minimization algorithm. $x^{(start)}$, $x^{(1)}$, $x^{(2)}$, $x^{(min)}$ represent successive coordinates during the minimization algorithm which starts at position $x^{(start)}$, goes through positions $x^{(1)}$ and $x^{(2)}$ until it reaches a minimum corresponding to $x^{(min)}$

Figure 10 is only a cartoon to illustrate the behavior of descent series minimization methods. Geometry optimization of real molecular systems involves simultaneous optimization of 3N cartesian coordinates, i.e, sometimes many thousands of variables. There is no general way of finding a global minimum (i.e., the minimum corresponding to the lowest possible value of the function). A different initial geometry will usually lead to a different final minimum. Only on very simple molecules will the single geometry optimization yield the global minimum on the first trial. However, It cannot be overemphasized that the result of a single minimization is usually a local minimum, not a global one. To find a global minimum (or at least, to be more confident about it) many minimizations need to be performed using different initial coordinates for each run.

All the structures that I simulated (see 2.8.1) were minimized using the conjugate gradient method before running MD simulations in order to remove any possible sterical clashes due to the low resolution of the crystal structure of the 15.5K-KtU4 complex.

2.5. Molecular dynamics simulations

In molecular dynamics, successive configurations of the system are generated by integrating Newton's laws of motion. The result is a trajectory that specifies how the positions and velocities of the particles in the system vary with time. Newton's laws of motion can be stated as follows: (i) a body continues to move in a straight line at constant velocity unless a force acts upon it, (ii) force equals the rate of change of momentum, (iii) to every action there is an equal and opposite reaction. The trajectory is obtained by solving the differential equations embodied in Newton's second law ($F = ma$):

$$\frac{d^2 x_i}{dt^2} = \frac{F_{x_i}}{m_i} \qquad (11)$$

This equation describes the motion of a particle of mass m_i along one corrdinate (x_i) with F_{xi} being the force exerted on the particle in that direction.

There are many algorithms for integrating the equations of motion, several of which are commonly used in MD calculations. All algorithms assume that the positions and dynamic properties (velocities, accelerations, etc.) can be approximated as Taylor series expansions:

$$r(t+\delta t) = r(t) + \delta t v(t) + \frac{1}{2}\delta t^2 a(t) + \frac{1}{6}\delta t^3 b(t) + \frac{1}{24}\delta t^4 c(t) + \ldots \qquad (12)$$

$$v(t+\delta t) = v(t) + \delta t a(t) + \frac{1}{2}\delta t^2 b(t) + \frac{1}{6}\delta t^3 c(t) \ldots \qquad (13)$$

$$a(t+\delta t) = a(t) + \delta t b(t) + \frac{1}{2}\delta t^2 c(t) \ldots \qquad (14)$$

$$b(t+\delta t) = b(t) + \delta t c(t) \ldots \qquad (15)$$

where v is velocity (the first derivative of positions with respect to time), a is acceleration (the second derivative), b is the third derivative, and so on. The 'Verlet' algorithm (126) is probably the most widely used method for integrating the the equations of motion in a MD simulation. It uses the positions and accelerations at time t, and the positions from the previous step, $r(t\text{-}\delta t)$, to calculate the new postions at $t+\delta t$, $r(t+\delta t)$. The following relationship can be written between these quantities and the velocities at time t:

$$r(t+\delta t)=r(t)+\delta t v(t)+\frac{1}{2}\delta t^2 a(t)+\ldots \qquad (16)$$

$$r(t-\delta t)=r(t)-\delta t v(t)+\frac{1}{2}\delta t^2 a(t)-\ldots \qquad (17)$$

Adding these two equations gives:

$$r(t+\delta t)=2r(t)-r(t-\delta t)+\delta t^2 a(t)-\ldots \qquad (18)$$

The velocities do not explicitly appear in the Verlet integration algorithm. The velocities can be calculated in a variety of ways; a simple approach is to devide the difference in positions at times $t+\delta t$ and $t\text{-}\delta t$ by $2\delta t$:

$$v(t)=[r(t+\delta t)-r(t-\delta t)]/2\delta t \qquad (19)$$

Alternatively, the velocities can be estimated at the half-step, $t+½\delta t$:

$$v\left(t+\frac{1}{2}\delta t\right)=[r(t+\delta t)-r(t)]/\delta t \qquad (20)$$

Implementation of the Verlet algorithm is straightforward and the storage requirements are modest, comprising two sets of positions [$r(t)$ and $r(t\text{-}\delta t)$] and the accelerations $a(t)$. One of its drawbacks is that the positions $r(t+\delta t)$ are obtained by adding a small term [$\delta t^2 a(t)$] to the difference of two much larger terms, $2r(t)$ and $r(t\text{-}\delta t)$. This may lead to loss of precision. The algorithm has some other disadvantages. The lack of an explicit velocity term in the equations makes it difficult to obtain the velocities, and indeed the velocities are not available until the positions have been computed at the next step. In addition, it is not a self-starting algorithm; the new positions are obtained from the current positions $r(t)$ and the positions from the previous time step, $r(t\text{-}\delta t)$. At $t=0$ there is obviously only one set of positions and so it is necessary to employ some other means to obtain positions at $t\text{-}\delta t$. One way to obtain $r(t\text{-}\delta t)$ is to use the Taylor series, Equation (17) truncated after the first term. Thus, $r(-\delta t)=r(0)-\delta t v(0)$.

Several variations on the Verlet algorithm have been developed. The 'leap frog' algorithm uses the following relationships:

$$r(t+\delta t)=r(t)+\delta t v\left(t+\frac{1}{2}\delta t\right) \qquad (21)$$

$$v\left(t+\frac{1}{2}\delta t\right)=v\left(t-\frac{1}{2}\delta t\right)+\delta t a(t) \qquad (22)$$

To implement this algorithm, the velocities $v(t+½\delta t)$ are first calculated from the velocities at time $t-½\delta t$ and the accelerations at time t. The positions $r(t+\delta t)$ are then deduced from the velocities just calculated together at the positions $r(t)$ using Equation (21). The velocities at time t can be calculated from:

$$v(t)=\frac{1}{2}\left[v\left(t+\frac{1}{2}\delta t\right)+v\left(t-\frac{1}{2}\delta t\right)\right] \qquad (23)$$

The velocities thus ‘leap-frog’ over the positions to give their values at $t+½\delta t$ (hence the name). The positions then leap over the velocities to give their new values at $t+\delta t$, ready for the velocities at $t+(3/2)\delta t$, and so on. The leap-frog method has two advantages over the standard Verlet algorithm: it explicitly includes the velocity and does not require the calculation of differences of large numbers. However, it has the obvious disadvantage that the positions and velocities are not synchronized. This means that it is not possible to calculate the kinetic energy contribution to the total energy at the same time as the positions are defined (from which the potential energy is determined).

The ‘velocity Verlet’ method gives positions, velocities and accelerations at the same time and does not compromise precision:

$$r(t+\delta t)=r(t)+\delta t v(t)+\frac{1}{2}\delta t^2 a(t) \qquad (24)$$

$$v(t+\delta t)=v(t)+\frac{1}{2}\delta t\left[a(t)+a(t+\delta t)\right] \qquad (25)$$

This method is actually implemented as three-stage procedure because, as can be seen from Equation (25), to calculate the new velocities requires the accelerations at both t and $t+\delta t$. Thus in the first step the positions at $t+\delta t$ are calculated according to Equation (24) using the velocities and the accelerations at time t. The velocities at time $t+½\delta t$ are then determined using:

$$v\left(t+\frac{1}{2}\delta t\right)=v(t)+\frac{1}{2}\delta t a(t) \qquad (26)$$

New forces are then computed from the current positions, thus giving $a(t+\delta t)$. In the final step, the velocities at time $t+\delta t$ are determined using:

$$v(t+\delta t)=v\left(t+\frac{1}{2}\delta t\right)+\frac{1}{2}\delta t a(t+\delta t) \qquad (27)$$

2.5.1. Choosing the time step

There are no hard and fast rules for calculating the time step to use in MD simulations; too small and the trajectory will cover only a limited portion of the phase space; too large and instabilities may rise in the integration algorithm due to high energy overlaps between atoms. Such instabilities would certainly lead to a violation of energy and linear momentum conservation and could result in a program failure due to numerical overflow. The total error is correlated with the time step, with the largest errors arising for large time steps. Of course, with a small time step much more computer time will be required for a given length of calculation; the aim is to find the correct balance between simulating the ‘correct’ trajectory and covering the phase step.

When simulating an atomic fluid (such as solvated biomolecules) the time step should be small comparing to the mean time between collisions. When simulating flexible molecules, a useful guide is that the time step should be approximately one-tenth the time of the shortest period of motion. In flexible molecules, the highest-frequency vibrations are due to bond stretches, especially those of bonds to hydrogen atoms. A C-H bond vibrates with a repeat period of approximately 10 fs.

The requirement that the time step is approximately one order of magnitude smaller than the shortest motion is clearly a severe restriction, particularly as these high-frequency motions are usually of relatively little interest and have a minimal overall behavior of the system. One solution to this problem is to ‘freeze out’ such vibrations by constraining the appropriate bonds to their equilibrium values while still permitting the rest of the degrees of freedom to vary under the intra- and inter-molecular forces present. This enables a longer time step to be used.

2.5.2. The SHAKE / RATTLE algorithm

As discussed above, the choice of the time step in MD simulations of biomolecules is restricted by the highest-frequency vibrations of the system (vibrations of X-H bonds). Therefore, to increase the time step by a small margin (generally from 1 fs to 2 fs) a procedure called ‘SHAKE’ [combined with standard Verlet algorithm; (127)] or RATTLE [combined with the velocity Verlet algorithm*;* (128)] is applied to constrain the bonds involving hydrogen atoms to a fixed length.

In the general case, the equations of motion for a constrained system involve two types of force: the ‘normal’ forces arising from the intra- and inter-molecular interactions, and the forces due to the constraints. The force due to the constraint σ_k that requires that the bond between atoms i and j remains fixed ca be written as follows:

$$F_{ckx} = \lambda_k \frac{\partial \sigma_k}{\partial x} \qquad (28)$$

where λ_k is the Lagrange multiplier and x represents one of the Cartesian coordinates of the two atoms. If an atom is involved in a number of constraints then the total constraint force equals the sum of all such terms. The nature of the constraint for a bond between atoms i and j is:

$$\sigma_{ij} = \left(r_i - r_j\right)^2 - d_{ij}^2 = 0 \qquad (29)$$

The constraint force lies along the bond all the times. For each constrained bond, there is an equal and opposite force on the two atoms that comprise the bond. The overall effect is that the constraint forces do not work. The constraint forces are obtained by differentiating the constraint with respect to the coordinates of atoms i and j and multiplying by an (as yet) undetermined multiplier:

$$\frac{\partial \sigma_k}{\partial r_i} = 2\left(r_i - r_j\right) \quad \Rightarrow \quad F_{ci} = \lambda\left(r_i - r_j\right) \qquad (30)$$

$$\frac{\partial \sigma_k}{\partial r_j} = -2\left(r_i - r_j\right) \quad \Rightarrow \quad F_{cj} = -\lambda\left(r_i - r_j\right) \qquad (31)$$

The factor of 2 that arises when the square term was differentiated has been incorporated into the Lagrange multiplier λ. The above expression for the forces can be incorporated into the Verlet algorithm as follows:

$$r_i\left(t + \delta t\right) = 2r_i\left(t\right) - r_i\left(t - \delta t\right) + \frac{\delta t^2}{m_i} F_i\left(t\right) + \sum_k \frac{\lambda_k \delta t^2}{m_i} r_{ij}\left(t\right) \qquad (32)$$

Recall that the positions that would be obtained from the Verlet algorithm without constraints are $r_i^{'}(t + \delta t) = 2r_i(t) - r_i(t - \delta t) + \delta t^2 F_i(t) / m_i$. The summation in Equation (32) is over all constraints k that affect atom i. These constraints perturb the positions that would otherwise have been obtained from the integration algorithm, and so the above expression can be written:

$$r_i\left(t + \delta t\right) = r_i^{'}\left(t + \delta t\right) + \sum_k \frac{\lambda_k \delta t^2}{m_i} r_{ij}\left(t\right) \qquad (33)$$

Equations (29) and (33) constitute a system of two equations with two unknowns [$r_i(t+\delta t)$ and λ_k]. When there are many constraints, the problem is equivalent to inverting a k x k matrix, even when quadratic terms of λ are ignored. The SHAKE method uses an approach in which each constraint is considered in turn and solved. Satisfying one constraint may cause another constraint to be violated, and so it is necessary to iterate around the constraints until they are all satisfied to within some tolerance. The tolerance should be tight enough to ensure that the fluctuations in the simulation due to the SHAKE algorithm are much smaller than the fluctuations due to other sources. Another important requirement is that the constrained degrees of freedom should be only weakly coupled to the remaining degrees of freedom, so that the motion of the molecule is not affected by the application of constraints.

In the case of the velocity Verlet algorithm, the method has been named RATTLE. When velocities appear in the integration algorithm, they must be corrected as well as the positions.

2.5.3. Periodic boundary conditions

In MD simulations, the solvent can be represented either implicitly or explicitly. In all the simulations that I performed the solvent was explicitly represented using the TIP3P water model (119). A cubic box of solvent was built around the solute and periodic boundary conditions (PBC) were applied during the simulations.

If PBC were not applied, the system would simply terminate, and atoms near the boundary would have fewer neighbors than atoms inside. In other words, the sample would be surrounded by surfaces. This situation is not realistic. No matter how large the simulated system is, its number of atoms N would be negligible compared with the number of atoms contained in a macroscopic piece of matter (of the order of 10^{23}), and the ratio between the number of surface atoms and the total number of atoms would be much larger than in reality, causing surface effects to be much more important than what they should.

When using PBC, particles are enclosed in a box, and this box is replicated to infinity by rigid translation in all the three cartesian directions, completely filling the space. In other words, if one of the particles is located at position r in the box, its assumed that this particle really represents an infinite set of particles located at: $r+la+mb+nc$, $(l,m,n=-\infty, \infty)$. where l, m, and n are integer numbers, and a, b, and c are the vectors corresponding to the edges of the box. All these 'image' particles move together, and in fact only one of them is represented in the computer program. The key point is that now each particle i in the box should be thought as interacting not only with other particles j in the box, but also with their images in nearby boxes. Thus, interactions can 'go through' box boundaries. In fact, one can easily see that: (i) the surface effects from the system were virtually eliminated, and (ii) the position of the box boundaries has no effect (that is, a translation of the box with respect to the particles leaves the forces unchanged).

2.5.4. The Particle Mesh Ewald method

The 'Particle Mesh Ewald' (PME) metod was developed by Darden et al. (129) for the evaluation of the electrostatic interactions of large periodic system. Those interactions that decay no faster than r^{-n}, where n is the dimensionality of the system can be a problem as their range is often greater than half the box length. The charge-charge interaction, which decays as r^{-1}, is particularly problematic in molecular simulations. There is much evidence that it is important to properly model these long-range forces, which are particularly acute when simulating charged species such as nucleic acids. The development of PME together with the explicit representation of the solvent and the increase of computer power had a great impact on the accuracy of biomolecular simulations (130).

The Ewald sum was first devised by Ewald in 1921 to study the energetics of ionic crystals. In this method, a particle interacts with all the other particles in the simulation box and with all of their images in an infinite array of periodic cells. The cell array is considered to have a spherical shape. The position of each image box (assumed for simplicity to be a cube of side L containing N charges) can be related to the central box by specifying a vector, each of whose components is an integer multiple of the length of the box, $(\pm iL, \pm jL, \pm kL)$; i,j,k = 0,1,2,3, etc. The charge-charge contribution to the potential energy due to all pairs of charges in the central simulation box can be written:

$$V = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \qquad (34)$$

where r_{ij} is the minimum distance between the charges i and j. There are six boxes at a distance L from the central box with coordinates (r_{box}) given by (0,0,L), (0,0,-L), (0,L,0), (0,-L,0), (L,0,0), and (-L,0,0). The contribution of the charge-charge interaction between the charges in the central box and all images of all particles in the six surrounding boxes is given by:

$$V = \frac{1}{2} \sum_{|n|=0} {}' \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{4\pi\varepsilon_0 \, | r_{ij} + n |} \qquad (35)$$

This expression includes the interaction between charges in the central box for which $|n|=0$. The prime on the first summation indicates that the series does not include the interaction $i=j$ for $n=0$. There is also a contribution from the interaction between the spherical array of boxes and the surrounding medium. The problem is that the summation in Equation (35) converges extremely slowly and in fact is conditionally convergent (contains a mixture of positive and negative terms which separately form divergent series). The sum of a conditionally convergent series depends on the order in which its terms are considered. An additional problem with the Coulomb interaction is that in can vary rapidly at small distances.

The trick when calculating the Ewald sum is to convert the summation into two series, each of which converges more rapidly. The mathematical foundation for this is the following identity:

$$\frac{1}{r} = \frac{f(r)}{r} + \frac{1 - f(r)}{r} \qquad (36)$$

The aim is thus to choose an appropriate function $f(r)$ which will deal with the rapid variation of $1/r$ at small r and the slow decay at long r. In the Ewald method each charge is considered to be surrounded by a neutralizing charge distribution of equal magnitude but of opposite sign. A Gaussian charge distribution of the following functional form is commonly used:

$$\rho_i(r) = \frac{q_i \alpha^3}{\pi^{3/2}} \exp\left(-\alpha^2 r^2\right) \qquad (37)$$

The sum over point charges is converted to a sum of the interactions between the charges plus the neutralizing contributions. This dual summation (the ‘real space’ summation) is given by:

$$V = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{|n|=0} {}' \frac{q_i q_j}{4\pi\varepsilon_0} \frac{erfc\left(\alpha \, | r_{ij} + n |\right)}{| r_{ij} + n |} \qquad (38)$$

$erfc$ is the complementary error function, which is:

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} \exp(-t^2) dt \qquad (39)$$

The Ewald method thus uses $erfc(r)$ for the function $f(r)$ in Equation (36). This new summation involving the error function converges very rapidly and beyond some cutoff distance its value can be considered negligible. The rate of convergence depends upon the width of the canceling Gaussian distributions; the wider the Gaussian, the faster the series converges. Specifically, α should be chosen so that the only terms in the series (38) are those for which $|n|=0$ (i.e. only pairwise interactions involving charges in the central box, or if a cutoff is used α is chosen so that so that only interactions with other charges within the cutoff are included). A second charge distribution is now added to the system which exactly counteracts the first neutralizing distribution:

$$V = \frac{1}{2} \sum_{k \neq 0} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{1}{\pi L^3} \frac{q_i q_j}{4\pi\varepsilon_0} \frac{4\pi^2}{k^2} \exp\left(-\frac{k^2}{4\alpha^2}\right) \cos\left(k \cdot r_{ij}\right) \qquad (40)$$

This summation is performed in the reciprocal space, the details of which I will not describe here. The vectors k are reciprocal vectors and are given by $k=2\pi n/L$. This reciprocal sum also converges much more rapidly than the original point-charge sum. However, the number of terms that must be included increases with the width of the Gaussians. There is thus a clear need to balance the real space and reciprocal space summations; the former converges more rapidly for large values of α, whereas the latter converges more rapidly for small α. A value of α of 5/L and 100-200 reciprocal vectors k, have been suggested as providing acceptable results. The reciprocal summation corresponds to the second term in Equation (36); the requirement for this term is that it is a slowly varying function for all r. As such its Fourier transform (which is what the summation is) can be represented by a small number of reciprocal vectors. The sum of Gaussian functions in the real space includes the interaction of each Gaussian with itself. A third self-term must therefore be substracted:

$$V = \frac{\alpha}{\sqrt{\pi}} \sum_{k=1}^{N} \frac{q_k^2}{4\pi\varepsilon_0} \qquad (41)$$

A fourth correction term may also be required, depending upon the medium that surrounds the sphere of simulation boxes. If the surrounding medium has an infinite relative permittivity (e.g. if it is a conductor) then no correction term is required. However, if the surrounding medium is a vacuum (with a relative permittivity of 1) then the following energy must me added:

$$V_{correction} = \frac{2\pi}{3L^3} \left| \sum_{i=1}^{N} \frac{q_i}{4\pi\varepsilon_0} r_i \right|^2 \qquad (42)$$

The final expression is thus:

$$V = \frac{1}{2} \sum_{i=0}^{N} \sum_{j=0}^{N} \left\{ \begin{aligned} & \sum_{|n|=0}^{\infty}{}' \frac{q_i q_j}{4\pi\varepsilon_0} \frac{erfc\left(\alpha \,|\, r_{ij} + n \,|\right)}{|\, r_{ij} + n \,|} + \\ & + \sum_{k \neq 0} \frac{1}{\pi L^3} \frac{q_i q_j}{4\pi\varepsilon_0} \frac{4\pi^2}{k^2} \exp\left(-\frac{k^2}{4\alpha^2}\right) \cos\left(k \cdot r_{ij}\right) - \\ & - \frac{\alpha}{\sqrt{\pi}} \sum_{k=1}^{N} \frac{q_k^2}{4\pi\varepsilon_0} + \frac{2\pi}{3L^3} \left| \sum_{k=1}^{N} \frac{q_k}{4\pi\varepsilon_0} r_k \right|^2 \end{aligned} \right\} \qquad (43)$$

The Ewald summation is computationally quite expensive to implement. The most promising way to tackle this difficulty is to modify the problem such as the fast Fourier transform (FFT) can be used to compute the reciprocal space summation. The FFT algorithm scales with $N \cdot lnN$ (as compared with N^2 scaling of the Ewald sum). If, in addition a sufficiently large value of α is chosen such that the inter-atomic interaction is negligible for r_{ij} greater than the cutoff then the real-space summation is reduced to order N and the order of the entire algorithm becomes $N \cdot lnN$.

The FFT method requires that the data are not continuous but are discrete values. In order to employ FFT in the Ewald summation the point charges with their continuous coordinates must be replaced by a grid-based charge distribution. Each of the atomic point charges must thus be distributed among the surrounding grid points in some fashion so as to reproduce the potential of the charge at the original location. The more surrounding points that are used the more accurately the potential of the charge at the original location can be approximated but the greater the computational cost per particle. The particle-mesh method of Hockney and Eastwood (131) uses the nearest 27 points in three dimensions. From this gridded charge density it is possible to calculate (through the use of the FFT algorithm) the potential due to the Gaussian distributions at the grid points, which by interpolation gives rise to the desired potential at (and thus the forces on) each of the particles. PME is a variant of the general scheme discussed in this chapter.

2.5.5. Constant temperature dynamics

In order to be able to simulate the real behavior of macro-molecules, it is required to control the temperature during the simulations. The simulations that I describe were run at room temperature (300 K). In order to achieve the desired temperature, a constant volume equilibration phase (first equilibration phase of 60 ps with 1 fs time step) was run in which the temperature was slowly increased from a low initial value (10 K) to 300 K in the first 30 ps and was maintained at 300 K for the rest of 30 ps. The solute was restrained from fluctuation, thus allowing only solvent rearrangements.

The temperature was controlled by coupling the system to an external bath that is fixed at the desired temperature (132). The bath acts as a source of thermal energy, supplying or removing heat from the system as appropriate. The velocities are scaled at each step, such as the rate of change of temperature is proportional to the difference in temperature between the bath and the system:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}(T_{bath} - T(t)) \qquad (44)$$

τ is a coupling parameter whose magnitude determines how tightly the bath and the system are coupled together. This method gives an exponential decay of the system towards the desired temperature. The change in temperature between successive time steps is:

$$\Delta T = \frac{\delta t}{\tau}(T_{bath} - T(t)) \qquad (45)$$

The scaling factor for the velocities is thus:

$$\lambda^2 = 1 + \frac{\delta t}{\tau}\left(\frac{T_{bath}}{T(t)} - 1\right) \qquad (46)$$

If τ is large, then the coupling will be week. If τ is small, the coupling will be strong and when the coupling parameter equals the time step ($\tau = \delta t$) then the algorithm is equivalent to the more simple velocity scaling method (which I did not explain).

During the equilibration phase, the coupling was maintained stronger, while during the production runs (actual simulations) the coupling was weakened to reflect more the real behavior of the system which fluctuates about the desired temperature.

However, it must be noted that the temperature coupling method does not generate rigorous canonical averages. Velocity scaling artificially prolongs any temperature differences among the components of the system. Other methods such as the stochastic collisions or the extended system methods shall be considered in the future.

2.5.6. Constant pressure dynamics

Just as the temperature can be specified and controlled during MD simulations, it may be important to maintain the system at constant pressure. This enables the behavior of the system to be explored as a function of the pressure, enabling the study of phenomena such as the onset of pressure-induced phase transitions. Many experimental measurements are performed under conditions of constant temperature and pressure, thus simulations in the isothermal-isobaric ensemble are most directly relevant to experimental data.

The pressure often fluctuates much more than quantities such as the total energy in a constant 'NVE' MD simulation. This is as expected because the pressure is related to the 'virial', which is obtained as the product of the positions and the derivative of the potential energy function. This product, $r_{ij} dV(r_{ij}) / dr_{ij}$, changes more quickly with *r* than does the internal energy, hence the greater fluctuation in the pressure.

A macroscopic system maintains constant pressure by changing its volume. A simulation in the isothermal-isobaric ensemble also maintains constant pressure by changing the volume of the simulation cell. The amount of volume fluctuation is related to the isothermal compressibility, k:

$$k = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T \qquad (47)$$

In an isobaric simulation, the volume can change in all directions, or just in one direction. The isothermal compressibility is related to the mean square volume displacement by:

$$k = \frac{1}{k_B T}\frac{\langle V^2 \rangle - \langle V \rangle^2}{\langle V^2 \rangle} \qquad (48)$$

The isothermal compressibility of an ideal gas is approximately 1 atm^{-1}. Thus, for a simulation in a cubic box of side 20 Å (volume 8000 Å^3) at 300 K, the root mean square change in the volume is approximately 18100 Å^3. This is larger than the initial size of the box! For a relatively incompressible substance as water (k = 44.75 x 10^{-6} atm^{-1}) the fluctuation is 121 Å^3 (about 0.1 Å in each direction). These values have clear implications for the appropriate size of the simulation system.

In the simulations described in this dissertation, a second equilibration phase of 140 ps was run in the isothermal-isobaric conditions to allow the system to reach a proper density. The restraints on the solute movement (applied in the first equilibration step) were gradually reduced, such as that for the last 70 ps the solute was allowed to fluctuate freely. The system was coupled to a “pressure bath”, analogues to the temperature bath (132). The rate of change of pressure is given by:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_P}(P_{bath} - P(t)) \qquad (49)$$

τ_P is the coupling constant, P_{bath} is the pressure of the bath (in my case 1 atm), and $P(t)$ is the actual pressure at time t. The volume of the simulation box was scaled by a factor λ, which is equivalent to scaling the atomic coordinates by a factor of $\lambda^{1/3}$. Thus:

$$\lambda = 1 - k\frac{\delta t}{\tau_P}(P - P_{bath}) \qquad (50)$$

The constant k can be combined with the relaxation constant τ_P as a single constant. This expression can be applied isotropically (i.e. such as the scaling factor is equal in all three directions) or anisotropically (the scaling factor is calculated independently for each of the three axes).

2.6. Locally enhanced sampling

As described in 2.5.1. and 2.5.2., the time step in MD simulations is restricted to 1-2 fs. Thus, the time scale that is accessible to standard simulations in restricted to nanoseconds. However, conformational transitions that are relevant for the biological function of diverse macromolecules occur on the microseconds to seconds time scale, depending on the amplitude of such transitions. Additionally, the trajectories obtained from MD simulations depend on the starting structure and it is often the case that an experimental determined structure (that corresponds to a minimum on the energy surface) is unable to undergo conformational transitions if the energy barriers required are too large.

Methods have been developed to increase the conformational sampling during MD simulations. For studying the dynamics of KtU4 RNA I employed the ‘locally enhanced sampling’ (LES) method. LES is a mean-field based theory (MFT) that can be coupled with explicit solvent treatment and PME in constant pressure simulations using periodic boundary conditions (84). It consists of dividing the system into separate regions and replacing them with multiple copies. In this way the energy potential surface is smoothened and the copies are allowed to sample more of the conformational space (85). Such a setup has been successfully employed before to explore structural diversity in RNA or protein loops (80-84).

If the coordinate vector of all atoms in the molecular system is X, the probability of finding the molecular coordinates between X and $X+dX$ is denoted by $\mu(X)dX$, where $\mu(X)$ is the probability density of the coordinates, normalized to 1. The total energy of the system is given by:

$$E_{eff} = \int V(X)\mu(X)dX \qquad (51)$$

where $V(X)$ is the potential energy function. If the system contains a single molecule with a unique conformation whose coordinates are X^0, Equation (51) becomes:

$$E^0 = V(X^0) \qquad (52)$$

The native conformation X_{nat} of the molecular system corresponds to the global minimum of E_{eff}. The search for this global minimum is hindered by the presence of many local minima. One way to alleviate this problem is to consider an effective, larger system as a computational tool to enhance the conformational sampling during simulations. This larger system is obtained by considering multiple copies of the molecule, or parts of the molecule.

According to the ‘Hartree’ approximation, the probability density μ is replaced by a product of independent probability densities of different systems, using a Hartree product:

$$\mu(X) = \prod_{j=1}^{J} \mu_j(X_j) \qquad (53)$$

As examples of such a partition into subsystems, a ligand and a protein can be considered separately, the partition of a protein into backbone and side-chains, or the partition of an RNA stem-loop into loop and stem. The effective system is then built by considering multiple copies of each subsystem j.

The probability density μ_j of each subsystem is expanded into a finite number of delta functions:

$$\mu_j(X_j) = \sum_{k_j=1}^{K_j} P(j,k_j)\delta(X_j - X^0_{k_j}) \qquad (54)$$

where k_j runs over all K_j copies of the subsystem j, $P(j,k_j)$ are normalization factors or probabilities veryfying:

$$\sum_{k_j=1}^{K_j} P(j,k_j) = 1 \qquad (55)$$

For practical reasons, K_j is always finite.

Substituting Equations (53) and (54) into Equation (51) and integrating over the spatial variables leads to the following functional form for the energy of the effective system:

$$E_{eff} = \sum_{j=1}^{J} \sum_{k_j}^{K_j} \left(\prod_{l=1}^{J} P(j,k_j) \right) V(X_{k_1}^0, \ldots, X_{k_j}^0, \ldots, X_{K_j}^0) \qquad (56)$$

The potential energy function V is assumed to be a one and two-body potential:

$$V(X) = \sum_{j=1}^{J} V_j(X_j) + \sum_{j=1}^{J} \sum_{i \geq j} V_{ij}(X_i, X_j) \qquad (57)$$

in which Equation (56) reduces to:

$$\begin{aligned} E_{eff} &= \sum_{j=1}^{J} \sum_{k_j}^{K_j} P(j,k_j) V_j(X_{k_j}^0) \\ &+ \sum_{j=1}^{J} \sum_{k_j}^{K_j} \sum_{i \geq j}^{J} \sum_{k_i=1}^{K_i} P(j,k_j) P(i,k_i) V_{ij}(X_{k_j}^0, X_{k_i}^0) \end{aligned} \qquad (58)$$

Equation (58) represents the effective energy function used in most MFT bio-applications. In essence, MFT replaces the problem of finding the global minimum energy given by Equation (52) by the problem of finding the minimum of the "effective" potential energy described by Equation (58). Its major advantage is a significant reduction of the variable space. For example, there are K_i x K_j alternative configurations of subsystems i and j that can be examined using a single configuration of the effective system.

In the LES protocol, the normalization factors P are kept constant (usually taken as $P(i,k_i) = 1/K_j$, where K_j is the number of copies of subsystem j). A smoothening of the potential energy surface is achieved (85) that will enable the system to sample more of the conformational space. However, the degree of sampling depends on the number of copies K_j that are employed for the system j. If two many copies are employed, random transitions might occur, while if too few copies are employed, transitions might not occur at all leading to similar trajectories as standard MD simulations. The behavior of the system is also influenced by the way the subsystems were defined and thus, it is recommended to test whether different choices of subsystems (LES regions) provide similar trajectories if the number of copies for each LES region is maintained constant.

2.7. Electrostatic potential surfaces

The electrostatic properties are crucial for macromolecular interactions, thus playing a role in all biological processes that involve molecular recognition. Therefore, calculation of the electrostatic potential on the molecular surfaces and/or around the molecules can give valuable insights into the structure-function relationship or can lead to the calculation of the free energy of solvation, a particularly important property of polar and charged molecules.

Among the methods available for calculating the electrostatic potential, I will discuss briefly the methods base upon the Poisson or the Poisson-Boltzmann equations. These methods have been particularly useful for investigating the electrostatic properties of DNA, RNA or proteins. The solute is treated as a body of constant low dielectric (usually between 2 and 4), and the solvent is modeled as a continuum of high dielectric. The Poisson equation relates the variation in the potential Φ within a medium of uniform dielectric constant ε to the charge density ρ.

$$\nabla^2\Phi(r) = \frac{\rho(r)}{\varepsilon_0\varepsilon} \qquad (59)$$

The Poisson equation is thus a second-order differential equation ($\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$). For a set of point charges in a constant dielectric the Poisson equation reduces to Coulomb's law. However, if the dielectric is not constant but varies with position, then Coulom's law is not applicable and the Poisson equation is:

$$\nabla \cdot \varepsilon(r)\nabla\Phi(r) = -4\pi\rho(r) \qquad (61)$$

If mobile ions are present in the medium, their distribution in response to the electric potential can be accounted by the Poisson Equation. The ions are prevented from congregating at the locations of extreme potential due to the repulsive interactions with each other and natural thermal motion in the solvent. The ion distribution in the solvent is described by the Boltzmann distribution of the following form:

$$n(r) = N\exp(-V(r)/k_BT) \qquad (62)$$

where $n(r)$ is the number density of ions at a particular location r, N is the bulk number density and $V(r)$ is the energy change to bring the ion from infinity to the position r, k_B is the Boltzmann constant and T is the temperature.

When these effects are incorporated into the Poisson equation the result is the Poisson-Boltzmann equation:

$$\nabla \cdot \varepsilon(r)\nabla\Phi(r) - k'\sinh[\Phi(r)] = -4\pi\rho(r) \qquad (63)$$

k' is related to the Debye-Hückel inverse length, *k*, by:

$$k' = \frac{k^2}{\varepsilon} = \frac{8\pi N_A e^2 I}{1000\varepsilon k_B T} \qquad (64)$$

where I is the ionic strength of the solution, N_A is the Avogadro's Number, and e is the electronic charge. Equation (63) is a non-linear differential equation and can be written in an alternative form by expanding the hyperbolic 'sin' function as a Taylor series:

$$\nabla \cdot \varepsilon(r)\nabla\Phi(r) - k'\Phi(r)\left[1 + \frac{\Phi(r)^2}{6} + \frac{\Phi(r)^4}{120} + \ldots\right] = -4\pi\rho(r) \qquad (65)$$

The linearized Poisson-Boltzmann Equation can written by taking only the first term in the Taylor expansion, as:

$$\nabla \cdot \varepsilon(r)\nabla\Phi(r) - k'\Phi(r) = -4\pi\rho(r) \qquad (66)$$

Equation (66) cannot be solved analytically for complex geometries. So the Poisson-Boltzmann equation is solved by the numerical methods ('finite difference method*')*, in which the macromolecule (solute) is put in a cubic grid along with the solvent. Values of the electrostatic potential, charge density, dielectric constant and ionic strength are assigned to each grid point. The atomic charges usually don't coincide with the grid points.

There is a problem at the borders of the grid, since the grid points at the border have less than six neighboring points. So if the grid is much larger than the molecule, the border grid points are far away from the molecule, such that Φ outside of the grid can be set to zero. If the grid is huge, the computational cost becomes very high. A solution to this problem can be the use of low resolution, huge grid and interpolation of the electrostatic potential from this calculation to a high resolution, small grid. Such focusing steps can be repeated, if necessary.

3. Materials and methods

3.1. RNA constructs

Five different RNA constructs shown in Figure 11 were used in the MD simulations. K1 RNA (Figure 11A) is the RNA taken from the crystallographic structure of its complex with 15.5K protein (pdbid 1E7K) (44). K2 RNA (Figure 11B) contains the naturally occurring UUUAU external loop attached to the NC-stem. Since a similar pentaloop structure was unavailable in the data bases, I applied a restrained simulated annealing (RSA) protocol to obtain an initial structure of the loop. A single stranded RNA (5'-GGUUUAUCC-3') was constructed using the NAB software (133). During the RSA trajectory the two guanines from the 5' end were forced to form base pairs with the two cytosines from the 3' end, thus adopting exactly the same conformation as in the crystal structure. The loop was allowed to move freely and explore the conformational space. The RSA protocol consists of 20 ps of vacuum MD simulation with distance-dependent dielectric constant. In the first ps the temperature was rapidly increased to 1000 K. For the next 4 ps the temperature was maintained at 1000 K, allowing the loop to sample the conformational space. The annealing step was achieved with a slow cooling phase followed by a fast cooling in the last 2 ps. The two G-C base pairs from the new stem-loop were overlaid on the two G-C base pairs from the NC-stem in the complex. Simulated annealing was then performed using the new complex structure with the loop attached. All the residues from the crystal structure were fixed, such as to allow only loop movement. The resulting structure was then used as input for MD simulations. K3 RNA (Figure 11C) has a UGAA tetraloop (pdbid 1AFX) in which U37 stacks on G38, C41 stacks on A40 and A40 stacks on A39 (134) while K4 RNA (Figure 11D) has a UUAAUU loop (pdbid 1HS3) in which the nucleotide U establishes a non Watson-Crick base pair (N3-H-O2 and N3-H-O4 hydrogen bonds, sugars in cys orientation) with U40 (135). K5 RNA (Figure 11E) has the C-stem extended with seven Watson-Crick base pairs and the NC-stem with six Watson-Crick base pairs.

3.2. Molecular dynamics simulations

Ten nanosecond MD trajectories were obtained for the complex, the unbound protein and the unbound RNA using all the RNA constructs described. The complexes and the unbound RNAs were neutralized with Na^+ ions. All the systems were dissolved in a box of TIP3P water with 12 Å distance between the edges of the box and the solute and were equilibrated as described in 2.5.5 and 2.5.6. Periodic boundary conditions were applied. The Particle Mesh Ewald (PME) method was used for the treatment of the electrostatic interactions (129). The SHAKE algorithm was applied to all the bonds involving hydrogen atoms thus allowing the use of a 2 fs integration time step (127). All the simulations were run at constant temperature of 300 K and at constant pressure of 1 atm using the Berendsen's coupling algorithm (132). The calculations were performed using the AMBER. 99 force field (113) and the NAMD program (136) on a p690 IBM cluster running on 16 CPUs. For the biggest system (the complex of 15.5K protein with K5 RNA) the calculation of one trajectory required ~500 hours of CPU time. For the smallest system (the unbound K1 RNA) the calculation required ~250 hours. For testing the reproducibility of the results I ran multiple trajectories using slightly different initial conditions and different software packages. The visualisation and analysis of the trajectories were performed with VMD (137) and different programs available in the distribution of the AMBER 7 software package (138,139)

3.3. Locally enhanced sampling

For the LES4 trajectory, I divided K1 RNA into four LES regions: (i) the two G-C base pairs from the NC-stem, (ii) the unpaired nucleotides, (iii) the two tandem-sheared G-A base pairs, and (iv) two of the three G-C base pairs from the C-stem. Each region was replaced by 4 identical copies. I applied LES coupled with PME method using explicit solvent and periodic boundary conditions. Ten nanosecond trajectories were obtained for the LES systems of the complex and of the unbound RNA using K1 and K2 RNAs. Shorter test trajectories were run for the unbound K3, K4 and K5 RNAs. The equilibrated non-LES structures have been used as input for the LES simulations.

For the LES1 trajectory, one region confined to the entire internal loop (A29, A30, U31, G32, A33, G43, A44) was replaced with 4 identical copies. For the LES2 trajectory, 2 different regions were defined in the RNA and each region was replaced with 4 identical copies: (i) A29, A30, U31, G32, A33, G43, A44 (the internal loop); and (ii) G34, G35, C41, C42 (the two G-C base pairs of the NC-stem). The LES1 and LES2 trajectories were obtained only for the naturally occurring RNA sequence both free and in the complex with the 15.5K protein.

The trajectories were decomposed into individual trajectories of each copy, further processed, visualized and analyzed using the VMD software and the 'ptraj' and ‘carnal’ modules from the AMBER7 distribution. In the final step, snapshots were taken every 10 ps for data visualization. For clarity reasons, I labeled the MD trajectories as ‘MD-LES’ when LES was not employed and ‘MD+LES’ when LES was applied.

3.4. Essential dynamics

Principal component analysis (PCA) of MD trajectory data, often called essential dynamics (ED), is frequently used to separate large-scale correlated motions from local harmonic fluctuations (140-143). ED analysis constructs a new orthogonal basis set for the atomic coordinates in a trajectory, such that the greatest variance occurs along the first vector, with decreasing variances along successive vectors. The eigenvalues from the decomposition of the eigenvectors represent the relative motion that occurs along each mode. The motion of a system during MD trajectories can be described by the displacement along the first few eigenvectors (144). There are a number of limitations associated with ED (145), but the method proved useful in identifying motions correlated with the large amplitude k-e motion of the RNA captured in the MD+LES trajectories.

To perform ED, coordinate data from each time-step was fitted to the initial structure to remove translational and rotational motion. The fitted trajectory data were used to construct a covariance matrix. The diagonalization of the matrix was performed to find the diagonal matrix of eigenvalues and the matrix of column eigenvectors. The trajectory can be reconstructed using the eigenvectors with significant contribution to the global motion. The matrix of the projections of each time-step onto each mode was calculated by multiplying the trajectory matrix by the matrix of column eigenvectors. I performed ED of the LES4 trajectory using the *ptraj* module from the AMBER8 distribution and visualized the modes with VMD using the IED program (144). I generated the covariance matrix for all the atomic coordinates in the RNA, diagonalized it, and wrote out the first 25 eigenvectors. The projections onto the first 3 modes were calculated and the motions along first and third modes were analyzed further.

3.5. Generating electrostatic potential surfaces

The electrostatic potential was calculated by solving the non-linear Poisson-Boltzmann equation using the APBS (Adaptive Poisson-Boltzmann Sover) software (146). ‘Pqr’ files (‘pdb’ files that include atomic charges and radii) were created from standard pdb file using the ‘ambpdb’ program from the AMBER distribution. A grid of 129^3 was used with 2 levels of focusing with 0.8125, 0.8125, 0.8696 reductions. The calculations were performed at 298 K in 150 mM NaCl.

The electrostatic potentials were visualized with the VMD software, version 1.8.3 by generating a molecular surface and coloring it according to the electrostatic potential (a feature which was newly implemented in VMD 1.8.3). I chose the RWB (red-white-blue) color scale with a range of -3.5 to 3.5.

3.6. Steady-state fluorescence spectroscopy

The influence of replacing different bases in KtU4 with fluorescent analogues was tested by “in silico” energy minimization of the labeled RNAs and investigation of the optimized structures. 2AP and different pteridine analogues for adenine, as well as several pteridine analogues for guanine were tested (data not shown). In order to apply energy minimizations on the resulting structures, AMBER force field parameters were derived applying the described procedure (see 2.3 – data not shown). This study predicted that the replacement of A29 with 2AP should not influence significantly the structure of the RNA. In the crystal structure of the 15.5K-KtU4 complex, A29 is involved in hydrophobic interactions with the protein residue ARG97 and in stacking interaction with the G45-C28 base pair of the C-stem. Thus, changes in the 2AP fluorescence in this position could presumably reflect differences in the number of stacking interactions involving A29.

Two oligonucleotides comprising the minimal binding site of 15.5K protein were ordered from Dharmacon: (i) an unlabeled oligo for control (seq: 5’-G.C.C.A.A.U.G.A.G.G.U.U.U.A.U.C.C.G.A.G.G.C-3’), and (ii) a labeled oligo with 2AP replacing A29 (seq: 5’-G.C.C.2AP.A.U.G.A.G.G.U.U.U.A.U.C.C.G.A.G.G.C-3’).

Steady-state fluorescence measurements were performed using a Photon Technologies International C-60SE spectrofluorometer, equipped with a thermostatted cuvette holder. Emission spectra were recorded at an excitation wavelength of 305 nm. 500 ml samples of 200 nM labeled RNA were measured in 5x5 mm cuvettes. All measurements, except for the temperature profiles were performed at 5℃. Protein titrations were performed by adding increasing amounts of 15.5K protein to a sample containing 200 nM labeled RNA. Two buffer systems were used: (i) buffer 1: 20 mM hepes pH=7.9, 150 mM KCl, and (ii) buffer 2: buffer 1 + 1.5 mM $MgCl_2$. In all graphs (Figures E1 to E4) 2AP spectra are shown after substraction of background (buffer) spectra.

4. Aims, motivation and relevance

I used an approach based on computer simulations to study the dynamics of KtU4. I simulated both the complex and the unbound RNA to investigate whether the folding of the RNA is assisted by binding to the 15.5K protein and whether such simulations can be correlated with data obtained by single molecule FRET and chemical RNA modifications experiments. Additionally, I aimed at: (i) characterizing the dynamics of the free RNA at atomic level, (ii) identifying specific flexible regions in the RNA that are responsible for the k-e motion, and (iii) predicting structural dynamics in the RNA and the 15.5K protein that might regulate the hierarchical assembly of U4/U6 snRNP.

I showed that if started from their bound structures, the simulation of RNAs undergoing protein-assisted folding requires the use of enhanced sampling techniques in combination with standard MD. This conclusion was strengthened by Rázga et al. in two complementary studies that appeared recently (147,148). Although they reported relatively long standard MD simulations of different K-turns (up to 74 ns), they observed only partial opening of K-turns while the G-A base pairs remained stable throughout the simulations. The behavior of the RNA in the LES-MD simulations was in excellent agreement with data obtained by chemical probing of RNA structure and by single molecule FRET investigation of the RNA K-turn motif.

The study presented in this dissertation constitutes a novel application of LES to study dynamics of RNA motifs undergoing protein-assisted RNA folding and to deduce local conformational flexibility that is required for large conformational transitions such as the k-e motion of KtU4. Using this approach, I provided the first atomic details of a relatively large k-e motion of the K-turn motif. However, the structures observed during the LES-MD simulations should not be regarded as predicted structures of the free RNA.

Although standard MD simulations in combination with LES proved to be very powerful in studying dynamical transitions in the U4 snRNA K-turn motif, it was crucial that the results obtained from such simulations were constantly compared with emerging experimental data on the system under investigation. Thus, in addition to experimental data available from other studies, I employed steady-state fluorescence spectroscopy to investigate the dynamics of KtU4 in the presence and absence of 15.5K protein. The RNA was labeled by introducing 2AP at position 29 and the results were compared with the simulations. However, the experimental data are only preliminary and do not provide sufficient information for an in-depth comparison with the simulations.

5. Results

5.1. Protein-assisted folding of KtU4

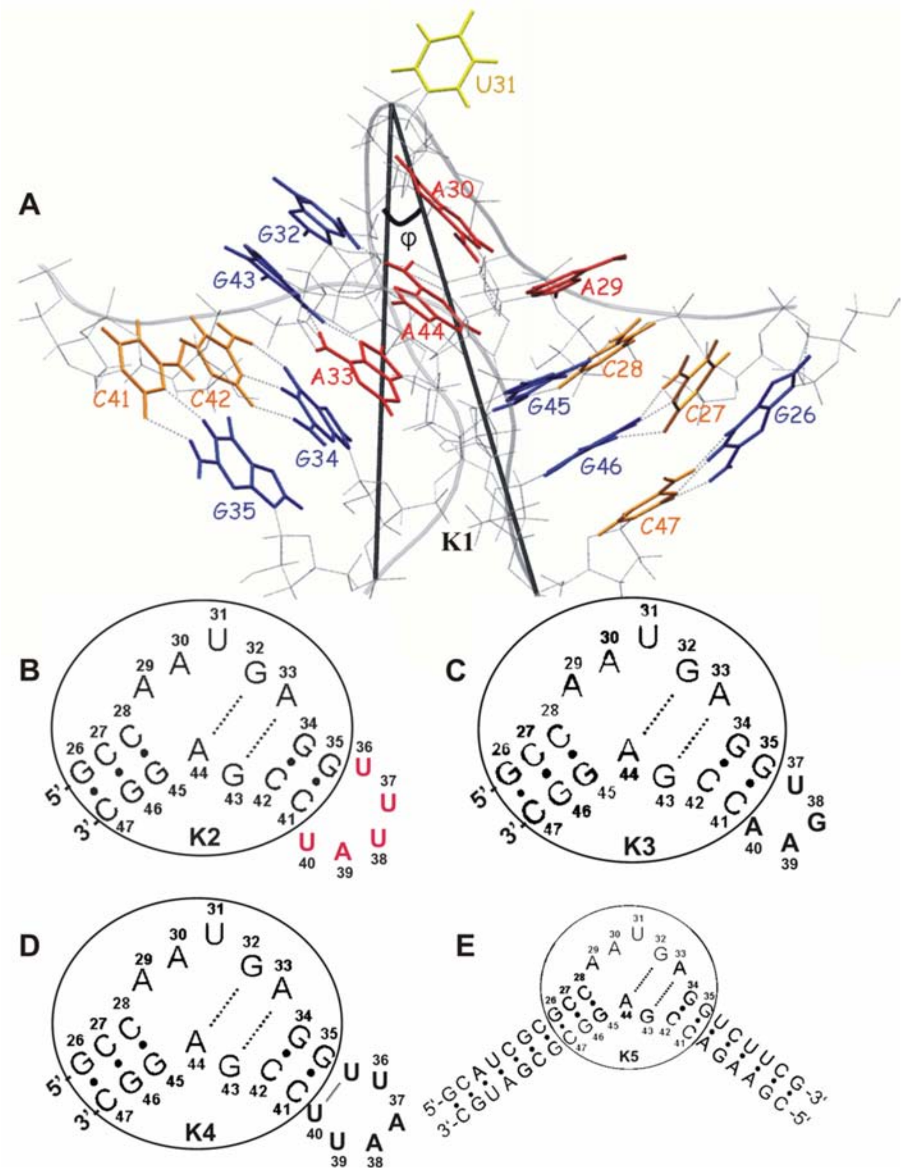


Figure 11: Simulated RNA constructs.

(A) The K-turn motif found in the crystal structure of 15.5K-KtU4 complex(K1 RNA); the C-stem has three Watson-Crick base pairs (G26-C47, G46-C27 and G45-C28), the internal loop contains three unpaired nucleotides (A29, A30, U31) and two tandem-sheared G-A base pairs (G32-A44, G43-A33), and the NC-stem consists of two Watson-Crick base pairs (G34-C42 and G35-C41). φ is the angle between the P atoms of C47, U31 and G35. (B) K2 RNA (naturally occurring): the pentaloop UUUAU (shown in red) is attached to the NC-stem of K1 RNA. (C) K3 RNA: the tetraloop UGAA was attached to the NC-stem of K1 RNA. (D) K4 RNA: the hexaloop UUAAUU was attached to the NC-stem of K1 RNA. (E) K5 RNA: the C-stem and the NC-stem of K1 RNA were extended with 7 and 6 Watson-Crick base pairs. The core of the K-turn RNA observed in the crystal structure is encircled in (B)-(E).

The simulations, in addition to the chemical RNA structure probing and the single molecule FRET studies described in the introduction provide conclusive evidence that the RNA K-turn motif, and in particular KtU4 is folded in a protein-assisted manner. In this subchapter, both standard MD simulations and MD+LES simulations (the LES4 trajectory) are described, the trajectories of the free RNA being compared with the equivalent trajectories of the bound RNA.

The increased flexibility in the free RNA and the observation of conformational transitions occurring only in the unbound RNA are proofs for the protein-assisted folding of KtU4. There is an excellent agreement between the simulations presented in this subchapter and the experimental data mentioned above (60). The behaviour of the system was independent on the choice of LES regions as long as they covered both the internal loop and the two stems.

5.1.1. General analysis of the simulations

Figure 11A shows the crystal structure of KtU4 RNA bound to 15.5K protein. I use the angle φ formed by the P atoms of C47, U31 and G35 to characterize the degree of kinking in the RNA (Figure 11A). The φ angle is sharper than the angle formed by the two helical axes (Φ) (~25º compared to ~68º) but it reflects the same property of the RNA (the sharp kinking in the sugar-phosphate backbone) and displays the same profile during the simulations. φ is independent of structural instabilities as opposed to Φ which cannot be defined if the NC-stem loses its helical properties. The natural occurring RNA (K2) is shown in Figure 11B. I used the crystallographic structure of the RNA (K1), lacking the external pentaloop, as a template to build and simulate five different RNA constructs (see 2.8.1): (i) K1 (Figure 11A), (ii) K2 (Figure 11B), (iii) K3 (Figure 11C), (iv) K4 (Figure 11D), and (v) K5 (Figure 11E). The constructs produced similar trajectories both in the bound and unbound forms, with the differences mainly lying in the stability of the Watson-Crick base pairs of the NC-stem and in the behavior of the external loop. All the trajectories described were 10 ns long. In the text I refer to the naturally occurring K2 RNA sequence unless indicated otherwise.

<u>MD-LES simulations</u>

The B-factors per residue were calculated for the bound and unbound RNA during the MD-LES trajectories (Figure 12A). The plot reveals that the unbound RNA fluctuates more than the bound RNA, indicating that the protein exerts a stabilizing effect on the

RNA. Significantly higher fluctuations in the unbound RNA are observed for the unpaired nucleotides A30 and U31 and for the G35-C41 and G34-C42 Watson-Crick base-pairs of the NC-stem. Root mean square deviations (rmsd) of the backbone for the core RNA structure (residues G26 to G35 and C41 to C47) were calculated using the initial structure as the reference (Figure 12B). I observed that the unbound RNA deviates more from the crystal structure that the bound RNA, with peak values for rmsd in the time interval between 1.5 ns and 3.5 ns, reflecting a slight opening of the K-turn. After 3.5 ns the K-turn closes back and stays closed until the end of the trajectory. The φ angle has a peak value of ~40° after ~3.2 ns of the trajectory of the unbound RNA compared to a relatively constant 20º in the bound RNA (Figure 12C).

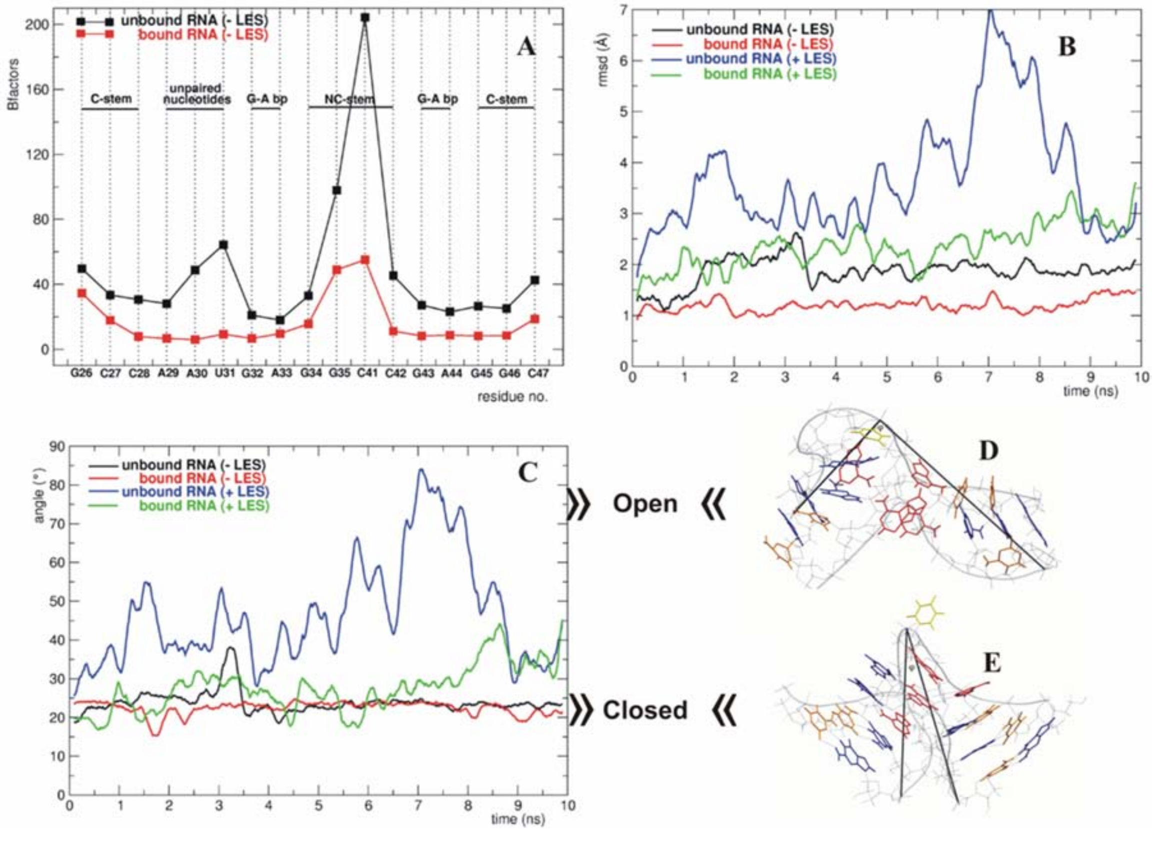


Figure 12: General analysis of MD simulations.

The case of K2 RNA. (A) B-factors per residue for the RNA structure during the MD-LES trajectories of the bound (red curve) and unbound RNA (black curve). (B) Root mean square deviations (rmsd) of the RNA backbone from the initial structure for the bound and unbound RNA during the MD-LES (red and black curves) and MD+LES trajectories (green and blue curves). (C) φ angle of the bound and unbound RNA during the MD-LES (red and black curves) and MD+LES trajectories (green and blue curves). (D) Closed conformation of the K-turn (φ = 20-40º). (E) Open conformation of the K-turn (φ = 50-80º). In (C*)* and (D*)* the loop attached to the NC-stem is not shown. In (C) and (D) φ is shown in black (Figure 11A)

MD+LES simulations (LES4 trajectory)

When LES was applied, a dramatic conformational transition occurred in the unbound RNA (referred to as the ‘k-e motion’) with rmsd values reaching peak values of about 7 Å (~4 Å greater than the peak deviation of the bound RNA) after ~7 ns (Figure 12B). Interestingly, the K-turn partially reforms towards the end of the trajectory. The open conformation is characterized by φ values of 50 - 90° (Figure 12C). A snapshot of the open conformation is shown in Figure 12D. Although the bound RNA opens slightly after ~8.5 ns of the LES trajectory, the K-turn conformation is not significantly altered (φ is relatively constant). Figure 12E shows a snapshot with the closed conformation of the K-turn. The k-e motion of the free RNA is accompanied by local conformational transitions summarized in table 1.

Table 1: Conformational transitions in the RNA (K2)

Conformational transition	Bound RNA (MD-LES)	Unbound RNA (MD-LES)	Bound RNA (MD+LES)	Unbound RNA (MD+LES)
Opening of the K-turn	no	partial	no	yes
Unpaired A30 syn > high anti/anti C2'endo > C3'endo	no	yes	no	yes
Loss of G-A base pairs	no	no	no	yes
Loss of inter-strand contacts	no	yes	no	yes

5.1.2. Flexibility in the purine-rich internal loop

The internal loop of KtU4 consists of two tandem-sheared G-A base pairs (G32-A44, G43-A33) and three unpaired nucleotides (A29, A30, U31). The four adenines establish a “3+1” stacking scheme in which A30 stacks on A44 which further stacks on A33. A29 stacks on the G45-C28 base pair of the C-stem and makes hydrophobic contacts and hydrogen bonds with ARG97 of 15.5K protein. The guanines G32 and G43 establish contacts with 15.5K protein and U31 is flipped out and trapped in a pocket of the protein.

Only A29 and A30 can be mutated to other purines without abolishing protein binding (29).

MD-LES simulations

During the MD-LES trajectories of the unbound RNA, the ‘3+1’ stacking scheme (Figure 13A) changed its conformational to a ‘2+2’ scheme (Figure 13B).

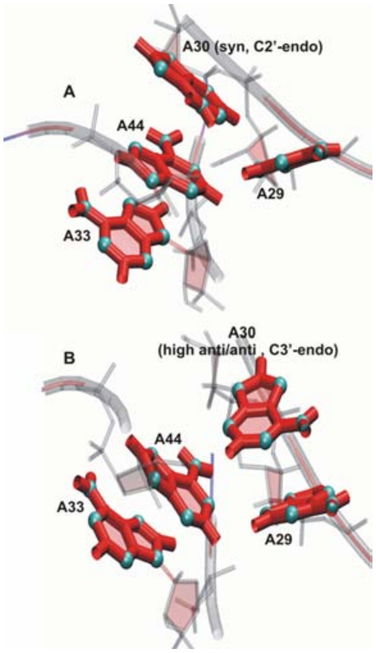


Figure 13: Conformational states of A30.
The case of K2 RNA. (A) ‘3+1’ stacking scheme formed by adenines A29, A30, A44 and A33 in the bound RNA, A30 being in a syn/C2'endo conformation. (B) ‘2+2’ stacking scheme formed by adenines A29, A30, A44 and A33 during the trajectories of the unbound RNA, A30 being in a high-anti, anti/C3'endo conformation.

The conformation of A30 evolved from a C2'-endo sugar pucker and a syn base-sugar orientation to a C3'endo sugar and a high-anti (or anti) base-sugar orientation. The percentages in which A30 is found in either conformation during different trajectories are presented in table 2. The transition occurs after ~8 ns of the MD-LES trajectory of the unbound RNA. In the complex, the A30 is held in its syn/C2'-endo conformation by LYS37, which makes a stable contact network between the N7 atom of A30 and the phosphate backbone of the RNA. The conformational transition of the stacking scheme does not induce the opening of the K-turn; during the last 2ns of the MD-LES trajectory of the unbound RNA the K-turn remains closed (φ=~20°) while the new stacking scheme forms.

Table 2: Conformational states of adenine A30

Trajectory	Sugar Pucker (A30) (%)		χ angle (A30) (%)	
	C2'-endo	C3'-endo	syn	high anti/anti
Bound K2 RNA (-LES)	80	0	93	0
unbound K2 RNA (-LES)	63	9	72	22
bound K2 RNA (+LES)	36	0	21	1
unbound K2 RNA (+LES)	16	24	10	84

The tandem sheared G-A base pairs were stable during the MD-LES trajectories of the bound and unbound RNA. Nevertheless, several differences were observed when comparing the trajectories of the unbound and bound RNAs. In particular, the displacement of the G32 and G43 was minimized in the unbound RNA, resulting in a better stacking of the two guanines, and triggering a slight increase in the propeller twist of the two G-A base pairs. In the complex, G32 establishes contacts with the protein residue GLU40 while G43 makes hydrogen bonds with ASN40 and LYS44. These contacts result in a displacement between the two guanines that was maintained during the trajectories of the complex.

Several inter-stem contacts contributing to the stability of the kinked conformation are formed in the region of the internal loop: (i) the 2' OH group of A30 is hydrogen-bonded with one oxygen atom from the phosphate group of U31, (ii) the 2' OH group of U31 is hydrogen-bonded with one oxygen atom from the phosphate group of A30, and (iii) the 2' OH group of A29 is hydrogen-bonded with the N1 atom of A44 (Figure 14A). These contacts are well preserved during the MD-LES trajectory of the bound RNA but are lost in the unbound RNA. The hydrogen-bonding distance between the 2' OH group of A29 and the N1 atom of A44 is plotted in Figure 14B.

MD+LES simulations

When applying LES, I observed the same conformational transition (‘3+1’→‘2+2’) of the stacking scheme in the unbound RNA on a shorter time scale. Due to the increased conformational sampling the number of intermediate states between the C2'-endo and C3'-endo conformations increased both for the bound and the unbound RNA.

Nevertheless there was a significant increase in the fractional population of the C3'-endo sugar pucker in the unbound RNA (table 2).

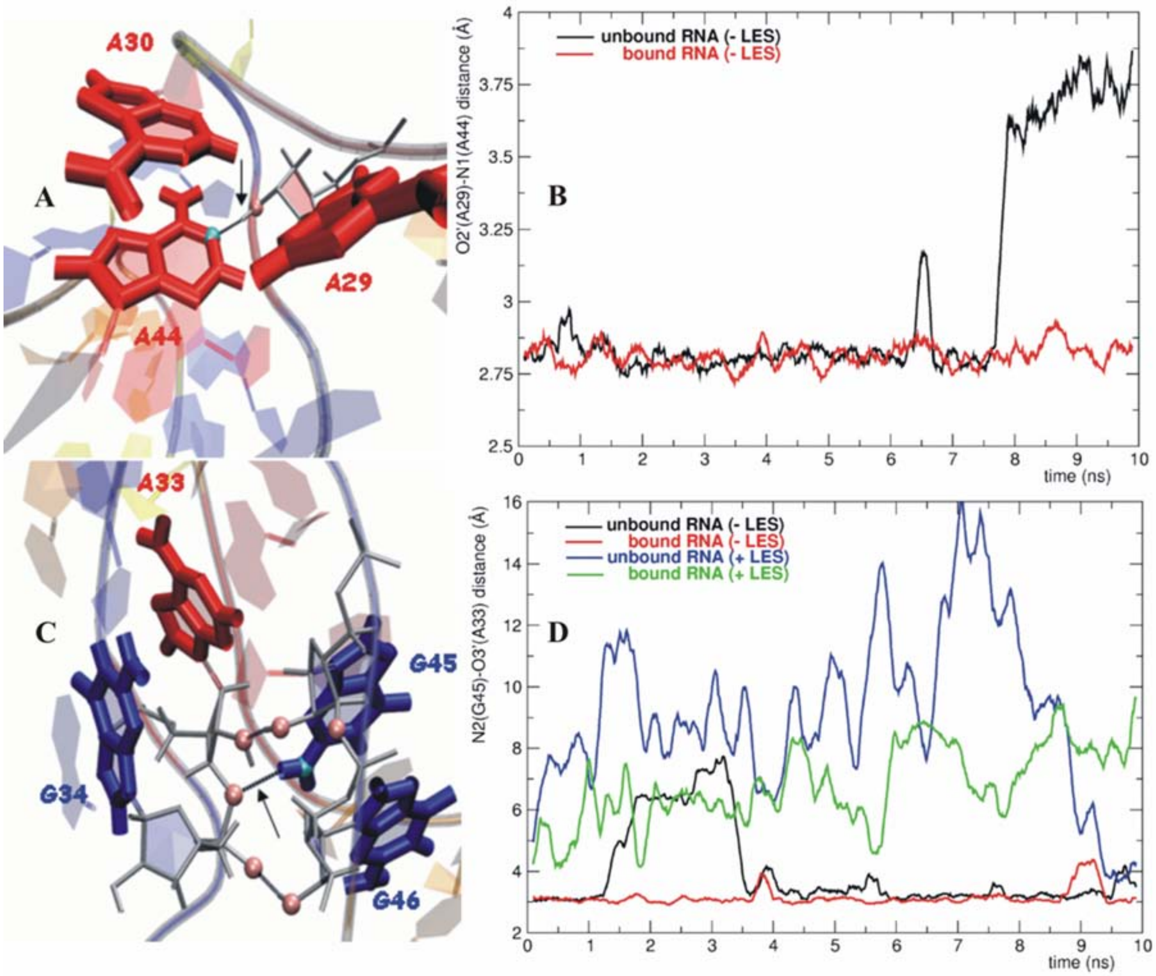


Figure 14: Inter-stem contacts established in the RNA
The case of K2 RNA. (A) The inter-stem contact formed between the O2' atom of A29 and the N1 atom of A44. (B) Hydrogen-bonding distance between the O2' atom of A29 and the N1 atom of A44 in the bound (red curve) and unbound RNA (black curve) during the MD-LES trajectories. (C) The inter-stem contact formed between the N2 atom of G45 and the O3' atom of A33. (D) Hydrogen-bonding distance between the N2 atom of G45 and O3' atom of A33 in the bound and unbound RNA during the MD-LES (red and black curves) and MD+LES (green and blue curves) trajectories.

Both G-A base pairs opened during the MD+LES trajectories of the unbound RNA. Figure 15 shows the hydrogen-bonding distance between the N2 atom of G32 and the N7 atom of A44. The transition to the open conformation of the K-turn triggered the loss of the G-A base pairs.

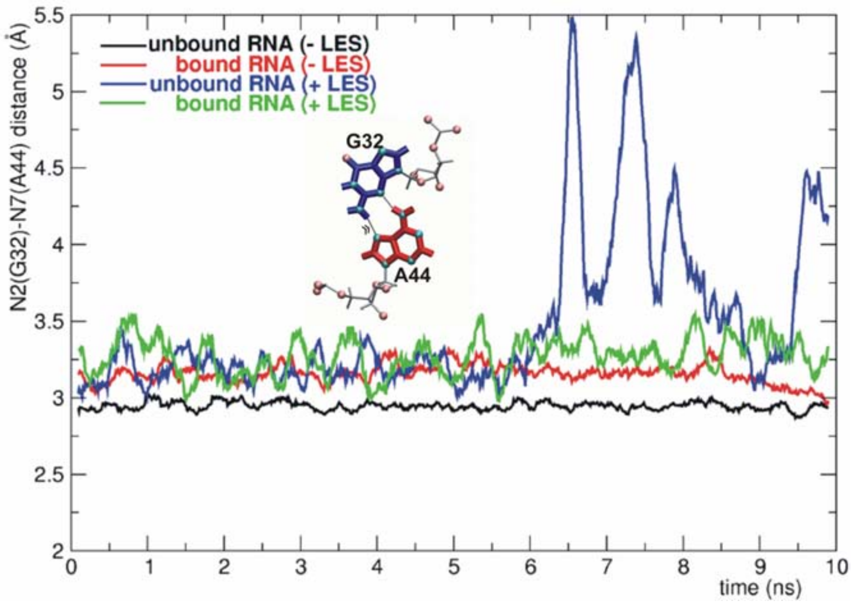


Figure 15: The G32-A44 base pair.
The case of K2 RNA; hydrogen-bonding distance between the N2 atom of G32 and the N7 atom of A44 in the bound and unbound RNA during the MD-LES (red and black curves) and MD+LES (green and blue curves) trajectories. The hydrogen bond is pointed with a double arrow head.

5.1.3. Flexibility in the NC-stem

To assess the stability of the G34-C42 and G35-C41 base pairs I calculated the percentage of trajectory frames in which the constituent hydrogen bonds were formed. The criteria chosen for hydrogen bond formation were: (i) maximal donor-acceptor distance of 3.15 Å and, (ii) maximal donor-hydrogen-acceptor angle of 60°. In table 3 I show the results for the standard MD trajectories of K1 RNA and the LES trajectories of K2 RNA in the bound and unbound forms.

MD-LES simulations

In the bound K1 RNA both base pairs were stable, whereas in the unbound K1 RNA the G35-C41 base pair opened and the G34-C42 base pair was less stable. However, during the MD-LES trajectories of K2, K3, K4 and K5 RNAs the NC-stem was stable (data not shown). The numbers in the first two rows of table 3 show that the opening of G35-C41 base pair occurred only in the unbound K1 RNA. Various inter-stem contacts are established between the NC-stem and the C-stem: (i) the 2' OH group of A33 is hydrogen-bonded with the 2' OH group of G45, (ii) the N2 atom of G45 is hydrogen-bonded with the O3' atom of the sugar of A33, and (iii) the 2' OH group of G46 is hydrogen-bonded with one of the oxygen atoms of the phosphate group of G34. All these contacts are well preserved during all the trajectories of the bound RNA. Figures 14C and 14D show that the hydrogen bond distance between the N2 atom of G45 and the O3' atom of A33 is strictly correlated with the partial opening of the K-turn during the MD-LES trajectories (the loss of the hydrogen bond is observed in the same interval as the increase in the value of φ).

Table 3: Stability of the Watson-Crick base pairs of the NC-stem

Trajectory	G34-C42 base pair (%**)			G35-C41 base pair (%**)		
	HB1*	HB2*	HB3*	HB1*	HB2*	HB3*
bound K1 RNA (-LES)	99	98	96	96	95	88
unbound K1 RNA (-LES)	96	83	63	32	24	20
bound K2 RNA (+LES)	70	62	53	52	44	44
unbound K2 RNA (+LES)	62	52	36	17	15	14

*The three hydrogen bonds of the G-C base pairs are: HB1=N2(G)-H(G)-O2(C); HB2=N1(G)-H(G)-N3(C); HB3=N4(C)-H(C)-O6(G).

** The percentage of trajectory frames in which the hydrogen bond is formed

MD+LES simulations

When applying LES, both base pairs remained relatively stable in the bound RNAs. The partial loss in stability compared to the non-LES trajectories reflects the increased sampling density. In both K1 and K2 unbound RNAs, the G35-C41 base pair opened and the G34-C42 base pair was less stable than in the bound RNAs, the opening of G35-C41 base pair occurring much faster in K1 RNA. The numbers in the last two rows of table 3 show that the stability of the NC-stem is dependent on protein binding, not only in K1 RNA but also in K2 RNA. Similar results were found for K3 and K4 RNAs, while in K5 RNA the NC-stem was stable regardless whether the RNA was bound or not to the protein. During the MD+LES trajectories the inter-stem hydrogen bond between the N2 atom of G45 and O3' atom of A33 was lost for both the bound and unbound RNAs. Nevertheless, the distance between the two atoms is significantly larger in the unbound RNA (Figure 14D).

5.2. Dynamics of the free KtU4

The behavior of the free RNA was not influenced by the number of LES regions as long as the conformational sampling was enhanced both in the internal loop and in the two stems. Only those LES setups under which the bound RNA was stable were considered for further processing. I chose to present in detail the LES4 trajectory because it was the longest (10 ns) and captured a significant reformation of the K-turn during the last 2 ns.

To study how the backbone flexibility in the stems influences the k-e motion of the K-turn, I restricted the degree of conformational freedom in the C and NC stems by confining LES to the internal loop (LES1) or to the internal loop and the NC stem (LES2). Simulations of the LES1 and LES2 systems were performed and the resulting trajectories were compared to the LES4 trajectory.

I present in this subchapter an in-depth analysis of simulations of the free K2 RNA (naturally occurring) performed with different LES setups, finding that the largest k-e motion was observed when LES regions were defined both in the internal loop and in the stems (LES4 trajectory). When LES was confined to the internal loop (LES1 trajectory), the RNA adopted an alternative conformation that is characterized by a sharper kink in the backbone, loss of G-A base pairs and modified stacking interactions. Essential dynamics of the free RNA in the LES4 trajectory show that the opening of G-A base pairs is correlated along the first mode but anti-correlated along the third mode with the k-e motion. The dynamics along the third mode were similar to that observed in the LES1 trajectory. Thus, loss of G-A base pairs is insufficient for achieving a large opening of the free RNA, the k-e motion of the K-turn being promoted by structural flexibility both in the internal loop and in the helical regions. Based on these findings, I propose that the G-A base pair formation occurs upon binding to the 15.5K protein, stabilizing a selective orientation of the stems rather than contributing to the k-e motion in the free RNA.

The results presented in this subchapter are part of the paper that I published in Nucleic Acids Research together with Reinhard Klement and Tom Jovin (149).

5.2.1. Opening-closing of KtU4

The conformation of KtU4 as observed in the crystal structure with the external loop modeled is shown in Figure 16A. During the LES4 trajectory the RNA underwent a large conformational transition to an extended structure. Figure 16B shows a snapshot from the LES4 trajectory taken at maximum φ. In the LES4 system, regions of enhanced sampling were defined both in the internal loop and in the stems (see 2.1.3). The large opening of the K-turn is accompanied by loss of G-A base-pairs, modification of the stacking pattern in the internal loop and opening of G-C base pairs in the NC-stem. In the LES1 trajectory, enhanced sampling was confined to the internal loop and the RNA adopted an alternative conformation (Figure 16C) characterized by loss of G-A base pairs and a distinct modification of the stacking pattern in the internal loop. However, no opening of G-C base pairs in the NC-stem was observed. The new stacking patterns formed in the LES4 and LES1 trajectories are indicated by the long arrows. The conformation observed in the LES1 trajectory had kinks in both strands (indicated by the short arrows), the fold resembling that of K-turn Kt58 (Figure 16D).

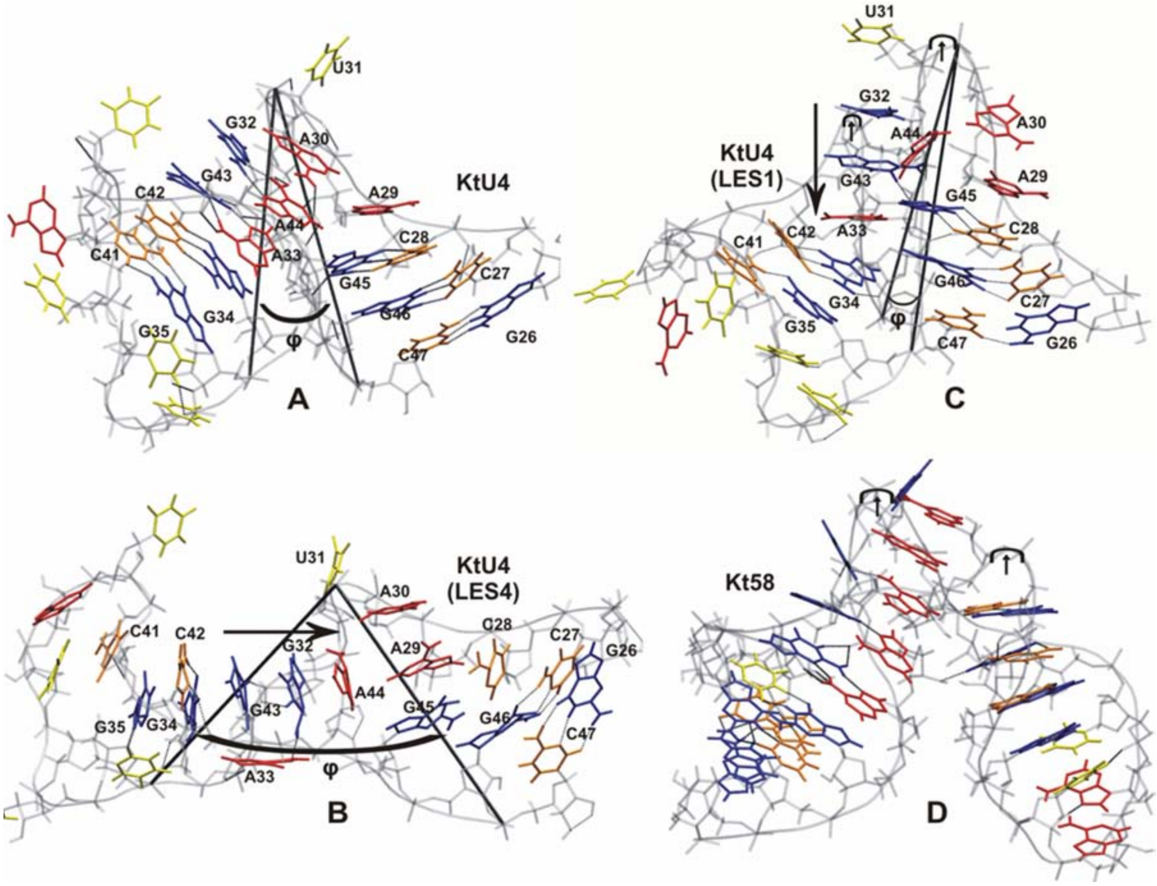


Figure 16: Different conformations of KtU4.
The long arrows indicate newly formed stacking patters. The short arrows indicate the kink in the backbone. φ is the angle between the P atoms of C47, U31 and G35. (A) KtU4 RNA structure from the crystallographic structure of KtU4-15.5K complex with the UUUAU external loop modeled; (B) Open conformation observed during the LES4 trajectory; (C) Tightly kinked alternative conformation captured in the LES1 trajectory; (D) Structure of the Kt58 for comparison with the structure in Figure 1C.

The LES2 trajectory was obtained by applying enhanced sampling to the internal loop and the G-C base pairs of the NC-stem but not to nucleotides from the C-stem.

Interestingly, opening of the K-turn occurred in the LES2 trajectory but the amplitude of the transition was far smaller than that observed in the LES4 trajectory. The φ angle for the three trajectories is shown in Figure 17A. In the LES4 trajectory the K-turn opening occurred rather quickly, reaching a peak after about 1.5 ns, after which the structure closed back such that after 3.7 ns the value of φ was close to the value observed in the crystal structure. After 4 ns the RNA opened gradually, reaching the highest φ (~90°) after approximately 7.1 ns. For about 1 ns, the structure remained open while during the last 2 ns, the K-turn partially reformed to a φ value of 30°. In the LES1 trajectory partial opening occurred only after 2.6 ns and φ reached a maximum value of 40° after ~3.2 ns. The partial open conformation was short-lived; after 3.5 ns the K-turn reformed while after 5 ns, φ decreased to lower values than those observed in the crystal structure and remained constant for the rest of the trajectory. In the LES2 trajectory, partial opening occurred quickly and the amplitude was comparable to that in the LES4 trajectory for the first 4.5 ns. However, after ~5 ns the RNA closed and remained closed until the last ns, when partial opening was observed again.

5.2.2. Opening of G-A base pairs

Tandem-sheared G-A base pairs are formed by establishing N2(G)-H-N7(A) and N6(A)-H-N3(G) hydrogen bonds. For studying the opening of G-A base pairs during the three different trajectories I calculated the donor-acceptor distance (d) between the N2 atom of guanines and N7 atom of adenines. Figure 17B represents the results obtained for the G32-A44 base pair, the G43-A33 base pair behaving in a similar manner. During the LES4 trajectory I observed large opening of the G32-A44 base pair only after 5.6 ns. In the interval 6-9 ns the base pair did not reform, but the relative movement of the two nucleotides showed a rather random distribution that did not correlate with the time-evolution of φ. The base pair reformed after ~9 ns but immediately opened again. In the LES1 trajectory, the same base pair opened partially after ~2ns and completely after ~4ns. Surprisingly, during the last 6 ns of the LES1 trajectory, d was larger than in the LES4 trajectory and the G32-A44 base pair never reformed. Also unexpectedly, the movement of A44 relative to G32 was much larger in the tightly kinked LES1 conformation than in the highly open LES4 conformation. The 2D plot of d against φ shows that d increased with φ for the most part of the LES4 trajectory. However, in a number of snapshots, the G-A base pair was open (large d) while the K-turn was relatively closed (small φ) (Figure 17C). In the LES1 trajectory there was an anti-correlation between the opening of G-A base pairs and the opening of the K-turn, very large values for d corresponding to a very tight K-turn (Figure 17D).

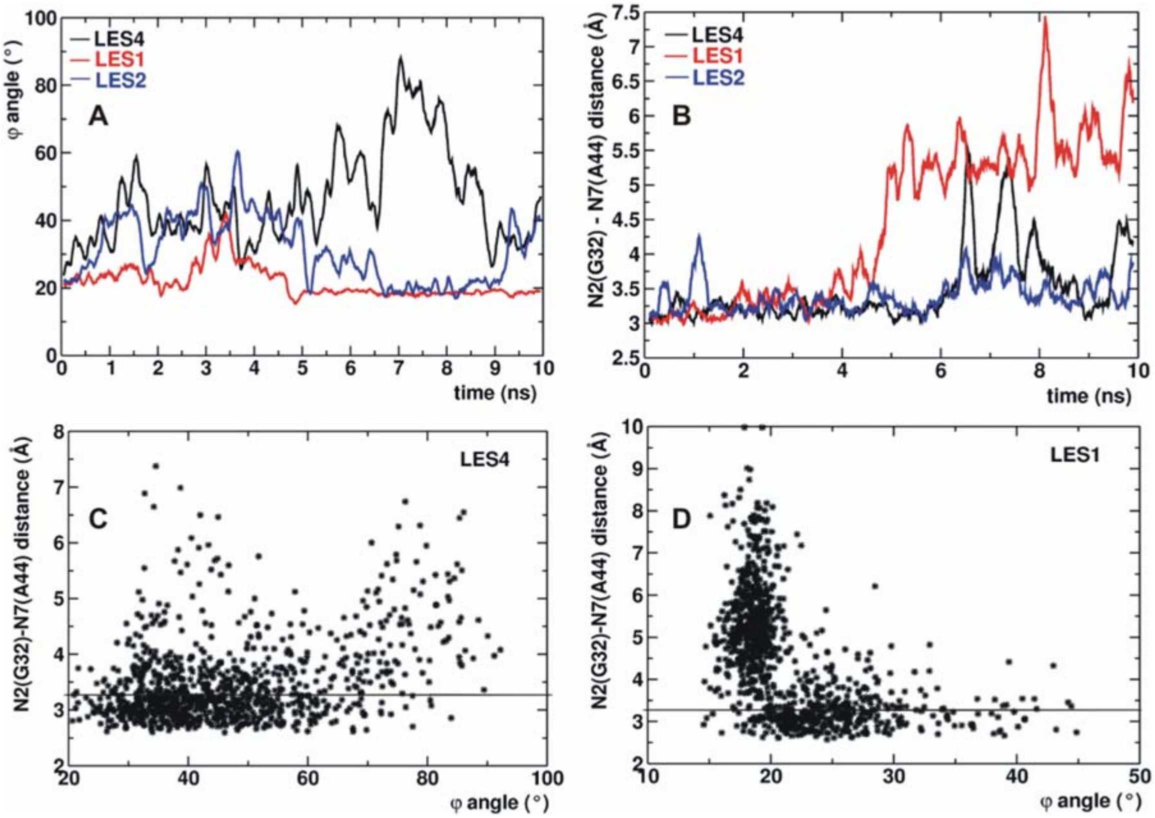


Figure 17: Correlation between opening of G-A base pairs and the k-e motion.
(A) The φ angle during the LES4 (black), LES1 (red) and LES2 (blue) trajectories; (B) The distance d between N2 of G32 and N7 of A44 during the LES4 (black), LES1 (red) and LES2 (blue) trajectories; (C) 2D diagram of d(φ) in the LES4 trajectory; (D) 2D diagram of d(φ) in the LES1 trajectory. In (C) and (D), the maximal donor-acceptor distance (3.2 Å) for hydrogen-bond interaction is indicated with a horizontal line.

In the crystallographic structure of the KtU4-15.5K complex, G32 stacks on G43 and A44 stacks on A33. The planes of the two guanines are displaced, owing to hydrogen bond contacts between the RNA and the protein. The O6 atom of G32 is hydrogen-bonded with the peptide backbone N atom of GLU41 and the O6 and N7 atoms of G43 are hydrogen-bonded with the side-chains of ASN40 and LYS44. There is no protein contact involving A44 and A33, therefore their planes are oriented such as to allow a maximized stacking interaction. To measure the stacking between two RNA bases, I calculated the distance between their geometrical centers (D) and the angle between their planes (θ). Maximum stacking is achieved when D is minimal and θ is either close to 0° or 180°. For

G32/G43 and A44/A33 stacking interactions, θ has a value close to 180° because the Watson-Crick edges of the stacked bases point in different directions. When a purine rotates about the N9-C1' bond, θ fluctuates between ~180° and ~0°.

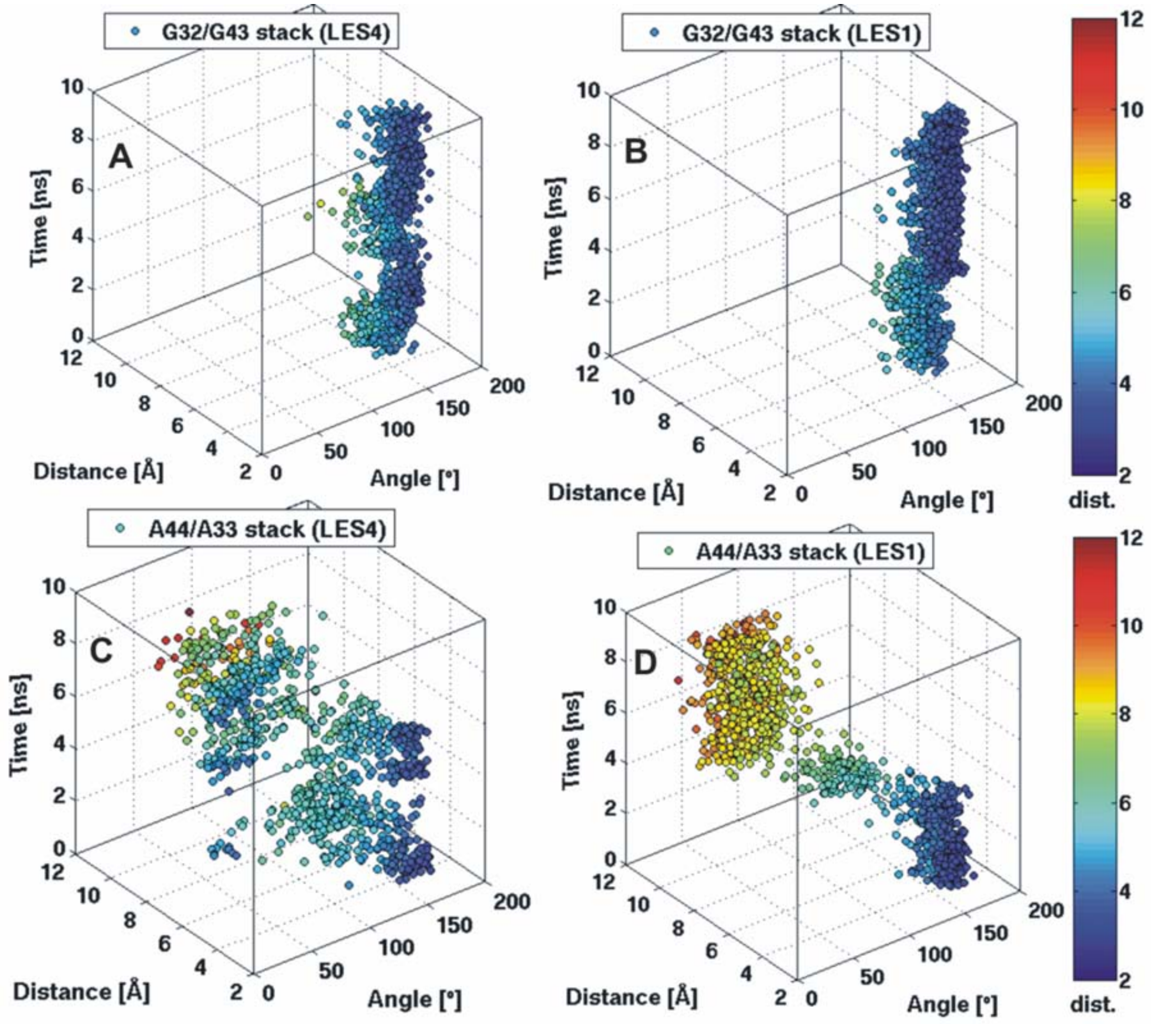


Figure 18: Scatter 3D plots of stacking interactions between: (A) G32/G43 in the LES4 trajectory. (B) G32/G43 in the LES1 trajectory. (C) A44/A33 in the LES4 trajectory. (D) A44/A33 in the LES1 trajectory. Stacking interaction occur when the distance (D) is minimal and the dihedral angle (θ) between the planes of the nucleotides is either ~0° or ~180°. Rotation of the base about the N9-C1' bond occurs when θ leaps towards intermediate values. The scatter points are colored according to their y value (D) with a color map ranging from blue to red.

I plotted in scatter 3D graphs, θ (x axis) against D (y axis) and time (z axis) and colored the points according to their y (D) values with a colormap ranging from blue (bases close to each other) to red (bases far from each other). In this way, I could monitor the transitions between different stacking patterns and identify rotations of bases about the

N9-C1’ bond. The stacking between the two guanines was preserved and maximized during both the LES4 (Figure 18A) and LES1 trajectories (Figure 18B) while the stacking between the two adenines did not persist in either trajectory (Figures 18C and 18D). Either A44 or A33 rotated sharply about the N9-C1’ bond in both LES4 and LES1 trajectories. In the LES4 trajectory, A33 rotated until its plane was almost perpendicular to the plane of G43, such as to increase the gap between the two stems, contributing to the large opening of the K-turn. During the last 2 ns of the LES4 simulation, the reformation of the kinked structure was accompanied by rotation of both A33 and A44 and a random distribution of the relative orientations of the two adenines was observed. In the LES1 trajectory three different structure clusters could be distinguished: (i) A44 stacked on A33 in the first 4 ns of the trajectory; (ii) A44 rotated, allowing the stems to achieve close proximity; (iii) the tightly kinked conformation formed, with the plane of A44 being almost perpendicular on the plane of G32.

In summary, the transition to the open conformation (LES4) was facilitated by rotation of A33 about its N9-C1’ bond, while the tightly kinked structure (LES1) was formed upon equivalent rotation of A44. Interestingly, rotation of A44 was also observed during the reformation of the K-turn in the LES4 trajectory.

5.2.3. Modified stacking patterns in the internal loop

The rotation about the N9-C1’ bond of A33 in the LES4 trajectory and of A44 in the LES1 trajectory was accompanied by formation of new stacking patterns in the region corresponding to the internal loop. In the LES4 trajectory the A44 stacked on G32 and further on G43 during the time of maximal K-turn opening (7-8 ns) (Figure 19A) while no such stacking interaction was formed in the LES1 trajectory (Figure 19B). The new stacking pattern formed by G43, G32 and A44 was interpolated horizontally between the two stems reflecting the maximal value of φ. In the LES1 trajectory, A33 formed a new stacking interaction with G43 (Figure 19D) and no stacking was formed between A44 and G32 (Figure 19C). The new stacking pattern formed by G32, G43 and A33 was extruded vertically from the stems, reflecting the minimal value of φ.

In both LES4 and LES1 trajectories one of the two adenines rotated about the N9-C1’ bond while the other formed a new stacking pattern with the guanines. A structural view of the transitions between different stacking patters is shown in Figure 20. Two significantly different conformations, one tightly kinked and one highly open were formed depending on the LES setup employed.

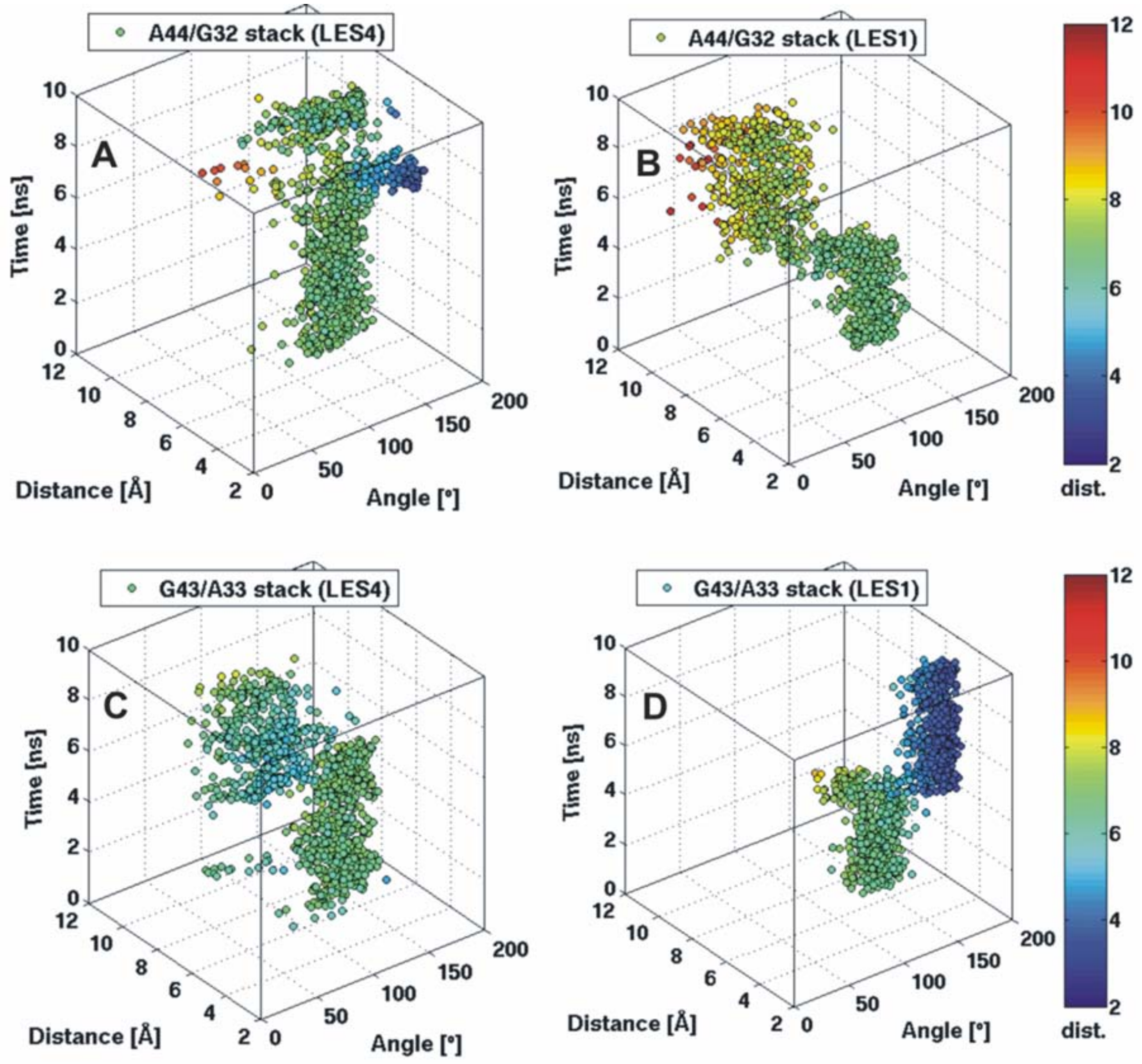


Figure 19: Scatter 3D plots of the stacking interactions between: (A) A44/G32 in the LES4 trajectory; (B) A44/G32 in the LES1 trajectory; (C) G43/A33 in the LES4 trajectory; (D) G43/A33 in the LES1 trajectory. For details clarifying the plots and coloring see Figure 18.

However, both conformations were characterized by loss of G-A base pairs. Because the difference between the two trajectories was the degree of backbone flexibility permitted, I analyzed the correlation between the K-turn opening and motions in the sugar-phosphate backbone.

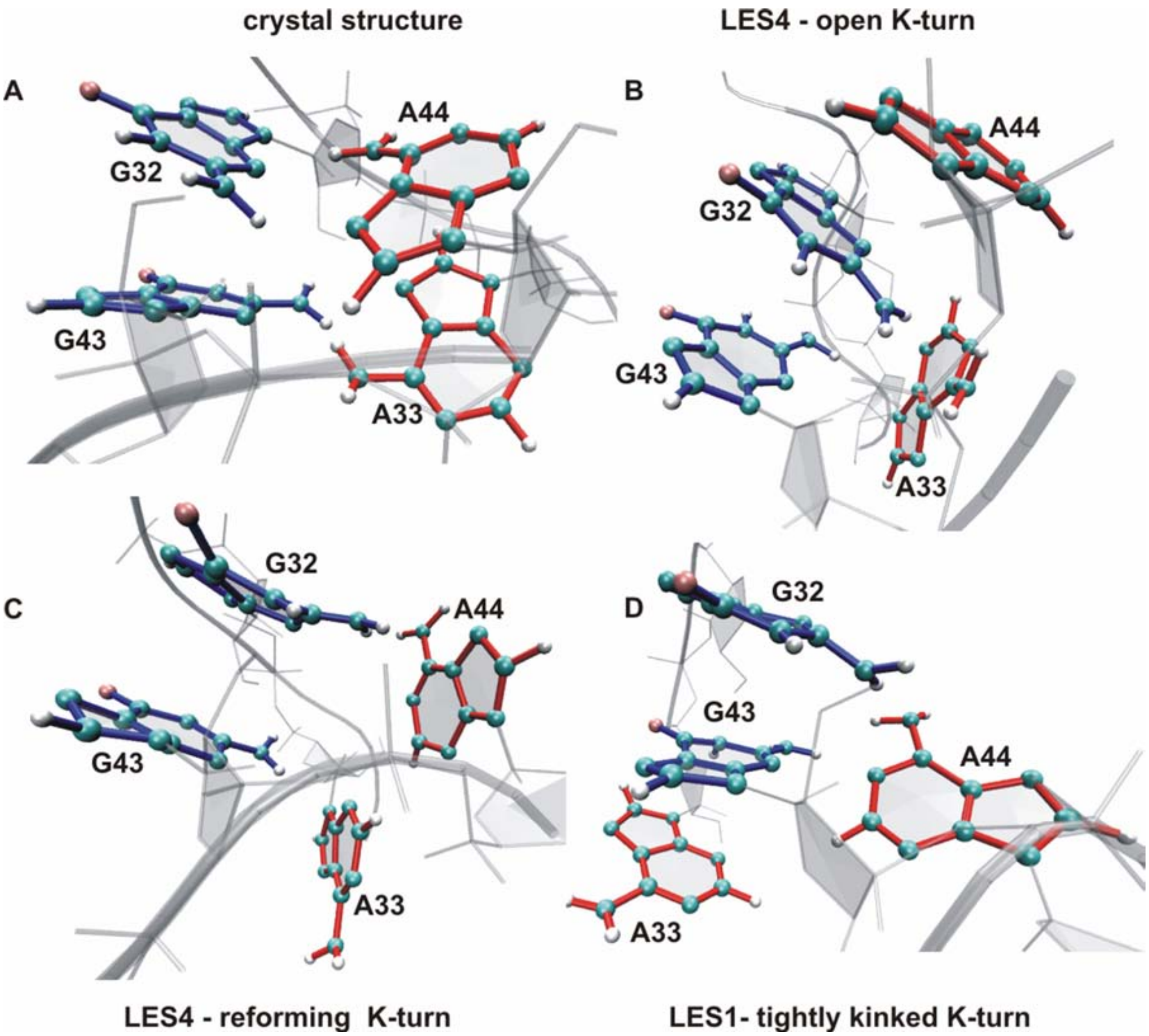


Figure 20: Structural view of the transitions between different stacking patterns in the free RNA. (A) G32/G43 and A44/A33 stacking interactions as observed in the crystal structure of KtU4-15.5K complex; (B) A44/G32/G43 stacking interactions formed in the LES4 trajectory after ~7 ns; (C) Rotation about the N9-C1' bond in A33 and A44 as observed in the LES4 trajectory after ~9ns; (D) G32/G43/A33 stacking interactions formed in the LES1 trajectory.

5.2.4. Opening of KtU4 is correlated with backbone flexibility in the stems

Local motions in the sugar rings and the phosphate groups of nucleotides at the point where the two stems branch revealed that certain conformations are required for the formation of the kinked structure. Stabilizing inter-stem contacts can form only if the interacting atoms are properly oriented for hydrogen bond formation. The following contacts are established between the C-stem and the NC stem: (i) the 2' OH group of A29 is hydrogen-bonded with the N1 atom of A44; (ii) the 2' OH group of A33 is hydrogen-bonded with the 2' OH group of G45; (iii) the N2 atom of G45 is hydrogen-bonded with the O3’ atom of A33; and (iv) the 2’ OH group of G46 is hydrogen-bonded with the phosphate group of G34. Applying LES to the stems loosens these contacts because the backbone has more conformational freedom, thus promoting the opening of the RNA. I investigated which local conformational transitions within different nucleotides were correlated with the K-turn opening, calculating: (i) the sugar pucker (the pseudorotation angle); (ii) the orientation between the base and the sugar (the χ angle) and (iii) the orientation about the C4’-C5’ bond (the γ angle).

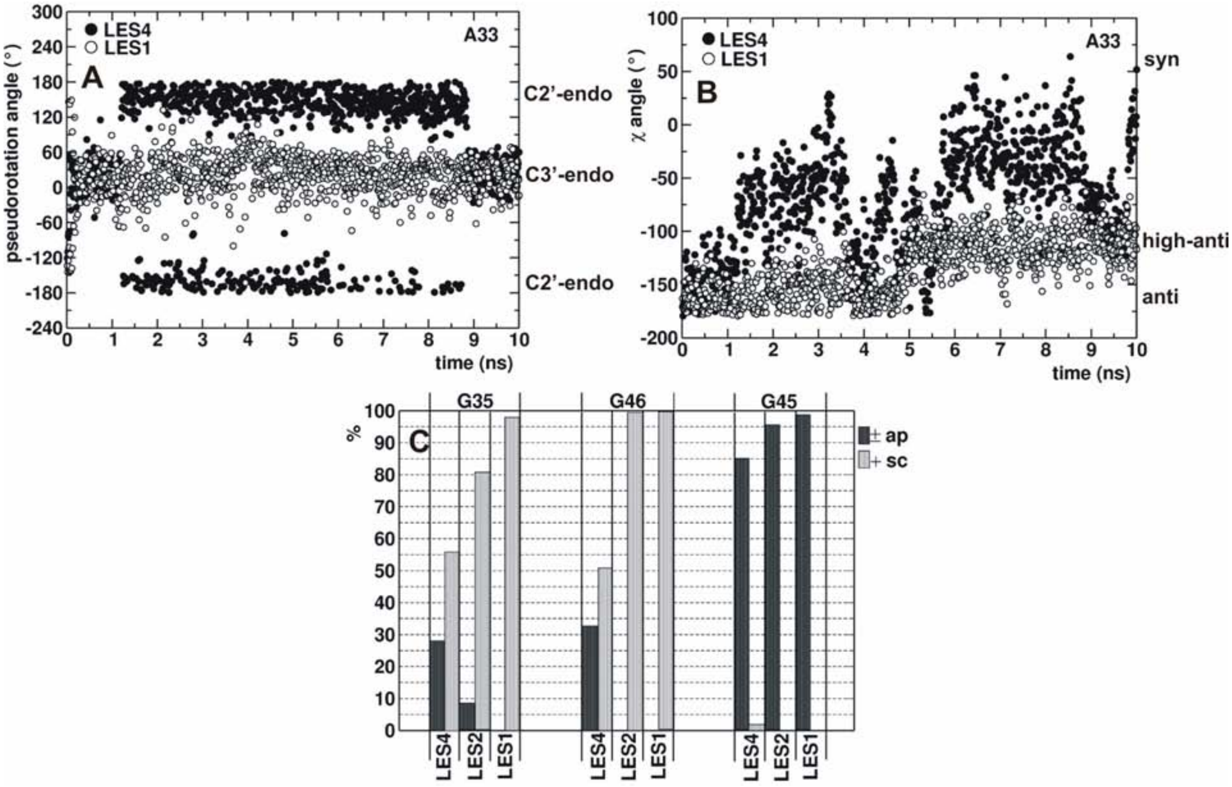


Figure 21: Correlation between backbone flexibility and the k-e motion.
(A) The pseudorotation angle describing the sugar pucker of A33 during the LES4 (filled circles) and LES1 (empty circles) trajectories; (B) The χ angle describing the rotation about N9-C1’ bond in A33 during the LES4 (filled circles) and LES1 (empty circles) trajectories; (C) relative populations of ±ap (black) and +sc (gray) configurations (describing the rotation about the C4’-C5’ bond).in G35, G46 and G45 in the LES4, LES1 and LES2 trajectories; the percentage of trajectory frames in which intermediate configurations were sampled is not shown.

For detailed description of the parameters see 1.7. A33 was very flexible in terms of the sugar pucker and the χ angle in the LES4 trajectory (Figures 21A and 21B). In the kinked conformation, A30 had a C3’-endo sugar pucker and an anti base-sugar orientation. When the RNA opened, the sugar pucker changed to C2’-endo and the χ angle varied towards a syn orientation describing the rotation of A33 about the N9-C1' bond. In the LES1 trajectory, although A33 was part of the defined LES region, its conformation was confined to a C3'-endo sugar pucker and an anti or high-anti base-sugar orientation. The high-anti configuration was adopted after 5 ns, corresponding to lower φ values.

The orientation about the C4'-C5' bond allows O5' to assume different positions relative to the furanose ring. If rotation about the C4'-C5' bond is permitted in a single nucleotide, the entire backbone can adopt a different orientation. In the LES4 trajectory, such a rotation occurred in the G34, G35 and G46 nucleotides (Figure 6C). Both ±ap [$\gamma \in (\pm150^\circ\ \pm180^\circ)$] and +sc [$\gamma \in (30^\circ\ 90^\circ)$] configurations were sampled by these nucleotides in the LES4 trajectory but not in the LES1 trajectory. Other nucleotides such as G45 showed no difference in the sampled configurations between LES4 and LES1 trajectories (Figure 21C). In the LES2 trajectory, the +sc configuration was less sampled than in the LES4 trajectory by G35 and G34, while the γ dihedral of G46 was restricted to values corresponding to the +sc configuration.

5.2.5. Essential dynamics of the LES4 trajectory

To study which types of motions are correlated with the k-e motion of KtU4 in the LES4 trajectory I performed ED analysis. First the motion was decomposed onto the 25 slowest modes (see 2.8.4), the contribution of the first three modes to the total motion being 68% (Figure 22A). The projection of the trajectory onto the first three modes showed that: (i) the first mode accounts for the large opening observed between 6 to 8 ns; (ii) the second mode accounts for the partial opening observed between 1 to 2 ns; and (iii) the third mode accounts for the reformation of the K-turn observed between 8 to 10 ns (Figure 22B).

The plot of φ along the first mode (Figure 22C) shows that the ED analysis captures the large opening of the RNA, separating it from partial opening that occurs due to higher frequency motions. The motions along the first mode included: (i) opening of G-A base pairs (Figure 22D); (ii) rotation of A33 about N9-C1' bond; (iii) formation of the A44-G32-G43 stacking pattern; and (iv) rotation of A30 about the N9-C1' bond. No correlated motion occurred in the NC-stem or in the external loop.

The plot of φ along the third mode (Figure 22E) shows that the motion along this eigenvector reflected the closing of the RNA observed after 8 ns. The G-A base pairs remained largely open although the K-turn reformed (Figure 7F).

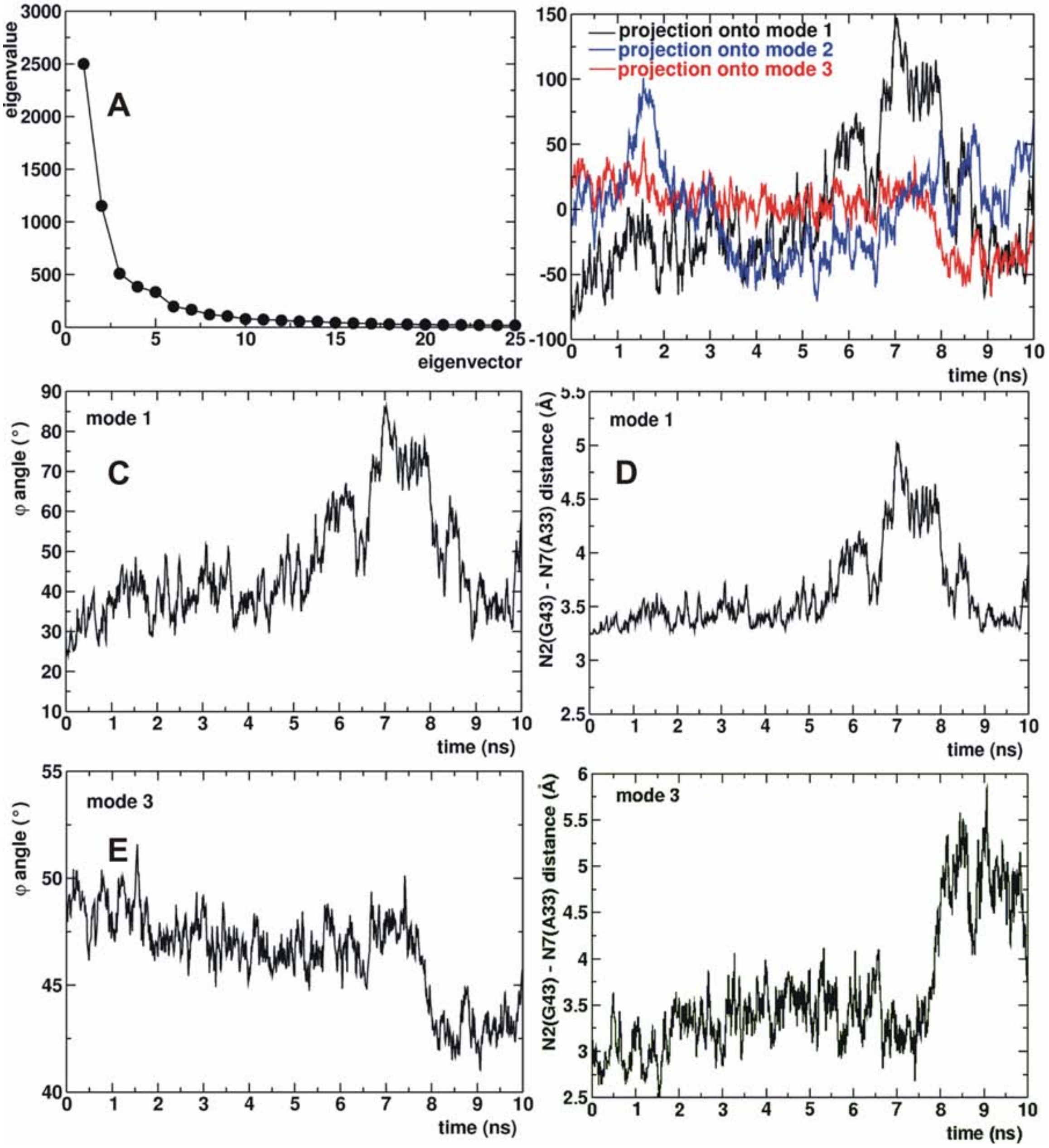


Figure 22: Principal component analysis of LES4 trajectory.
(A) Decomposition of the trajectory along the first 25 modes. (B) Projection of the LES4 trajectory onto first (black line), second (blue line) and third (red line) modes. (C) φ angle along the first mode; (D) Distance between N2 of G43 and N7 of A33 along the first mode. (E) φ angle along the third mode; (F) Distance between N2 of G43 and N7 of A33 along the third mode.

This anti-correlation was similar to that observed during the LES1 trajectory. The motions along the third slowest mode included: (i) closing of the K-turn; (ii) rotation of A33 about the N9-C1' bond; (iii) rotation of A44 about the N9-C1' bond; and (iv) motions in the NC-stem and external loop. The similarity with the LES1 trajectory was striking with the exception that the G32-G43-A33 stacking pattern was not formed and the RNA did not kink as tightly.

5.3. The role of the external UUUAU pentaloop

The external UUUAU was not revealed in the crystallographic structure of the KtU4-15.55K complex although it was present in the construct used for crystallization (44). It is very likely that the loop is very flexible even in the presence of the 15.5K protein, thus not permitting the construction of a reasonable electron density map at 2.9 Å resolution. This external loop was shown to cross-link with the 61K protein, thus suggesting that it might play a role in the binding of 61K to the U4 snRNA. However, the exact function of the external pentaloop is not yet understood.

Because no structural data was available for this loop, modeling was required to obtain an RNA construct for simulations that represents the naturally occurring sequence (K2). The loop was modeled as described (see 2.8.1) and the resulting trajectories showed a very interesting behavior of the UUUAU loop. The loop adopted a specific orientation, its backbone turning over C41 in the trajectory of the bound K2 RNA. This orientation was not stable in the unbound RNA and could not be established if different loops were attached to the NC-stem. The establishment of this orientation allows an optimization of the electrostatic interactions between the loop backbone and the protein residues ASN40 and LYS44. A similar orientation was observed in the structure of the K-turn Kt15 in complex with the ribosomal proteins L7AE and L15E.

The behavior of the external loop described here was observed in the standard MD simulations, different trajectories of the bound and unbound K2 RNAs showing similar results. The application of LES leads to a destabilization of the specific orientation of the UUUAU pentaloop in the bound K2 RNA probably due to the increased flexibility in the proximity of the loop. It is very likely that applying an increased sampling procedure adjacent to a very flexible loop (as in the LES4 trajectory) resulted in an artificial behavior of the loop, thus making the MD+LES trajectories useless in the study of the external loop.

5.3.1. Trajectories of the external loop in the bound and unbound RNAs

During the non-LES trajectories of the bound K2 RNA, the sugar-phosphate backbone has a turn between the residues U40 and C41 and forms a groove that encloses the G-C base pairs of the NC-stem (Figure 23A). This structure remains relatively stable during the simulations of the bound RNA but is lost in the unbound RNA after only hundreds of picoseconds. The turn of the backbone is also observed in the crystal structure from the orientation of the phosphate group of C41.

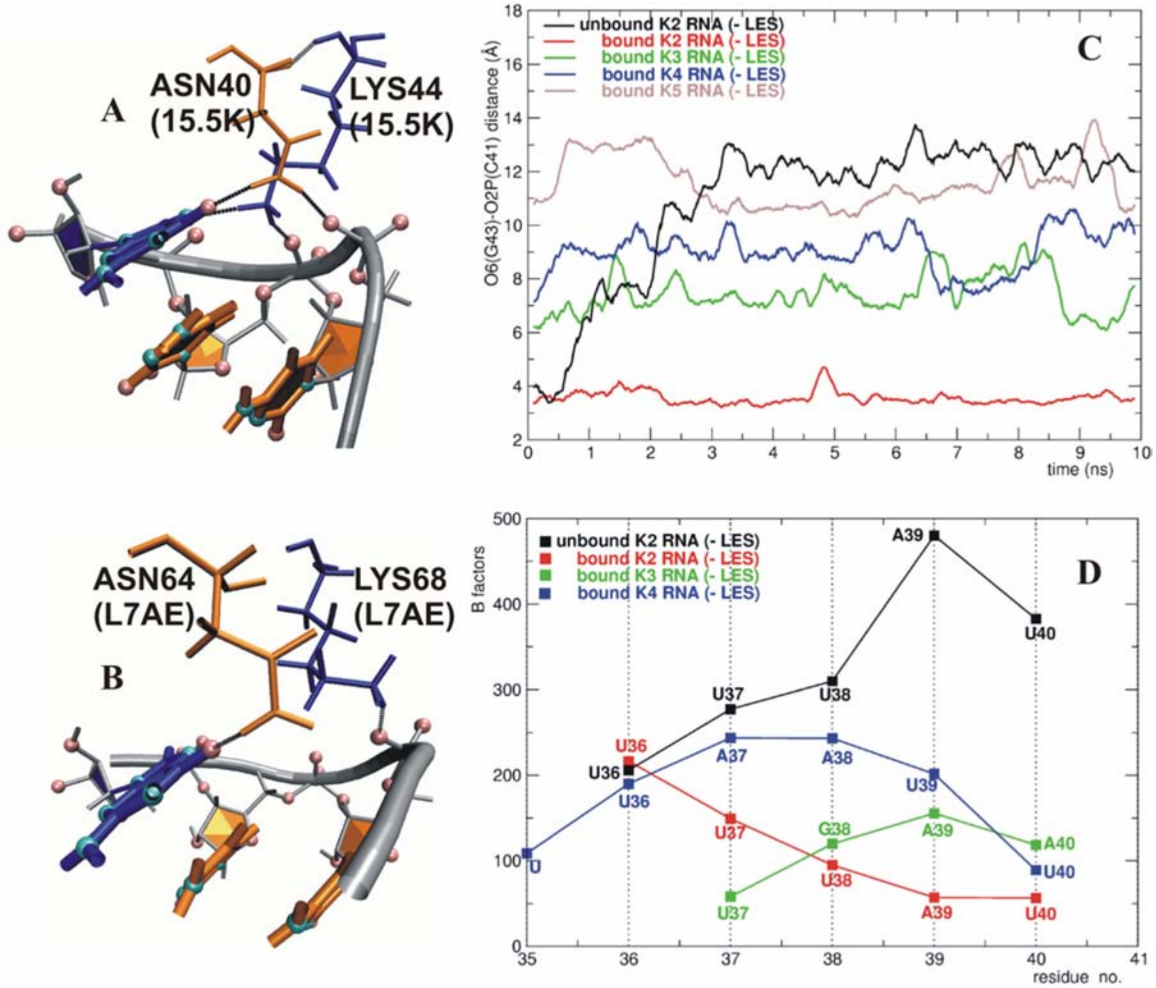


Figure 23: Interactions between the 15.5K protein and the external loop.
(A) ASN40 and LYS44 of the 15.5K protein establish a bridge between G43 and the sugar-phosphate backbone of the external loop of the 5' stem-loop of U4snRNA. (B) Bridging distance between the O6 atom of G43 and the O2P atom of C41 for the unbound K2 RNA (black curve) and for the bound K2 (red curve), K3 (green curve), and K4 (blue curve) and K5 (brawn curve) RNAs during the MD-LES trajectories; the bridge is shown in figure 6A. (C) B factors per residue for the nucleotides of the external loop for the unbound K2 RNA (black curve) and for the bound K2 (red curve), K3 (green curve), and K4 (blue curve) RNAs during the MD-LES trajectories. (D) ASN64 and LYS68 of the ribosomal L7AE protein establish a similar bridge between G43 and the sugar-phosphate backbone of the Kt15 RNA.

Two amino-acid residues, ASN40 and LYS44, are important in maintaining this structure, forming a bridge between G43 and the backbone of the external loop. The NH2 group of ASN40 establishes hydrogen bonds with the O6 of G43 and with one oxygen atom of the phosphate group of C41, and the $NH3^+$ group of LYS44 establishes hydrogen bonds with the N7 of G43 and an oxygen atom of the phosphate group of C42 (Figure 23A). These protein-RNA interactions were stable during all trajectories of 15.5K-KtU4 complexes (for the K2 construct). However, in the crystallographic structure ASN40 has a different orientation most probably due to the relatively low resolution to which the crystal diffracted. Figure 23B shows that the bridging distance between the O6 atom of G43 and the phosphate group of C41 significantly increases during the trajectories of the unbound RNA when comparing the naturally occurring K2 RNA in its bound and unbound forms, while a comparison of the trajectories of the bound K2 RNA with the bound K3, K4 and K5 RNAs reveals that the bridging distance is dependent on the loop flexibility (Figure 23B). The UGAA tetra-loop has a fold in which the uracil stacks on the guanine and the two adenines stack together (134). The UUAAUU hexa-loop has a further U-U base pair lengthening the NC-stem to three base pairs (135). Both structures prevent the backbone from turning over C41 and forming hydrogen bonds with ASN40 and LYS44, as observed in the case of the UUUAU pentaloop. The pentaloop is the only loop in which the nucleotide next to C41 lacks stacking interactions with C41. To analyze the flexibility of the three different loops in the bound and unbound forms of RNA I calculated the B-factors per residue during the non-LES trajectories. From a comparison of the fluctuations of the UUUAU loop in the bound and unbound RNA, there is a significant decrease in flexibility from U36 to U40 in the bound RNA (Figure 23C, black and red curves), suggesting that the 15.5K protein stabilizes the orientation of the backbone of the loop in the vicinity of C41. Nevertheless, the absolute values of the B-factors indicate that the pentaloop is still very flexible, suggesting that other factors might contribute to its stability. In the RNAs bound to the 15.5K protein, the flexibility is lower at the end nucleotides and higher at the core nucleotides of the UGAA and UUAAUU loops (Figure 23C, blue and green curves). This behaviour is expected because at the end nucleotides are covalently linked to a stable structure (C41 and G35, respectively). Although the UUUAU loop was modelled and the structures of UGAA and UUAAUU loops have been experimentally determined, the flexibility of the nucleotide N40 is lowest in the pentaloop.

5.3.2. Trajectories of the UGAA and UUAAUU loops in the bound RNAs

To investigate whether different loops attached to the NC-stem show a similar behaviour as the UUUAU loop, I modelled two RNA constructs: K3 and K4 RNAs. As described in 2.8.1, K3 RNA has a UGAA tetraloop in which U37 stacks on G38, C41 stacks on A40 and A40 stacks on A39 while K4 RNA has a UUAAUU loop (pdbid 1HS3) in which the nucleotide U establishes a non Watson-Crick base pair (N3-H-O2 and N3-H-O4 hydrogen bonds, sugars in cys orientation) with U40. The common feature of these structures is a stacking interaction between the last nucleotide of the NC-stem (C41) and the adjacent nucleotide in the loop (A40 in K3 RNA and U in K4 RNA) which is not present in K2 RNA. As shown in 3.3.1, the orientation adopted by the UUUAU loop in the bound K2 RNA was not adopted by the UGAA and UUAAUU loops in the bound K3 and K4 RNAs.

However, a very interesting behaviour of the loops was observed during the 10 ns standard MD simulations of the bound K3 and K4 RNAs, the stacking interaction between C41 and N40 being destabilized. This destabilization was not observed in the unbound RNAs and was accompanied in the bound RNAs by a movement of the loop backbone towards the protein due to attractive electrostatic forces. However, the amplitude of the movement was much smaller than that observed for the UUUAU loop in the simulations of the bound K2 RNA.

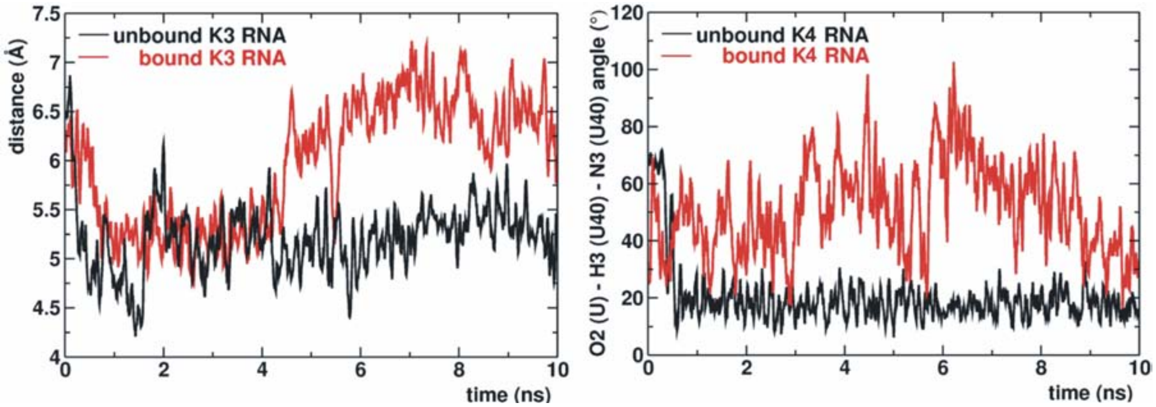


Figure 24: Stacking interactions between C41 and N40.
(A) The distance between the geometrical centers of C41 and A40 is shown for the unbound K3 RNA (black curve) and for the bound K3 RNA (red curve) (B) The hydrogen bonding angle between the O2 atom of U, the H3 atom of U40 and the N3 atom of U40 is shown for the unbound K4 RNA (black curve) and for the bound K4 RNA (red curve).

In the bound K3 RNA the stacking between C41 and A40 is destabilized after about 4.2 ns as shown in Figure 24A by plotting the distance between the geometric centers of the two bases. A40 is pushed towards outside by the movement of the backbone towards the protein. In the bound K4 RNA, the U-U40 base pair is slightly destabilized by an increase in the buckle between the two bases that is also a result of the electrostatic attraction between the phosphate groups in the loop backbone and protein residues

(Figure 24B). The C41/N40 stacking interactions are preserved in the unbound RNAs. Snapshots taken from the trajectories of the bound K3 and K4 RNA are shown in Figure 25 illustrating the transitions described above.

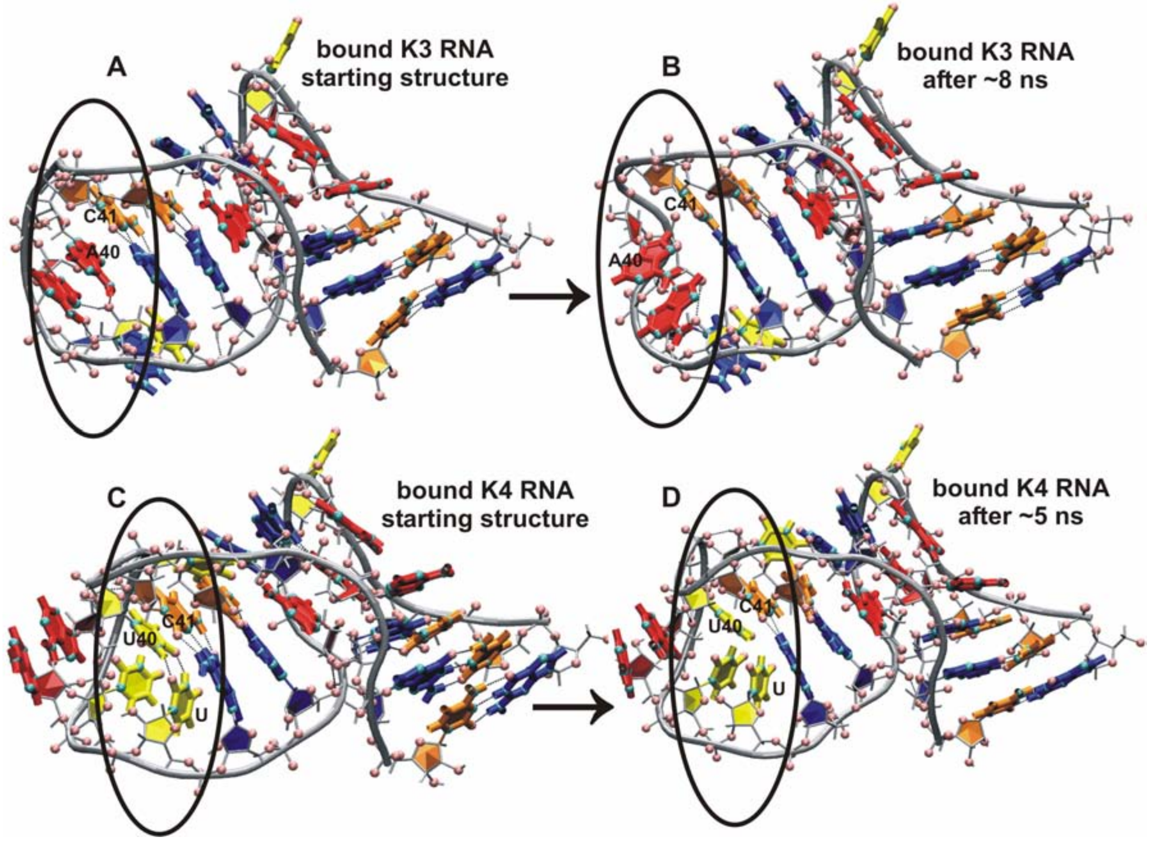


Figure 25: Snapshots from the simulations of bound K3 and K4 RNAs:
(A) The bound K3 RNA at the start of the simulations. (B) The bound K3 RNA after ~8 ns. (C) The bound K4 RNA at the start of the simulations. (D) The bound K4 RNA after ~5 ns. The external loop is encircled and the C41, A40 (in K3 RNA), U, and U40 (in K4 RNA) are highlighted

5.3.3. Trajectories of the bound and unbound Kt15

To find whether my model for the external loop shares structural features with other K-turns, I extracted the K-turns and their associated proteins from the big ribosomal subunit of *Haloarcula marismortui* and investigated their 3D structures.

Interestingly, the K-turn Kt15 shares many features with KtU4. The only difference is that instead of the G43-A33 base pair (in KtU4), Kt15 has an A-U-G triplex. The NC-stem has two Watson-Crick base pairs extended with a loop-like structure different from the UUUAU loop of U4 snRNA. Kt15 binds to the ribosomal L7AE protein that has a similar fold to 15.5K protein. The loop-like structure of Kt15 starts with a distorted A-U base pair adjacent to last base pair of the NC-stem (Figure 26A). In the distorted A-U base pair, the propeller twist between the two bases is ~90°, permitting an orientation of the backbone similar to that observed for the UUUAU loop during the trajectories of the bound K2 RNA. In the L7AE protein two residues (ASN64 and LYS68) establish a bridge between the G nucleotide of the A-U-G triplex and the backbone of the loop-like structure (Figure 23D) in an almost identical manner as the ASN40 and LYS44 residues from 15.5K protein,

I simulated the Kt15 RNA both in the unbound form and in the complex with the ribosomal L7AE protein, finding that the distorted A-U base pair evolves rapidly into a Watson-Crick base pair in the unbound RNA (Figure 26B) while remaining relatively (but not throughout the entire trajectory) stable in the bound Kt15.

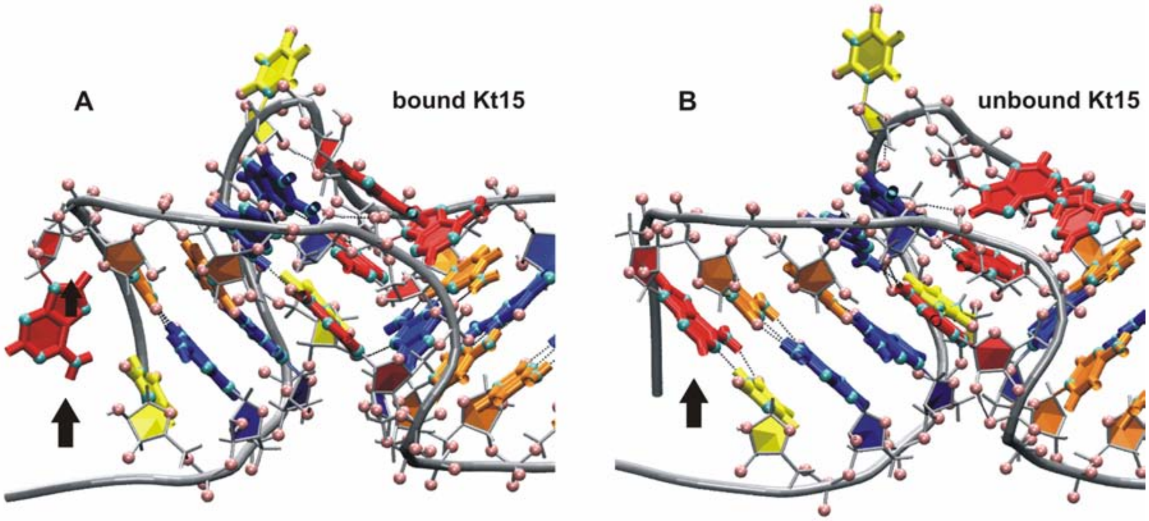


Figure 26: Snapshots from the simulations of Kt15.
The black arrow indicates: (A) the distorted A-U base pair as observed in the crystal structure of the L7AE-Kt15 complex (structure maintained during the simulation of the complex). (B) the formation of the Watson-Crick A-U base pair in the unbound Kt15 RNA.

A comparison between the orientations observed during the trajectories of the bound and unbound KtU4 and Kt15 is shown in Figure 27. The orientation of the loop backbone creates a groove in which ASN and LYS residues contact the RNAs. However, in the bound Kt15 the groove is much narrower than that observed in bound KtU4 (compare Figure 27A with Figure 27C). In the unbound Kt15, upon the formation of the A-U base pair adjacent to the NC-stem, the groove becomes a typical major groove of an A-type helix (Figure 27B), while in the unbound KtU4, the was rapidly flattened (Figure 27D).

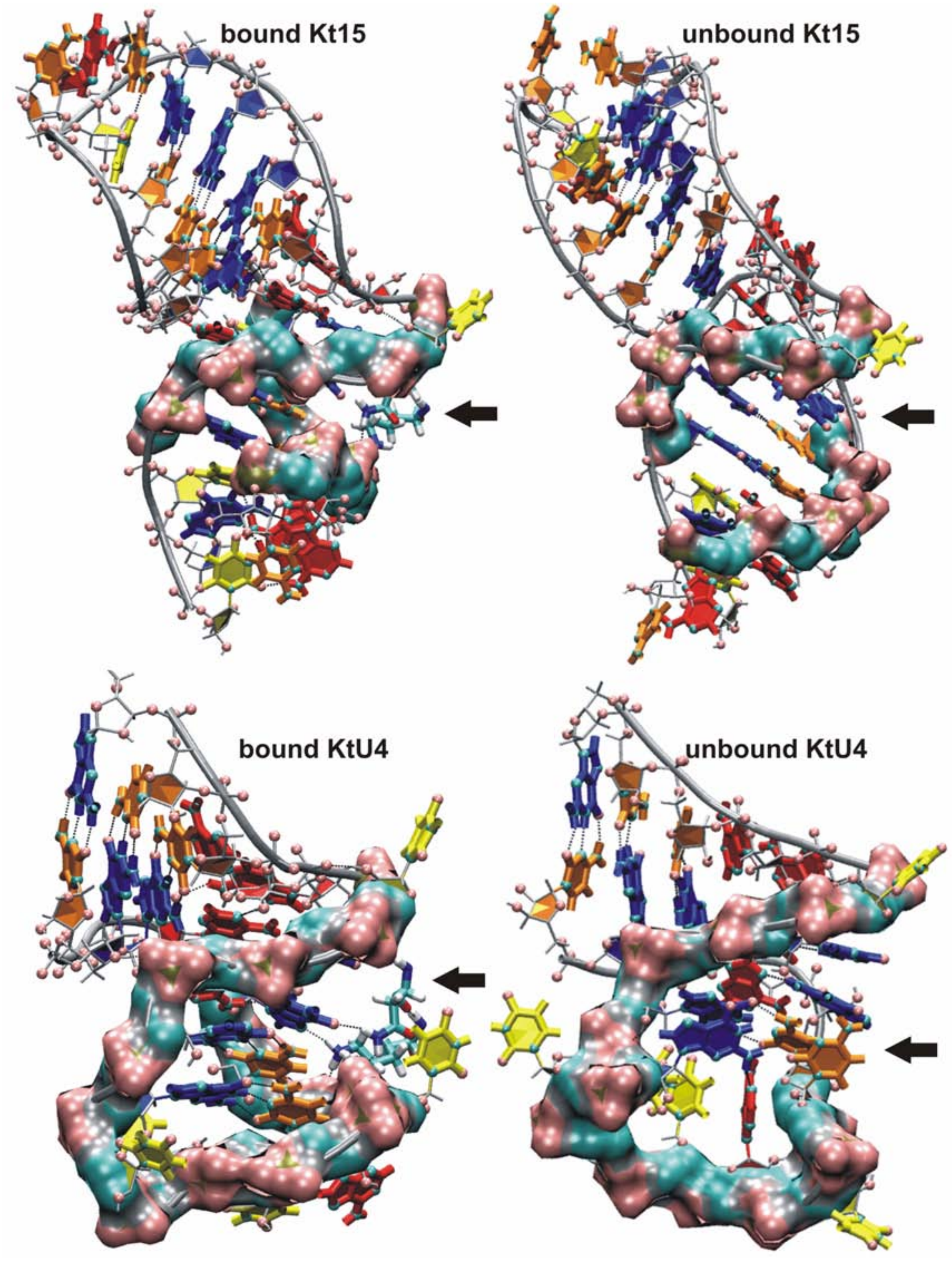


Figure 27: Orientations of the external loop in Kt15 and KtU4
The backbone of the external loop is shown as a molecular surface colored by atom name (P yellow, O pink and N cyan). The black arrows indicate the groove formed by the RNA backbone. (A) The very narrow groove in Kt15 RNA if bound to L7AE protein. (B) The groove becomes a regular minor groove of an A-type helix during the simulations of the unbound Kt15. (C) The groove formed in the simulations of the naturally occurring K2 KtU4 RNA bound to the 15.5K protein. (D) Flattening of the groove during the trajectories of the unbound K2 RNA.

5.4. Flexibility of the 15.5K protein

Existing experimental data on the dynamics of the free K-turn RNA motif provided the opportunity to compare such data with the results obtained by computer simulations, validating in this way the simulations. Unfortunately, the experimental data on the dynamics of the free protein is scarce, and thus no such validation is possible. It is thus far not understood to what extent the 15.5K protein is folded upon binding to its cognate RNA or whether binding to a specific RNA influences the dynamics of the 15.5K protein. Additionally, it is likely that the 61K protein or other proteins contact the 15.5K protein (already bound to KtU4) upon binding to U4 snRNA. However, the precise interaction interface between 15.5K protein and other proteins is not known. The standard MD simulations (all 10 ns long) performed on the unbound 15.5K and L7AE proteins, on the complexes between 15.5K protein and K1, K2, K3, K4 and K5 RNAs, and on the L7AE-Kt15 complex revealed that: (i) in the unbound 15.5K protein, the residues at the interface are more flexible but overall the free protein does not deviate significantly from the crystal structure of the protein in the complex, (ii) the 61K protein might contact a hairpin-like structure formed by PRO62 in the 15.5K protein, (iii) the flexibility of helix α5 depends on the nature of the RNA bound to the 15.5K protein, the largest flexibility being observed for the naturally occurring K2 RNA, (iv) the helix α5 of L7AE is also flexible in the trajectory of the L7AE-Kt15 complex, and (v) different flexible regions in the 15.5K protein that might play a role in the interaction between the 15.5K protein and other U4/U6-specific proteins. However, these results have not been tested experimentally, thus providing only predictive information about interactions based on the dynamics of 15.5K protein.

5.4.1. Flexible regions in the 15.5K protein

The B-factors per residue during four distinct trajectories of the 15.5K protein are plotted in Figure 28. The plot shows that, except for the five residues at the N-terminal and the C-terminal, several other flexible regions can be identified in the 15.5K protein: (i) residues 13 to 18 at the start of helix α1 (R1 shown in Figure 29A), (ii) residues 29 to 33 at the end of helix α1 (R2), (iii) the residues 33 to 44 of helix α2 and sheet β1 in the trajectory of the unbound protein (R3; protein-RNA interface), (iv) residues 46 to 50 at the end of helix α2 (R4), (v) residues 58 to 68 of the unstructured region between sheet β2 and helix α3 (R5; shown in Figure 29B), (vi) the residue 73 to 78 at the end of helix α3 (R6; shown in Figure 29C), (vii) residues 93 to 98 of the unstructured region between helix α4 and sheet β4 in the trajectory of the unbound protein (R7; protein-RNA inteface), (viii) residues 105 to 110 of the unstructured region between sheet β4 and helix α5 (R8; shown in Figure 29D), (ix) the helix α5 in the trajectory of the bound K2 RNA (R9).

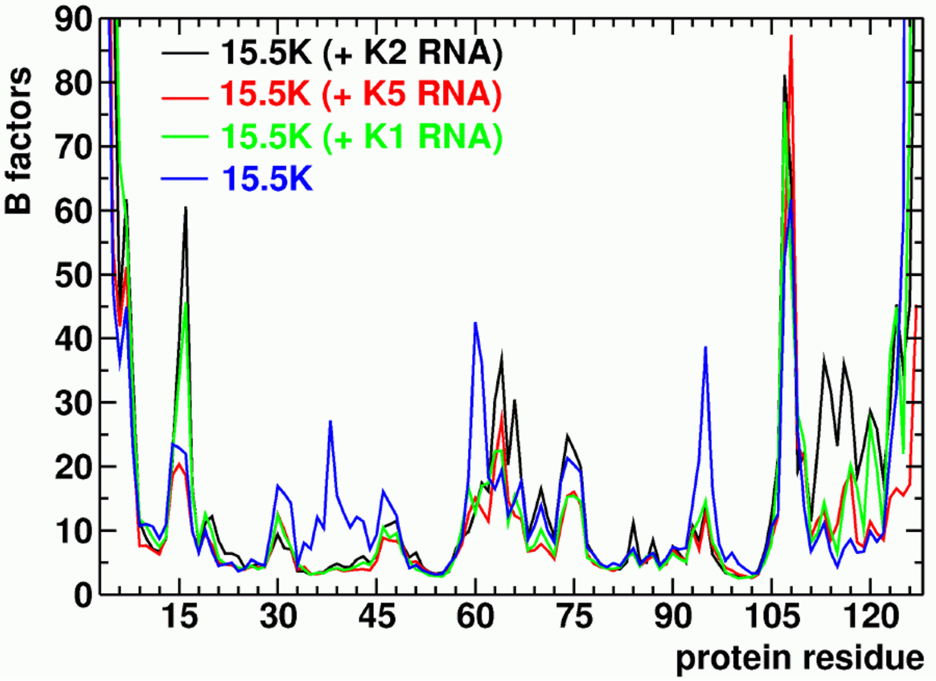


Figure 28: B-factors per protein residue in different simulations.
during the simulations of: 15.5K-K2 RNA complex (black curve), 15.5K-K5 RNA complex (red curve), 15.5K-K1 RNA (green curve) and free 15.5K protein (blue curve)

The flexibility of some of the regions does not show any dependence on: (i) whether the protein is bound or unbound to RNA, or (ii) on the RNA construct bound to the protein (R1, R2, R5, R6 and R8). All of these regions are unstructured, thus their flexibility might not have any functional relevance. However, I found that R5 and perhaps R6 (its flexibility might be related to that of R5) display several interesting features, R5 being also present in the ribosomal protein L7AE (see 3.4.3). Other regions such as R3, part of R5 and R7 constitute the protein-RNA interface and are flexible only in the free protein while R9 is of particular interest because its flexibility depends on the nature of the RNA structure bound to the protein (the helix α5 is flexible only in the naturally occurring 15.5K-K2 RNA complex). The reproducibility of these findings was tested by running multiple trajectories of the same systems. The only system that produced different trajectories in terms of behaviour was the unbound protein which had a very flexible α5 in one trajectory and a rather stiff one in another trajectory (different software was used and slightly different initial conditions). However all the other flexible regions of the unbound protein were found in both trajectories.

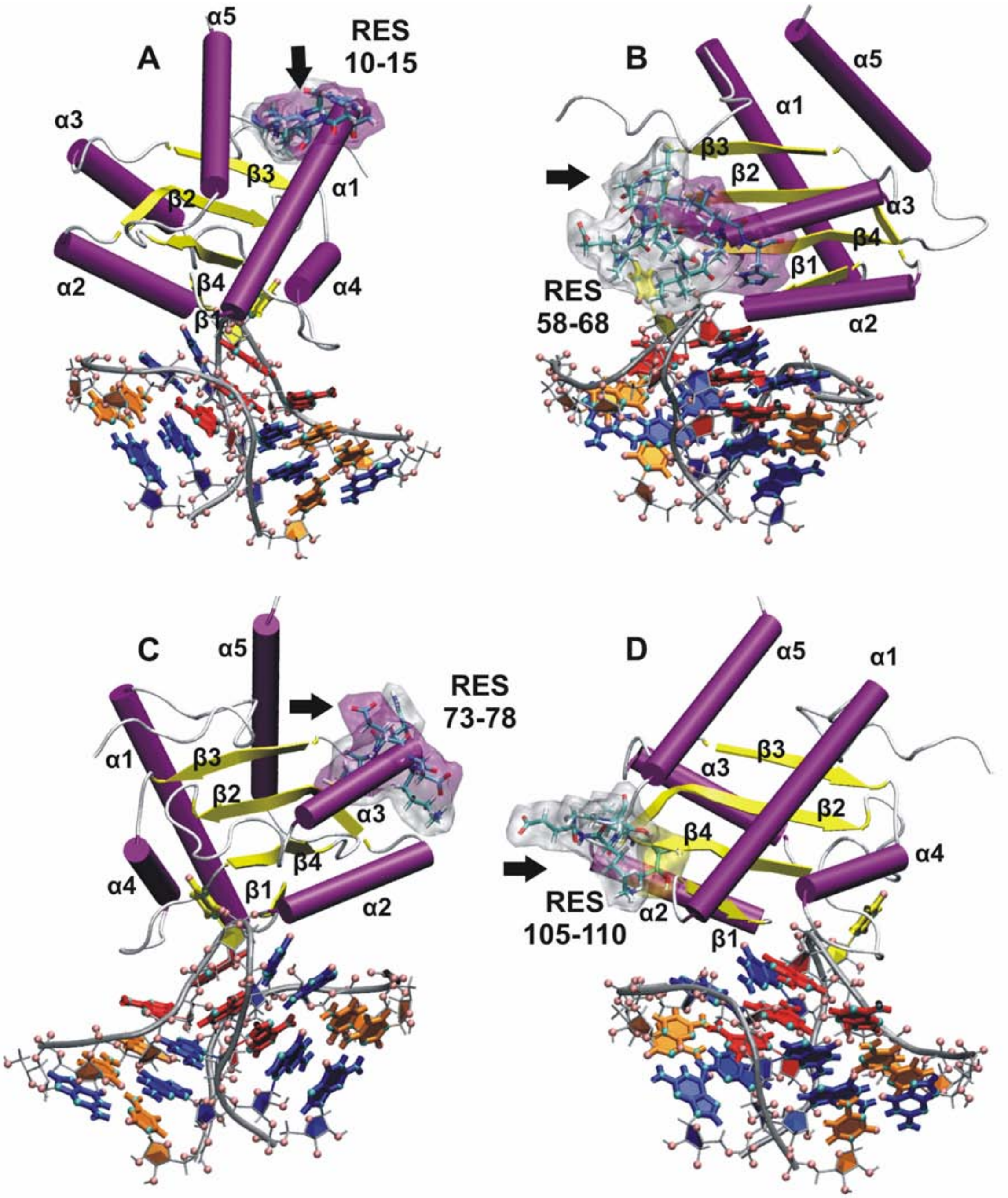


Figure 29: Flexible regions in the 15.5K protein (in all simulated system).
The flexible regions are indicated by a transparent molecular surface colored by the secondary structure and by bond representations of the residues colored by atoms (A) R1 region (RES 10-15), (B) R5 region (RES 58-68), (C) R6 region (RES 73-78), and (D) R8 region (RES 105-110)

A comparison between the B-factors per residue during the trajectories of the 15.5K-K2 RNA complex and the L7AE-Kt15 complex revealed that the flexible regions of 15.5K protein overlap with the flexible regions of L7AE (Figure 30). Interestingly, the flexibility of helix α5 was also observed in the trajectories of both complexes.

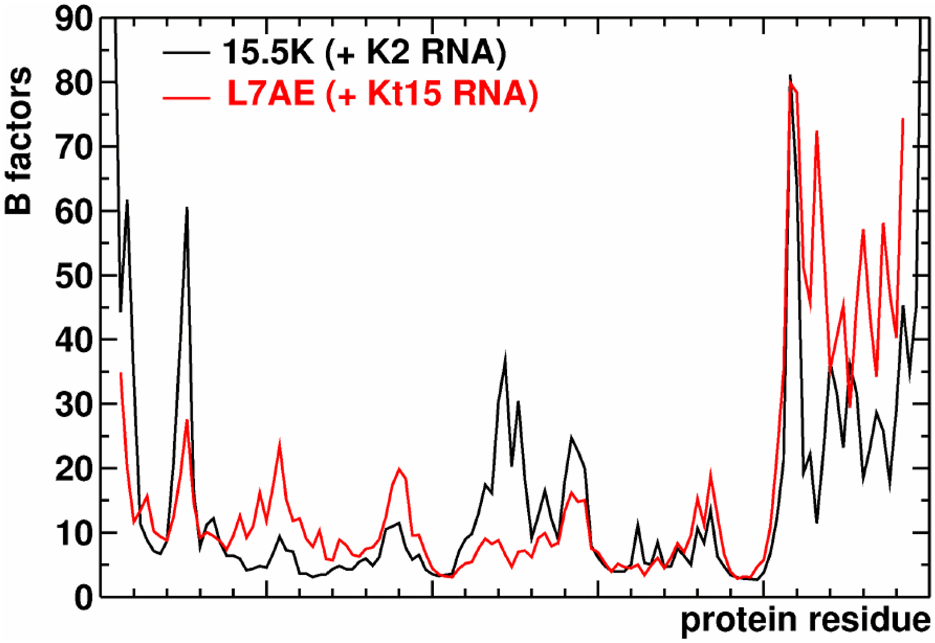


Figure 30: Comparison between the flexibilities of 15.5K and L7AE proteins bound to their cognate RNAs B-factors per protein residue are shown for: (A) the 15.5K protein in the simulation of the 15.5K-K2 RNA complex (black curve), and (B) the L7AE protein in the simulation of the L7AE-Kt15 RNA complex (red curve)

5.4.2. The protein-RNA interface

The protein-RNA interface is shown in Figure 31 and is composed of the entire regions R3 and R7 and part of the region R5. For detailed description of the contacts that are established between the protein and the RNA see 1.3. In the unbound protein, several inter-residue hydrogen bonds (stabilized in the complex by a network of hydrogen bonds involving also RNA bases) become instable. In the bound protein, ARG36 establishes hydrogen bonds with GLU41, the interaction being stabilized by a complete network involving the RNA nucleotides, G32 and G43 (Figure 32A). In the trajectory of the free protein, the hydrogen bonds formed between ARG36 and GLU41 are not stable and the residues become flexible and adopt random orientations in space (a snapshot is shown in Figure 32B). However, the only influence observed in the overall protein structure due to such local instabilities is a shortening of β1.

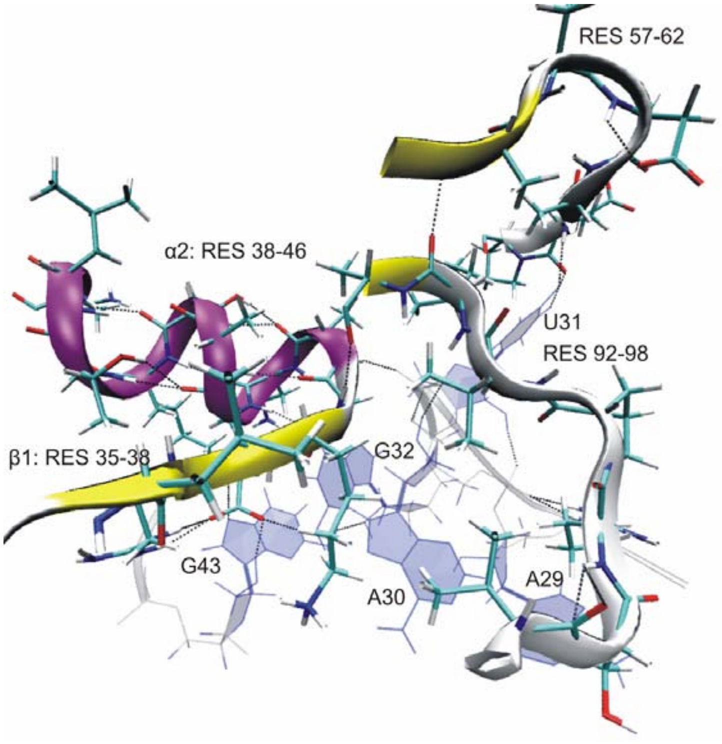


Figure 31: Protein-RNA interface.
The protein residues are shown in two representations: (i) cartoon (α-helices purple, β-sheets yellow and unstructured regions white), (ii) bonds colored by atoms (carbons cyan, oxygens pink, nitrogens blue and hydrogens white. The RNA bases are shown in blue

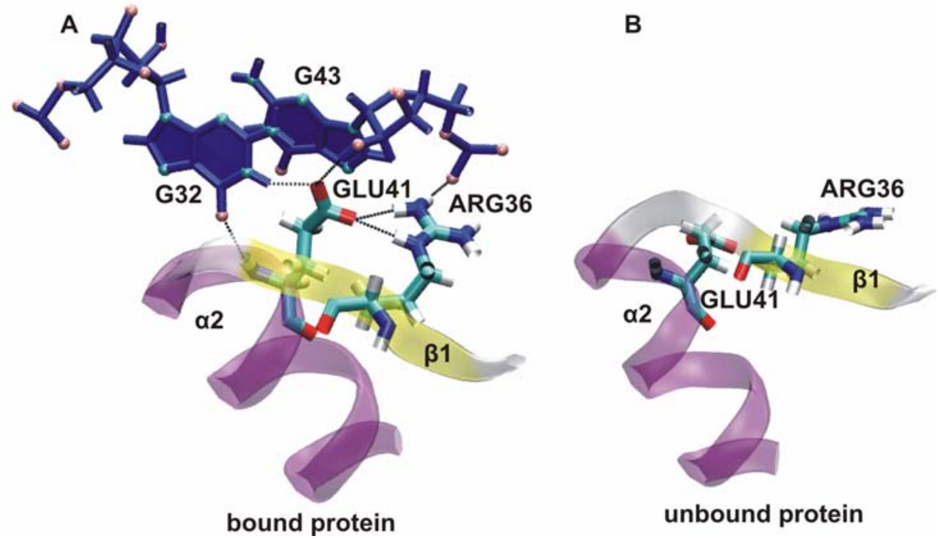


Figure 32: Example of protein-RNA contact.
(A) as observed in the 15.5K-KtU4 complex; (B) as observed during the trajectory of the free 15.5K protein. Coloring is consistent with Figure 31. Hydrogen bonds are shown as dotted black lines

5.4.3. PRO62 induces a hairpin-like structure

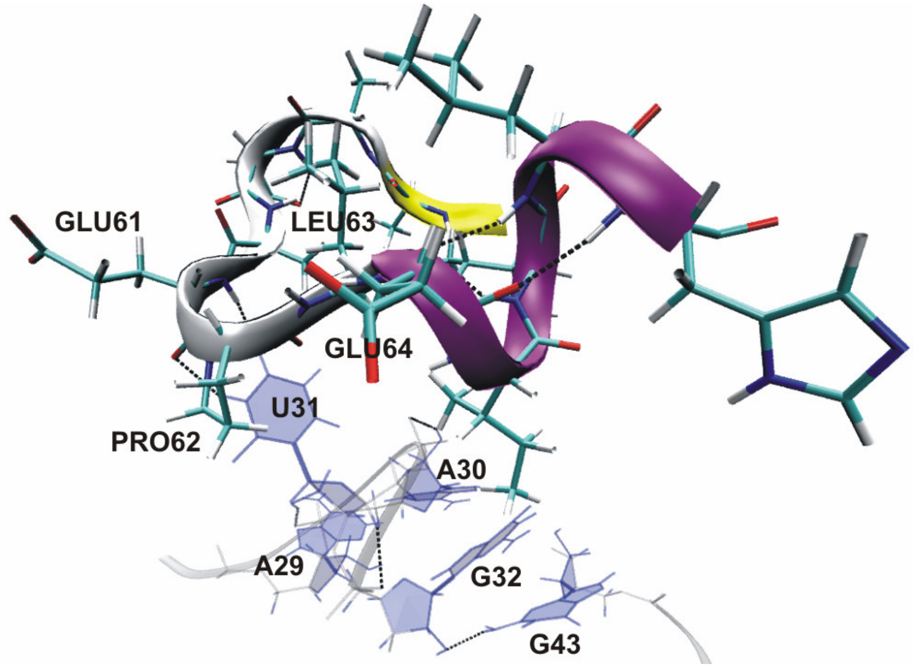


Figure 33: The hairpin-like structure in the 15.5K protein.
Coloring is consistent with Figure 31

R5 is the unstructured region between β2 and α3. ALA60 and GLU61 are contacting the RNA nucleotide U31, thus contributing to the protein-RNA interface. Although this region is clearly flexible in all the trajectories analyzed, its flexibility is somehow restricted by a hairpin-like structure induced by PRO62 (Figure 33). Interestingly, although it is located in the middle of an unstructured coil, this hairpin-like structure is relatively stable during all the simulations. Furthermore, the same structure is present in the ribosomal L7AE protein where it is part of the binding pocket of the L15E protein to L7AE.

PRO62 is surrounded by GLU61 and GLU64 in the 15.5K protein. However, ILE 63 interposes between PRO62 and GLU64 as opposed to L7AE where two glutamates analogues to GLU61 and GLU64 are located in the immediate vicinity of the proline.

The electrostatic potential represented on the molecular surfaces of the two proteins shows a negative area on a relatively positive wheel-like structure (Figure 34; white arrows). There are several differences between the two proteins: (i) the wheel-like structure is larger in the 15.5K protein, (ii) the pit on the molecular surface that is surrounded by the wheel-like structure is deeper in the 15.5K protein, (iii) the two negative charges (the two glutamates) are adjacent in the L7AE protein and are facing each other in the 15.5K protein, and (iv) the positive charge around the glutamates is more pronounced in the 15.5K protein. However, in spite of these differences, there is a clear similarity between the R5 regions in the two proteins.

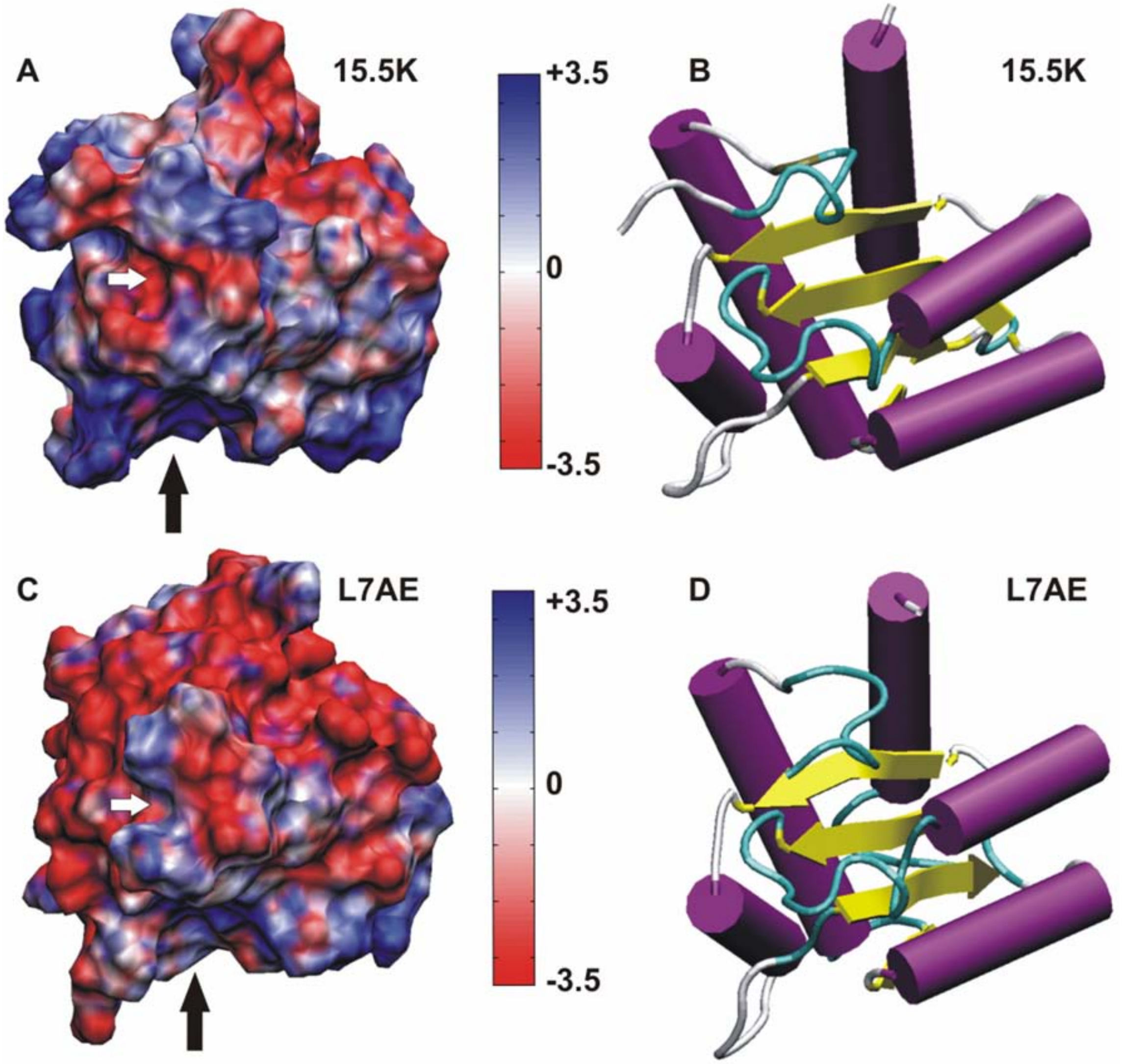


Figure 34: Electrostatic potential on molecular surfaces.
(A) Electrostatic potential on the molecular surface of 15.5K protein. (B) Electrostatic potential on the molecular surface of 15.5K protein. The black arrows indicate the RNA binding pocket while the white arrows indicate the location of the hairpin-like structure on the surface. The orientation of the proteins inside the surfaces is shown in (C) and (D).

5.4.4. The flexibility of α-helix 5

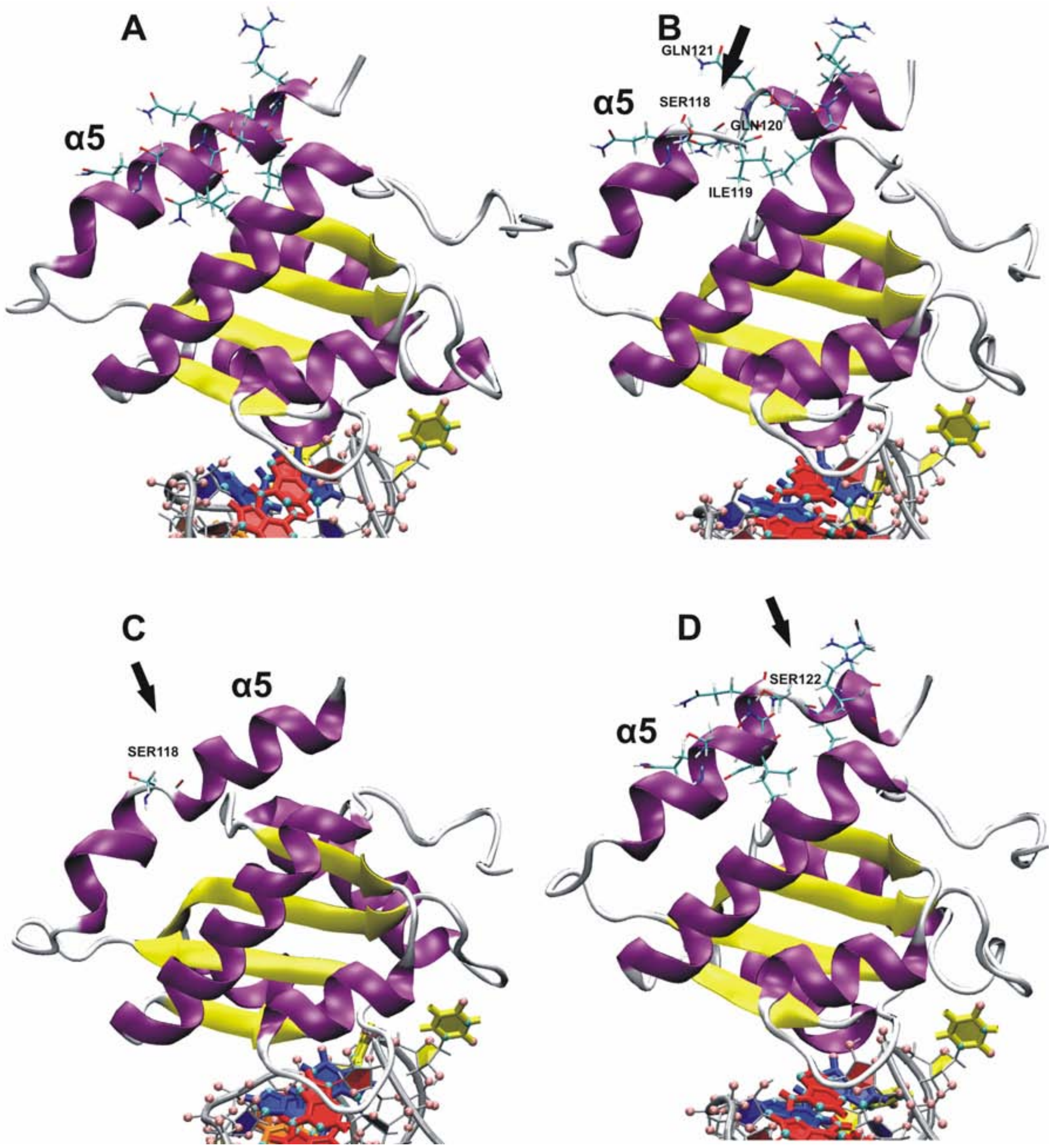


Figure 35: Transitions in the α-helix 5

The residues forming the inter-helix coil are shown as bonds colored by atoms. The black arrows indicate the locations of the coil (A) The helix α5 is shown as observed in the crystal structure of the 15.5K-KtU4 complex. (B) Snapshot in which α5 is split into two short helices of approximately the same length by an inter-helix coil formed by residues SER118, ILE119, GLN120, GLN121. (C) Snapshot in which the coil is formed only by residue SER118. (D) Snapshot in which α5 is split towards its C-terminal end into two helices that form a sharp angle between them; the coil is formed by residue SER122

The helix α5 is very flexible both in the 15.5K-K2 RNA complex and the L7AE-Kt15 complex. In the 15.5K-K1 RNA complex, its flexibility is reduced, while in all the other complexes, α5 is relatively rigid. The trajectories of the unbound protein were not conclusive regarding the flexibility of α5, in some trajectory the helix being flexible while in others rigid. The transitions occurring in the helix during the simulation of the 15.5K-K2 RNA complex are shown in Figure 35. The long α-helix breaks into two short α-helices with a coil between them. Different states were observed depending on the residues constituting the coil. In the most abundant state, the helix was split approximately in the middle, separating two short helices oriented at a variable (but generally small) angle against each other. Figure 35B shows a snapshot in which the coil is four residues (SER118, ILE119, GLN120, GLN121) long while Figure 35C shows a snapshot in which only one residue (SER 118) forms the coil. An alternative state was also observed in which the length of the two short helices were significantly different (Figure 35C) and the angle between them was sharper. However, the latter state was short-lived. Such transition were not observed or were very short-lived in the protein bound to other RNA constructs, providing an example of protein motions that are dependent on the nature of the bound RNA.

5.5. Steady-state fluorescence spectroscopy

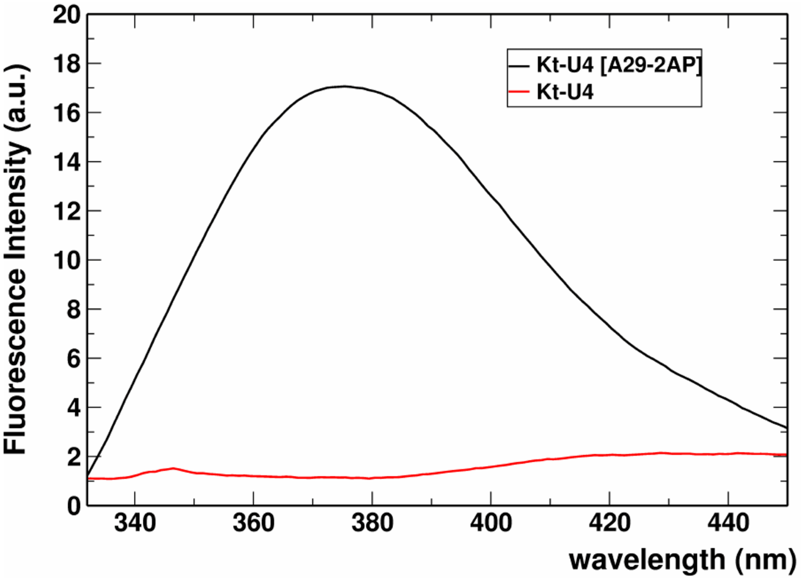


Figure 36: Spectrum of 2-aminopurine in KtU4
The black line shows the fluorescence spectrum of 2AP when incorporated in the RNA at position 29. The red curve shows a fluorescent spectrum of the unlabeled RNA as a control. The spectra are shown after background substraction (buffer spectrum).

A typical spectrum of 2-aminopurine (2AP) incorporated in KtU4 is shown in Figure 36. The emission had a maximum at 375 nm. Upon titration with 15.5K protein, the fluorescence of 2AP was quenched both in the absence (Figure 37A) and presence (Figure 37B) of Mg^{2+} ions in solution.

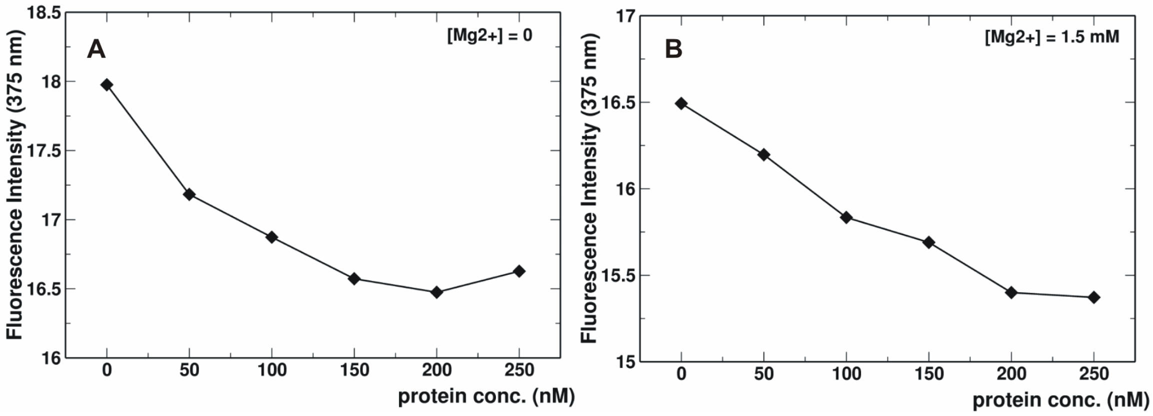


Figure 37: Titration of labeled KtU4 with 15.5K protein
in buffer 1 (A) - no Mg^{2+} ions present in solution - and in buffer 2 (B) - in the presence of 1.5 mM Mg^{2+}

The melting profiles of the RNA in the two buffer systems are shown in Figure 38. In the absence of Mg^{2+} (Figure 38A), a steady decrease in fluorescence intensity was observed until 30℃. Between 30 and 40℃, the fluorescence intensity decreased more sharply, while after 50℃, the fluorescence showed a substantial increase. In the presence of Mg^{2+}, the melting profile was similar (Figure 38B). However, the predicted melting temperature (T_m) was higher.

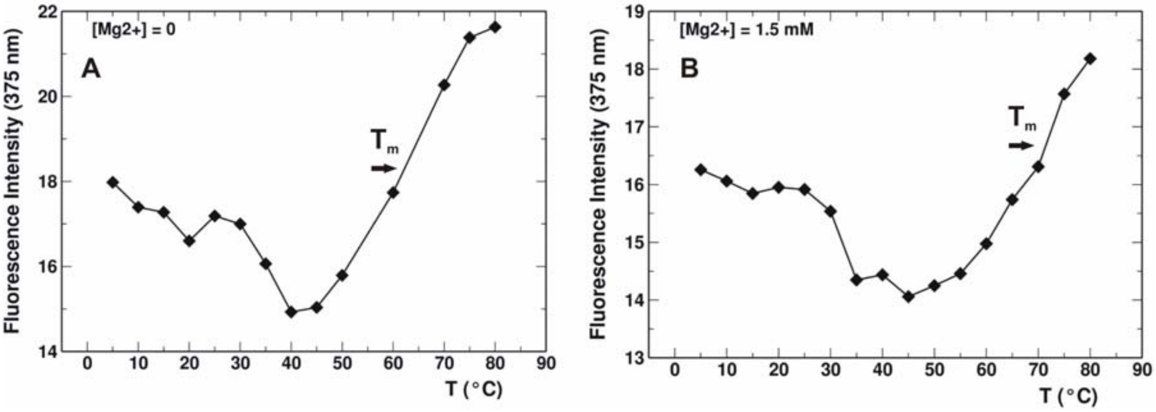


Figure 38: Melting profiles of labeled KtU4
in buffer 1 (A) - no Mg^{2+} ions present in solution - and in buffer 2 (B) - in the presence of 1.5 mM Mg^{2+}.
The predicted melting temperatures are indicated with a black arraw.

The influence of Mg^{2+} on the structure of KtU4 was tested by titration of the labeled RNA in buffer 1 with increasing amounts of $MgCl_2$ at 5℃ (Figure 39). A sharp decrease in fluorescence intensity was observed upon addition of up to 3 mM $MgCl_2$, while increasing the concentration of cations above this value lead to a slight increase in fluorescence.

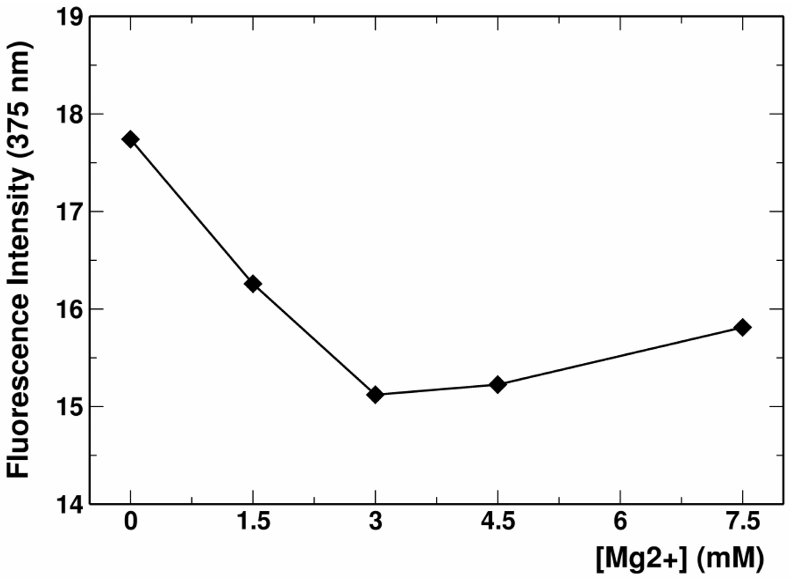


Figure 39: Titration with Mg^{2+} of labeled KtU4 at 5°C

6. Discussion

6.1. Protein-assisted RNA folding

I propose that the folding of KtU4 is assisted by the protein binding. The folding pathway is largely dependent on the RNA sequence and local flexibility. My model is based on MD simulations indicating significantly different trajectories for the bound and unbound RNAs and on the susceptibility to chemical modifications of different bases of the unbound and bound RNA. The protein cannot bind to the RNA and promote its folding unless the RNA has the inherent flexibility conferred upon it by the particular nucleotide sequence.

By probing the chemical reactivity of the N1 and N2 positions of guanines (using Kethoxal) and the N1 position of adenines (using DMS) in the presence and in the absence of 15.5K protein, Stephanie Nottrott showed that certain interactions in the RNA arise only upon protein binding (60). Conversely, the simulations indicate that conformational transitions with loss of specific secondary structure elements occur only in the unbound RNA. These findings, summarized in table 4, suggest that the RNA folding is assisted by the binding of 15.5K protein. The K-turn opens and sometimes reforms in the absence of the protein, suggesting that in addition to protein binding, the local flexibility of the RNA plays a crucial role in the folding process. In most of the trajectories the K-turn does not reform from the open conformation within the 10 ns time limit, only the local conformational transitions being reversible. The transition between the closed and open conformations was observed using FRET for the K-turn Kt7, which contains the most of the consensus sequence of K-turn motifs (58). The opening-closing process of the K-turn observed by FRET may be quite different and occur on a longer time scale. Nonetheless, the present simulations show a significant degree of opening for the K-turn in the absence of the protein.

Table 4: Interactions in the unbound RNA

Interaction	Unbound RNA (MD-LES)	Unbound RNA (MD+LES)	Unbound RNA (experiment)
G35-C41 base pair	lost (only K1)	lost (K1, K2)	lost
G32-A44 base pair	preserved	lost	lost
G43-A33 base pair	preserved	lost	no data
A30-A44 stack	lost	lost	no data
A30-A29 stack	formed	formed	no data
O2' (A29)-H-N1(A44)	lost	lost	lost
N2(G45)-H-O3' (A33)	preserved	lost	no data

The K-turn formed by the 5' stem-loop of U4 snRNA is of particular interest because the binding of 15.5K protein to this RNA nucleates the assembly of U4/U6 snRNP (37). Several other studies devoted to the mechanism of K-turn RNA folding suggest that factors such as the association with proteins and the presence of metal ions in solution are important (58,59).

Previously, interactions such as Tat-TAR interaction of HIV-1 or the binding of U1A protein to its RNA substrate have been assigned to the induced fit category (54-57,76,150,151). In these cases the conformation of the unbound RNA is stable but significantly differs from that of the bound RNA. The conformational capture, [as described by Leulliot and Varani (53)], is more difficult to locate due to the large flexibility of the RNA, which adopts an equilibrium between different conformations in its unbound state such that structure determination is not feasible. The MD simulations do not sample continuous inter-conversions between conformations since such events occur in the microsecond to millisecond time scale. Nevertheless, the FRET study by Goody et al. (58) and the partial reforming of the kinked conformation during the simulations of the unbound RNA might suggest a 'conformational capture' mechanism for the folding of KtU4. In addition, the folding process is influenced by other factors such as ionic strength and magnesium ions, factors that were not considered in the current study. However, localized Mg^{2+} ions were not found in the crystal structure of the complex, suggesting that they influence the RNA folding by a 'diffusely bound' rather than 'site bound' mechanism (71,72).

From the six K-turns found in the 23S rRNA of *Haloarcula marismortui* (47), one (Kt38) is not associated with proteins, suggesting that the folding of the K-turn motif is not always assisted by the protein binding but it also depends on the structural context in which the K-turn forms. The K-turns from the ribosome are segments of large RNA structures and therefore might have a folding pathway different from the 5' stem-loop of U4 snRNA, which is an isolated small RNA with a very flexible external loop attached to the NC-stem.

6.2. The role of G-A base pair formation in the folding of KtU4

It has been shown that mutation of any of the purines forming the G-A base pairs in KtU4 abolishes binding of the 15.5K protein (29). Therefore, the formation of G-A base pairs in the internal loop is required for the correct folding of the RNA. However, their precise role is not understood. I provide here the first insights at atomic detail into the correlation between G-A base pair formation and K-turn folding. The G-A base pairs are unstable in the absence of 15.5K protein. Thus, a reasonable assumption would be that they are required for the formation of the kinked structure. However, the opening of G-A base pairs is not tightly correlated with the k-e motion of the K-turn. When increased sampling was applied to both the internal loop and parts of the stems, the k-e motion was accompanied by loss of G-A base pairs (LES4 trajectory). Strikingly, when LES was confined to the internal loop, the RNA adopted a relatively stable alternative conformation in which the G-A base pairs did not form while the RNA was a more tightly kinked structure resembling that of K-turn Kt58 (Figure 16D), having backbone kinks on both strands. I investigated whether this alternative conformation was stabilized due to LES artifacts by comparing the dynamics observed in the LES1 and LES4 trajectories. In the absence of experimental data at atomic level describing the dynamics of the free RNA, the identification of convergent motions between the LES1 and LES4 trajectories provides a reasonable argument that at least some of the dynamics observed in the LES1 trajectory occur during the k-e motion of the free RNA. The behavior of the RNA in the LES4 trajectory was in excellent agreement with data obtained by single molecule FRET experiments (58) and by chemically probing the RNA structure in the presence and absence of the 15.5K protein (60). To characterize the concerted motions in the free K-turn, I applied ED analysis of the LES4 trajectory. The slowest mode captured the k-e motion that is correlated with the opening of G-A base pairs. Surprisingly, along the third slowest mode, the degree of kinking in the RNA increased without the formation of G-A base pairs. Furthermore, the movement of the adenines relative to the guanines was even larger than that observed along the first mode and similar to the motions observed in the LES1 trajectory. The projection of the LES4 trajectory along different modes shows that the third eigenvector reflected the reformation of the K-turn in the last 2 ns of the trajectory. Therefore, the two simulations showed convergent motions, suggesting that the RNA is capable of adopting intermediate conformations characterized by a sharp angle between the two stems without the formation of G-A base pairs. The stabilization of the alternative conformation in the LES1 trajectory enabled the search for local structural dynamics contributing to the large opening of the RNA, such dynamics being identified by analyzing the divergent motions between the LES4 and LES1 trajectories.

From these findings, we infer that the loss G-A base pair is insufficient for achieving the k-e transition. The opening of the K-turn caused the G-A base pair to open, thus accounting for the correlation between these two motions along the slowest mode of the LES4 trajectory. The stability of the stacking interaction between G32 and G43 suggests that G32 and G43 together with the flipped out U31 and a particular orientation of A30 could be the recognition motif for the 15.5K protein. The protein has a tight cavity (Figure 40) formed by residues 37 to 41, 44 and 95 to 99, which recognizes the RNA and selects a specific orientation of the two stems. I propose that the role of the G-A base pairs is to stabilize the orientation of the two stems at a precise angle which is selected during protein recognition.

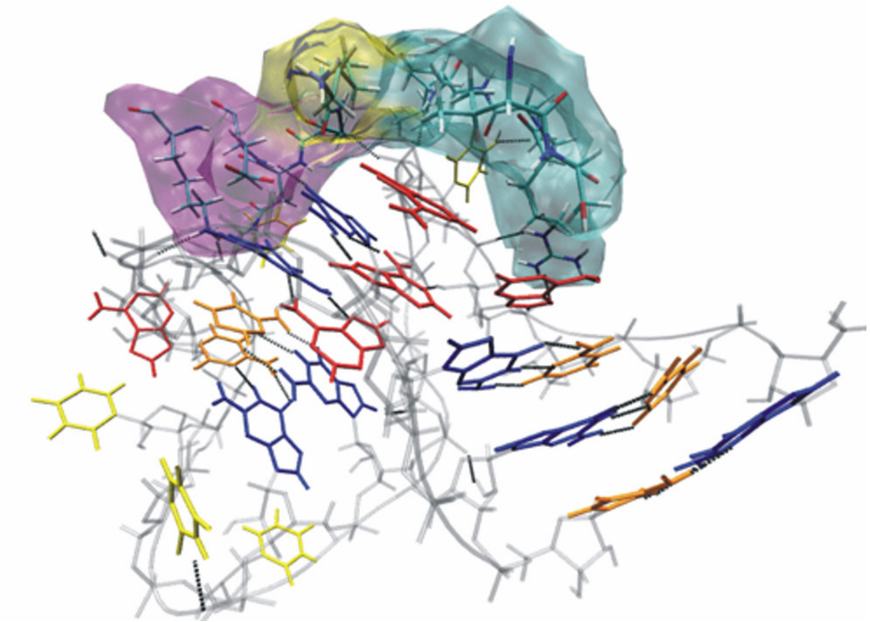


Figure 40: Protein cavity recognizing the RNA.
Coloring of the RNA is consistent with Figure 11. The protein cavity is shown as a transparent molecular surface colored by secondary structure (α-helix purple, β-sheet yellow and unstructured region cyan

The ribosomal K-turns vary in the angle between the stems and in the number and nature of non-canonical interactions, extending the NC-stem into the internal loop. Kt46 and Kt58 have the highest degree of kinking (~21° and ~41° between the helical axes of the two stems). They contain three G-A base pairs followed by a G-U base pair while K-turns with a lower degree of kinking bear two G-A base pairs, sometimes followed by a C-C base pair. This distribution suggests that there might be a connection between the orientation of the two stems and the number of G-A base pairs formed in the internal loop. The major function of the K-turns is to orient the two stems such as to permit binding of proteins that contact both stems or to provide the frame for large RNAs such as the ribosomal RNA to compact in the structures observed in the ribosomes. Therefore, it is plausible to suggest that the RNA folding in large structures requires specific orientations of helices that are regulated by non-Watson Crick interactions in flexible motifs such as the K-turns. It is very challenging to experimentally verify this hypothesis because adding or subtracting G-A base pairs from the RNA abolishes binding to their cognate proteins. Computer simulations are also limited due to the flexibility of the free RNAs which has to be accounted for to obtain an accurate representation of the simulated systems.

It was suggested by one of the reviewers of the paper that I published in Nucleic Acids Research (149) that the instability of the G-A base pairs in the free KtU4 was incompatible with thermodynamics experiments [(152) and refs therein] and may thus have been due to the limitations of LES and the force fields However, chemical RNA structure probing experiments have shown that both G32 and G43, as well as G34 and G35, are accessible to modification by kethoxal in the absence of 15.5K protein, indicating that the G-A base pairs and the G-C base pairs in the NC-stem are not formed in the unbound RNA (60). The stability of G-A base pairs during standard MD simulations of the free KtU4 might be due to the limited sampling accessible by standard MD. I believe that LES has proved useful in simulating the opening of G-A base pairs in the unbound KtU4. Nevertheless, I do not question the stability of sheared G-A base pairs in other RNAs or even in other K-turns. Additional studies are required to assess the stability of the G-A base pairs in the related ribosomal K-turns by considering both their bound proteins and attached RNA fragments.

6.3. The role of backbone flexibility in the folding of KtU4

If the loss of G-A base pairs was insufficient for achieving a large opening of the K-turn, it is likely that other factors contribute to the k-e transition. Capturing the LES1 conformation enabled me to characterize the motions in the backbone of nucleotides situated at the branching origin of the stems that are relevant for the opening of the K-turn. The rotational freedom about backbone bonds in A33, G34, G35 and G46 drives the k-e motion of the RNA. When this flexibility is restricted, the ability of the RNA to open is diminished (as observed in the LES1 and LES2 trajectories), suggesting that the formation of the K-turn is not a selective property of the internal loop but also involves the stems. A33 plays a special role in the transitions, being the most flexible nucleotide during the K-turn opening. Its pucker and χ angle are highly correlated with the global RNA motions, suggesting that the degree of conformational freedom in A33 is of crucial importance for the K-turn dynamics. I show that the flexibility of A33 is influenced by the degree of rotational freedom about the C4’-C5’ bond in G46. The 2’ OH group of G46 must adopt a correct orientation to form a hydrogen bond with the phosphate group of G34 that bridges the gap between the two stems at their branching origin. G45 does not show the same type of flexibility as G46 although it is closer to the internal loop, suggesting that the flexibility of G46 backbone directs the movement of the C-stem relative to the rest of the RNA. In the LES2 trajectory, the amplitude of the k-e motion is significantly lower than in the LES4 trajectory because conformational sampling in the C-stem was not enhanced. The backbone of G46 does not have a sufficient degree of conformational freedom to promote large RNA opening. Nevertheless, a certain degree of opening was observed in comparison to the LES1 trajectory, suggesting that flexibility in the backbone of G34 and G35 is equally important.

6.4. RNA folding and U4/U6 snRNP assembly

The 15.5K protein is required for the recruitment of other specific proteins to the U4/U6 snRNP such as the 61K or the 20/60/90K protein complex (37). I propose that the assembly of U4/U6 snRNP is driven by the hierarchical folding of the 5' stem-loop of U4 snRNA. First, the RNA is folded into a K-turn conformation upon binding of the 15.5K protein, while the pyrimidine-rich external loop remains flexible. The simulations of K2 RNA show that the external loop adopts different conformations in the presence and absence of the protein, suggesting that the 15.5K protein creates a pre-folded structure of the external loop. It is very likely that the complete folding of the UUUAU loop is achieved only after recruitment of 61K protein to the U4 snRNA. When replacing the UUUAU loop with the UGAA, the UUAAUU loops, or seven Watson-Crick base pairs, the backbone cannot adopt the orientation required for establishing hydrogen bonds with ASN40 and LYS44, suggesting that this orientation might be the first step in the folding of the external loop.

The naturally occurring RNA sequence (K2) is the only simulated construct in which C41 does not stack on the nucleotide N40 of the loop. Furthermore, electrostatic attractive forces between the protein helix α2 and the RNA backbone triggered structural changes in the external loops of the bound K3 and K4 RNAs: (i) loss of stacking between C41 and A40 in the UGAA loop (K3 RNA) and, (ii) instability of the U-U40 base pair in the UUAAUU loop (K4 RNA). These findings together with the experimental observation that the addition of one Watson-Crick G-C base pair to the NC-stem inhibits the binding of the 61K protein *in vitro* (data not shown), suggest that the lack of stacking interactions between C41 and N40 may be critical for the recruitment of 61K protein to the U4 snRNA.

When bound to the ribosomal protein L7AE, the backbone of the K-turn Kt15 shows a similar orientation with that observed in the trajectory of the bound K2 RNA. However, the UUUAU pentaloop from K2 RNA is replaced by a loop-like structure in Kt15. This structure has a distorted A-U base pair adjacent to the NC-stem, A not stacking on the nucleotide analogous to C41. Furthermore, in the trajectories of the free Kt15 this distorted base-pair becomes a canonical Watson-Crick base pair, indicating that the canonical base pair is destabilized upon binding of L7AE to Kt15. However, the loop-like structure in Kt15 is stabilized by the binding of the ribosomal protein L15E and by a Mg^{2+} ion. The cation might be needed because the stabilization of such a structure requires more stabilizing factors than does the flexible pentaloop in KtU4. The similarities observed between the orientation of the external loop backbone in K2 RNA and the backbone of the loop-like structure in Kt15 indicate a possible common folding path of KtU4 and Kt15, in which the L7AE/15.5K protein stabilizes a pre-orientation of the structure adjacent to the NC-stem such as to create a groove which is further stabilized by the L15E/61K protein. The stabilization of the distorted A-U base pair in Kt15 might indicate that the lack of stacking between the twisted A and the nucleotide analogue to C41 is important for the formation of the ternary complex L7AE-L15E-Kt15 such as the lack of stacking between U40 and C41 in KtU4 might be important for the formation of the ternary complex 15.5K-61K-KtU4.

During the non-LES simulations, the NC-stem proved less stable in the unbound K1 RNA than in the bound K1 RNA, the external loops (K2, K3, K4 RNAs) or subsequent Watson-Crick base pairs (K5 RNA) stabilizing the stem in the absence of the protein. The LES simulations have shown that protein binding plays a role in the stabilization of the G34-C42 and G35-C41 base pairs when the NC-stem is attached to flexible external loops. Furthermore, the chemical probing experiments have shown that the NC-stem is unstable in the unbound RNA. These findings suggest that a non-rigid external loop is required to keep the NC-stem relatively flexible in the unbound KtU4. Therefore, it is likely that the NC-stem folds together with the G-A base pairs and the external loop upon binding of the 15.5K protein.

6.5. Computer simulations of protein-assisted RNA folding

Structural data are rarely available for both the free and bound RNAs in cases of protein-assisted RNA folding. Computer simulation techniques provide the means for exploring the structure of the RNA in its unbound form and drawing conclusions about flexible regions contributing significantly to the folding process. For MD simulations, the starting structure is essential and dictates the evolution of the system. They are easily trapped in local minima on the potential energy surface. The time scale of conformational transitions important for the RNA folding is often much greater than that available for the simulations. MD simulations have been applied previously to U1A-RNA binding (75). In two recent studies that appeared almost in the same time as my article in the RNA journal, Rázga et al. reported 35-40 ns MD simulations of three different K-turns and derived conclusions about the dynamics of the K-turn around the conformation observed in the ribosomal crystal structure (147,148). Neither their standard MD simulations nor mine demonstrated the transition between the closed and open conformations of the K-turn, only partial opening being observed. Thus, I opted for LES to increase the conformational sampling during MD trajectories. A major advantage of this method is that it can be applied in combination with PME and explicit solvent (84). Nevertheless, smoothing the potential energy surface might permit transitions that do not occur in reality. Therefore, the choice of the LES regions and the number of copies is crucial.

Considering the benefits of applying LES despite the drawbacks apparent in the study of DNA G-quartets (86), I propose that in the absence of atomic resolution structures of unbound RNAs, such theoretical approaches are suitable for studying the dynamics of RNAs undergoing protein-assisted folding starting from structures of protein-RNA complexes. However, they do not provide structural predictions. Thus, the structures observed during the LES-MD trajectories should only be regarded as possible, perhaps even highly plausible, transition states during dynamic motions rather than assumed as unique 3D arrangements of the free KtU4. In reality, the unbound RNA may undergo transitions of much grater amplitudes that cannot be simulated even with more extensive conformational sampling.

For characterizing the k-e motion of the free RNA, in this dissertation I presented trajectories with LES systems containing four copies for four different regions. When the entire core structure of the K-turn RNA (K1 RNA) was replaced with three identical copies, the transition between the closed and open conformations was not observed, while with five copies the transition occurred for the unbound RNA on a shorter time scale but also for the bound RNA. The crystal structure of the complex was preserved during all viable simulations, its stability being the criterium for accepting trajectories for further processing. Using different combinations of LES regions I obtained different trajectories but observed the same local conformational transitions in the unbound RNA. Nevertheless, the transition from a closed to an open K-turn took place only when different LES regions were applied for the internal loop, the NC-stem and parts of the C-stem, suggesting that the kinked conformation is an intrinsic property of the entire RNA. LES proved very useful for studying large conformational transitions such as the opening of the K-turn, while the standard MD simulations provided a more accurate description of fine contacts such as the inter-stem hydrogen bonds

6.6. Perspectives for simulating protein-assisted RNA folding

The main limitation of LES is that it uses a rather artificial way to smoothen the potential energy barriers. The behavior of the RNA depends on the number of LES regions and the number of copies replacing each region. Nevertheless, using LES I obtained excellent agreement with experimental data in characterizing the behavior of KtU4. However, it was necessary to simulate both the free and bound RNAs under the same conditions to ensure that I was observing protein-assisted RNA folding and not random transitions in the RNA. This resulted in a sharp increase in CPU time required to obtain a viable trajectory. The next step in developing a protocol that can be extensively applied to study the mechanism of protein-assisted RNA folding would be to test whether the behavior of the K-turn motif can be reproduced by employing other enhanced sampling techniques such as replica exchange molecular dynamics simulations (153-155). Ideally, such a standardized protocol would not require the simulations of the bound RNAs and could predict induced fit or conformational capture mechanisms (19,53) only from the simulations of the free RNAs starting from their bound structures. Such a protocol is justified for studying systems for which structural data is very difficult to obtain due to their flexibility. The K-turn motif constitutes a typical example of such a system.

Understanding the dynamics of RNA motifs such as the K-turn is required to elucidate the mechanisms of large RNA folding with the ultimate goal of characterizing the folding of structures such as the ribosome, the single recognition particle, the telomerase or the dynamical events occurring during the assembly of the functional spliceosome.

6.7. Dynamics of 15.5K protein and U4/U6 snRNP assembly

I showed that upon binding of the 15.5K protein to KtU4, the RNA folds in a protein-assisted manner. The lack of structural information on the free protein raises the possibility that the fold of 15.5K protein is also stabilized upon binding to its cognate RNA. The simulations of the free protein and the complex revealed that only the protein-RNA interface is stabilized upon binding to its cognate RNA suggesting that the conformational changes in the protein upon RNA recognition are minor. However, networks of hydrogen bonds involving both protein and RNA residues were broken in the absence of the RNA even though this did not results in major secondary structure rearrangements in the protein. Nevertheless further experimental evidence is required to test whether conformational transitions on larger time scale do not occur.

The binding of the15.5K protein is required for the recruitment of the 61K protein or the 20/60/90 complex to the U4 snRNA. Thus, the dynamics of the 15.5K protein bound to KtU4 might play a role in regulating its interactions with other U4/U6 snRNP-specific proteins. Additionally, the 15.5K protein was the first protein found to be sheared by different RNPs such as U4/U6 snRNP and box C/D snoRNP. In different RNPs, the protein binds slightly different RNAs that are all proposed to be kink turns. I tested whether motions in the protein are dependent on the nature of the bound RNA. Surprisingly, quite far away from the protein-RNA interaction site, the flexibility of helix α5 depended on the nature of the bound RNA. In the trajectory of the complex with the natively occurring K2 RNA, the helix was very flexible while a very small degree of flexibility was observed in the trajectories of other complexes. α5 was also very flexible in the ribosomal protein L7AE when bound to its cognate Kt15 RNA, providing yet another common feature between the L7AE-Kt15 and 15.5K-KtU4 complexes. These findings suggest that a certain RNA type might regulate structural dynamics in the 15.5K protein to facilitate further interactions with other associated proteins. However, further simulations and/or experimental data are required to prove this supposition. In particular, structural data on the complex between the 15.5K protein and its binding sites within snoRNPs would be useful to study further the dynamics of the protein when bound to different RNAs.

Besides the protein RNA interface and the helix α5, a flexible region was identified in the 15.5K protein in the unstructured region between the sheet β2 and the helix α3 (R5). Based on the analogy with the binding of the L7AE and L15E proteins to Kt15, I propose that the R5 region of the 15.5K is contacted by the 61K upon binding to the U4 snRNA. The ribosomal L7AE protein exposes a hairpin-like structure formed around a proline residue with two adjacent glutamates to a docking site on the surface of the L15E protein (Figure 41A). Experimentally, it was shown that the 61K protein cross-links with both stems of KtU4 (37). Additionally, the hairpin-like structure is also present in the 15.5K protein (Figure 41B), suggesting that the 61K protein might recognize the 15.5K-KtU4 complex in a similar fashion as the L15E recognizes the L7AE-Kt15 complex. However, these findings provide only predictive information since experimental data on the 15.5K-61K protein complex is thus far unavailable and the 61K protein does not resemble the L15E protein.

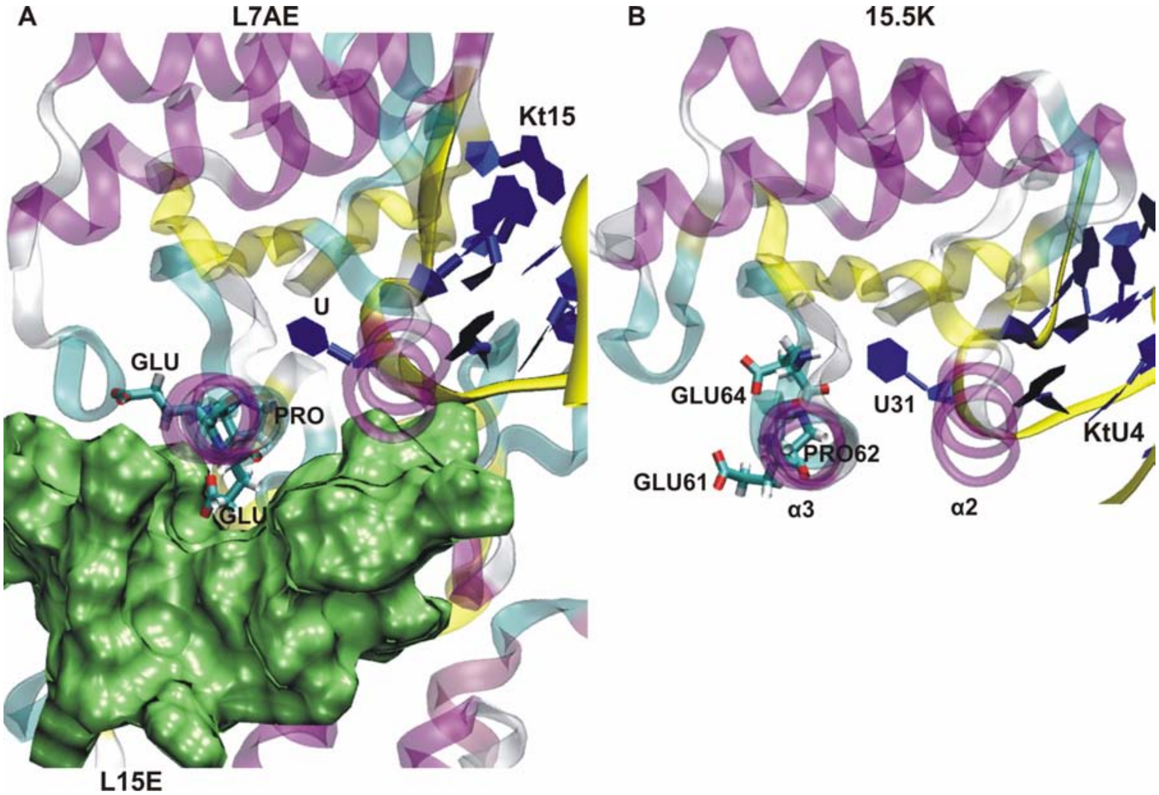


Figure 41: Proposed similarities between the L7AE-L15E-Kt15 and 15.5K-61K-KtU4 complexes. L7AE (A) and 15.5K (B) proteins are shown in cartoon representation (coloring as in Figure 6). The RNA bases are shown in blue and the RNA backbone in yellow. The L15E protein (A) is shown as a green molecular surface. The residues forming the hairpin-like structure (PRO62, GLU61, GLU64 in the 15.5K protein) are shown as bonds colored by atoms (carbons cyan, oxygens red, nytrogens blue and hydrogens white).

6.8. Structural changes of KtU4 monitored by 2AP fluorescence

The quenching of fluorescence upon protein binding both in the presence and in the absence of Mg^{2+} indicates that the cations are not required for the binding of 15.5K protein to KtU4 (result in agreement with available experimental data). The quenching of fluorescence in the presence of 15.5K protein may be attributed to the interaction between A29 and the protein residue ARG 97. However, it may also reflect a conformational transition in the RNA upon protein binding that leads to different stacking interactions of 2AP, consistent with the simulations and the chemical RNA structure probing suggesting a protein-assisted RNA folding mechanism. Both effects may be operative.

The temperature dependent quenching of 2AP fluorescence observed in the melting profiles up to 30℃ was in agreement with previous studies of 2AP-labeled nucleic acids (156). The sharper decrease observed between 30 to 40℃ in the absence of Mg^{2+} (or between 30 and 45°C in the presence of Mg^{2+}) may possibly reflect a conformational transition that involves an increase in the number of stacking interactions between A29 and adjacent bases, interpretation in agreement with the simulations showing the transition between the “3+1” stacking scheme (A33-A44-A30 + A29) to the “2+2” stacking scheme (A33-A44 + A30-A29) in the unbound RNA. The gradual increase in fluorescence at temperatures above 50°C in the absence of Mg^{2+} suggests that the 2AP stacking on the C-stem is lost upon melting of the latter. In the presence of Mg^{2+}, the melting profile was similar. However, the difference between the predicted T_m in the two buffer systems suggests that the RNA is more structured in the presence of Mg^{2+}. This effect is also suggested by the titration of the RNA with Mg^{2+}at 5°C that showed a sharp decrease in fluorescence upon addition of up to 3 mM Mg^{2+}.

These preliminary steady-state fluorescence measurements indicate that 2AP may be a good reporter for structural changes in the K-turn motif. Studies of the 2AP fluorescence lifetimes, and the kinetics of thermal unfolding and of the binding to the 15.5K protein will be required to obtain a more detailed view of the transitions occurring in the K-turn motif and compare them with those observed in the simulations.

7. Summary and conclusions

Throughout the dissertation I provided evidence that: (i) the 5'-stem loop formed by U4 snRNA is folded only upon protein binding, (ii) the free RNA undergoes a dramatic conformational transition between a kinked and an extended conformation, (iii) the loss of G-A base pairs in the internal loop is not sufficient for achieving a large opening of the RNA, (iv) local backbone motions contribute significantly to the opening of the K-turn, (v) the backbone of the external UUUAU pentaloop adopts an orientation that is specific for the naturally occurring K2 RNA and is sheared between KtU4 and Kt15, (vi) hydrogen bonding networks between the protein residues at the interface are instable in the absence of the bound RNA without significantly altering the overall protein fold, (vii) the helix α5 is flexible only when the 15.5K protein is bound to the natively occurring K2 RNA, and (viii) the flexible region R5 is sheared by the 15.5K protein and the ribosomal L7AE protein. L15E binds to L7AE by recognizing the R5 region, thus the analogue region in the 15.5K protein might constitute the binding site for the 61K protein.

Based on these findings, I propose the following model incorporating the motions occurring at the 5' stem-loop of U4 snRNA (Figure 42): (i) the free KtU4 samples different conformations during a continuous inter-conversion between a partially kinked structure and an extended conformation, the structures of which remain unknown, (ii) upon binding to the 15.5K protein, a certain orientation between the two stems of the RNA is selected and the flexible external loop is oriented to form a groove in which protein residues can establish contacts with the RNA, (iii) formation of G-A and G-C base pairs in the NC-stem stabilizes the pre-selected orientation between the two stems, (iv) the binding of 61K to the U4 snRNA completes the folding of KtU4 by stabilizing the groove formed by the backbone of the external loop. Although experimental evidence is still scarce for the last step, the simulations and the analogy with the binding of L7AE and L15E to the Kt15 RNA show that this might be a possible path of the early U4/U6 snRNP assembly.

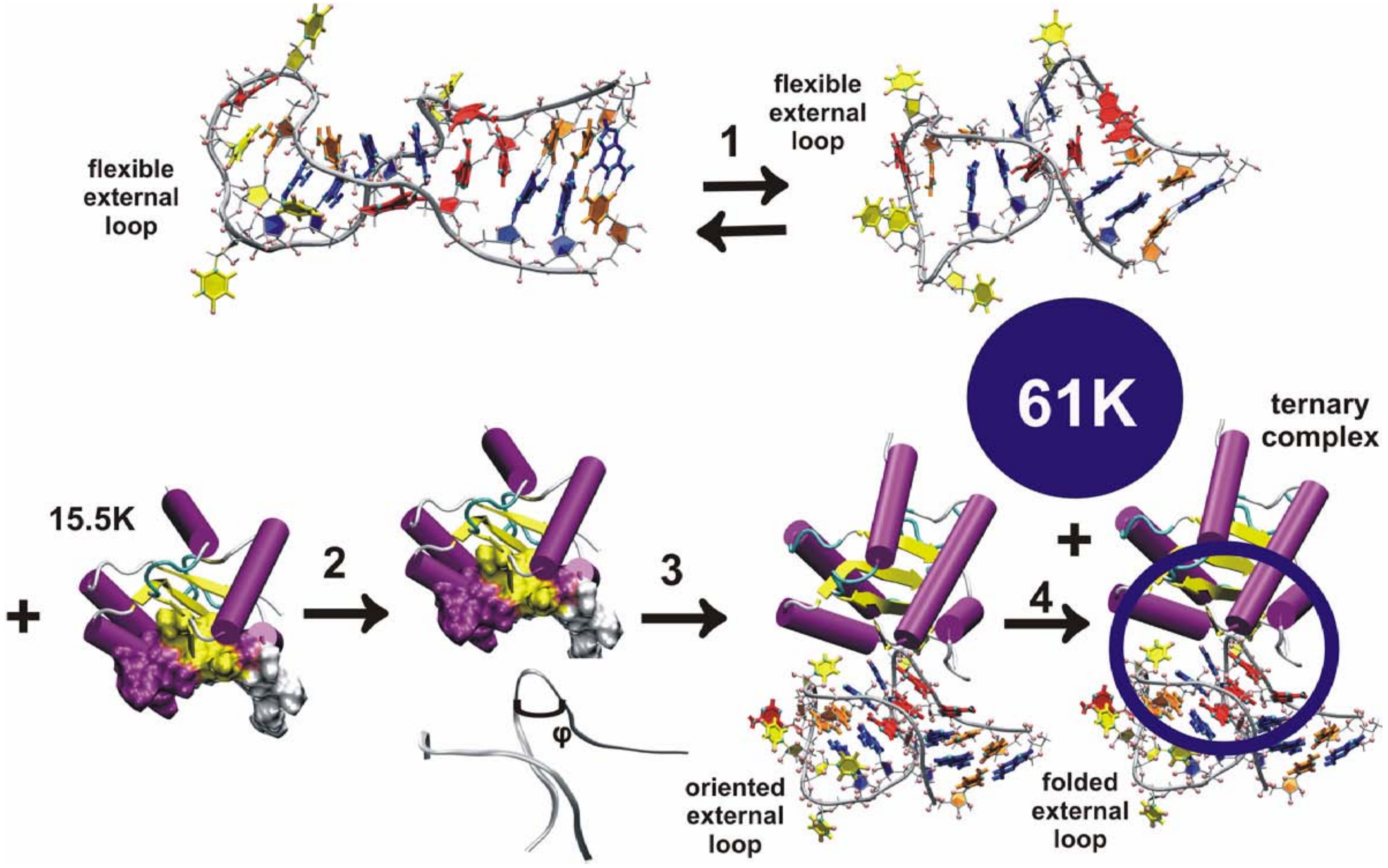


Figure 42: Model of dynamic events occurring at the 5’ stem-loop of U4 snRNA during U4/U6 snRNP assembly.

The RNA is flexible before binding to the 15.5K protein, undergoing transitions between the kinked and extended structures (step 1). G-A base pair formation does not occur in neither of the two states sampled by the free RNA. The external loop and the NC-stem form one flexible entity. Upon binding to the 15.5K protein a certain orientation of the two stems is selected (step 2). This orientation is stabilized by the formation of the tandem-sheared G-A base pairs. The G-C base pairs of the NC-stem form also upon binding to the 15.5K protein. The external loop adopts a specific orientation , forming a groove in which ASN40 and LYS44 contact the RNA, but it remains flexible. (step 3). The 61K binds to the 15.5K-KtU4 complex by contacting both stems of the RNA and the hairpin-like structure of the 15.5K protein in a similar way as the L15E recognizes the L7AE-KtU4 complex. The external loop is folded upon binding of 61K protein.

8. References

1. Burge, C.B., Tuschl, T.H., and Sharp, P.A. 1999. Splicing of precursors to mRNAs by the spliceosome. In *The RNA World* (eds. R. F. Gesteland et al.), pp. 525-560. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
2. Nilsen, T.W. 1999. RNA/RNA interactions in nuclear pre-mRNA splicing. In *RNA Structure and Function* (eds. R. F. Gesteland et al.), pp. 297-307. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
3. Staley, J.P., and Guthrie, C. 1998. Mechanical devices of the spliceosome: Motors, clocks, springs, and things. *Cell* **92**:315-326.
4. Will, C.L., and Lührmann, R. 2001. Spliceosomal UsnRNP biogenesis, structure and function. *Curr. Opin. Cell Biol.* **13**:290-301.
5. Collins, C.A., and Guthrie, C. 2000. The question remains: Is the spliceosome a ribozyme? *Nat. Struct. Biol.* **7**:850-854.
6. Nilsen, T.W. 2000. The case for an RNA enzyme. *Nature* **408**:782-783.
7. Michel, F., and Ferat, J.L. 1995. Structure and activities of group-II introns. *Annu. Rev. Biochem.* **64**:435-461.
8. Padgett, R.A., Podar, M., Boulanger, S.C., and Perlman, P.S. 1994. The stereochemical course of group-II intron self-splicing. *Science* **266**:1685-1688.
9. Pyle, A.M. 1996. Catalytic reaction mechanisms and structural features of group II intron ribozymes. In *Catalytic RNA* (eds. F. Eckstein and D. M. J. Lilley), pp. 75-107. Springer, Berlin, New York.
10. Weiner, A.M. 1993. Messenger-rna splicing and autocatalytic introns - distant cousins or the products of chemical determinism. *Cell* **72**:161-164.
11. Teigelkamp, S., Newman, A.J., and Beggs, J.D. 1995. Extensive interactions of Prp8 protein with the 5' and 3' splice sites during splicing suggest a role in stabilization of exon alignment by U5 snRNA. *EMBO J.* **14**:2602-2612.
12. Dix, I., Russell, C.S., O'Keefe, R.T., Newman, A.J., and Beggs, J.D. 1998. Protein-RNA interactions in the U5 snRNP of *Saccharomyces cerevisiae*. *RNA* **4**:1239-1250.
13. Vidal, V.P.I., Verdone, L., Mayes, A.E., and Beggs, J.D. 1999. Characterization of U6 snRNA-protein interactions. *RNA* **5**:1470-1481.
14. Beggs, J.D., Teigelkamp, S., and Newman, A.J. 1995. The role of PRP8 protein in nuclear pre-mRNA splicing in yeast. *J. Cell Sci. Suppl.* **19**:101-105.

15. Newman, A.J. 1997. The role of U5 snRNP in pre-mRNA splicing. *EMBO J.* **16**:5797-5800.

16. Weeks, K.M. 1997. Protein-facilitated RNA folding. *Curr. Opin. Struct. Biol.* **7**:336-342.

17. Weeks, K.M., and Cech, T.R. 1995. Protein facilitation of group-I intron splicing by assembly of the catalytic core and the 5'-splice-site domain. *Cell* **82**:221-230.

18. Frankel, A.D., and Smith, C.A. 1998. Induced folding in RNA-protein recognition: More than a simple molecular handshake. *Cell* **92**:149-151.

19. Williamson, J.R. 2000. Induced fit in RNA-protein recognition. *Nat. Struct. Biol.* **7**:834-837.

20. Bringmann, P., Appel, B., Rinke, J., Reuter, R., Theissen, H., and Lührmann, R. 1984. Evidence for the existence of snRNAs U4 and U6 in a single ribonucleoprotein complex and for their association by intermolecular base-pairing. *EMBO J.* **3**:1357-1363.

21. Brow, D.A., and Guthrie, C. 1988. Spliceosomal RNA U6 is remarkably conserved from yeast to mammals. *Nature* **334**:213-218.

22. Hashimoto, C., and Steitz, J.A. 1984. U4 and U6 RNAs coexist in a single small nuclear ribonucleoprotein particle. *Nucleic Acids Res.* **12**:3283-3293.

23. Rinke, J., Appel, B., Digweed, M., and Lührmann, R. 1985. Localization of a base-paired interaction between small nuclear RNAs U4 and U6 in intact U4 U6 ribonucleoprotein-particles by psoralen cross-linking. *J. Mol. Biol.* **185**:721-731.

24. Tarn, W.Y., and Steitz, J.A. 1997. Pre-mRNA splicing: The discovery of a new spliceosome doubles the challenge. *Trends Biochem. Sci.* **22**:132-137.

25. Raghunathan, P.L., and Guthrie, C. 1998. RNA unwinding in U4/U6 snRNPs requires ATP hydrolysis and the DEIH-box splicing factor Brr2. *Curr. Biol.* **8**:847-855.

26. Laggerbauer, B., Achsel, T., and Lührmann, R. 1998. The human U5-2OOkD DEXH-box protein unwinds U4/U6 RNA duplices *in vitro*. *PNAS* **95**:4188-4192.

27. Horowitz, D.S., Kobayashi, R., and Krainer, A.R. 1997. A new cyclophilin and the human homologues of yeast Prp3 and Prp4 form a complex associated with U4/U6 snRNPs. *RNA* **3**:1374-1387.

28. Teigelkamp, S., Achsel, T., Mundt, C., Gothel, S.F., Cronshagen, U., Lane, W.S., Marahiel, M., and Lührmann, R. 1998. The 20kD protein of human [U4/U6·U5] tri-

snRNPs is a novel cyclophilin that forms a complex with the U4/U6-specific 60kD and 90kD proteins. *RNA* **4**:127-141.

29. Nottrott, S., Hartmuth, K., Fabrizio, P., Urlaub, H., Vidovic, I., Ficner, R., and Lührmann, R. 1999. Functional interaction of a novel 15.5kD [U4/U6·U5] tri-snRNP protein with the 5' stem-loop of U4 snRNA. *EMBO J.* **18**:6119-6133.
30. Watkins, N.J., Segault, V., Charpentier, B., Nottrott, S., Fabrizio, P., Bachi, A., Wilm, M., Rosbash, M., Branlant, C., and Lührmann, R. 2000. A common core RNP structure shared between the small nucleoar box C/D RNPs and the spliceosomal U4 snRNP. *Cell* **103**:457-466.
31. Makarova, O.V., Makarov, E.M., Liu, S.B., Vornlocher, H.P., and Lührmann, R. 2002. Protein 61K, encoded by a gene (PRPF31) linked to autosomal dominant retinitis pigmentosa, is required for U4/U6·U5 tri-snRNP formation and pre-mRNA splicing. *EMBO J.* **21**:1148-1157.
32. Gautier, T., Berges, T., Tollervey, D,. and Hurt, E. 1997. Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis. *Mol. Cell. Biol.* **17**:7088-7098.
33. Schneider, C., Will, C.L., Makarova, O.V., Makarov, E.M., and Lührmann, R. 2002. Human U4/U6·U5 and U4atac/U6atac·U5 tri-snRNPs exhibit similar protein compositions. *Mol.Cell. Biol.* **22**:3219-3229.
34. Banroques, J., and Abelson, J.N. 1989. Prp4 - a protein of the yeast U4/U6 small nuclear ribonucleoprotein particle. *Mol.Cell. Biol.* **9**:3710-3719.
35. Stevens, S.W., and Abelson, J. 1999. Purification of the yeast U4/U6·U5 small nuclear ribonucleoprotein particle and identification of its proteins. *PNAS* **96**:7226-7231.
36. Weidenhammer, E.M., Ruiz-Noriega, M., and Woolford, J.L. 1997. Prp31p promotes the association of the U4/U6·U5 tri-snRNP with pre-spliceosomes to form spliceosomes in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 17:3580-3588.
37. Nottrott, S., Urlaub, H., and Lührmann, R. 2002. Hierarchical, clustered protein interact ions with U4/U6 snRNA: a biochemical role for U4/U6 proteins. *EMBO J.* **21**:5527-5538.
38. Frilander, M.J., and Steitz, J.A. 2001. Dynamic exchanges of RNA interactions leading to catalytic core formation in the U12-dependent spliceosome. *Mol. Cell* **7**:217-226.

39. Vankan, P., McGuigan, C., and Mattaj, I.W. 1992. Roles of U4 and U6 snRNAs in the assembly of splicing complexes. *EMBO J.* **11**:335-343.

40. Wersig, C., and Bindereif, A. 1992. Reconstitution of functional mammalian U4 small nuclear ribonucleoprotein - Sm protein-binding is not essential for splicing *in vitro*. *Mol.Cell. Biol.* **12**:1460-1468.

41. Filipowicz, W., Pelczar, P., Pogacic, V., and Dragon, F. 1999. Structure and biogenesis of small nucleolar RNAs acting as guides for ribosomal RNA modification. *Acta Biochim. Pol.* **46**:377-389.

42. Venema, J., and Tollervey, D. 1999. Ribosome synthesis in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.* **33**:261-311.

43. Weinstein, L.B., and Steitz, J.A. 1999. Guided tours: from precursor snoRNA to functional snoRNP. *Curr. Opin. Cell Biol.* **11**:378-384.

44. Vidovic, I., Nottrott, S., Hartmuth, K., Lührmann, R., and Ficner, R. 2000. Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. *Mol. Cell* **6**:1331-1342.

45. Heus, H.A., Wijmenga, S.S., Hoppe, H., and Hilbers, C.W. 1997. The detailed structure of tandem G-A mismatched base-pair motifs in RNA duplexes is context dependent. *J. Mol. Biol.* **271**:147-158.

46. Santalucia, J., and Turner, D.H. 1993. Structure of (rGGCGAGCC)2 in solution from NMR and restrained molecular dynamics. *Biochemistry* **32**:12612-12623.

47. Klein, D.J., Schmeing, T.M., Moore, P.B., and Steitz, T.A. 2001. The kink-turn: a new RNA secondary structure motif. *EMBO J.* **20**:4214-4221.

48. Charron, C., Manival, X., Clery, A., Charpentier, B., Marmier-Gourrier, N., Branlant, C., and Aubry, A. 2004. The archaeal snoRNA binding protein L7Ae has a 3D structure very similar to that of its eukaryal counterpart while having a broader RNA-binding specificity. *J. Mol. Biol.* **342**:757-773.

49. Hamma, T., and Ferre-D'Amare, A.R. 2004. Structure of protein L7Ae bound to a K-turn derived from an archaeal box H/ACA snoRNA at 1.8 Å resolution. *Structure* **12**:893-903.

50. Moore, T., Zhang, Y.M., Fenley, M.O., and Li, H. 2004. Molecular basis of box C/D RNA-protein interactions: Cocrystal structure of archaeal L7Ae and a box C/D RNA. *Structure* **12**:807-818.

51. Chao, J.A., and Williamson, J.R. 2004. Joint x-ray and NMR refinement of the yeast L30e-mRNA complex. *Structure* **12**:1165-1176.

52. Leontis, N.B., and Westhof, E. 2003. Analysis of RNA motifs. *Curr. Opin. Struct. Biol.* **13**:300-308.

53. Leulliot, N., and Varani, G. 2001. Current topics in RNA-protein recognition: Control of specificity and biological function through induced fit and conformational capture. *Biochemistry* **40**:7947-7956.

54. Avis, J.M., Allain, F.H.T., Howe, P.W.A., Varani, G., Nagai, K., and Neuhaus, D. 1996. Solution structure of the N-terminal RNP domain of U1A protein: The role of C-terminal residues in structure stability and RNA binding. *J. Mol. Biol.* **257**:398-411.

55. Gubser, C.C., and Varani, G. 1996. Structure of the polyadenylation regulatory element of the human U1A pre-mRNA 3'-untranslated region and interaction with the U1A protein. *Biochemistry* **35**:2253-2267.

56. Oubridge, C., Ito, N., Evans, P.R., Teo, C.H., and Nagai, K. 1994. Crystal structure at 1.92 Å resolution of the RNA binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature* **372**:432-438.

57. Varani, L., Gunderson, S.I., Mattaj, I.W., Kay, L.E., Neuhaus, D., and Varani, G. 2000. The NMR structure of the 38 kDa U1A protein - PIE RNA complex reveals the basis of cooperativity in regulation of polyadenylation by human U1A protein. *Nat. Struct. Biol.* **7**:329-335.

58. Goody, T.A., Melcher, S.E., Norman, D.G., and Lilley, D.M.J. 2004. The kink-turn motif in RNA is dimorphic, and metal ion-dependent. *RNA* **10**:254-264.

59. Matsumura, S., Ikawa, Y., and Inoue, T. 2003. Biochemical characterization of the kink-turn RNA motif. *Nucleic Acids Res.* **31**:5544-5551.

60. Cojocaru, V., Nottrott, S., Klement, R., and Jovin, T.M. 2005. The snRNP 15.5K protein folds its cognate K-turn RNA: A combined theoretical and biochemical study. *RNA* **11**:197-209.

61. Wang, W., Donini, O., Reyes, C.M., and Kollman, P.A. 2001. Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu. Rev. Bioph. Biom.* **30**:211-243.

62. Cheatham, T.E., and Kollman, P.A. 2000. Molecular dynamics simulation of nucleic acids. *Annu. Rev. Phys. Chem.* **51**:435-471.

63. Cheatham, T.E., and Brooks, B.R. 1998. Recent advances in molecular dynamics simulation towards the realistic representation of biomolecules in solution. *Theor. Chem. Acc.* **99**:279-288.
64. Cheatham, T.E., Miller, J.L., Fox, T., Darden, T.A., and Kollman, P.A. 1995. Molecular dynamics simulations on solvated biomolecular systems - the Particle Mesh Ewald method leads to stable trajectories of DNA, RNA, and proteins. *J. Am. Chem. Soc.* **117**:4193-4194.
65. Noy, A., Perez, A., Lankas, F., Luque, F.J., and Orozco, M. 2004 Relative flexibility of DNA and RNA: a molecular dynamics study. *J. Mol. Biol.* **343**:627-638.
66. Réblová, K, Špačková, N, Šponer, J.E., Koča, J, and Šponer, J. 2003. Molecular dynamics simulations of RNA kissing-loop motifs reveal structural dynamics and formation of cation-binding pockets. *Nucleic Acids Res.* **31**:6942-6952.
67. Réblová, K, Špačková, N, Štefl, R., Csaszar, K, Koča, J, Leontis, N.B., and Šponer, J. 2003. Non-Watson-Crick basepairing and hydration in RNA motifs: Molecular dynamics of 5S rRNA loop E. *Biophys. J.* **84**:3564-3582.
68. Li, W., Ma, B.Y., and Shapiro, B.A. 2001. Molecular dynamics simulations of the denaturation and refolding of an RNA tetraloop. *J. Biomol. Struct. Dyn.* **19**:381-396.
69. Guo, J.X., Daizadeh, I., and Gmeiner, W.H. 2000. Structure of the Sm binding site from human U4 snRNA derived from a 3 ns PME molecular dynamics simulation. *J. Biomol. Struct. Dyn.* **18**:335-344.
70. Auffinger, P., Bielecki, L., and Westhof, E. 2004. Anion binding to nucleic acids. *Structure* **12**:379-388.
71. Auffinger, P., Bielecki, L., and Westhof, E. 2004. Symmetric K^+ and Mg^{2+} ion-binding sites in the 5 S rRNA loop E inferred from molecular dynamics simulations. *J. Mol. Biol.* **335**:555-571.
72. Auffinger, P., Bielecki, L., and Westhof, E. 2003. The Mg^{2+} binding sites of the 5S rRNA loop E motif as investigated by molecular dynamics simulations. *Chem. Biol.* **10**:551-561.
73. Auffinger, P., and Westhof, E. 2000. Water and ion binding around RNA and DNA (C,G) oligomers. *J. Mol. Biol.* **300**:1113-1131.

74. Réblová, K, Špačková, N, Štefl, R., Csaszar, K, Koča, J, Leontis. N.B., and Šponer, J. 2004. Long-residency hydration, cation binding, and dynamics of loop E/helix IV rRNA-L25 protein complex. *Biophys. J.* **87**:3397-3412.
75. Pitici, F., Beveridge, D.L., and Baranger, A.M. 2002. Molecular dynamics simulation studies of induced fit and conformational capture in U1A-RNA binding: Do molecular substrates code for specificity? *Biopolymers* **65**:424-435.
76. Reyes, C.M., Nifosi, R., Frankel, A.D., and Kollman, P.A. 2001. Molecular dynamics and binding specificity analysis of the bovine immunodeficiency virus BIV Tat-TAR complex. *Biophys. J.* **80**:2833-2842.
77. Blakaj, D.M., McConnell, K.J., Beveridge, D.L., and Baranger, A.M. 2001. Molecular dynamics and thermodynamics of protein-RNA interactions: Mutation of a conserved aromatic residue modifies stacking interactions and structural adaptation in the U1A-stem loop 2 RNA complex. *J. Am. Chem. Soc.* **123**:2548-2551.
78. Reyes, C.M., and Kollman, P.A. 2000. Structure and thermodynamics of RNA-protein binding: Using molecular dynamics and free energy analyses to calculate the free energies of binding and conformational change. *J. Mol. Biol.* **297**:1145-1158.
79. Elber, R., and Karplus, M. 1990. Enhanced sampling in molecular dynamics - Use of the time-dependent Hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *J. Am. Chem. Soc.* **112**:9161-9175.
80. Simmerling, C., and Elber, R. 1994. Hydrophobic collapse in a cyclic hexapeptide - computer simulations of CHDLFC and CAAAAC in water. *J. Am. Chem. Soc.* **116**:2534-2547.
81. Simmerling, C., Fox, T., and Kollman, P.A. 1998. Use of locally enhanced sampling in free energy calculations: Testing and application to the α→β anomerization of glucose. *J. Am. Chem. Soc.* **120**:5771-5782.
82. Simmerling, C., and Kollman, P. 1996. Improved simulation of flexible systems through locally enhanced sampling. *Abstracts of Papers of the American Chemical Society* **212**:153-COMP.
83. Simmerling, C., Lee, M.R., Ortiz, A.R., Kolinski, A., Skolnick, J., and Kollman, P.A. 2000. Combining MONSSTER and LES/PME to predict protein structure from amino acid sequence: Application to the small protein CMTI-1. *J. Am. Chem. Soc.* **122**:8392-8402.

84. Simmerling, C., Miller, J.L., and Kollman, P.A. 1998. Combined locally enhanced sampling and Particle Mesh Ewald as a strategy to locate the experimental structure of a nonhelical nucleic acid. *J. Am. Chem. Soc.* **120**:7149-7155.
85. Stultz, C.M., and Karplus, M. 1998. On the potential surface of the locally enhanced sampling approximation. *J. Chem. Phys.* **109**:8809-8815.
86. Fadrná, E., Špačková, N., Štefl, R., Koča, J., Cheatham, T.E.3rd, and Šponer, J. 2004. Molecular dynamics simulations of guanine quadruplex loops: advances and force field limitations. *Biophys. J.* **87**:227-242.
87. Saenger, W. 1984. *Principles of Nucleic Acids Structure*. Springer-Verlag New York Inc.
88. Ward, D.C., Reich, E. and Stryer, L. 1969. Fluorescence studies of nucleotides and polynucleotides .I. formycin 2-aminopurine riboside 2,6-diaminopurine riboside and their derivatives. *J. Biol. Chem.* **244**:1228.
89. Shchyolkina, A.K., Kaluzhny, D.N., Borisova, O.F., Hawkins, M.E., Jernigan, R.L., Jovin, T.M., Arndt-Jovin, D.J. and Zhurkin, V.B. 2004. Formation of an intramolecular triple-stranded DNA structure monitored by fluorescence of 2-aminopurine or 6-methylisoxanthopterin. *Nucleic Acids Res.* **32**:432-440.
90. Jiao, Y.G., Stringfellow, S. and Yu, H.T. 2002. Distinguishing "looped-out" and "stacked-in" DNA bulge conformation using fluorescent 2-aminopurine replacing a purine base. *J. Biomol. Struct. Dyn.* **19**:929-934.
91. Nordlund, T.M., Andersson, S., Nilsson, L., Rigler, R., Graslund, A. and McLaughlin, L.W. 1989. Structure and dynamics of a fluorescent DNA oligomer containing the Ecori recognition sequence - fluorescence, molecular-dynamics, and NMR-studies. *Biochemistry* **28**:9095-9103.
92. Nordlund, T.M., Xu, D.G. and Evans, K.O. 1993. Excitation-energy transfer in DNA - duplex melting and transfer from normal bases to 2-aminopurine. *Biochemistry* **32**:12090-12095.
93. Rachofsky, E.L., Seibert, E., Stivers, J.T., Osman, R. and Ross, J.B.A. 2001. Conformation and dynamics of abasic sites in DNA investigated by time-resolved fluorescence of 2-aminopurine. *Biochemistry* **40**:957-967.
94. Stivers, J.T. 1998. 2-aminopurine fluorescence studies of base stacking interactions at abasic sites in DNA: metal-ion and base sequence effects. *Nucleic Acids Res.* **26**:3837-3844.

95. Guest, C.R., Hochstrasser, R.A., Sowers, L.C. and Millar, D.P. 1991. Dynamics of mismatched base-pairs in DNA. *Biochemistry* **30**:3271-3279.

96. Menger, M., Tuschl, T., Eckstein, F. and Porschke, D. 1996. Mg^{2+}-dependent conformational changes in the hammerhead ribozyme. *Biochemistry* **35**:14710-14716.

97. Allan, B.W., Reich, N.O. and Beechem, J.M. 1999. Measurement of the absolute temporal coupling between DNA binding and base flipping. *Biochemistry* **38**:5308-5314.

98. Baliga, R., Baird, E.E., Herman, D.M., Melander, C., Dervan, P.B. and Crothers, D.M. 2001. Kinetic consequences of covalent linkage of DNA binding polyamides. *Biochemistry* **40**:3-8.

99. Bandwar, R.P. and Patel, S.S. 2001. Peculiar 2-aminopurine fluorescence monitors the dynamics of open complex formation by bacteriophage T7 RNA polymerase. *J. Biol. Chem.* **276**:14075-14082.

100. Beechem, J.M., Otto, M.R., Bloom, L.B., Eritja, R., Reha-Krantz, L.J. and Goodman, M.F. 1998. Exonuclease-polymerase active site partitioning of primer-template DNA strands and equilibrium Mg^{2+} binding properties of bacteriophage T4 DNA polymerase. *Biochemistry* **37**:10144-10155.

101. Holz, B., Klimasauskas, S., Serva, S. and Weinhold, E. 1998. 2-Aminopurine as a fluorescent probe for DNA base flipping by methyltransferases. *Nucleic Acids Res.* **26**:1076-1083.

102. Jia, Y.P., Kumar, A. and Patel, S.S. 1996. Equilibrium and stopped-flow kinetic studies of interaction between T7 RNA polymerase and its promoters measured by protein and 2-aminopurine fluorescence changes. *J. Biol. Chem.* **271**:30451-30458.

103. Raney, K.D., Sowers, L.C., Millar, D.P. and Benkovic, S.J. 1994. A fluorescence-based assay for monitoring helicase activity. *PNAS* **91**:6644-6648.

104. Ujvari, A. and Martin, C.T. 1996. Thermodynamic and kinetic measurements of promoter binding by T7 RNA polymerase. *Biochemistry* **35**:14574-14582.

105. Hochstrasser, R.A., Carver, T.E., Sowers, L.C. and Millar, D.P. 1994. Melting of a DNA helix terminus within the active site of a DNA polymerase. *Biochemistry* **33**:11971-11979.

106. Jean, J.M. and Hall, K.B. 2001. Effect of base stacking on the excited state structure and fluorescence dynamics of 2-aminopurine. *Biophys. J.* **80**:8A-8A.

107. Jean, J.M. and Hall, K.B. 2001. 2-Aminopurine fluorescence quenching and lifetimes: role of base stacking. *PNAS* **98**:37-41.

108. Jean, J.M. and Hall, K.B. 2002. 2-Aminopurine electronic structure and fluorescence properties in DNA. *Biochemistry* **41**:13152-13161.

109. Jean, J.M., Showalter, S.A. and Hall, K.B. 2002. Electronic structure and dynamics of 2-aminopurine in DNA. *Biophys. J.* **82**:355A-355A.

110. Menger, M., Eckstein, F. and Porschke, D. 2000. Dynamics of the RNA hairpin GNRA tetraloop. *Biochemistry* **39**:4500-4507.

111. Schlick, T. 2002. *Molecular modeling and simulation: an interdisciplinary guide*. Springer-Verlag New York, Inc.

112. Leach, A.R. 2001 *Molecular modelling: principles and applications - second edition*. Pearson Education Limited, Harlow.

113. Cheatham, T.E., Cieplak, P., and Kollman, P.A. 1999. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* **16**:845-862.

114. Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. (1996) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**:5179-5197.

115. Weiner, S.J., Kollman, P.A., Nguyen, D.T., and Case, D.A. 1986. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.* **7**:230-252.

116. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. 1984. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**:765-784.

117. Hopfinger, A.J., and Pearlstein, R.A. 1984. Molecular mechanics force field parameterization procedures. *J. Comput. Chem.* **5**:486-499.

118. Halgren, T.A. 1990. Maximally diagonal force constants in dependent angle bending coordinates: Implications for the design of empirical force fields. *J. Am. Chem. Soc.* **112**:4710-4723.

119. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**:926-935.

120. Singh, U.C., and Kollman, P.A. 1984. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* **5**:129-145.

121. Reynolds, C.A., Essex, J.W., and Richards, W.G. 1992. Atomic charges for variable molecular conformations. *J. Am. Chem. Soc.* **114**:9075-9079.

122. Berendsen, H.J.C., Grigera, J.R., and Straatsma, T.P. 1987 The missing term in effective pair potentials. *J. Phys. Chem.* **91**:6269-6271.

123. Bayly, C.I., Cieplak, P., Cornell, W.D., and Kollman, P.A. 1993. A well behaved electrostatic potential based method using charge restraints for deriving atomic charges - the RESP model. *J. Phys. Chem.* **97**:10269-10280.

124. Cornell, W.D., Cieplak, P., Bayly, C.I., and Kollman, P.A. 1993. Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *J. Am. Chem. Soc.* **115**:9620-9631.

125. Cieplak, P., Cornell, W.D., Bayly, C., and Kollman, P.A. 1995 Application of the multimolecule and multiconformational RESP methodology to biopolymers - charge derivation for DNA, RNA, and proteins. *J. Comput. Chem.* **16**:1357-1377.

126. Verlet, L. (1967) Computer experiments on classical fluids: thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **159**:98-103.

127. Ryckaert, J.P., Ciccotti, G., and Berendsen, H.J.C. 1977. Numerical integration of cartesian equations of motion of a system with constraints - molecular dynamics of N-alkanes. *J. Comput. Phys.* **23**:327-341.

128. Miyamoto, S., and Kollman, P.A. 1992. SETTLE - an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13**:952-962.

129. Darden, T., and Pedersen, L. 1996. Use of fast Ewald summation in molecular dynamics simulations. *Abstracts Of Papers Of The American Chemical Society* **212**:6-COMP.

130. Darden, T., Perera, L., Li, L.P., and Pedersen, L. 1999. New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Struct. Fold. Des.* **7**:55-60.

131. Eastwood, J.W., and Hockney, R.W. 1974. Shaping force law in 2-dimensional particle mesh models. *J. Comp. Physiol.* **16**:342-359.

132. Berendsen, H.J.C., Postma, J.P.M., Vangunsteren, W.F., Dinola, A., and Haak, J.R. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**:3684-3690.

133. Macke, T., and Case, D.A. 1998. Modeling unusual nucleic acid structures. In *Molecular Modeling of Nucleic Acids* (eds. N. B. Leontes, and J. Santalucia). pp. 379-393. American Chemical Society, Washington, DC.

134. Butcher, S.E., Dieckmann, T., and Feigon, J. 1997. Solution structure of the conserved 16 S-like ribosomal RNA UGAA tetraloop. *J. Mol. Biol.* **268**:348-358.

135. Zhang, H., Fountain, M.A., and Krugh, T.R. 2001. Structural characterization of a six-nucleotide RNA hairpin loop found in *Escherichia coli*, r(UUAAGU). *Biochemistry* **40**:9879-9886.

136. Kale, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., and Schulten, K. 1999. NAMD2: Greater scalability for parallel molecular dynamics. *J. Comput. Phys.* **151**:283-312.

137. Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**:33-38.

138. Case, D.A., Pearlman, D.A., Caldwell, J.W., Cheatham, T.E., Wang, J., Ross, W.S., Simmerling, C.L., Darden, T.A., Merz, K.M., Stanton, R.V. *et al.* (eds.) 2002. AMBER 7. University of California, San Francisco.

139. Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.S., Cheatham, T.E., Debolt, S., Ferguson, D., Seibel, G., and Kollman, P. 1995. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **91**:1-41.

140. Alakent, B., Doruker, P., and Camurdan, M.C. 2004. Time series analysis of collective motions in proteins. *J. Chem. Phys.* **120**:1072-1088.

141. Doruker, P., Atilgan, A.R., and Bahar, I. 2000. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to alpha-amylase inhibitor. *Proteins-Structure Function and Genetics* **40**:512-524.

142. Kitao, A., and Go, N. 1999. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* **9**:164-169.

143. Ota, N., and Agard, D.A. 2001 Enzyme specificity under dynamic control II: Principal component analysis of α-lytic protease using global and local solvent boundary conditions. *Protein Sci.* **10**:1403-1414.

144. Mongan, J. (2004) Interactive essential dynamics. *J. Comput. Aid. Mol. Des.* **18**:433-436.

145. Balsera, M.A., Wriggers, W., Oono, Y., and Schulten, K. 1996. Principal component analysis and long time protein dynamics. *J. Phys. Chem.* **100**:2567-2572.

146. Baker, N.A., Sept, D., Joseph, S., Holst, M.J., and McCammon, J.A. 2001. Electrostatics of nanosystems: Application to microtubules and the ribosome. *PNAS* **98**:10037-10041.

147. Rázga, F., Koča, J., Šponer, J., and Leontis, N.B. 2005 Hinge-like motions in RNA kink-turns: The role of second A-minor motif and nominally unpaired bases. *Biophys. J.* **88**:3466-3485.

148. Rázga, F., Špačková, N., Réblová, K., Koča, J., Leontis, N.B., and Šponer, J. 2004. Ribosomal RNA kink-turn motif - A flexible molecular hinge. *J. Biomol. Struct. Dyn.* **22**:183-193.

149. Cojocaru, V., Klement, R. and Jovin, T.M. 2005. Loss of G-A base pairs is insufficient for achieving a large opening of U4 snRNA K-turn motif. *Nucleic Acids Res.* **33**:3435-3446.

150. Puglisi, J.D., Chen, L., Blanchard, S., and Frankel, A.D. 1995. Solution structure of a bovine immunodeficiency virus Tat-Tar peptide-RNA complex. *Science* **270**:1200-1203.

151. Puglisi, J.D., Tan, R.Y., Calnan, B.J., Frankel, A.D., and Williamson, J.R. 1992. Conformation of the Tar RNA-arginine complex by NMR Spectroscopy. *Science* **257**:76-80.

152. Chen, G., Znosko, B.M., Kennedy, S.D., Krugh, T.R. and Turner, D.H. 2005. Solution structure of an RNA internal loop with three consecutive sheared GA pair. *Biochemistry* **44**:2845-2856.

153. Rhee, Y.M., and Pande, V.S. 2003. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J.* **84**:775-786.

154. Sanbonmatsu, K.Y., and Garcia, A.E. 2002. Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins-Structure Function and Genetics* **46**:225-234.

155. Zhou, R.H. 2004. Exploring the protein folding free energy landscape: coupling replica exchange method with P3ME/RESPA algorithm. *J. Mol. Graph. Model.* **22**:451-463.

156. Law, S.M., Eritja, R., Goodman, M.F. and Breslauer, K.J. 1996. Spectroscopic and calorimetric characterizations of DNA duplexes containing 2-aminopurine. *Biochemistry* **35**:12329-12337.

List of figures

Figure 1: Overview of the splicing process. .... - 12 -
Figure 2: RNA rearrangements during spliceosome activation. .... - 14 -
Figure 3: Secondary structure of U4/U6 snRNA. .... - 16 -
Figure 4: Minimal binding site for 15.5K protein. .... - 18 -
Figure 5: RNA motifs capable of binding the 15.5K protein. .... - 20 -
Figure 6: Crystal structure of 15.5K protein bound to U4 snRNA. .... - 21 -
Figure 7: Structure of the K-turn motif. .... - 24 -
Figure 8: Experimental data on the K-turn motif. .... - 27 -
Figure 9: Conformational parameters in the RNA .... - 29 -
Figure 10: simplified view of the energy landscape. .... - 41 -
Figure 11: Simulated RNA constructs. .... - 64 -
Figure 12: General analysis of MD simulations. .... - 66 -
Figure 13: Conformational states of A30. .... - 68 -
Figure 14: Inter-stem contacts established in the RNA .... - 70 -
Figure 15: The G32-A44 base pair. .... - 71 -
Figure 16: Different conformations of KtU4. .... - 74 -
Figure 17: Correlation between opening of G-A base pairs and the k-e motion. ...... - 76 -
Figure 18: Scatter 3D plots of stacking interactions .... - 77 -
Figure 19: Scatter 3D plots of the stacking interactions .... - 79 -
Figure 20: Structural view of the transitions between different stacking patterns in the free RNA. .... - 80 -
Figure 21: Correlation between backbone flexibility and the k-e motion. .... - 81 -
Figure 22: Principal component analysis of LES4 trajectory. .... - 83 -
Figure 23: Interactions between the 15.5K protein and the external loop. .... - 85 -
Figure 24: Stacking interactions between C41 and N40. .... - 87 -
Figure 25: Snapshots from the simulations of bound K3 and K4 RNAs: .... - 88 -
Figure 26: Snapshots from the simulations of Kt15. .... - 89 -
Figure 27: Orientations of the external loop in Kt15 and KtU4 .... - 90 -
Figure 28: B-factors per protein residue in different simulations. .... - 92 -
Figure 29: Flexible regions in the 15.5K protein (in all simulated system). .... - 93 -
Figure 30: Comparison between the flexibilities of 15.5K and L7AE proteins bound to their cognate RNAs .... - 94 -

Figure 31: Protein-RNA interface.. - 95 -
Figure 32: Example of protein-RNA contact.. - 95 -
Figure 33: The hairpin-like structure in the 15.5K protein...................................... - 96 -
Figure 34: Electrostatic potential on molecular surfaces.. - 97 -
Figure 35: Transitions in the α-helix 5.. - 98 -
Figure 36: Spectrum of 2-aminopurine in KtU4.. - 99 -
Figure 37: Titration of labeled KtU4 with 15.5K protein.. - 100 -
Figure 38: Melting profiles of labeled KtU4... - 100 -
Figure 39: Titration with Mg^{2+} of labeled KtU4 at 5℃... - 101 -
Figure 40: Protein cavity recognizing the RNA.. - 105 -
Figure 41: Proposed similarities between the L7AE-L15E-Kt15 and 15.5K-61K-KtU4 complexes.. - 112 -
Figure 42: Model of dynamic events occurring at the 5' stem-loop of U4 snRNA during U4/U6 snRNP assembly…………………………………………….- 115 -

List of tables

Table 1: Conformational transitions in the RNA…………………………………………- 67 -
Table 2: Conformational states of adenine A30…………………………………………- 69 -
Table 3: Stability of the Watson-Crick base pairs of the NC-stem…………………….- 72 -
Table 4: Interactions in the unbound RNA……………………………………………...- 102 -

Curriculum vitae

Personal Details

Institution	Max Planck Institute for Biophysical Chemistry		
Work Address	Am Fassberg 11, 37077 Göttingen, Germany		
Home Address	Albrecht-Thaer Weg 10a/04, 37075 Göttingen, Germany (till 30.06.05)		
Telephone	**Work:** ++49-551-2011327	**Home:** +49-551-9963204	**Mobile:** +49-179-6851586
Nationality	Romanian	**Marital Status:** Single	
Date of Birth:	15.07.1976	**Place of Birth:** Arad, Romania	
E-mail	Vlad.Cojocaru@mpi-bpc.mpg.de		

Education

1) October 2000 - Present

International Max Planck Research School, Göttingen, Germany

(International MSc/PhD Program); URL: http://www.gpmolbio.uni-goettingen.de/

1.1) October 2001 – Present

Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

Laboratory of **Dr. Thomas Jovin**

PhD thesis; URL: http://www.mpibpc.gwdg.de/abteilungen/060/

1.2) October 2000 – September 2001

Preparatory Year: Graduated with a total grade of 2.0 (very good)

Lectures and Method Courses covering different topics:

- Biochemistry and Structural Biology,
- Molecular Genetics,
- Functional Organization of the Cell,
- Model Systems of Molecular Biology

Three Laboratory Rotations (see Research Experience)

2) September 1999 – September 2000

West University of Timişoara, Romania

Master Program: “The chemistry of biologically active compounds” (interrupted after I joined the MSc/PhD program in Göttingen). Reference: Prof. Dr. Tudor Oprea

3) October 1995 – June 1999

West University of Timişoara, Romania

Physics-Chemistry Faculty (Graduated with a total grade of 9.64 out of 10.00)

BSc Thesis: “Chemical analysis of milk and milk products”

Research Experience

Oct. 2001 – present **Max Planck Institute for Biophysical Chemistry** Laboratory of **Dr.Thomas Jovin** Göttingen, Germany **PhD thesis** (PhD expected in July 2005) References: Dr. Thomas Jovin and Reinhard Klement	**Title**: “Molecular motions at the 5’ stem-loop of U4 snRNA: Implications for U4/U6 snRNP assembly” Methods*:* Molecular dynamics simulations, locally enhanced sampling, essential dynamics, force field parameter development, simulated annealing, electrostatic calculations, molecular modeling, steady-state fluorescence.
January 2001 – March 2001 **Max Planck Institute for Biophysical Chemistry** Laboratory of **Dr.Thomas Jovin** Göttingen, Germany **Laboratory Rotation**	**Title:** “Modeling of mixed DNA duplexes composed of one anti-parallel B or Z DNA flanked by two parallel stranded DNA in right-handed or left-handed conformation" Methods*:* Molecular modeling
March 2001 – May 2001 **Georg August University** Laboratory of **Prof. Dr. George Scheldrick** Göttingen, Germany **Laboratory Rotation**	**Title:** “Structure determination of the sweet protein thaumatin using X-ray crystallography” Methods*:* crystallization, structure determination by molecular replacement, crystal soaking, phasing via anomalous signal of iodide

May 2001 – July 2001 **Georg August University** Laboratory of **Prof. Dr. Hans-Joachim Fritz** Göttingen, Germany **Laboratory Rotation**	**Title:** "Uracil-DNA glycosylases in *Thermus Thermophylus* : differences in primary sequence may lead to a different repair mechanism" Methods: site-directed mutagenesis, PCR, cloaning, gel electrophoresis
January 1999 – June 1999 **West University** Laboratory of **Prof. Dr. Dumitru Tita** Timişoara, Romania **BSc Thesis**	**Title:** "Chemical analysis of milk and milk products"

Additional Expereince

Teaching Experience

Teaching students of the International MSc/PhD Program in Goettingen:

- Tutorials in "DNA and chromatin structure"
- Methods course in "Molecular Modeling"

Scientific Meeting Organization

Member of the organizing committee and chair of the Structural Biology section of the "Horizons in Molecular Biology" PhD student's symposium in Göttingen, Germany (last meeting "Decoding Nature: Hierarchy of Interactions", 17.03.05-19.03.05)

Scientific Conferences Attendance (Poster or oral presentation)

- Horizons in Molecular Biology PhD student's symposium, Göttingen, Germany, December 2003, March 2005 – **poster presentation**
- Biannual meeting of the International Society of Quantum Biology and Pharmacology in Como, Italy, June 2004 – **poster presentation**
- EMBO conference on Structures in Biology, Heidelberg, Germany, Nov. 2004 – **poster presentation**

- RNA Structure and Function meeting, Edinburgh, Scotland, December 2004 – **poster presentation**
- Computer simulation & Theory of Macromolecules, Hühfeld, Germany, 22-24 April 2005 – **oral presentation**

Skills

Computer Knowledge

	Operating Systems	SUSE LINUX 8.0++, UNIX, WINDOWS 95++
	Molecular Modeling Software	AMBER, NAMD, VMD, GOPENMOL, GAUSSIAN, NAMOT, SYBYL, NAB, CURVES, RED, APBS, PDB and NDB databases
	Other Software	MS OFFICE, XMGRACE, CORREL Suite, ENDNOTE, MATLAB (beginner level)
	Programming Languages	PERL, TCL, CSH

Languages

	fluent	Romanian, English, Spanish
	good	German
	average	French, Italian

Scholarships and Awards

1996-1999: Scholarship awarded for exceptional results during undergraduate studies, West University Timişoara, Romania

1999-2000: Scholarship for M.Sc. studies, West University Timişoara, Romania

2000-2001: Stipend, International M.Sc./Ph.D. Program, Georg August University Göttingen

2001-present: PhD stipend, Max Planck Institute for Biophysical Chemistry, Göttingen

June 2004: Special award, 2nd ISQBP President's meeting, Como, Italy

Other Interests

Mountains, Football, Tennis, Rock Music

References

Dr. Thomas Jovin
Max Planck Institute
for Biophysical Chemistry,
Dept. of Molecular Biology
Am Fassberg 11
37077 Goettingen, Germany
Tel: ++49-551-2011382
Fax: ++49-551-2011467
E-mail: tjovin@gwdg.de

Prof. Dr. Tudor Oprea
Office of Biocomputing
MSC08 4560 / BMSB B61
1 University of New Mexico
Albuquerque, New Mexico, USA 87131
Tel: +1 (505) 272-6950
Fax: +1 (505) 272-8738
Email: toprea@salud.unm.edu

Reinhard Klement
Max Planck Institute
for Biophysical Chemistry,
Dept. of Molecular Biology
Am Fassberg 11
37077 Goettingen, Germany
Tel: ++49-551-2011389
Fax: ++49-551-2011467
E-mail: rklemen@gwdg.de

Prof. Dr. Reinhard Lührmann
Max Planck Institute
for Biophysical Chemistry,
Dept. of Cellular Biochemistry
Am Fassberg 11
37077 Göttingen,Germany
Tel: ++49-551-2011407
Fax: ++49-551-2011197
E-mail:
Reinhard.Luehrmann@mpi-bpc.mpg.de

Declaration of originality:

Hereby I declare that I have written independently the entire content of the present PhD thesis using no other sources and aids than quoted. The chemical probing experiments presented in figures 8B and 8C were performed by Stephanie Nottrott and were included in the paper that we published together in the RNA journal in February 2005.