Università degli Studi di Napoli "Federico II"
Facoltà di Farmacia

Dottorato di Ricerca in "*Scienza del Farmaco*" XXIV Ciclo

Indirizzo
Chimica delle Sostanze Naturali Bioattive

Dipartimento di Chimica delle Sostanze Naturali

Insight into macromolecular interactions involving nucleic acids: a study of the interaction of G-quadruplexes with pharmacologically active molecules and investigations of the RNA binding properties of LARP4 protein.

Coordinatore
Chiar.ma Prof.ssa M. Valeria D'Auria

Tutor:
Chiar.mo Prof. Antonio Randazzo

Dottoranda
Dott.ssa Roberta Trotta

INDICE

Abstract……..6

CHAPTER I

G-Quadruplex structures……..…7

G-Quadruplex: functions……..12

Telomeres and Telomerase……...12

Transcription regulation……...14

References……...19

Aptamers…….20

References……....23

CHAPTER II

Targeting G-Quadruplex……...24

Introduction……24

References……...28

Structural and conformational requisites in DNA Quadruplex groove binding……29

Results and discussio…….31

Synthesis of compound 3……...31

Nuclear Magnetic Resonance Experiments…….33

Structure calculations……....43

Isothermal Titration Calorimetry Measurements……..50

Material and method……54

Oligonucleotide synthesis…….54

Synthesis of N-[5-({[5-({[3-Amino-3-(N-methylcarbamoyl)-ethyl]amino}carbonyl)-1-methyl-1H-pyrrol-3-yl]amino}carbonyl)-1-methyl-1H-pyrrol-3-yl]-4-(formylamino)-1-methyl-1H pyrrole-2-carboxamide (2)…….....55

Synthesis of 4-Amino-N-[5-({[5-({[3-amino-3-(N-methylcarbamoyl)ethyl]amino}carbonyl)-1-methyl-1H-pyrrol-3-yl]amino}carbonyl)-1-methyl-1H-pyrrol-3-yl]-1-methyl-1H-pyrrole-2-carboxamide (3)……56

Nuclear Magnetic Resonance Experiments……..…....56

Isothermal Titration Calorimetry…….....57

Structure calculation……58

References……..62

Tandem application of Virtual Screening and NMR experiments in the discovery of brand new DNA Quadruplex groove binders……...64

Results and discussion……...76

NMR and molecular modelling studies……76

Isothermal Titration Calorimetry Measurements……..95

Material and method……101

Oligonucleotide synthesis…….101

Virtual Screening calculation……..102

Nuclear Magnetic Resonance Experiments……103

Molecular modelling…….105

Isothermal Titration Calorimetry……...106

References…….....108

“Fine tuning” of the selected compounds……....111

References……...117

Conclusions…….118

CHAPTER III

La protein……119

Structural analysis of La proteins…….....122

Poly-U binding of La protein…….....128

References……132

CHAPTER IV

La-related proteins (LARPs)…….....139

Role of LARP4……140

References…….....145

CHAPTER V

Materials and methods……...147

Plasmid and bacterial cell strains used in this study……...147

Transformation of competent cells……....147

Expression studies……...148

Large scale expression of unlabelled proteins……..148

Large scale expression of isotope-labelled proteins…….....149

Purification of proteins……...149

Chromatography procedures…….150

Ni-affinity column……...150

TEV cleavage……...150

Heparin column……...151

NMR samples preparation…….152

NMR measurements……...153

References…….....154

Results……..155

Purification of LARP4(111-196)……156

Purification of LARP4(196-287)……159

NMR analysis of LARP4(111-196) and LARP4 (196-287)……..171

Assignment of NMR spectra……..172

^{1}H, ^{13}C and ^{15}N backbone assignment of LARP4(111-196) and LARP4(196-287)……..175

Secondary structure prediction…….180

^{1}H and ^{13}C side chain assignment of LARP4(111-196) and LARP4(196-287)……..182

Structure determination…….184

References……188

CHAPTER VI

Discussion……189

Conclusions…….198

ABSTRACT

This dissertation can be divided in two parts.

In the first part, my efforts have been focused on the study of the G-quadruplex structures.

In chapter one, I describe the unusual arrangement of the DNA to form the G-quadruplex structures and their biological role, underlying the reason why in the last period a lot of research groups focused their studies on this topic.

In chapter two, after a brief introduction of the first molecules capable to bind the G-quadruplex (Distamicyn A and its derivates), I describe the new organic molecule capable to interact with the DNA quadruplex, which I have discovered applying some innovative approaches during my PhD.

In the second part of this thesis I introduce the research conducted in the laboratory of Dr. Maria R. Conte at the King's College of London. In this stimulating environment I started studying the structure of the protein LARP4 and its interaction with a poly (A) RNA using the NMR.

In the course of my research I have learnt the techniques allowing to do the transformation, the expression and the purification of the protein. A summary presentation of the techniques I have applied is reported in the second part in chapter five.

CHAPTER I

G-Quadruplex structures

The double helix structure of duplex DNA is well known. The two antiparallel strands are held together by complementary base pairing between adenine (A) and thymine (T), and between guanine (G) and cytosine (C), using two hydrogen bonds in AT, and three in GC. However, this is not the only base pairing arrangement that can occur between bases, and alternative base pairing leads to alternative structures, including triple-stranded and four-stranded structures called G-quadruplex.

G-quadruplex, as the name suggests, have a core that is made up of guanine bases only, with four guanines arranged in a rotationally symmetric manner, making hydrogen bonds from N1-O6 and N2-N7 around the edges of the resulting square (see figure 1).

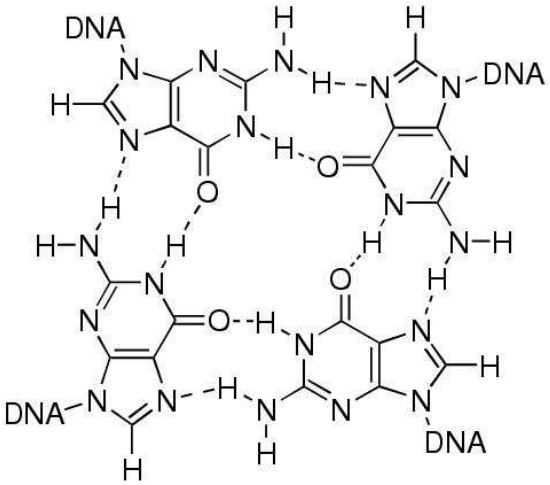


Figure 1: Four guanines can hydrogen bond in a square arrangement to form a G quartet. There are two hydrogen bonds on each side of the square.

These planar structures are called G-quartets, and are stabilized by monovalent cations, in particular K^+ and to a lesser extent $NH4^+$ and Na^+, which interact with the alone pairs on the O6 atoms surrounding the central core.

These G-quartets have large π-surfaces, and hence tend to stack on each other due to π–π stacking, as well as to enable cations to intercalate between the G-quartets. In particular, oligonucleotides with contiguous runs of guanine, such as d(TGGGT) can form stacked structures with the G-quartets linked by the sugar–phosphate backbone. These are called G-quadruplexes and can form from DNA or RNA strands. They are helical in nature due to the constraints of π–π stacking, although for convenience they are often depicted without the helicity, as shown in figure 2.

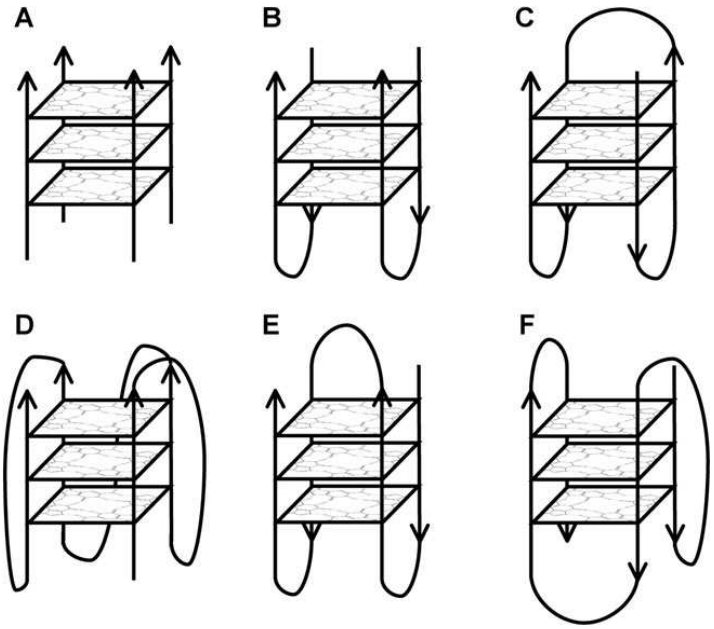


Figure 2: G-quartets can stack on top of each other, forming G-quadruplex structures. These are held together by π-stacking and the sugar–phosphate backbone. G-quadruplex can adopt a range of different stoichiometries and folding patterns. **(A)** Tetramolecular structure with all strands parallel; **(B)** bimolecular antiparallel structure with adjacent parallel strands; **(C)** unimolecular antiparallel structure with alternating parallel strands; **(D)** unimolecular parallel structure with three double chain reversal loops; **(E)** unimolecular antiparallel structure with adjacent parallel strands and a diagonal loop; **(F)** unimolecular mixed structure with three parallel and one antiparallel strands. All three structures (D), (E) and (F) have been observed for the human telomeric repeat.

Since there is a directionality to the strands, customarily described as from the 5' end to the 3' end, there are topological variants possible for these four strands. All four strands may be parallel, three parallel and one in the opposite direction (antiparallel), or there may be two in one direction and two in the other, either with the parallel pairs adjacent to each other or opposite each other. A shorthand has arisen which describes all the arrangements with at least one antiparallel strand as 'antiparallel', although

this does not give a full description of the structures. These are depicted in figure 2.

At a molecular level, the different directionality of the strands relates to the conformational state of the glycosidic bond between the guanine base and the sugar. This may be either syn or anti, as depicted in figure 3.

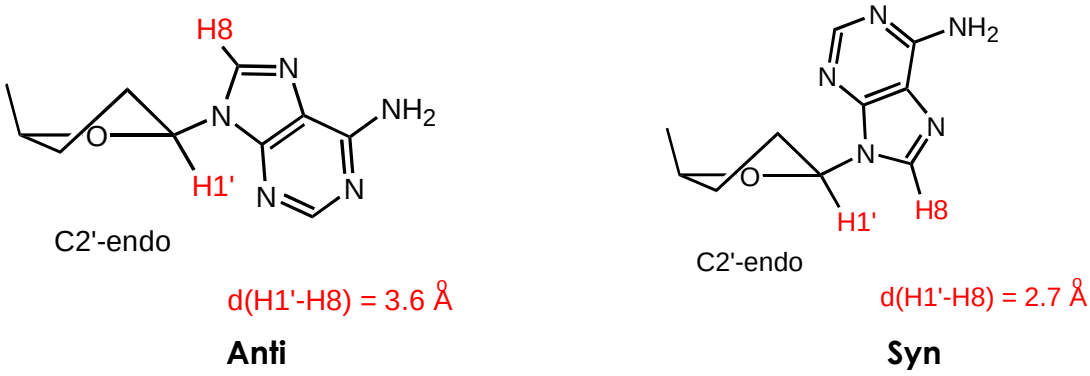


Figure 3: The bond between the base and sugar can rotate. It has two preferred conformations, syn and anti.

When all four strands are parallel, all the bases are in the anti conformation and the grooves between the backbones are all of equal size, the system is entirely C4 symmetric. When any of the strands are antiparallel, the bases must be in the syn form in order for the hydrogen bonds to be formed correctly. This then affects the orientation of the backbone relative to the G-quartets, and hence results in grooves of different sizes. When successive guanines (starting with the guanine contributing N1 and N2) are both anti or both syn, the groove is medium in size; if the first is anti and the second syn, the groove is wider, and if the first is syn and the second anti, then the groove is narrower. Thus a G-quadruplex with adjacent parallel strands will be

arranged with glycosidic bonds anti–syn–syn–anti, and will have grooves that are wide, medium, narrow and medium. In contrast, a structure with alternating strands will have glycosidic bonds anti–syn–anti–syn, with grooves wide, narrow, wide and narrow. G-quadruplexes may be comprised of four separate strands, as in the example above, forming tetramolecular G-quadruplexes, which are always found in the all antiparallel form. Alternatively, they may be formed from two strands, each with two sets of contiguous guanines, or just from one strand, folding back on itself to form an intramolecular structure. In either of these cases, there will be loops that serve to connect the strands of the structure together. Depending on which strands are connected, these loops may cross diagonally across the top of the structure, joining diagonally opposed antiparallel strands; go across a side, linking adjacent antiparallel strands; or may loop around the side of the structure linking parallel strands and forming a double-strand reversal loop. Some examples of these are shown in figure 2.

G-quadruplexes: functions

The important role of these structures in biological systems lies mainly on three features:

- Architecture of telomeres of many organisms
- G-rich sequences that are present within a wide range of genes
- Scaffolds of several oligonucleotide aptamers.

Telomeres and telomerase

A fundamental difference in the behavior of normal cells compared to tumour cells in culture is that, normal cells divide for a limited number of times (exhibit cellular senescence) whereas tumour cells usually have the ability to proliferate indefinitely (are immortal). There is substantial experimental evidence that cellular aging is dependent on cell division and that the total cellular lifespan is measured by the number of cell generations, not by chronological time.

This means there is an intrinsic molecular counting process occurring during cell growth that culminates in the cessation of cell division.

Human telomeres contain long stretches of the repetitive sequence TTAGGG which are bound by specific proteins. With each cell division, telomeres shorten by 50-200 base pairs, primarily because the lagging strand of DNA synthesis is unable to replicate the extreme 3' end of the chromosome, known

as the end replication problem (figure 4).

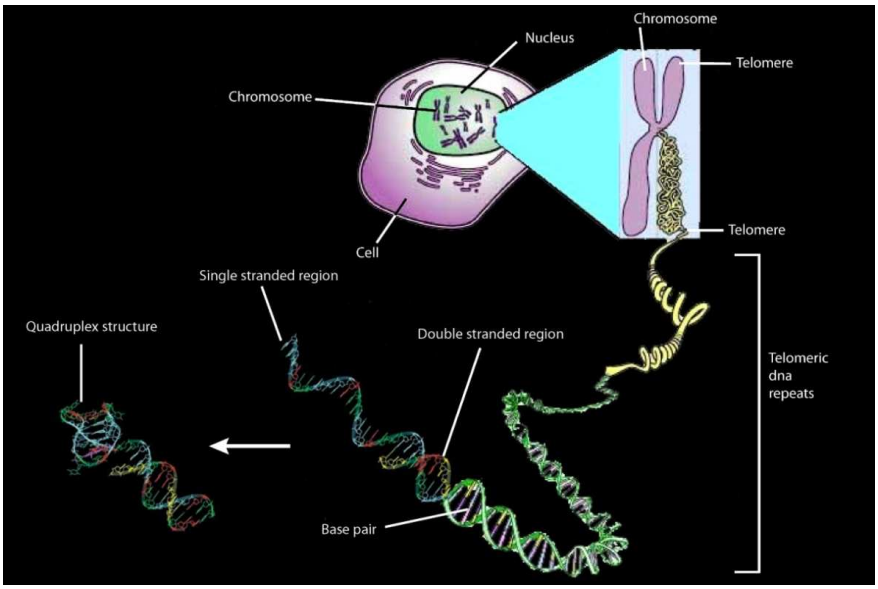


Figure 4: A telomere is the region on ends of chromosomes, composed of repeating units of 6 nucleotides (humans = TTAGGG, repeating 1500 times),without genetic value, capable of forming G-quadruplex structures.

When telomeres become sufficiently short, cells enter an irreversible growth arrest called cellular senescence. In most instances cells become senescent before they can accumulate enough mutations to become cancerous, thus the growth arrest induced by short telomeres may be a potent anti-cancer mechanism.

Telomerase, an eukaryotic ribonucleoprotein (RNP) complex, helps to stabilize telomere length in human stem cells, reproductive cells and cancer cells by adding TTAGGG repeats onto the telomeres using its intrinsic RNA as a template for reverse transcription. Telomerase activity has been found in almost all human tumours but not in adjacent normal cells. The most prominent hypothesis is that maintenance of telomere stability is required for

the long-term proliferation of tumours. Thus, escape from cellular senescence and becoming immortal by activating telomerase, or an alternative mechanism to maintain telomeres, constitutes an additional step in oncogenesis that most tumours require for their ongoing proliferation. This makes telomerase a target not only for cancer diagnosis but also for the development of novel anti-cancer therapeutic agents. One approach to telomerase inhibition involves sequestering its substrate, single-stranded telomeric DNA, by inducing it to form G-quadruplex structures. There has been considerable work on developing G-quadruplex ligands, especially to target the human telomeric repeat and hence blocking the action of telomerase.

Transcription regulation

Gene transcription is tightly regulated, by a variety of methods. One method that is used in some cases is based on the presence of G-quadruplexes located in the promoter region of a gene, broadly speaking the kilobase upstream of the transcription start site (TSS). This model was originally demonstrated by Hurley and co-workers[1] for the oncogene c-myc, an important transcription factor involved in regulating around 15% of all human genes. As a result of this, overexpression of c-myc has been implicated in a wide range of cancers including colorectal cancer. Within its promoter there is a region, 115–142 base pairs upstream of the TSS, which is highly sensitive to

nucleases, suggesting that it forms an accessible structure free from histone proteins. This region controls the vast majority of the transcription of the gene, and studies in vitro of the sequence d(GGGGAGGGTGGGGAGGGTGGGGAAGG) show that it is capable of forming into a family of polymorphic G-quadruplex, using various combinations of the guanine runs underlined. It has further been shown that the G-quadruplex ligand TMPyP4 binds to this element leading to down regulation of c-myc expression. This clear proof of principle led to the proposal that this may be a general mechanism for gene regulation. The simplest form of the model (figure 5) suggests that there is an equilibrium between two forms of the DNA. On one side the equilibrium is double helix DNA, and the transcription occurs as normal; on the other side, one strand is separated, and has folded up into a G-quadruplex. This structure then acts as a steric block to the transcription. Addition of a G-quadruplex ligand, whether a small molecule or a protein, will energetically favour the G-quadruplex form, and hence move the equilibrium towards that side reducing the transcriptional activity. This dynamic equilibrium has been experimentally demonstrated. Although this model is presented in terms of steric blockage leading to a reduction in transcriptional activity, as was found for c-myc, it is also possible that the G-quadruplex form could be an activating domain, either because of putative protein recognition of the G-quadruplex, or because the accessibility of the other strand leads to increased transcriptional activity.

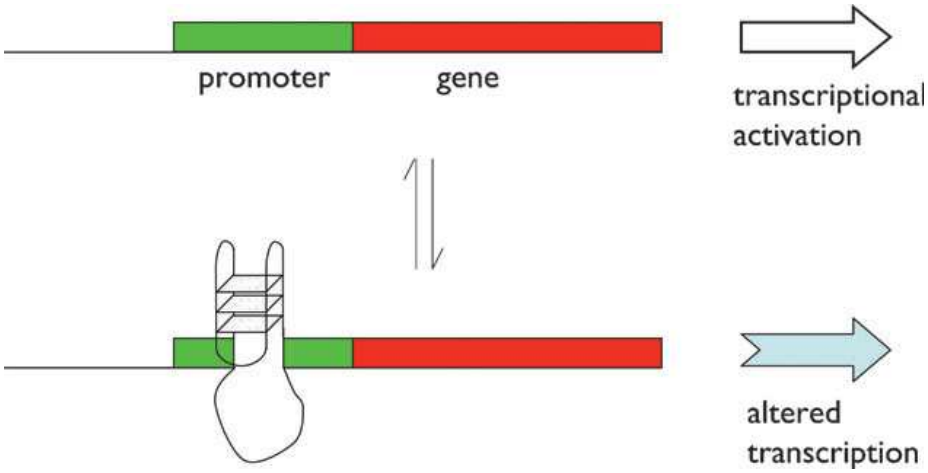


Figure 5: The formation of a G-quadruplex in a promoter can affect the level and nature of transcription from that gene. At the simplest level, it may act as a steric block to the transcription machinery.

However, although a wide variety of different genes have now been shown to have promoter G-quadruplex, such as VEGF, HIF-1a, Bcl-2, Ret, c-kit and KRAS, none have yet had G-quadruplex formation leading to increased transcriptional activity. Although in figure 5 the complementary C-rich strand is drawn as an unstructured sequence, it is possible that it could form an alternative four-stranded structure called an i-motif (figure 6). An i-motif has four strands forming a structure somewhat like two interleaved ladders, with pairs of strands held together by diagonal CRC^+ bonds. These rely on the protonation of the N3 of cytosine, which has a pKa of 4.2. As a result, these structures are generally only stable under acidic conditions, but the stability will clearly be different in the context of chromosomal DNA, especially if a G-quadruplex structure forms and holds the ends of the i-motif together in the correct orientation.

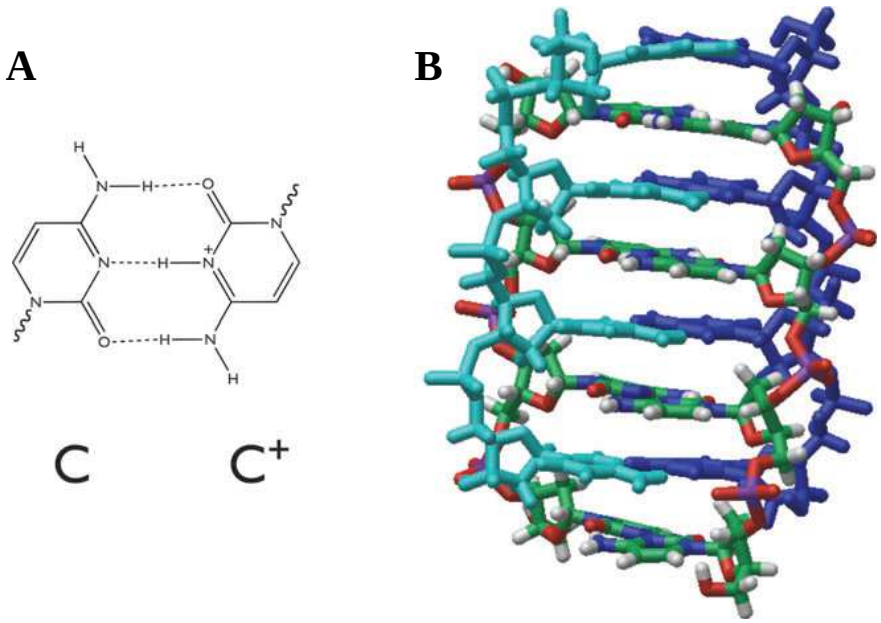


Figure 6: A) Two cytosines can pair up, forming three hydrogen bonds, when one of them has been protonated. **B)** Four cytosine containing strands can form an i-motif, with interleaved C:C+ bonds. One pair of binding strands is shown in CPK colouring, the other in blue/cyan.

Nonetheless, there is still controversy as to the biological relevance of the i-motif structure. Using the predictive algorithm quadparser, it is possible to investigate how many human genes contain G-quadruplex motifs in their promoter regions, both to identify novel genes for experimental testing, but also to see if there is evidence of over- or under-representation of such genes. This analysis has shown that almost half of all genes (43%) contain putative G-quadruplex structures in their promoters, which is considerably more than would be expected by chance, based on the rest of the genome. The enrichment of G-quadruplex motifs occurs increasingly nearer the TSS, with the first few hundred bases seeming to be particularly important for G-quadruplex motifs to be present. Interestingly, genes involved in cancer are

even more likely to have such promoter G-quadruplex, with 67% having such sequences. Using the online Gene Ontology database, which describes the functions of every human gene, it is possible to see if there is any general bias for types of genes that have promoter G-quadruplex. This has revealed that genes involved in tightly-regulated processes such as development, neurogenesis and cell differentiation are more likely than other classes of genes to have promoter G-quadruplex, whereas genes involved in processes such as protein biosynthesis, olfaction and immune response are much less likely to have promoter G-quadruplex.[2]

References

1. Han H., Langley D.R., Rangan A., Hurley L.H. **J. Am. Chem. Soc.** 2001, 123, 8902-8913.
2. Julian Leon Huppert. **Chem. Soc. Rev.**, 2008, 37, 1375–1384.

Aptamers

Aptamers[1] are nucleic acid macromolecules that bind to molecular targets, including proteins, with high affinity and specificity. Aptamers have typically from 15 to 40 nucleotides in length and can be composed of DNA and RNA. Base composition defines aptamer secondary structure, consisting primarily of helical arms and single-stranded loops. Stable tertiary structure, resulting from combinations of these secondary structures, allows aptamers to bind to targets via van der Waals, hydrogen bonding and electrostatic interaction.

The conceptual framework and process of aptamer generation emerged from pioneering experiments published in the early 90s.[2-3] Particularly, Gold et al.[3] described a process of in vitro aptamer selection, dubbed SELEX (Systematic Evolution of Ligand by Exponential enrichment). SELEX is a combinatorial chemistry methodology based on oligonucleotide libraries which are screened for high affinity binding to a given target. High affinity ligands can be isolated from the library using iterative rounds affinity-based enrichment, alternating with oligonucleotide amplification.

The aptamer drug TBA (Thrombin Binding Aptamer), also named ARC183 (Archemix Corp) is a consensus DNA 15-mer, namely 5'GGTTGGTGTGGTTGG3',[4,5] discovered with this technique (figure 7).

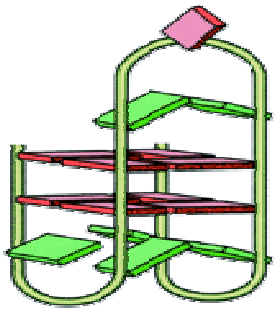


Figure 7: Pictorial model of TBA

The three dimensional solution structure of TBA was solved using NMR and X-ray techniques (figure 8).[6,7] TBA is characterized by a chair-like quadruplex structure consisting of two G-tetrads connected by two TT loops and a single TGT loop. The aptamer is a thrombin inhibitor in development for use as an anticoagulant during coronary artery bypass graft procedures. Currently, the only approved anticoagulant for coronary artery bypass graft is heparin. TBA exhibits a K_d of 2 nM for thrombin, 50 nM for prothrombin, and binding to other serum proteins or proteolytic enzymes is essentially undetectable.[2] As suggested by others,[8] TBA binds at the anion exosite I of thrombin in a conformation little changed from that of the unbounded species. TBA is a strong anticoagulant in vitro, and inhibits thrombin-catalyzed activation of fibrinogen, and thrombin induced platelet aggregation. TBA has key advantages in that it avoids heparin use and the risk of associated thrombocytopenia, is a specific inhibitor with rapid onset, is effective at inhibiting clot-bound thrombin, and has a short in vivo half-life of approximately 2 min which allows for rapid reversal of its effects and the avoidance of dose-adjusting complications of heparin and protamine.

Neither significant toxicities nor excessive bleeding intraoperatively has been observed.

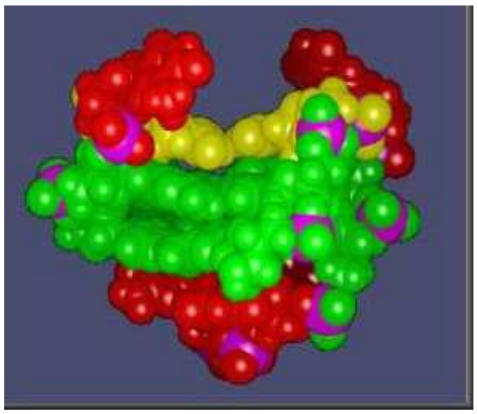


Figure 8: Three dimensional solution structure of TBA.

References

1. Nimjee S. M., Rusconi C. P., Sullenger B. A. **Annual Review of Medicine**, 2005, 56, 555.
2. Block L. C., Griffin L. C., Latham J. A., Vermaas, E. H., Toole J. J. **Nature**, 1992, 355, 564.
3. Gold L., Polisky B., Uhlenbeck O., Yarsus M. **Annu. Rev. Biochem.**, 1995, 64, 763.
4. Griffin L. C., Tidmarsh G. F., Bock L. C., Toole J. J., Leung L. K. **Bloods**, 1993, 81, 3271.
5. Li W. X., Kaplan A. V., Grant G. W., Toole J. J., Leung L. L. **Blood**, 1994, 83, 677.
6. Wang K. Y., McCurdy S., Shea R. G., Swaminathan S., Bolton P. H. **Biochemistry**, 1993, 32, 1899.
7. Macaya R. F., Schultze P., Smith F. W., Roe J. A., Feigon J. **Proc. Natl. Acad. Sci. USA**, 1993, 90, 3745-3749.
8. Paborsky L. R., McCurdy S. N., Griffin L. C., Toole J. J., Leung L. K. **J. Biol. Chem.**, 1993, 268, 2080

CHAPTER II

Targeting G-quadruplex

Introduction

In the recent years studies on molecules that stabilize the DNA quadruplex are becoming increasingly popular, due to the fact that these molecules can be used as anticancer agents.

G-quadruplexes are four-stranded helical DNA or RNA structures, comprising stacks of G-tetrads, which are the planar association of four guanines in a cyclic Hoogsteen hydrogen-bonding arrangement.[1] These structures are formed by the folding of one DNA or RNA strand or by the association of two or more strands. This results in different combinations of relative strand orientations, with consequent formation of grooves of different widths, and a number of loop arrangements. From a biological point of view, G-quadruplexes are widespread in genome and they seem to play a role in a number of processes, such as replication, recombination transcription, and translation.[2] Furthermore, quadruplexes are also found in telomeric DNA.

Telomers consist of an ensemble of proteins and specialized noncoding DNA sequences which protect the ends of the chromosome from damage and recombination, and their shortening is implicated in cellular senescence.

It has been demonstrated that the elongation of telomeric DNA operated by

the enzyme telomerase leads cancer cells to an infinite lifetime. Hence, the inhibition of telomerase, which is overactive in 85% of tumors, is expected to move into the forefront of research for new effective anticancer drugs.

Since this enzyme requires a single stranded telomeric primer, the formation of G-quadruplex complexes by telomeric DNA inhibits the telomerase activity.

In this respect, it has been found that small molecules that stabilize G-quadruplex structures are effective telomerase inhibitors.[3] Quadruplex structures offer several recognition sites, most of the interacting molecules discovered so far have been found to interact with the wide π-stacking surface of the G-tetrads at the 5′ and/or 3′ edges of the quadruplex.[4]

The first groove-binder that has been discovered is the Distamicyn A, small molecule with antibiotic propriety.[5] This molecule is capable to interacts with the grooves of the quadruplex $[d(TGGGGT)]_4$ as antiparallel dimers that bind simultaneously two opposite grooves of the quadruplex, adopting a particular crescent shape. The complex is also stabilized by four drug/DNA H-bonds with guanine bases, furthermore there are also strong Coulombic interaction between the positively charged amidinium moiety and the phosphate groups of the quadruplex.

Classical intercalation, as observed in the duplex DNA structures, has not been demonstrated to date, most probably due to the presence of cations like K^+ or Na^+ in the very center cavity of the quadruplex structures, that prevents such a binding mode. Generally, stacking interaction and groove-

binding modes are characterized by very different specificities. Groove-binding recognition generally offers a higher extent of selectivity, since it can more easily recognize different DNA sequences and, in the case of quadruplex structures, can also discriminate among several quadruplex topologies, taking advantage of their different groove widths.

In nature there are a lot of derivatives of Distamicyn A and in this regard some molecules have been studied which modify the crescent shape's extensions, changing the number of the pyrrole unitis in the Distamicyn A.[6] These studies demonstrated that the presence of one pyrrole unit more than the Distamycyn A, (is the case of **compound 1**)[6] changes both its binding mode and its stoichiometry, whilst the addition of two pyrrole units more than the Distamicyn A (**compound 2**)[6] causes the minor affinity to the quadruplex $[d(TGGGGT)]_4$ (figure 1).

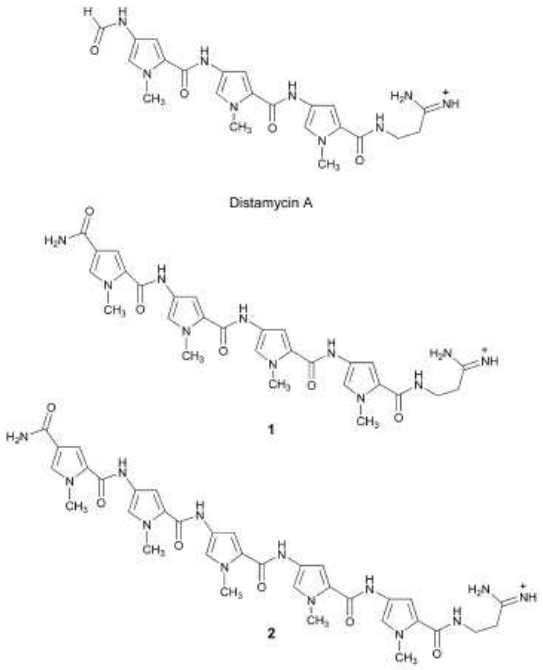


Figure 1: Chemical structures of Distamycin A and its analogues: compound 1 and compound 2.

Molecules that we can define groove-binders, if compared with intercalate agents, have an additional worth: more selectivity and specifity of binding to G-quadruplex structures.

In this chapter I report the work that I have made with my research group which has led to the discovery of a new organic molecule capable to interact and to stabilize the DNA quadruplex, which probably can become an anti-cancer drug because, like other molecules capable to interact and stabilize the DNA quadruplex, it could be used as a telomerase inhibitor.

References

1. Burge S., Parkinson G. N., Hazel P., Todd A. K., Neidle S. **Nucleic Acids Res.** 2006, 34, 5402–5415.
2. Johnson J. E., Smith J. S., Kozak M. L., Johnson F. B. **Biochimie** 2008, 90, 1250–1263.
3. Kelland L. R. Eur. J. **Cancer** 2005, 41, 971–979.
4. Ou T., Lu Y., Tan J., Huang Z., Wong K., Gu L. **ChemMedChem** 2008, 3, 690–713.
5. Martino L., Virno A., Pagano B., Virgilio A., Di Micco S., Galeone A., Giancola C., Bifulco G., Mayol L., Randazzo A. **J. Am. Chem.Soc.** 2007 129: 16048–16056
6. Pagano B., Virno A., Mattia Carlo A., Mayol L., Randazzo A., Giancola C. **Biochimie** 2008.90: 1224-1232.

Structural and conformational requisites in DNA Quadruplex groove binding

Small molecules able to stabilize G-quadruplex have been shown to be effective telomerase inhibitors.[1,2] These findings have brought a great momentum in the study of G-quadruplex binders. So far, most of the reported G-quadruplex binding agents bind to DNA by interacting with the wide π-surface of the G-tetrads at the edges of the quadruplex. Groove-binding mode to quadruplex structure was first proposed by Randazzo's group for the Distamycin A,[3] and up to now, this is the only molecular entity for which a pure groove binding mode has been experimentally proven. The interest in the groove binding recognition is mainly fuelled by the chemical and conformational differences existing between quadruplex and duplex grooves. Unfortunately, the understanding of the chemical, structural, and conformational features responsible for G-quadruplex groove binding is still in its infancy.

In this chapter I report a study of a Distamycin A analogue: **coumpound 3** (figure 1), where the major change into the structure is the replacement of amidinium group by an uncharged N-methyl amide moiety, to probe the real importance of the unique Coulombic interaction in the Distamycin A/$[d(TGGGGT)]_4$ complex formation.

We have used a NMR study to understand the binding mode of the

compound, in particular we have performed a NMR titration of compound 3 with the quadruplex $[d(TGGGGT)]_4$. All the results were confirmed with an Isothermal Titration Calorimetric (ITC) study. A molecular and dynamic study was developed to have the three-dimensional structure of $3/[d(TGGGGT)]_4$ complex.

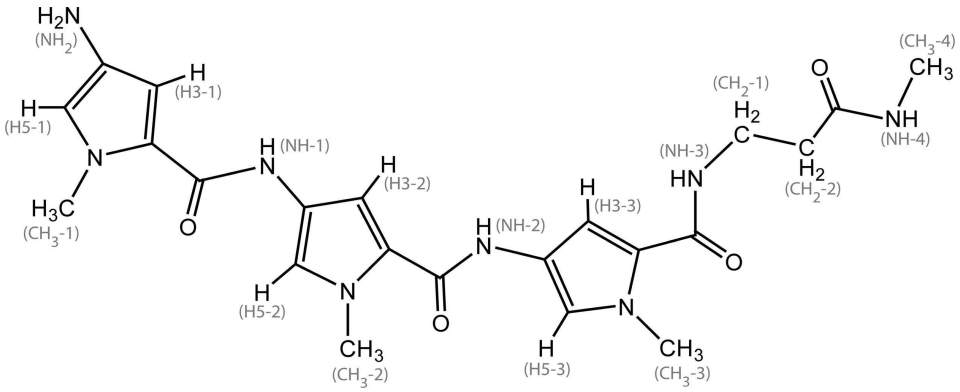


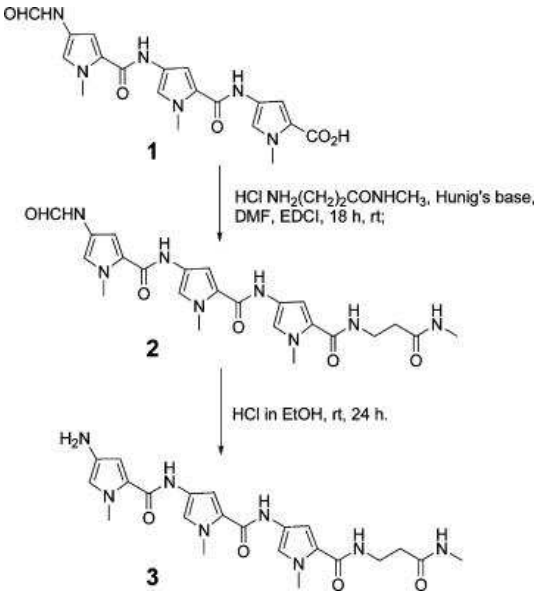
Figure 1: Chemical structure of Coumpound 3

Results and discussion

Synthesis of compound 3.

The novel compound 3 was obtained following the procedure reported in the Scheme 1.

Scheme 1: Chemical synthesis of compound 3.



Starting from derivative 1, which was prepared following the methodology reported in the section Material and method at page 55, we have condensed it with 3-amino-N-methylpropanamide hydrochloride[4] using an excess (2 equiv) of 1-ethyl-3-[3-(dimethylamino) propyl]carbodiimide hydrochloride (EDCI) as coupling agent, in DMF and in presence of Hunig's base at room temperature for 18 h.

This process has led to the N-methyl carbamoyl derivative 2 in good yield after purification by silica gel flashchromatography. The amido modified Distamycin A derivative 2 was then transformed in the corresponding desformyl compound 3 by treatment of HCl in a mixture of ethanol and water. The reason for this last synthetic step was to obtain the hydrochloride compound 3, that turned out to be more soluble in pure H_2O than derivative 2 at the concentration used for the preparation of the NMR samples.

Nuclear Magnetic Resonance Experiments.

To evaluate the binding properties of compound 3, and to perform a direct comparison with the binding behaviour of Distamycin A, $[d(TGGGGT)]_4$ has been titrated with 3 at the same experimental conditions (buffer, temperature, DNA concentration) used for Distamycin A.[3]

As far as Distamycin A is concerned, below 2:1 ligand/quadruplex stoichiometry, the addition of Distamycin A to $[d(TGGGGT)]_4$ caused gradual changing in chemical shift of the signal of the quadruplex, whereas further addition of drug caused the appearance of a new set of proton signals, whose intensities rose by increasing the amount of drug with the concomitant falling off of the original signals which completely disappeared at a ratio of 4:1 drug-DNA.

The NMR titration of compound 3 with the quadruplex is clearly different from that observed in the case of Distamycin A (figure 2).

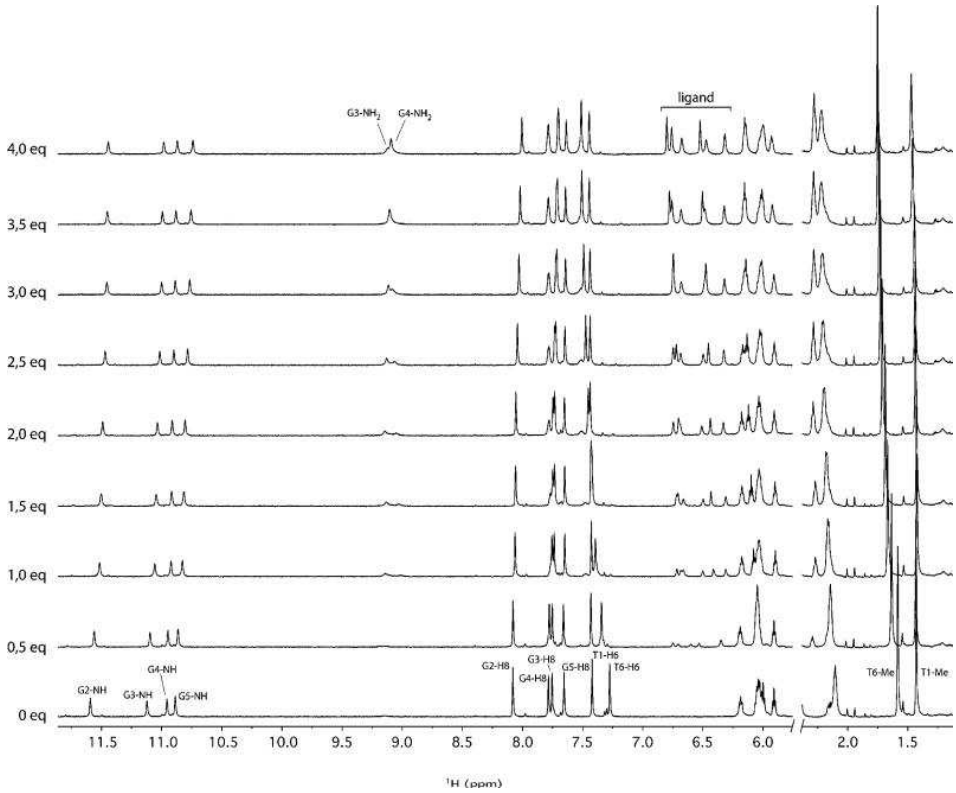


Figure 2: NMR titration of $[d(TGGGGT)]_4$ with compound 3 (700 MHz, T = 25 °C). Equivalents of the drug are reported on the left of each spectrum.

In fact, the addition of 3 to $[d(TGGGGT)]_4$ caused only gradual changes in the chemical shift of DNA proton resonances.

At ligand/DNA ratio of 4:1, the titration was virtually completed. The four strands resulted to be magnetically equivalent throughout the titration, and no splitting of resonances was observed at any stage. Moreover, we have observed that during the whole NMR titration, a single set of signals was present for compound 3 protons, which only grew in intensity and did not show any significant change in chemical shift values by increasing ligand concentration. These observations suggest that compound 3 binds the quadruplex in a fast process on the NMR time scale.

To preliminarily evaluate the binding site of compound 3, a comparison of resonances of some protons of the uncomplexed DNA and the complexed one has been carried out. In particular, I have reported the Δδ values (chemical shifts of the complex minus free DNA) of aromatic, methyl and imino protons in figure 3.

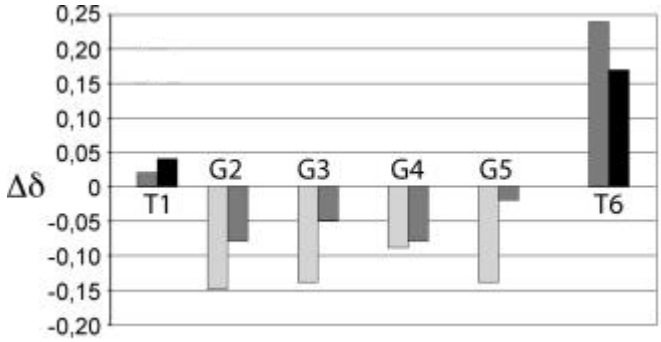


Figure 3: Difference in chemical shifts (Δδ) of DNA upon binding of 3 (complex 3:DNA minus DNA alone) to $[d(TGGGGT)]_4$. Δδ values are reported for imino (light gray), aromatic (dark gray) and methyl (black) protons.

All the analysed resonances shifted led us to conclude that for the signal of the protons of T1 residue there is the least shift, whilst for those of residue T6 the most.

In any case, a general shift of the aromatic and imino signals was observed also for the G2, G3, G4, and G5; this means that compound 3 is able to recognize most of the molecule, even if it recognizes preferentially the 3' edge of the quadruplex.

An almost complete assignment of the nonexchangeable/exchangeable protons of the complex has been accomplished by means of a combination of the analysis of 2D NOESY and TOCSY spectra (700 MHz, T = 25°C), in the

Tables 1 and 2 we can see the NMR assignment, in figure 4 I have reported an expanded region of the NOESY spectrum of the 4:1 3/$[d(TGGGGT)]_4$ complex.

Table 1: ^{1}H-NMR assignment of $[d(TGGGGT)]_4$ complexed with 3 at 25°C (700 MHz).

	H8/H6	H1'	H2'	H2"	H3'	H4'	"H5'	"H5"	H2/Me	NH2	NH	P
T1	7.34	5.84	2.10	2.36	4,62		3,96	3,65	1.36			-1,02
G2	7.91	5.91	2.52	2.83	4,85	4,27	4,26				11.34	-0,97
G3	7.60	5.93	2.46	2.68	4,92	4,13	4,04	3,97		9.00	10.88	-1,30
G4	7.61	6.05	2.54		4,91	4,38	4,12	4,33		8.97	10.77	-1,29
G5	7.54		5.91	2.39	2.67	4,90	4,13	4,19	4,37		10.64	-1,17
T6	7.41		6.05	2.12	2.12	4,84	4,40	4,11	4,36	1.66		

Table 2: ^{1}H-NMR assignment of 3 complexed with $[d(TGGGGT)]_4$ at 25°C (700 MHz).

Proton	δH (ppm)
H3-1	6.42
H3-2	6.37
H3-3	6.22
NH-1	
NH-2	
NH-3	
NH-4	7.70
H5-1	6.70
H5-2	6.66
H5-3	6.57
CH_3-1	3.52
CH_3-2	3.40
CH_3-3	3.37
CH_3-4	2.46
CH_2-1	3.15
CH_2-2	2.16
NH_2	

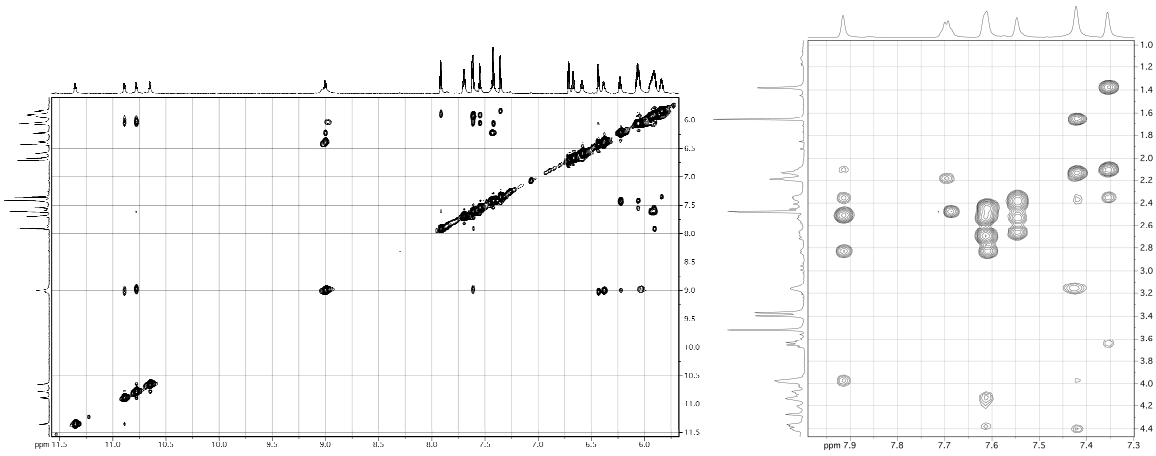


Figure 4: Expanded region of the NOESY (700 MHz, 25°C, mt=100 ms) spectrum of the 4:1 3/$[d(TGGGGT)]_4$ complex.

Using a 2D TOCSY experiment it was possible to identify, for the DNA, the ^1H resonances within each deoxyribose; following the analysis of NOEs among base protons and H1', H2', and H2'' protons allowed us to assign all base protons. The direct comparison of the intensities of the NOESY cross-peaks (700 MHz, T = 25°C, mixing time 100 ms) between the H8 proton bases and sugar H1' resonances, and among H8 proton bases and sugar H2'/H2'' resonances, indicates that all Gs residues of the complexed DNA adopt an anti glycosidic conformation. Then, all bases have classical H8/H2'-H2'' sequential connectivities to 5' neighboring, indicating that the four strands are involved in the formation of a helical structure. The observation of the entire pattern of NOEs has confirmed that the backbone conformation resembles closely to conformation of the uncomplexed $[d(TGGGGT)]_4$ possessing a right-handed B-form helix structure. Each exchangeable proton signal was then assigned to the pertinent hydrogen by an in-depth analysis of

the NOESY spectra. In particular, the NOE contact between H1 proton of DNA at δH 11.34 and the methyl protons at δH 1.34 of the T1 residue led us to identify the imino protons belonging to the tetrad that is in proximity of the 5' edge of the quadruplex. Similarly, the identification of the imino protons belonging to the tetrad in proximity of the 3' edge was confirmed by the NOE between H1 proton at δH 10.64 and the methyl group at δH 1.66 ppm of the thymine T6. The other four H1 resonances of the remaining two central tetrads were identified by analysing the NOE connections with the adjacent tetrads.

The resonances of compound 3 were assigned, first, identifying the resonances belonging to the hydrogen on the convex side of the molecule. Particularly, the H5 pyrrole protons (δH 6.70, 6.66, and 6.57 ppm for H5-1, H5-2 and H5-3, respectively) have been assigned taking into account that they both exhibit scalar and dipolar coupling with methyls that are linked to the same pyrrole ring (δH 3.52, 3.40, and 3.37 ppm for CH3-1, CH3-2, CH3-3, respectively). As for the protons on the concave side of compound 3, the H3 pyrrole protons (δH 6.42, 6.37, and 6.22 for H3-1, H3-2 and H3-3, respectively) have been assigned by virtue of very small scalar coupling with H5 protons of the same pyrrole ring. Unfortunately, we were not able to assign NH-1, NH-2 and NH-3 protons due to the presence of ambiguous NOEs with the adjacent H3s. The NOESY spectrum of the complex contains a number of intermolecular ligand-ligand and ligand-DNA NOEs, in addition to intramolecular ones.

From this careful analysis of NMR spectra we identified ligand-ligand

contacts, 12 head to tail NOEs (for each molecule of 3, 24 for each dimer) (Table 3), that unambiguously indicates that the ligand molecules bind to the quadruplex, two by two, with each term of the dimeric pairs with an antiparallel orientation and in close contact to its partner (as observed for the binding of Distamicyn A to both duplex and quadruplex DNA).[3,5]

Table 3: Intermolecular Head to Tail Ligand-Ligand NOE Contacts.

ligand proton	ligand proton
H3-1	CH_3-4,CH_3-3, H5-3
H5-1	CH_2-1,CH_3-4, CH_3-3
H3-2	CH_3-3, H3-3, H5-3
H5-2	CH_3-4
H3-3	CH_3-1
H5-3	CH_3-1

In addition, 14 NOEs were observed between each molecule of compound 3 and $[d(TGGGGT)]_4$, in particular we found a lot of contact with residues G4, G5, and T6 of DNA. Analogously to that observed for Distamycin A, also in this case, there are no contacts between compound 3 and the imino protons of the quadruplex (since these point inside the quadruplex core) (Table 4).

Table 4: Intermolecular Ligand-DNA NOE Contacts.

ligand proton	DNA proton
H3-1	T6:H2′, T6:H1′,G5:H2′, G3:NH2
H5-1	G5:H1′, G5:H2′, G4:H2′
CH_3-1	T6:H1′
H3-2	T6:H1′
H5-2	G5:H2′
H3-3	T6:H6, G5:H1′, G4:H2′
H5-3	G4:H2′, G5:H1′

All this suggests that compound 3 actually does not stack on the edges of the

quadruplex, even if it prefers to interact with the 3' edge of the quadruplex.

To further confirm that compound 3 is able to bind the grooves of the quadruplex, as its analogue Distamicyn A, we have designed and synthesized a modified oligonucleotide, namely, d(TGG^{Br}GGT), where dG^{Br} is 8-bromo-2'-deoxyguanosine, potentially capable to form quadruplex structure and possessing a bulky group (bromine) at the very center of the grooves.

We have performed a titration of compound 3 with the modified quadruplex d(TGG^{Br}GGT) and then analyzed the titration profile.

If compound 3 interacts with the portion of the grooves of the quadruplex where the bromine is present, this should prevent (or at least should limit) the insertion of molecules of compound 3 into the grooves.

At the first stage of the analysis we have tested the capability of d(TGG^{Br}GGT) to form a quadruplex structure, preparing a NMR sample of d(TGG^{Br}GGT) at a concentration of 2 mM, in 0.6 mL (H_2O/D_2O 9:1) buffer solution having 10 mM KH_2PO_4, 70 mM KCl, 0.2 mM EDTA, pH 7.0. The sample was annealed for 5-10 min at 80°C and slowly cooled down to room temperature, then ^{1}H NMR spectra were recorded using DPFGSE pulse sequence for H_2O suppression.[6,7] The ^{1}H NMR spectrum (700 MHz, T = 25°C) of d(TGG^{Br}GGT) shows the presence of four well-defined singlets in the region 11-12 ppm, attributable to imino protons involved in Hoogsteen hydrogen bonds of G-quartets, and the presence of five signals (figure 5), belonging to three guanine H8 and two thymine H6 protons in the aromatic region.

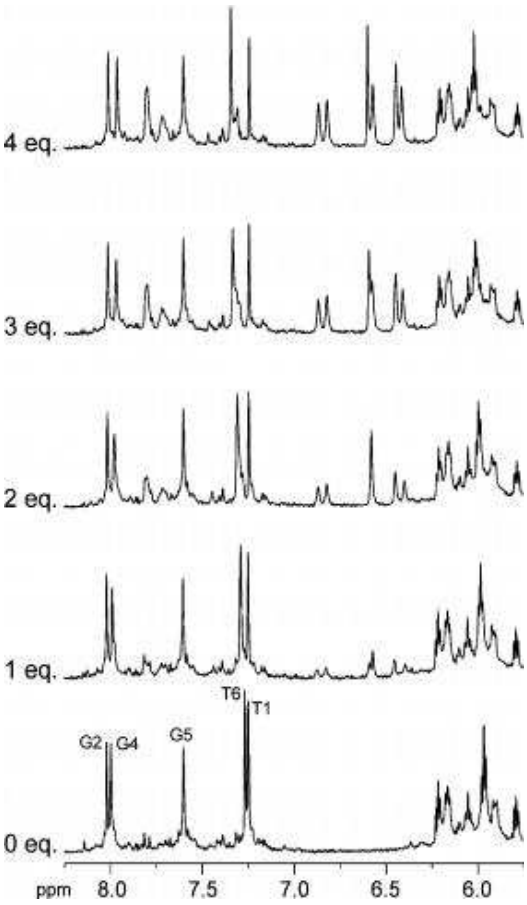


Figure 5: Expanded regions of NMR titration of $[d(TGGBrGGT)]_4$ with compound 3 (700 MHz, T=25 °C). Equivalents of the drug are reported on the left.

In figure 5 is clearly indicated that a single well-defined quadruplex species is present in solution, consisting of four G-tetrads and possessing a 4-fold symmetry with all strands parallel to each other. In a quadruplex like this, each Br group of the quadruplex faces right into the grooves, pointing outward the quadruplex.

The quadruplex $[d(TGG^{Br}GGT)]_4$ has been titrated with compound 3 and the results are reported in figure 5. The progressive increase of concentration of compound 3 up to 4 mol equiv caused only a slight drift of DNA signals and

drug resonances to gradually grow in intensity (between 6 and 7 ppm). Further additions of drug did not lead to significant changes. As suggested also by the Δδ values of DNA signals (figure 6), the extent of chemical shift perturbation is lighter in comparison to the unmodified quadruplex.

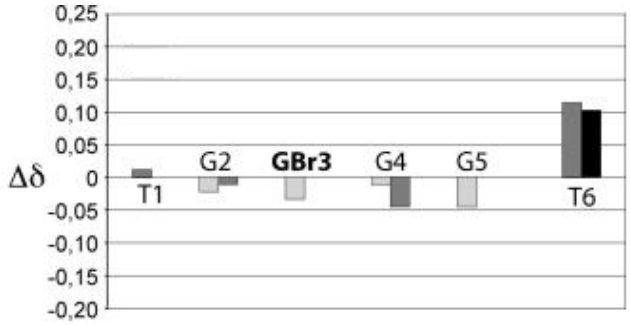


Figure 6: Difference in chemical shifts (Δδ) of DNA upon binding of 3 (complex 3:DNA minus DNA alone) to $[d(TGG^{Br}GGT)]_4$. Δδ values are reported for imino (light gray), aromatic (dark gray) and methyl (black) protons. No aromatic hydrogens are present for brominated residues.

This strongly indicates that, actually, the presence of the bromines at the very center of the grooves does affect the recognition process, and that compound 3 binds the grooves of $[d(TGGGGT)]_4$. This also suggests that, in spite of compound 3 which does not show any NOE contacts with the 5' edge of the groove, it is actually able to span the entire groove.

Structure Calculations.

To obtain the three-dimensional structure of the 4:1 complex at atomic level, an estimation of proton-proton distances has been retrieved from cross-peak intensities in 2D NOESY experiments (700 MHz, T = 25°C). A total of 278 experimental distance restraints were used for the calculations, and as suggested by the presence of eight guanine imino protons in the 1D ^{1}H NMR spectrum, 32 supplementary distance restraints (HN1-O6, HN2-N7) for 32 hydrogen bonds corresponding to the four G-quartets were also incorporated during the computations (Table 5).

Table 5: Structural data for the NMR Restraints and Statistics for the 10 best structures.

total no. of experimental NOE restraints	278
Quadruplex intraresidual restraints	96
Quadruplex interresidual restraints	68
Intermolecular ligand-quadruplex restraints	62
Intermolecular ligand-ligand restraints	52
Dihedral angles (α, β, γ, δ ε and χ) restraints	96
Quadruplex H-bonds restraints	32
Planarity restraints for G bases	56
Average rmsd (all atoms excluding terminalT bases)	0.84 Å
Energy (kcal/mol)	
Mean amber energy	-1610.42 (62.29
NOE distance restraints violation energy	84.34 (4.82
Torsion angle restraints violation energy	184.10 (18.29
Restraint Violations	
Distance (>0.3 Å)	7
Experimental dihedral angles (>10 Å)	2

We have collected also restraints for a number of backbone torsion angles; the combined analysis of the PE-COSY and zTOCSY experiments revealed that all measurable $J_{H1',H2'}$ and $J_{H1',H2''}$ ranged from 7 to 9 Hz and from 5 to 6.5 Hz, respectively, and that a number of cross-peaks for H3'/H4' correlations

were missing. Furthermore, the measurable $J_{H3', H4'}$ values were very low.

This data indicates a predominant S-type nature of sugar ring conformations. Therefore, the relative δ angles constraints have been consistently constrained into the range 95°/175°.[8] Very interesting information could be also retrieved from the analysis of the sum of H4′ couplings constant ($\Sigma J_{H4'} = J_{H4',H3'} + J_{H4',H5'} + J_{H4',H5''}$). All measured $\Sigma J_{H4'}$ turned out to be <10 Hz. This, along with the lack of intense NOEs between H8 and any H5′/H5″ protons, suggested that γ angles could be constrained in the range 20°/100°.[8] As far as β and ε angles are concerned, their estimation have been performed measuring the scalar coupling in the 2D proton-detected heteronuclear ^{1}H-^{31}P COSY and by using the semiempirical Karplus equation.[8] Thus, since for all $J_{P,H5'}$ and $J_{P,H5''}$ were <8 Hz, β backbone torsion angles were restricted to -230°/ -110°.[8] On the contrary, $J_{P,H3'}$ for residues 2, 3, 4, and 5 were larger than 8 Hz, and the corresponding ε angles were restricted to -230°/-110°.[8] Furthermore, since α (O3′-P-O5′-C5′) and ζ (C3′-O3′-P-O5′) dihedral angles do not involve any protons, it has been impossible to directly retrieve structural information for these angles. Nevertheless, since in a structured DNA α and ζ angles cannot be considered independent from β, γ, δ and ε, we thought opportune to restraint at least one of these angles; thus, α angles have been constrained in the wide range -150°/-30°. Finally, due to the presence of very weak NOEs between H8 and H1′ hydrogens, glycosidic torsion angles χ were fixed in the anti-domain (-155°/-75°). Therefore, 3D structures which satisfy NOEs and dihedral angles were constructed by simulated annealing (SA)

calculations. A total of 100 structures were generated, and among them, the 10 best models were selected based on the value of the overall potential energy and NMR restraint violations for further simulations. In figure 7 is shown the representative ligand/DNA conformations (it is the ones with the lowest rmsd value with respect to the average structure calculated over the whole production run).

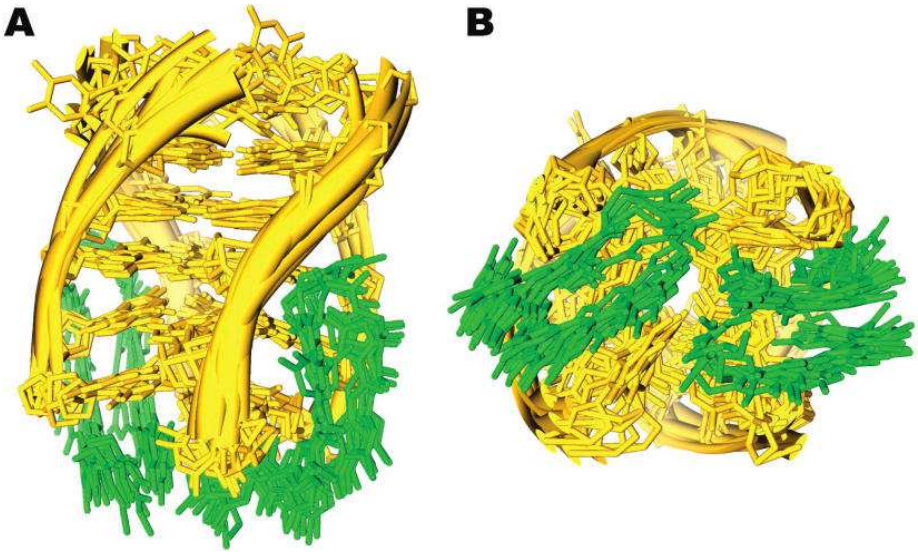


Figure 7: A) Side view and B) bottom view of the superimposition of the 10 best structures of the 4:1 3/[d(TGGGGT)]$_4$ complex. compound 3 is reported in green; DNA is colored in yellow.

As expected, differently from what demonstrated for Distamycin A, compound 3 dimers are shifted toward the 3' end of the quadruplex. The two staggered antiparallel 3 molecules overlap for their total length with each N-methylpyrrole ring of one molecule contacting the terminal amide bond of the other, thus, making synchronized twisting to easily fit the curvature of the DNA quadruplex grooves. Moreover, one molecule of the dimer of compound 3 is more solvent exposed, while the other is sandwiched

between the first one and the quadruplex structure (herein referred as 3ex and 3in, respectively). The preference for such an orientation with respect to the G-quadruplex could be ascribable to the absence of a positively charged anchoring point that in Distamycin A stabilizes each monomer so as to project the methyl groups toward the solvent and the opposite amide protons toward the crevice of the groove.

Compound 3 interact with the quadruplex in an antiparallel fashion, as its analogue Distamycin A, this is a very interesting result because, in theory, two possible orientations could be predicted for a dimer of 3: parallel and antiparallel. This behavior could be imputed to the charge-transfer π-π interactions and to the energetically favorable dipole-dipole interactions established between the monomers as demonstrated by the dipole moments of the two staggered structures which are pointing in opposite directions (figure 8).

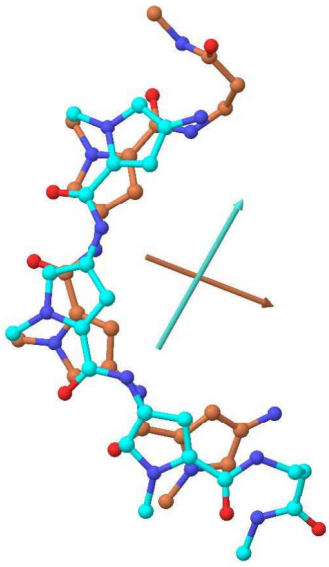


Figure 8: NMR structure of the compound 3 dimer. Dipole moments are displayed as arrows.

The crescent shape of the ligand allows the formation of favorable interactions with the quadruplex by following the curvature of the grooves in the G4-G5 region and taking contact with T6 at the 3′ end.

Comparing our structural results with the dates of the Distamycin A, is clear that in both compounds a crescent shape is adopted, but the conformation of compound 3 alone does not guarantee a pure groove binding mode in the $[d(TGGGGT)]_4$; this is an important difference with the Distamycin A because in the Distamycin A the crucial anchoring positively charged point (amidinium group) and the crescent shape conformation are both responsible for a pure groove binding mode.

Calculation of the Delphi electrostatic potential surface reveals that the terminal methylcarboxamide groups of the compound 3 dimer are hosted in rather polar regions of the quadruplex while the terminal pyrrole ring is

exposed to the solvent in the G4 region in 3ex or embedded in a less hydrophilic environment in the T6 region in 3in (figure 9).

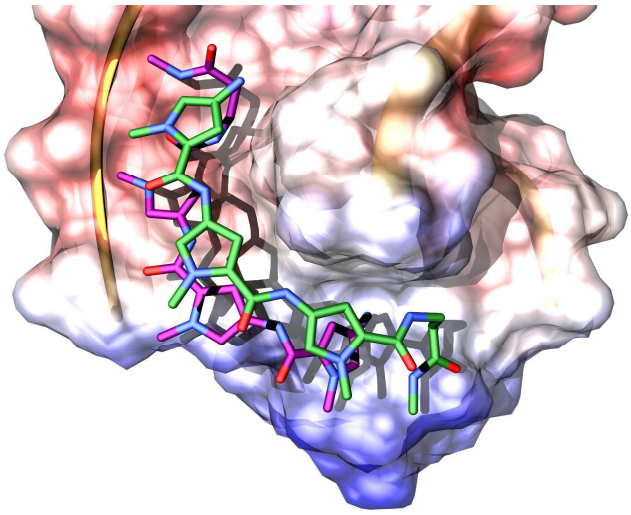


Figure 9: Electrostatic potential surface of the representative structure. The surface of the quadruplex is colored according to electrostatic potential, more positive potentials are colored in blue and more negative potentials are colored in red. 3in and 3ex are represented as magenta and green sticks, respectively.

A close-up view of the best binding pose resulting from NMR constrained SA experiments reveals that the aforementioned terminal groups of 3 are involved in a network of H-bonds with the quadruplex (figure 10).

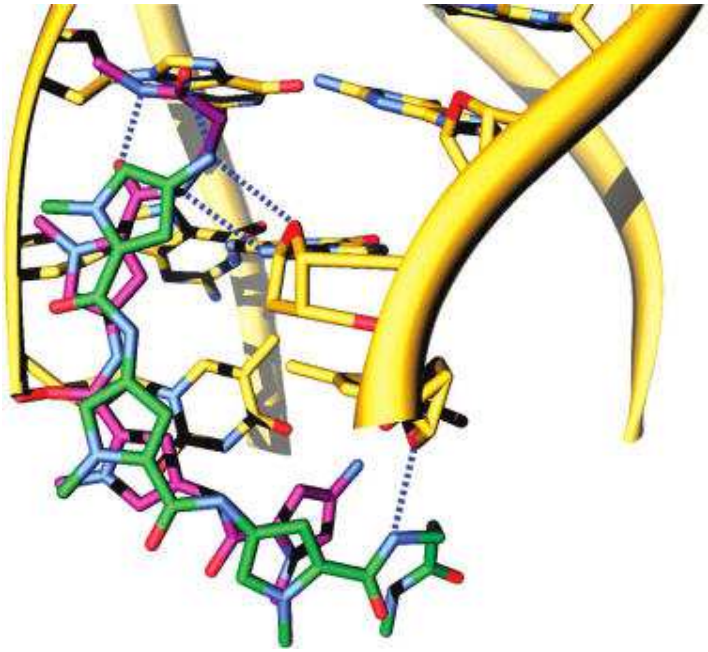


Figure 10: Side view representation of the 4:1 complex 3/[d(TGGGGT)]$_4$. 3in, 3ex, and [d(TGGGGT)]$_4$ are depicted in magenta, green, and yellow sticks, respectively.

Most precisely, the flexible carboxamide terminal group of the 3in molecule is kept in place by an intramolecular H-bond between its NH-4 proton and the carbonyl oxygen attached to the third pyrrole ring and by an intermolecular one established between the terminal carbonyl oxygen of 3in and the terminal NH2 group of 3ex. Moreover, the G5 base is contacted by the dimer through two H-bonds between its O4′ and its N3 atoms with the 3ex NH2 and 3in NH-4 groups, respectively. While polar contacts mainly govern the recognition between the compound 3 dimer and the center of the quadruplex groove, rather hydrophobic interactions are established with the 3′ end of the [d(TGGGGT)]$_4$ structure. Notably, the terminal pyrrole ring of the 3in molecule is embedded in a sort of aromatic cage formed by the two T6 bases of the groove and the third pyrrole of the 3ex molecule.

Isothermal Titration Calorimetry Measurements.

Thanks to the great advances in the sensitivity and reliability of the calorimeter, ITC has become an important tool for the direct measurement of thermodynamic parameters in several biological interactions. Moreover, it has been proven to be useful to study the energetic aspects of interactions between G-quadruplexes and other molecules, including small ligands.[9] Therefore, we have utilized ITC to characterize the thermodynamics of binding of compound 3 to $[d(TGGGGT)]_4$ quadruplex.

Figure 12 shows the binding isotherm resulting from the integration of raw data after correction for the heat of ligand dilution. The calorimetric data indicates an endothermic interaction.

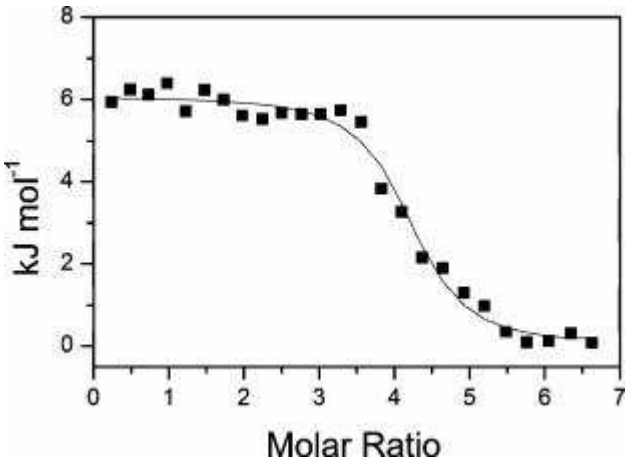


Figure 12: Normalized heat of interaction between compound 3 and $[d(TGGGGT)]_4$. The squares represent the experimental data obtained by integrating the raw data and subtracting the heat of ligand dilution into the buffer. The line represent the best fit obtained by a nonlinear least-squares procedure based on a single set of identical sites model.

After a number of injections of ligand, increasingly less heat uptake was observed until constant values were obtained (corresponding to the heat of dilution), reflecting a saturable process. The values of K_b and ΔH° derived by ITC enable us to complete the thermodynamic binding profiles by calculating corresponding values of TΔS° and ΔG°. The resulting thermodynamic parameters are showed in Table 6.

Table 6: Thermodynamic Parameters for the Interaction of Compound 3 with $[d(TGGGGT)]_4$.Determined by ITC at 25 °C[a]

Drug	*n*	K_b (M^{-1})	$\Delta H°$ (kJ mol^{-1})	$T\Delta S°$ (kJ mol^{-1})	$\Delta G°298K$ (kJ mol^{-1})
3	4	9×10^5	6	40	-34

[a] The experimental error on each thermodynamic property is <10%.

The binding affinity measured from ITC is 9 x 10^5 M^{-1} at 25°C. The binding enthalpy was found to be positive (ΔH°= 6 kJ mol^{-1}) and the stoichiometry of interaction was found to be 4:1 ligand-quadruplex.

Inspection of the data reveals that the entropic contribution (TΔS° = 40 kJ mol^{-1}) to the binding of compound 3 to the quadruplex provides the driving force for the ligand-quadruplex binding reaction, while the positive heat of formation of the complex indicates that this interaction is an enthalpically unfavorable process. The results from the calorimetric titration data alone do not allow us to demonstrate with certainty the binding mode of this ligand. Usually, small molecules bind to duplex DNA by two predominant binding modes, intercalation and groove binding. Calorimetric studies have determined the enthalpic and entropic contributions to the DNA binding of representative DNA binding compounds. Generally, the binding of an

intercalator to DNA is driven by a large favorable enthalpy change and by an unfavorable entropy decrease, while the binding of a groove binder to DNA is characterized by a large favorable increase of entropy and a small favorable or unfavorable enthalpy change.[10]

In the case of the binding of compound 3 to $[d(TGGGGT)]_4$, the thermodynamic data determined by ITC are in agreement with the groove binding profile and are confirmed by the structure of the complex. The thermodynamic parameters for the interaction of compound 3 are slightly different from the ones determined by ITC for the binding of Distamycin A to the same quadruplex molecule. In both cases, the driving force for the binding reaction is the entropy change. However, the binding of compound 3 shows a small unfavorable enthalpy change, while in the case of Distamycin A, a small favorable enthalpy change was observed. This interesting difference is probably due to structural features of the two ligands, probably the positively charged amidinium moiety of Distamycin A interacts with the phosphate groups of the quadruplex, providing a favourable (although small) enthalpy contribution, while in the compound 3, the amidinium group is replaced by the uncharged N-methyl amide moiety that cannot give this contribution.

In summary, the structural data herein reported unveil the role of the Coulombic interaction engaged by Distamicyn A and the quadruplex. Such an interaction not only influences the strength of ligand/quadruplex formation, but surprisingly governs the orientation of the ligand with respect

to the DNA. In fact, the removal of the positively charged terminal group results in an unprecedented ligand binding position in which both the groove and the 3' end of the DNA are occupied. Furthermore, the lack of charge in the ligand does not affect the relative orientation of two molecules of compound 3 forming a dimer. All this opens up the tempting opportunity for a fine-tuning of the drug-quadruplex interaction mode. Moreover, with this further contribution, we provided the scientific community with an enhanced knowledge of the structural and conformational demands of the quadruplex groove that will serve as a platform for a rational design of novel groove binding agents.

Material and method

Oligonucleotide synthesis

The oligonucleotide d(TGGGGT) was synthesized on a Millipore Cyclone Plus DNA synthesizer using solid phase β-cyanoethyl phosphoramidite chemistry at 15 μmol scale. Commercially available 5'-DMT-aminoprotected-8-bromodeoxyguanosine- 3'-phosphoramidite was used for the preparation of the modified oligonucleotide d(TGG^{Br}GGT). The latter was assembled using the standard solid phase β-cyanoethylphosphoramidite chemistry. The oligomers were detached from the support and deprotected by treatment with concentrated aqueous ammonia at 55 °C for 12 h. The combined filtrates and washings were concentrated under reduced pressure, redissolved in H_2O, analyzed, and purified by high-performance liquid chromatography (HPLC) on a Nucleogel SAX column (Macherey-Nagel, 1000-8/46), using buffer A, 20 mM KH_2PO_4/K_2HPO_4 aqueous solution (pH 7.0), containing 20% (v/v) CH_3CN; buffer B, 1 M KCl, 20 mM KH_2PO_4/K_2HPO_4 aqueous solution (pH 7.0), containing 20% (v/v) CH_3CN; a linear gradient from 0 to 100% B for 30 min and flow rate 1 mL/min were used. The fractions of the oligomer were collected and successively desalted by Sep-pak cartridges (C-18). The isolated oligomers proved to be >98% pure NMR.

Synthesis of N-[5-({[5-({[3-Amino-3-(N-methylcarbamoyl)-ethyl]amino}carbonyl)-1-methyl-1H-pyrrol-3-yl]amino}carbonyl)-1-methyl-1H-pyrrol-3-yl]-4-(formylamino)-1-methyl-1H pyrrole-2-carboxamide (2).

A solution of 3-amino-N-methylpropanamide hydrochloride (123 mg, 1.2 mmol) in dry DMF (8 mL) was cooled to 5 °C and N, N'-diisopropylethylamine (206 mL, 1.2 mmol) was added. After 10 min, the acid 1 (412 mg, 1 mmol.) and then EDCI (384 mg, 2 mmol) were added. The reaction mixture was stirred at room temperature for 18 h, the solvent evaporated in Vacuo, and the crude residue purified by flash chromatography. The crude product purified by chromatography on a silica gel column, using DCM/MeOH (8:2) as eluent, furnished the compound 2 as a white solid (367 mg, 74% yield). Mp = 156-158 °C; IR (KBr): 3420, 3310, 1665, 1554, 1312, and 1212 cm-1; ^{1}H NMR (DMSO-d6) δ: 2.42 (m, 2H), 2.77 (m, 3H), 3.34 (m, 2H), 3.78 (s, 3H), 3.82 (s, 3H), 3.83 (s, 3H), 6.65 (d, J) 1.6 Hz, 1H), 6.83 (d, J) 1.6 Hz, 1H), 6.82 (d, J) 1.6 Hz, 1H), 7.01 (d, J) 1.6 Hz, 1H), 7.17 (d, J) 1.6 Hz, 1H), 7.22 (d, J) 1.6 Hz, 1H), 7.89 (bs, 1H), 8.08 (s, 1H), 8.13 (s, 1H), 8.18 (s, 1H); 9.91 (s, 1H), 9.94 (s, 1H). FABMS (MALDI-TOF): 497.8 $[M + 1]^{+}$.

Synthesis of 4-Amino-N-[5-({[5-({[3-amino-3-(N-methylcarbamoyl)ethyl]amino}carbonyl)-1-methyl-1H-pyrrol-3-yl]amino}carbonyl)-1-methyl-1H-pyrrol-3-yl]-1-methyl-1H-pyrrole-2-carboxamide (3).

To 5 mL of a solution obtained by the addition of 24 mL of ethanol (95%) to 1 mL of 36% HCl in water was added compound 2 (248 mg, 0.5 mmol), and the mixture stirred at room temperature for 24 h. After this time, the reaction was filtered, the residue washed with dry ethyl ether (5 mL), and then dried on P_2O_5, to furnish compound 3 as brown solid. Mp = 262-264 °C; ^{1}H NMR (DMSO-d6) δ: 2.41 (m, 2H), 2.56 (m, 3H), 3.24 (m, 2H), 3.78 (s, 3H), 3.80 (s, 3H), 3.81 (s, 3H), 6.75 (d, J) 1.6 Hz, 1H), 6.82 (d, J) 1.6 Hz, 1H), 6.84 (d, J) 1.6 Hz, 1H), 6.96 (d, J) 1.6 Hz, 1H), 7.04 (d, J) 1.6 Hz, 1H), 7.20 (d, J) 1.6 Hz, 1H), 7.88 (bs, 1H), 8.12 (s, 1H), 9.92 (s, 1H), 9.96 (s, 1H), 10.20 (bs, 3H). FAB-MS (MALDI-TOF): 469.4 [M + 1]$^+$.

Nuclear Magnetic Resonance Experiments.

The quadruplex NMR sample was prepared at a concentration of 2 mM (8 mM single strand concentration), in 0.6 mL (H_2O/D_2O 9:1) buffer solution having 10 mM KH_2PO_4, 70 mM KCl, 0.2 mM EDTA, pH 7.0. For D_2O experiments, the H_2O was replaced with D_2O by drying down the sample, followed by

lyophilisation and redissolution in D_2O alone. NMR spectra were recorded with Varian ^{Unity}INOVA 700 MHz spectrometer. ^{1}H chemical shifts were referenced relative to external sodium 2,2-dimethyl-2-silapentane-5-sulfonate (DSS). 1D proton spectra of the sample in H_2O were recorded using pulsed-field gradient DPFGSE[6,7] for H_2O suppression. Phase-sensitive NOESY spectra[11] were recorded with mixing times of 100 and 200 ms (T = 25 °C). Pulsed-field gradient DPFGSE[6,7] sequence was used for NOESY experiments in H_2O. TOCSY[12] and zTOCSY[13] spectra with mixing times of 100 ms were recorded with D_2O solution. All experiments were recorded using STATES-TPPI[14] procedure for quadrature detection. In all 2D experiments, the time domain data consisted of 2048 complex points in t2 (16K in the zTOCSY) and 400-512 fids in t1 dimension. The relaxation delay was kept at 3s for NOESY experiments used in the structure determination. A relaxation delay of 1.2 s was used for all other experiments. The NMR data were processed on a SGI Octane workstation using FELIX 98 software (Accelrys, San Diego, CA) and on iMAC running iNMR software (www.inmr.net).

Isothermal Titration Calorimetry.

ITC measurements were conducted at 25 °C on a CSC 4200 Calorimeter from Calorimetry Science Corporation (Lindon, UT). In these experiments, 10 µL aliquots of a 1250 µM ligand solution were injected into a sample cell containing 1300 mL of 40 µM quadruplex solution, with stirring at 297 rpm, for

a total of 25 injections. The delay between injections was 400 s. Each ligand injection produced a heat burst curve. The areas under these heat burst curves were determined by integration to yield the associated injection heats. The heats of dilution were determined in parallel experiments by injecting a ligand solution of the same concentration in the same buffer. The heat of dilution was subtracted from the interaction heats prior to curve fitting. The corrected heat values are plotted as a function of the molar ratio, to give the corresponding binding isotherms. The resulting isotherms were then fitted to a single set of identical sites model using the Bindwork program supplied with the instrument, to give the binding enthalpy (ΔH°), equilibrium binding constant (K_b), and stoichiometry (n). The remaining thermodynamic parameters, ΔG° and TΔS°, were derived using the standard relationships ΔG° = -RT ln K_b and TΔS° = ΔH° - ΔG°. The buffer conditions used for ITC measurements were the same of NMR experiments.

Structure Calculations

Restrained Simulated Annealing (SA) calculations were performed using the AMBER 10.0 package,[15] with DNA described by the latest AMBER force field for nucleic acids (ff99bsc0),[16,17] while compound 3 was parameterized using the General AMBER Force Field.[18] The Jaguar software package (Schrodinger) was used to calculate atomic charges using DFT with an HF/6-31G** basis set at the B3LYP level. For the initial AMBER model, compound 3

was manually docked into two opposite grooves of the $[d(TGGGGT)]_4$ using as a reference structure the Distamycin A/$[d(TGGGGT)]_4$ solution structure (PDB code 2JT7).[12] In fact, given the presence of many head to tail ligand-ligand NOEs, as happened for Distamycin A, also compound 3 is expected to bind the quadruplex structure in a dimeric antiparallel form. Each 3 dimer was then placed into the quadruplex groove consistently with what recorded in NMR experiments. For annealing simulations, the General Born solvation (igb = 2) with monovalent salt concentration corresponding to 0.1 M was used. The complex was heated to 900 K in the first 5 ps, cooled to 100 K for the next 13 ps, and then cooled to 0 K for the last 2 ps. The temperature of the system was maintained with a varying time constant: 0.4 ps during heating, 4 ps during cooling to 100 K, 1 ps for the final cooling stage, and then reduced from 0.1-0.05 for the last picosecond. The force constants for NOE constraints were increased from 3 to 30 kcal mol^{-1} Å-2 during the first 5 ps and then maintained constant for the rest of the simulation. These force constants were applied in the form of a parabolic, flat-well energy term where r is the model distance or torsion angle and k is the respective force constant.

$$\mathbf{E_{constraint}} = k(r2 - r)2 \qquad r1 \le r \le r2$$

$$\mathbf{E_{constraint}} = 0 \qquad r2 \le r \le r3$$

$$\mathbf{E_{constraint}} = k(r3 - r)2 \qquad r3 \le r \le r4$$

The values for r1 and r4 represent upper and lower distance bounds, defining

the linear energetic penalty before and after the flat-well energy term. The applied distance NOE constraints were retrieved by cross-peak volume integrations performed with the program FELIX 98 (Accelrys, San Diego, CA), using the NOESY experiment collected at mixing time of 100 ms. The NOE volumes were then converted to distance restraints after they were calibrated using known fixed distances. The NOE restraints were generated with three distance classifications as follows: strong NOEs (r2 = 1.0 Å; r3 = 3.5 Å), medium NOEs (r2 = 3.0 Å; r3 = 4.5 Å), and weak NOEs (r2 = 4.0 Å; r3 = 6.0 Å). Distance constraints between base pairs, set at half the NOE distance constraint strength, were applied to maintain hydrogen bonding (as indicated by the observation of imino proton resonances). Hydrogen bonds constraints were used in the range of 1.7 Å (r2)/2.3 Å (r3). Planarity force constraints of 200 kcal mol^{-1} rad^{-2} were applied to G bases throughout the simulations. In agreement with NMR data, the backbone torsion angles α, β, γ, δ and ε were restrained in the range -150°/-30°, -230°/-110°, 20°/100°, 95°/ 175°, and -230°/-110°, respectively.[8] Further, glycosidic torsion angles χ were fixed in the anti domain (-155°/-75°). Interactions within the system were calculated. An unrestrained energy minimization step completed the simulated annealing run. This simulated annealing/energy minimization procedure was repeated 100 times. Ten best models were selected based on the value of the overall potential energy and NMR restraint violations for further analysis. After charge neutralization by the addition of 20 K^+ ions, the complexes were solvated with 16758 water molecules in a TIP3P pre-

equilibrated box.[19] Several equilibration steps were performed comprising minimization of the solvent molecules with the DNA and ligand fixed, minimization of the whole system, and slow heating to 300 K with weak positional restraints on DNA and drug atoms under constant-volume conditions. The following 5 ns production runs were applied in the NPT ensemble. The particle mesh Ewald method[20] was used to evaluate the electrostatic interactions with a direct space sum cut off of 10 Å. With the bond lengths involving hydrogen atoms kept fixed with the SHAKE algorithm, a time step of 2 fs was employed.[21] Related conformational substates populated during the molecular dynamics simulation were analyzed with the AMBERS's PTRAJ module.[22] For the 10 trajectories, the average configuration was taken as a reference for subsequent mass-weighted rmsd calculations considering all atoms excluding terminal T bases. The single snapshot of each trajectory with the lowest rmsd was taken as the representative dynamic structure of that simulation.

The final set of coordinates has been deposited in the Protein Data Bank (accession code: 2KVY).

References

1 Neidle S. **FEBS J.** 2010, 277, 1118–1125.

2 Pagano B., Giancola C. **Curr. Cancer Drug Targets** 2007, 7, 520–540.

3 Martino L., Virno A., Pagano B., Virgilio A., Di Micco S., Galeone A., Giancola C., Bifulco G., Mayol L., Randazzo A. **J. Am. Chem. Soc.** 2007, 129, 16048–16056.

4 Juaristi E., Quintana D., Lamatsch B., Seebach D. **J. Org. Chem.** 1991, 56, 2553–2557.

5 Pelton J. G., Wemmer D. E. **J. Am. Chem. Soc.** 1990, 112, 1393–1399.

6 Hwang T. L., Shaka A. J**. J. Magn. Reson.** 1995, A112, 275–279.

7 Dalvit C. J. Biomol. **NMR** 1998, 11, 437–444.

8 Kim S., Lin L., Reid B. R. **Biochemistry** 1992, 31, 3564–3574.

9 Pagano B., Mattia C. A., Giancola C. **Int. J. Mol. Sci.** 2009, 10, 2935–2957.

10 Chaires J. B. **Arch. Biochem. Biophys.** 2006, 453, 26–31.

11 Jeener J., Meier B., Bachmann H. P., Ernst R. R. **J. Chem. Phys.** 1979, 71, 4546–4553.

12 Braunschweiler L., Ernst R. R. **J. Magn. Reson.** 1983, 53, 521–528.

13 Trippleton M. J., Keeler J. **Angew. Chem., Int.** Ed. 2003, 42, 3938– 3941.

14 Marion D., Ikura M., Tschudin R., Bax A. **J. Magn. Reson.** 1989, 85, 393–399.

15 Case D. A. AMBER; 10th ed.; University of California: San Francisco,CA, 2008.

16 Wang J. M., Cieplak P., Kollman P. A. **J. Comput. Chem.** 2000, 21, 1049–1074.

17 Perez A., Marchan I., Svozil D., Sponer J., Cheatham T. E. 3rd, Laughton C. A., Orozco M. **Biophys.** J. 2007, 92, 3817–3829.

18 Wang J. M., Wolf R. M., Caldwell J. W., Kollman P. A., Case D. A. **J. Comput. Chem.** 2004, 25, 1157–1174.

19 Jorgensen W. L., Chandrasekhar J., Madura J. D., Impey R. W., Klein M. L. **J. Chem. Phys.** 1983, 79, 926–935.

20 (a) Darden T., York D., Pedersen L. **J. Chem. Phys.** 1993, 98, 10089–10092. (b) Essmann U.,Perera L., Berkowitz M. L., Darden T., Lee H., Pedersen L. G. **J. Chem. Phys.** 1995, 103, 8577–8593.

21 van Gunsteren W. F., Berendsen H. **J. C. Mol. Phys.** 1977, 34, 1311– 1327.

22 Shao J., Tanner S. W., Thompson N., Cheatham T. E. **III. J. Chem. Theory Comput.** 2007, 3, 2312–2334.

Tandem application of Virtual Screening and NMR experiments in the discovery of brand new DNA Quadruplex groove binders.

To find new quadruplex groove binding agents, the chemical nature of the quadruplex grooves must be considered. Interestingly, all known quadruplex structures are characterized by grooves that are chemically and conformationally very different from the minor groove of the duplex DNA. This means that searching for new quadruplex groove binding agents among duplex minor groove binders may not be the most successful strategy. For this reason, a number of research groups have recently focused their attention on finding alternative molecular scaffolds able to recognize the groove of the quadruplex.[1,2] In this work, I attempted to search for brand new molecular scaffolds able to interact with the groove of quadruplex structures by means of a structure-based virtual screening (VS) approach.[3]

Actually, whereas there are numerous studies using these screening methods for targeting proteins, only few VS campaigns have been undertaken targeting nucleic acids, and, to the best of our knowledge, none targeting the groove of the quadruplex. In a recent review, Trent and co-workers[4] have outlined that the software Autodock optimally balances docking accuracy and ranking. Thus, as a starting point, this program was used in VS experiments aimed at targeting a very simple quadruplex, namely $[d(TGGGGT)]_4$ (PDB code 1S45).[5] This quadruplex possesses a 4-fold symmetry

with all strands parallel to each other, which afford four grooves of identical medium width, and all nucleosides in an anti glycosidic conformation. Thus, the docking software Autodock4 (AD4) was used to dock a diversity set from the commercially available Life Chemicals database (6000 compounds). To avoid finding redundant information, a search area large enough to enclose only one of the four identical grooves was used (figure 1).

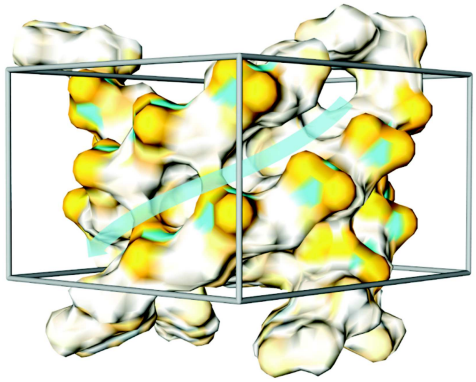


Figure 1: Surface representation of the quadruplex structure. Orange, white and cyan areas represent negatively, neutrally and positively charged regions, respectively. The grey box represents the searched area in VS calculations. The cyan transparent arrow outlines the groove area taken into consideration.

The VS results were sorted on the basis of their predicted binding free energies (ΔG_{AD4}) which ranged from -0.95 to -9.55 kcal/mol. Solutions with a predicted binding free energy greater than -6.0 kcal/mol and a cluster size lower than 10 out of 100 individuals were discarded. Based on these criteria only 137 individuals were retained for further consideration. The binding poses calculated for these compounds were then visually inspected to discard all the individuals which were not predicted to establish tight interactions with

the groove of the quadruplex structure. More precisely, compounds that were not able to form H-bonds with any of the guanine bases and/or to establish an electrostatic interaction with the backbone phosphate groups were not considered. After this final step, 30 compounds (figure 2) corresponding to 0.5% of the original Life Chemicals database were selected and purchased for further analysis.

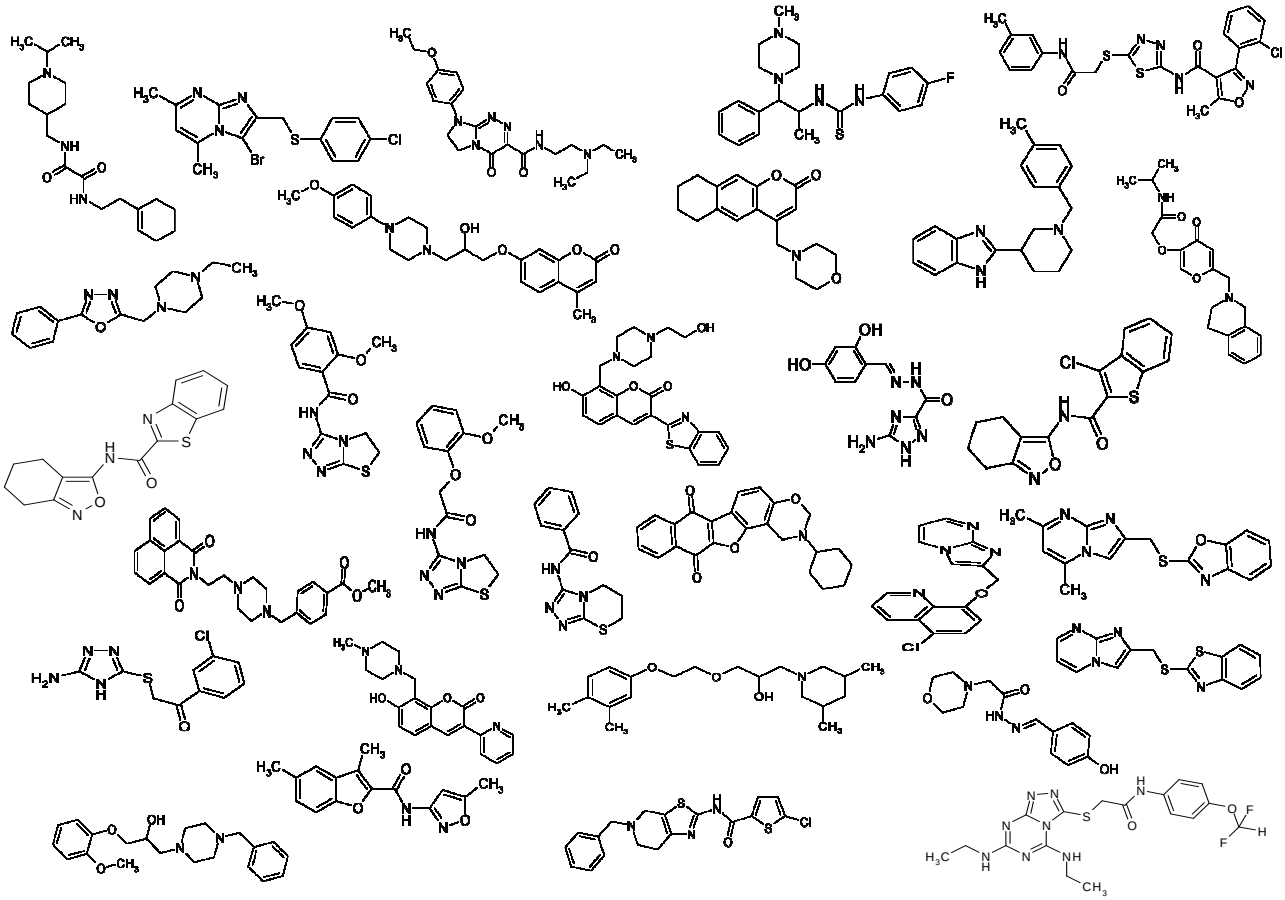


Figure 2: 30 compounds selected.

The experimental test of the top computational "hits" for binding has been performed by NMR, which has the significant advantages of being able to

detect weak binders and readily identify the ligand binding sites, consequently verifying the binding specificity.[6] Thus, we prepared 30 identical DNA samples containing 6.8 mM of d(TGGGGT) in 0.2 mL (H_2O/D_2O 9:1) of buffer solution having 10 mM KH_2PO_4, 70 mM KCl, 0.2 mM EDTA, at pH 7.0.

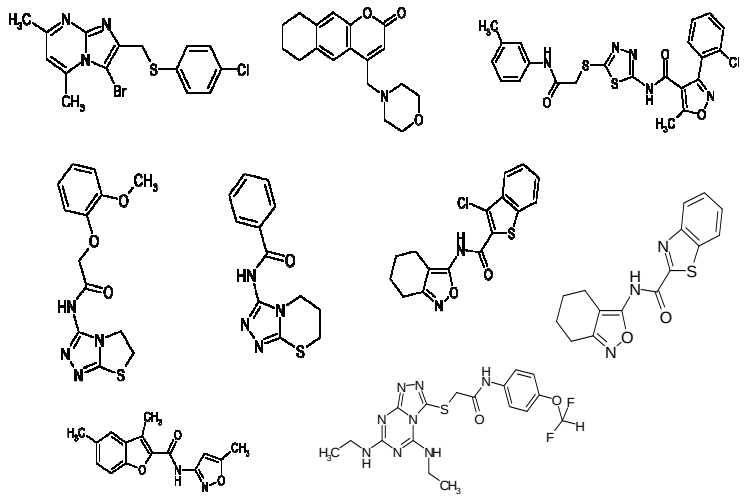
The first problem that we encountered in testing the selected compounds was the solubility (most of them were not soluble in water). Nevertheless, all samples turned out to be soluble in DMSO.

So far each titration of a new compound with a quadruplex structure was done in water; we had to understand if the quadruplex $[d(TGGGGT)]_4$ maintain the same features in DMSO, for this reason we have tested the stability of the quadruplex $[d(TGGGGT)]_4$ in buffers containing different percentages of DMSO. The quadruplex turned out to be perfectly structured even in buffer solutions containing 30% of DMSO.

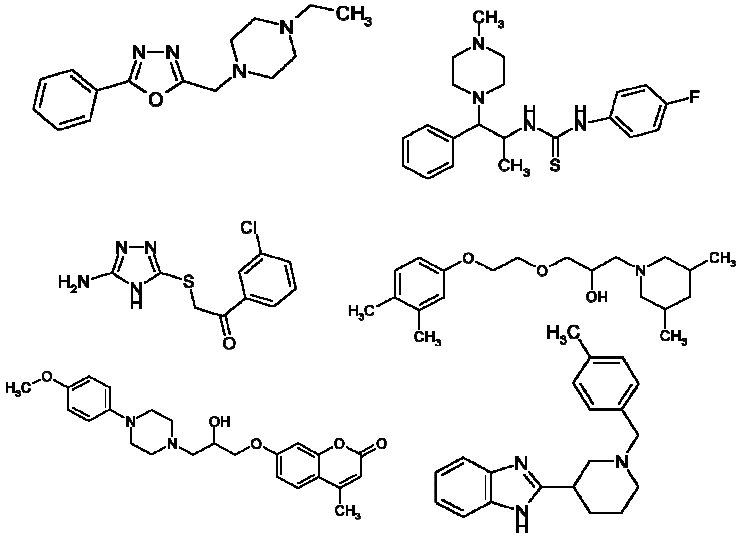
After this results, we dissolved each equivalent of the selected compounds in 5 µL of DMSO, in such a way to gain an overall percentage of DMSO at the end of the titration not higher than 15%. The NMR titrations (700 MHz, T = 25°C) were carried out monitoring resonance chemical shift changes of DNA, which were used to estimate whether a given compound is able to interact with the quadruplex and to determine the binding site. In line with the expected results of a VS campaign, a number of false positives were found (figure 3):

- 9 molecules did not significantly interact with the quadruplex
- 6 molecules caused a shift of the signal of Ts (T1 and T6)
- 9 molecules caused a shift of the resonances belonging to external bases (T1 and G2, T6 and G5), most probably due to an end-stacking interaction
- 6 molecules were found to cause an appreciable shift, among the others, of the G3 and G4 resonances signals, thus suggesting, as expected, a groove binding interaction.

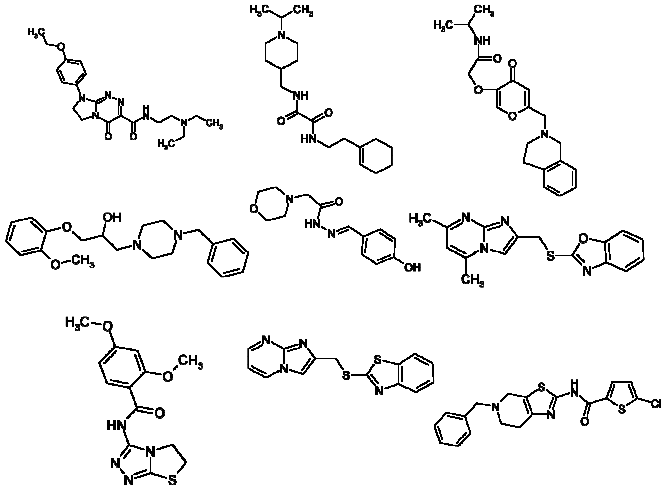
Molecules which did not interact with the quadruplex



Molecules interacting with Ts



Molecules interacting as end-stackers



Molecules interacting as groove-binders

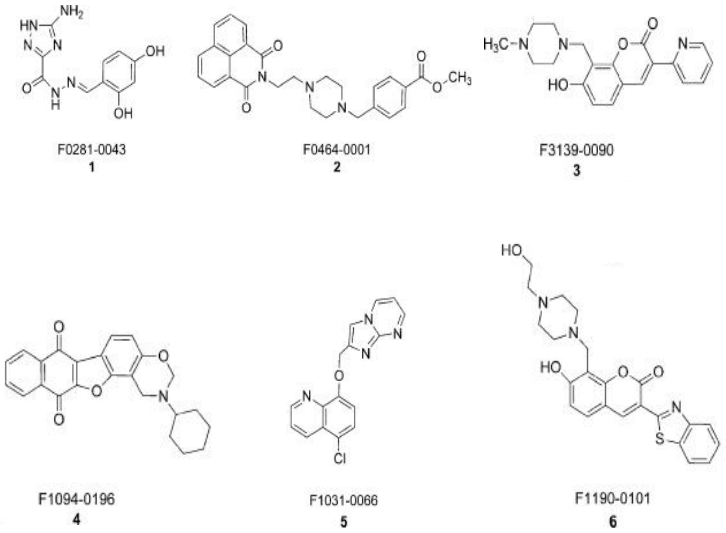
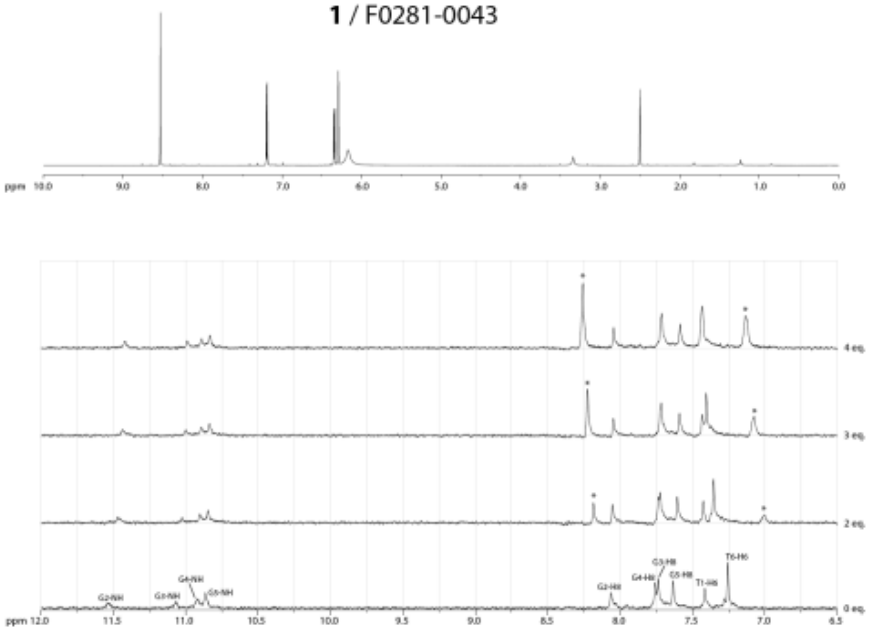


Figure 3: Structures of the newly identified groove binders. I have divided the 30 compounds selected by Virtual Screening in the several groups: Molecules not interacting with the quadruplex; Molecules only interacting with the Ts; Molecules interacting as end-stakers.; Molecules interacting as groove binders.

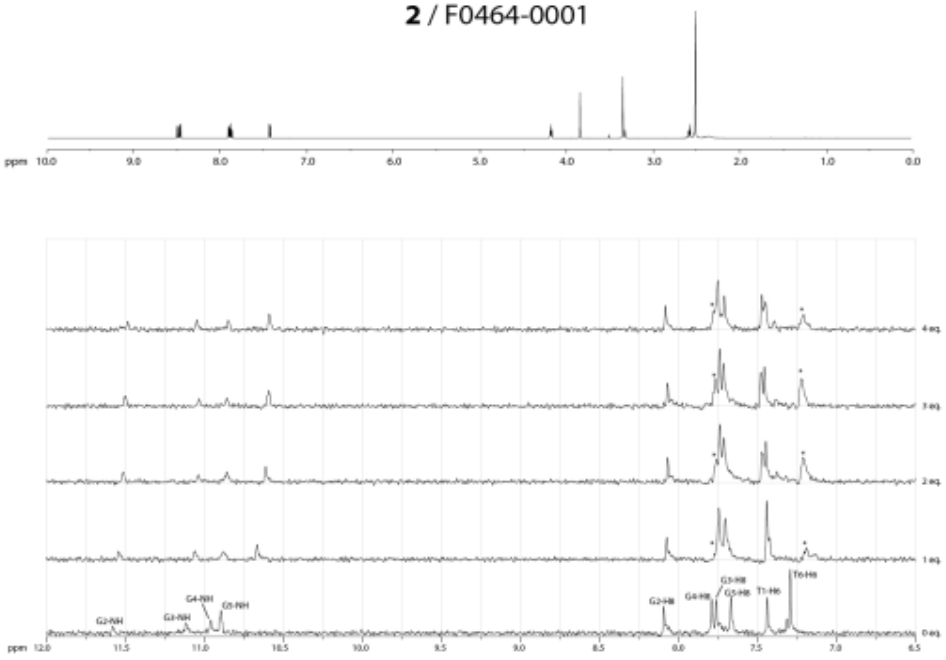
The Life Chemicals codes have been reported only for the six groove binders, numbered from 1 to 6.

More in details, compounds 1-5, belonging to the group of molecules interacting as groove binders, provided NMR titration profiles very similar to each other, causing mainly drifting of the signals of G3, G4, G5, and T6, and so indicating that the recognition process involves mostly the 3' side of the grooves (figure 4).

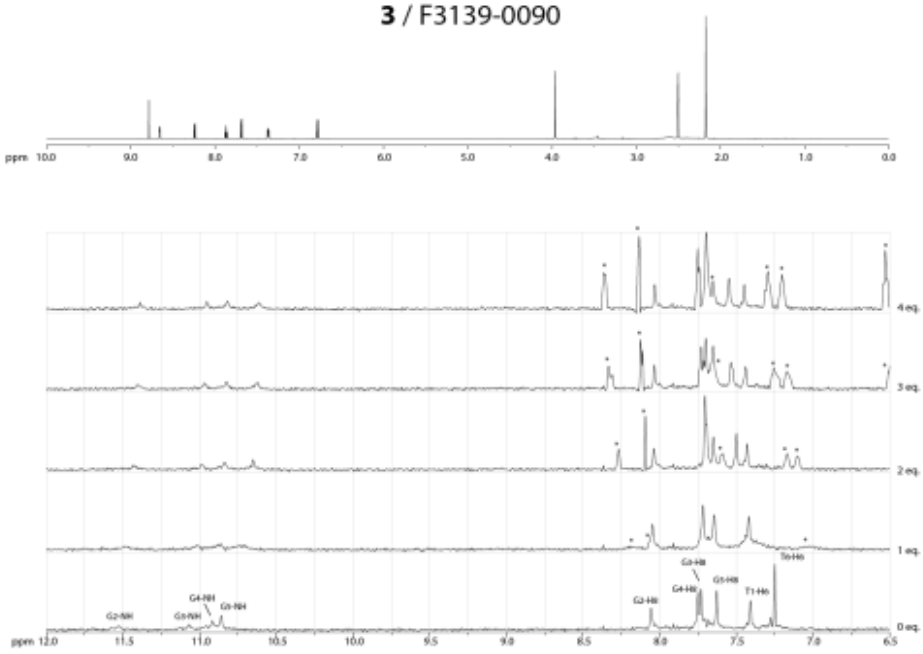
1 / F0281-0043



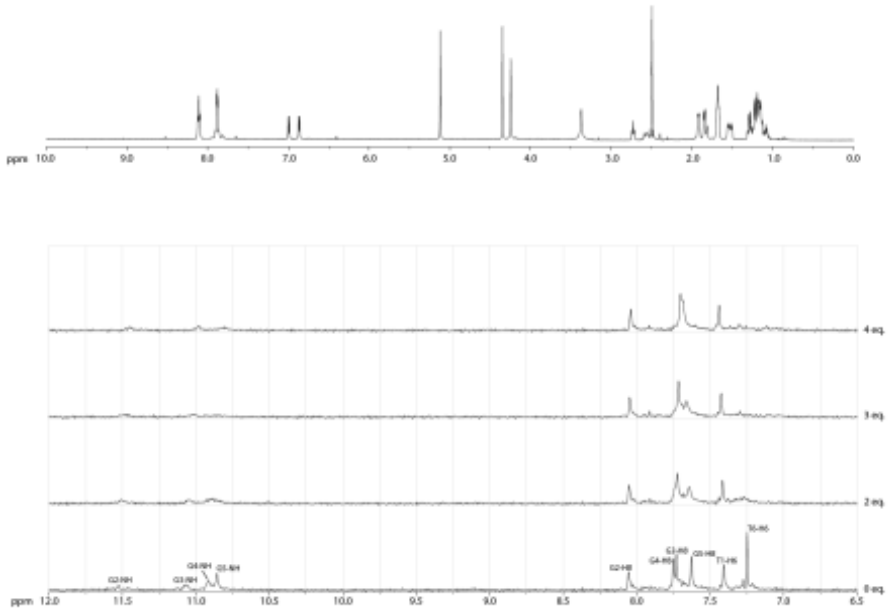
2 / F0464-0001



3 / F3139-0090



4 / F1094-0196



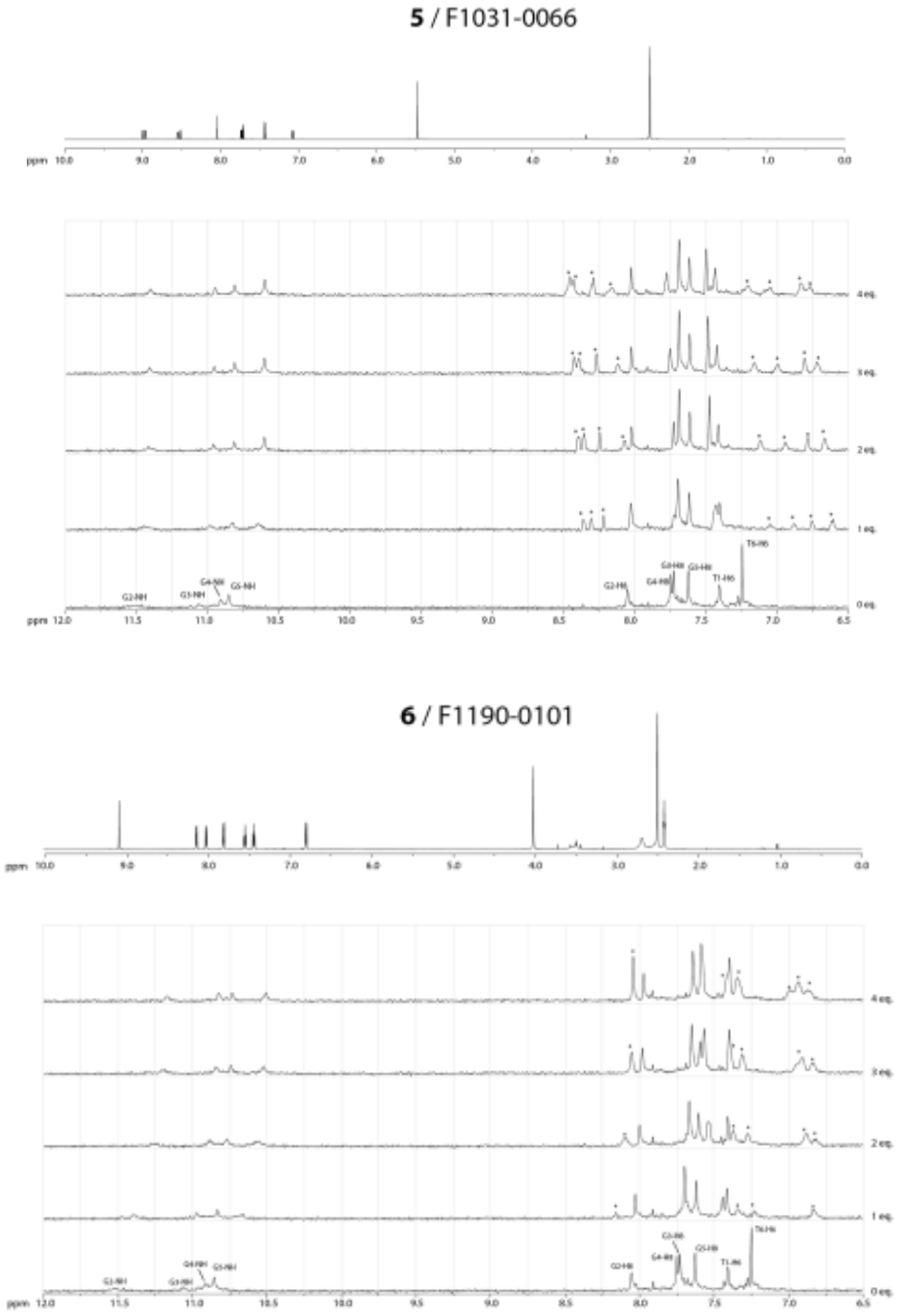


Figure 4: NMR titration profiles of compounds 1-6. We can see the compounds 1-5 causing mainly drifting of the signals of G3, G4, G5, and T6, the recognition process involves mostly the 3' side of the grooves. The compound 6 causes a major change of the resonances of the quadruplex, suggesting a higher affinity. On the top of each NMR titration profile I have reported the 1D ^{1}H-NMR spectrum of the compound alone in DMSO-d6 and on the bottom I have reported the NMR titration of $[d(TGGGGT)]_4$ with the compound.

On the other hand, the compound 6 seems to entirely span the grooves, perturbing more uniformly all the residues of the quadruplex, in addition it causes a major change of the resonances of the quadruplex, suggesting a higher affinity.

This is a very interesting result, since it is surprisingly consistent with the mode of binding calculated by the VS (figure 5).

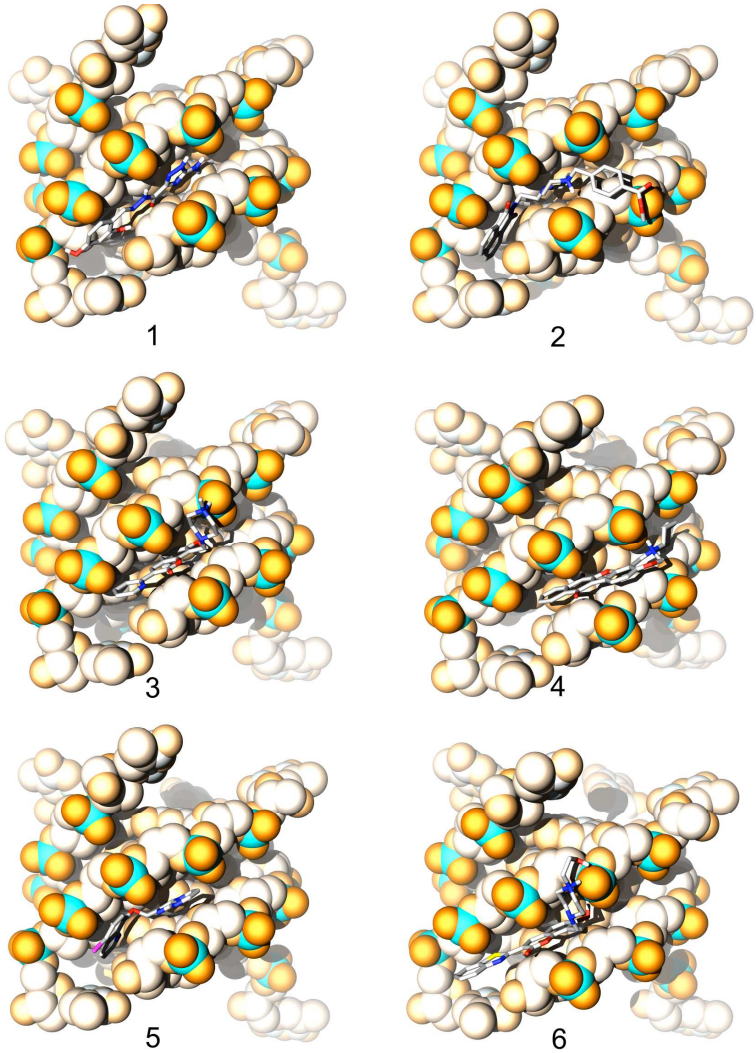


Figure 5: Binding poses calculated by AD4 for compounds 1-6 in the considered quadruplex structure. DNA is represented as spheres with the 5' at the top and the 3' at the bottom. Orange, white and cyan spheres represent negatively, neutrally and positively charged regions, respectively. Ligands are depicted as white sticks.

Interestingly, while compounds 1-4 and 6 are positively charged as most of the already known groove binders, compound 5 does not possess any charge, so that the binding might be more driven by the presence of a number of H-bond acceptor heteroatoms. The titration of all 6 molecules turned out to be virtually completed at 4 equiv.

In order to confirm the results obtained with the new approach of Virtual Screening and to find new molecular scaffolds able to interact with the groove of the DNA quadruplex $[d(TGGGGT)]_4$, we decided to make a novel NMR titration experiments where we used a modified quadruplexes, namely $[d(TGG^{BrG}GT)]_4$ and $[d(TGGGG^{BrT})]_4$. The novel NMR spectroscopy experiments combined with molecular modelling studies, allows for a more detailed pictures of the interaction between each binder and the quadruplex DNA, also an Isothermal Titration Calorimetry (ITC) measurements was used to confirm the results

Results and discussion

NMR and molecular modelling studies

Using NMR, six molecules (1-6, figure 3) were found to be potential groove binders: they cause an appreciable shift, amongst others, of the signals of G3, G4, G5 and T6 of the parallel quadruplex $[d(TGGGGT)]_4$, indicating that the recognition process involves mostly the 3' side of the grooves, a result also confirmed by the virtual screening calculations.[7]

In order to get further insights into the binding mode of compounds 1-6, we have acquired a number of NOESY experiments of the complexes of the six compounds with the quadruplex $[d(TGGGGT)]_4$. Unfortunately, as no diagnostic NOE cross-peak could be retrieved for any complex, none of the three-dimensional structures at atomic level could be determined.

Thus, we decided to get a more detailed picture of the interactions from the NMR titration of modified quadruplexes and from molecular docking calculations. As a result, we designed and synthesised modified oligonucleotides, namely $d(TGG^{Br}GGT)$, $d(TGGG^{Br}GT)$ and $d(TGGGG^{Br}T)$, where dG^{Br} is 8-bromo-2'-deoxyguanosine, potentially capable of forming quadruplex structures and possessing a bulky group (bromine) at different positions of the grooves (G3, G4 and G5), as this portion of the DNA should be involved in 1-6 recognition.[8]

All the new six scaffolds individuated were titrated with the modified quadruplexes . An analysis of the NMR titration's results profile was then carried out.

The rationale behind these experiments relies on the assumption that if these compounds interacted with the quadruplex groove region, the presence of the bromine group should have prevented (or at least limited) the ligand/DNA interactions.

Prior to these experiments, we tested the capability of d(TGG^{Br}GGT), d(TGGG^{Br}GT) and d(TGGGG^{Br}T), to form a quadruplex structure.[9]

Their NMR samples were prepared at a concentration of 2 mM, in 0.6 ml (H_2O/D_2O 9:1) buffer solution having 10 mM KH_2PO_4, 70 mM KCl, 0.2 mM EDTA, pH 7.0. These samples were then annealed for 5-10 min at 80 °C and slowly cooled down at room temperature, then ^{1}H NMR spectra were recorded by using DPFGSE pulse sequence for H_2O suppression.[10,11] The ^{1}H NMR spectra (700 MHz, T = 25 °C) of d(TGG^{Br}GGT) and d(TGGGG^{Br}T) show the presence of four well defined singlets in the region 11-12 ppm, ascribable to imino protons involved in Hoogsteen hydrogen bonds of G-quartets, as well as the presence of five signals belonging to three guanine H8 and to two thymine H6 protons in the aromatic region. This indicates that a single well defined quadruplex species is present in solution for both molecules, consisting of four G-tetrads and possessing a four fold symmetry with all strands parallel to each other. On the other hand, the 1D ^{1}H NMR spectrum of d(TGGG^{Br}GT) shows the presence of a great number of signals in the regions of imino and aromatic

protons, suggesting that d(TGGG^{Br}GT) is affected by structural heterogeneity, thus preventing its use in our experiments. Hence, only the quadruplexes [d(TGG^{Br}GGT)]$_4$ and [d(TGGGG^{Br}T)]$_4$ were titrated with compounds 1-6 and the titrations monitored by NMR.

A comparison of resonances of protons of the uncomplexed quadruplex and the complexed one has been performed, in particular we have studied the Δδ values (chemical shifts of the complex minus free DNA) of aromatic, methyl and imino protons.

As far as compound 1 is concerned, the titration with unmodified quadruplex [d(TGGGGT)]$_4$ led to a general shift of the monitored signals (white bars in figure 6).

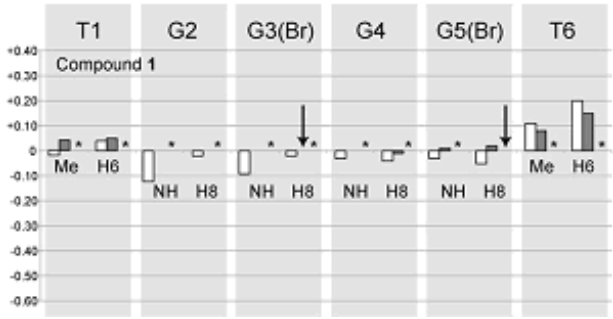


Figure 6: Δδ values of aromatic (H6/H8), methyl (Me) and imino protons (NH) for [d(TGGGGT)]$_4$ (white bars), [d(TGG^{Br}GGT)]$_4$ (light-grey bars) and [d(TGGGG^{Br}T)]$_4$ (dark-grey bars). Asterisks indicate a severe line broadening of the monitored signals. Arrows indicate the lack of bars due to the presence of bromine atoms.

On the other hand, the titration of [d(TGG^{Br}GGT)]$_4$ causes a slight shift (light-grey bars in figure 6) of the residues at the 3' edge of the quadruplex, namely G5-H8, T6-H6/Me, and only of T1-H6/Me at 5' edge, whilst no appreciable shift

can be measured for the other signals. This means that the bromine atom at the very centre of the groove did affect the binding of compound 1, and that it can be confidently considered a groove binder. Furthermore, the titration of $[d(TGGGGBrT)]_4$ led to a severe line broadening of all signals, making impossible to retrieve any information from the spectra. This NMR phenomenon can be interpreted assuming that the ligand is changing its binding pose on the NMR time scale.

It is noteworthy that compounds 1-6 have been discovered from the virtual screening approach and in that study Autodock4 program was used with a search area large enough to enclose only one of the four identical grooves to avoid redundant information.[8] However, this could not cast out the possibility that one or more of the selected molecules can also be able to bind to other part of the target.

Thus, herein, I report results of new docking calculations where the search area has been enlarged to comprise the entire surface of the quadruplex $[d(TGGGGT)]_4$ with the purpose of checking for the capability of the ligands to extend their binding out of the groove. Regarding compound 1, docking calculations showed that all solutions in the lowest energy families (ΔG = -6,2 kcal mol^{-1}), were found to be anchored to the 3' side of the groove in line with the above-mentioned NMR data (figure 7), compound 1 was predicted in a groove binding mode 75 out of 100 time. As shown in figure 8, the amine group of compound 1 engages a H-bond with the 4' O of the T6 nucleoside, while the ligand carbonyl group together with the ortho-hydroxyl group of the

phenyl moiety forms H-bonds with G5 base. Three H-bonds have been also detected between compound 1 and the phosphate backbone.

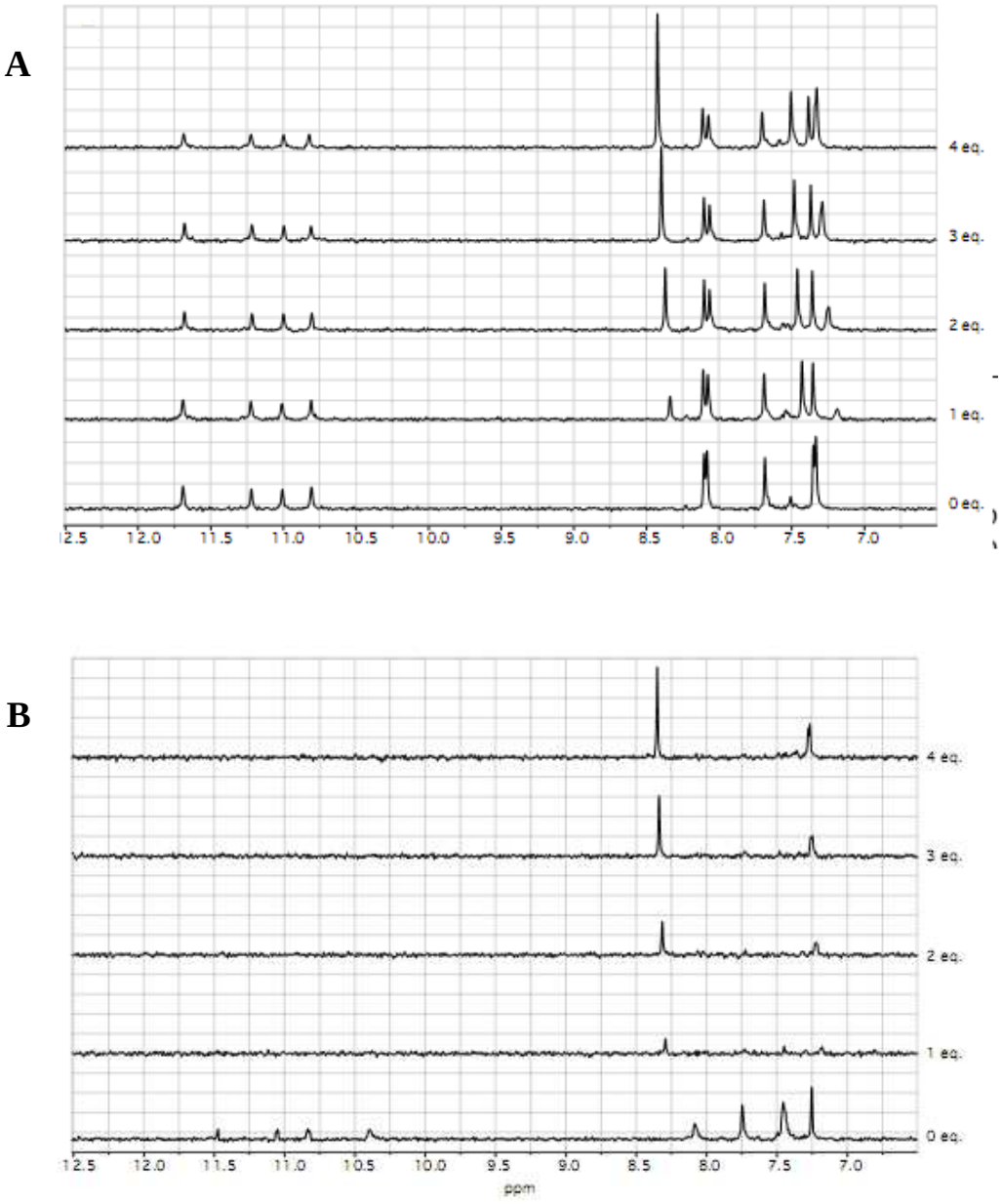


Figure 7: NMR titrations of compounds 1 with the modified quadruplex **A)** [d(TGG^{Br}GGT)]$_4$ and **B)** [d(TGGGG^{Br}T)]$_4$.

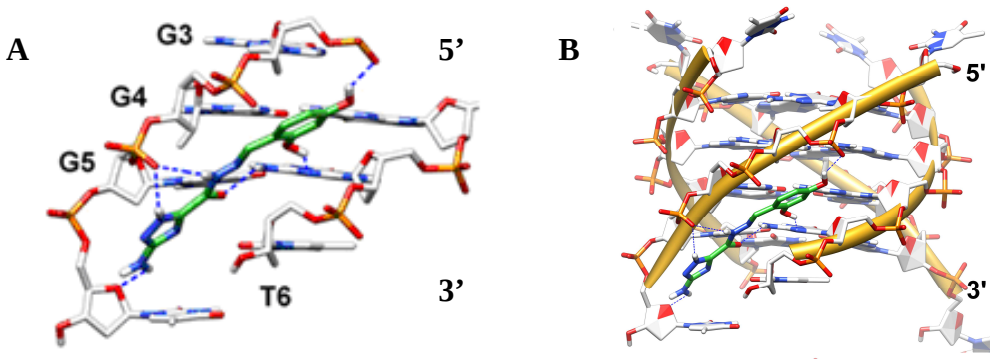


Figure 8: Binding poses calculated by AD4 for compounds 1 in the quadruplex structure. DNA backbone is represented as **A)** white stick bonds and **B)** in smooth gold ribbons and stick bonds, special filled rings representations are used for the sugars and bases.. Ligands are depicted as green sticks. H-bonds are represented as dashed blue lines.

Surprisingly, compound 2 generates almost identical Δδ values in the titration with $[d(TGGGGT)]_4$ and $[d(TGG^{Br}GGT)]_4$ (white and light-grey bars, respectively, (figure 9), having a general perturbation of the signals belonging to the 3' side of the groove.

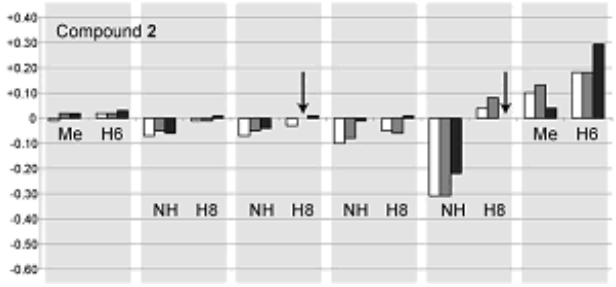


Figure 9: Δδ values of aromatic (H6/H8), methyl (Me) and imino protons (NH) for $[d(TGGGGT)]_4$ (white bars), $[d(TGG^{Br}GGT)]_4$ (light-grey bars) and $[d(TGGGG^{Br}T)]_4$ (dark-grey bars). Arrows indicate the lack of bars due to the presence of bromine atoms.

This can be interpreted by assuming that the bromine atom does not affect the binding of compound 2 suggesting that it should not preferentially bind

the very centre of the groove. On the other hand, the titration of $[d(TGGGG^{BrT})]_4$ (dark-grey bars, figure 9) led to a marked shift of T6-H6/Me signals whereas only a slight shift of all the other monitored signals can be observed. This confirm that compound 2 prefers to interact with the 3' edge of the quadruplex. A clearer picture of the ligand and quadruplex interaction could be obtained from docking calculations using again the unmodified quadruplex $[d(TGGGGT)]_4$ as target. The calculations highly converged towards one family of conformations (ΔG = -7.4) (figure 10) in which the benzoisoquinolinedione ring stacks on the surface of the 3' terminal quartet, particularly between the T6 and G5 rings, and the rest of the molecule inserts into the groove, where the main anchor point is represented by the protonated nitrogen of the piperazine ring, which establishes a charge reinforced H-bond with the phosphate backbone.

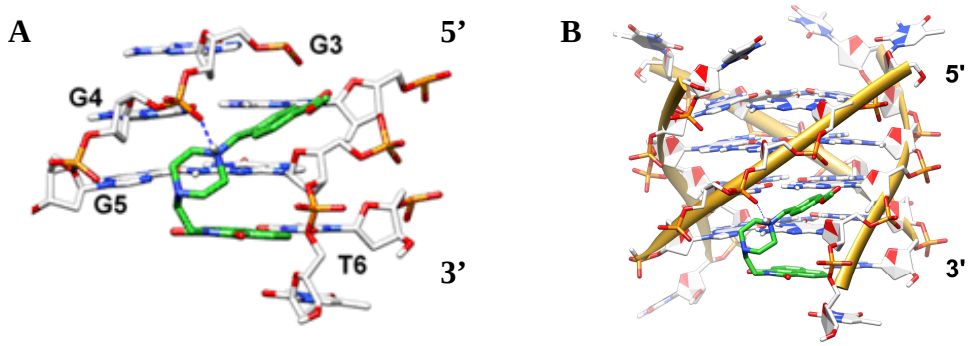


Figure 10: Binding poses calculated by AD4 for compounds 2 in the quadruplex structure. DNA backbone is represented as **A)** white stick bonds and **B)** in smooth gold ribbons and stick bonds, special filled rings representations are used for the sugars and bases.. Ligands are depicted as green sticks. H-bonds are represented as dashed blue lines.

This binding mode is consistent with the new NMR titrations (figure 11)

performed in this investigation indicating that compound 2 is actually characterised by a mixed binding mode, providing both stacking and groove binding interactions.

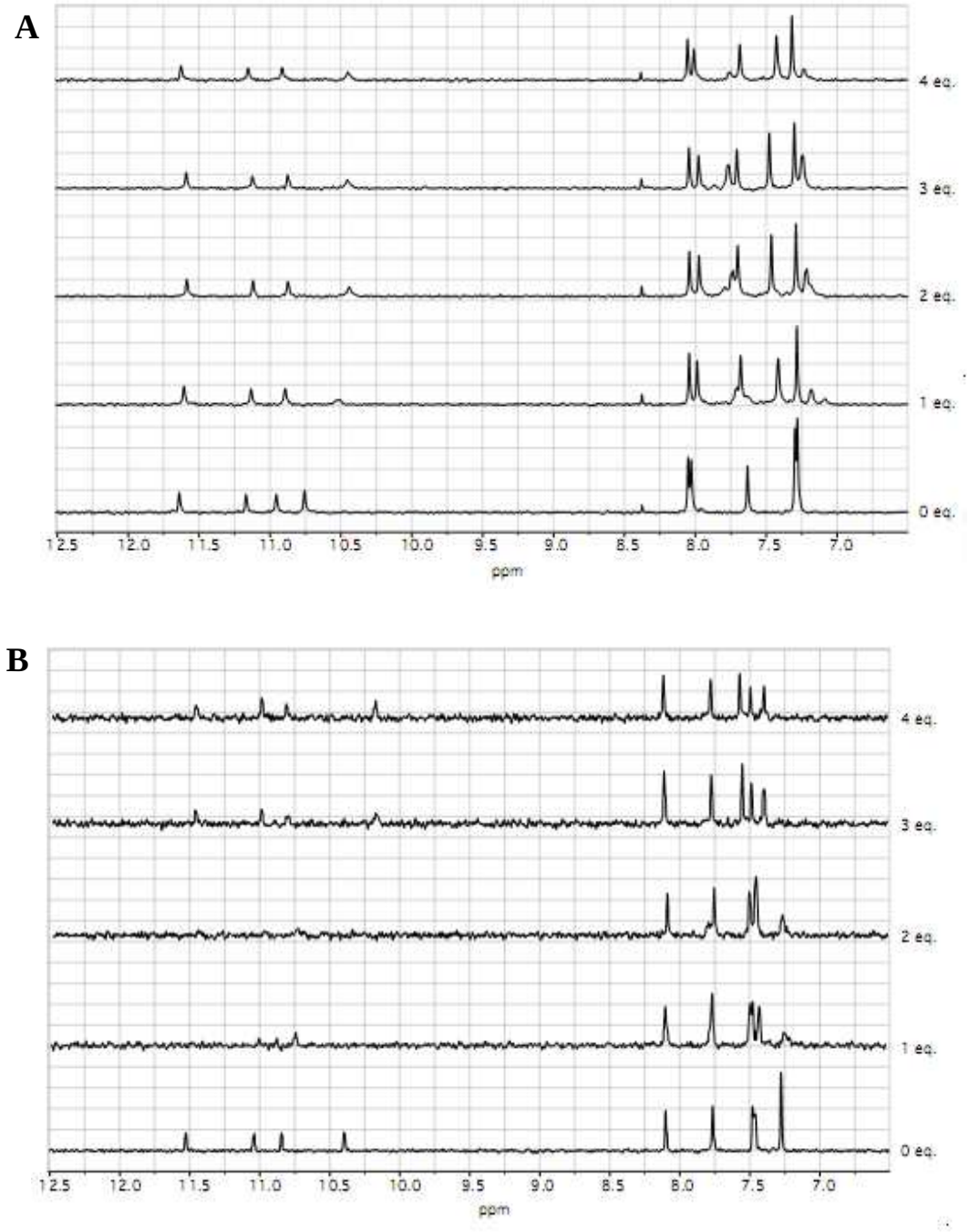


Figure 11: NMR titrations of compounds 2 with the modified quadruplex **A)** $[d(TGG^{BrG}GT)]_4$ and **B)** $[d(TGGGG^{BrT})]_4$.

Differently from 1 and 2, titration of compound 3 with $[d(TGGGGT)]_4$, $[d(TGG^{Br}GGT)]_4$ and $[d(TGGGG^{Br}T)]_4$ led to a general shift of almost all signals (figure 12).

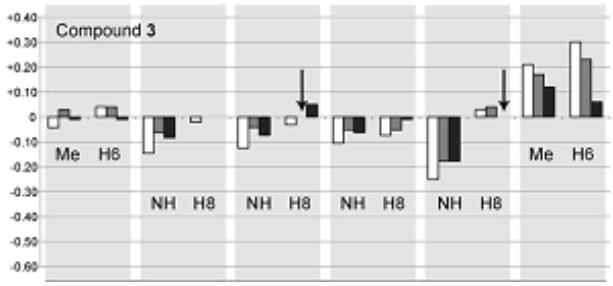


Figure 12: Δδ values of aromatic (H6/H8), methyl (Me) and imino protons (NH) for $[d(TGGGGT)]_4$ (white bars), $[d(TGG^{Br}GGT)]_4$ (light-grey bars) and $[d(TGGGG^{Br}T)]_4$ (dark-grey bars). Arrows indicate the lack of bars due to the presence of bromine atoms.

However, the resonances of $[d(TGGGGT)]_4$ (white bars) shifted more and those of $[d(TGG^{Br}GGT)]_4$ and $[d(TGGGG^{Br}T)]_4$ less (light- and dark-grey bars, respectively). This indicates that compound 3 does not possess a well defined binding mode, so that, when the groove is unavailable because of the presence of the bromine atoms, compound 3 slides towards the 3' edge groove of the quadruplex; vice versa, in the case of $[d(TGGGG^{Br}T)]_4$, the molecule is able to interact only weakly with the available part of the groove. Docking calculations on $[d(TGGGGT)]_4$ suggest that the ligand binds the 3'-end of the groove (ΔG = -8.2) (compound 3 was predicted in a groove binding mode 87 out of 100 time). As shown in figure 13, two H-bonds were found, the first between the pyridine nitrogen of compound 3 and the T6 3'-OH group and the second between the hydroxyl group of the chromenone core and the G3NH2.

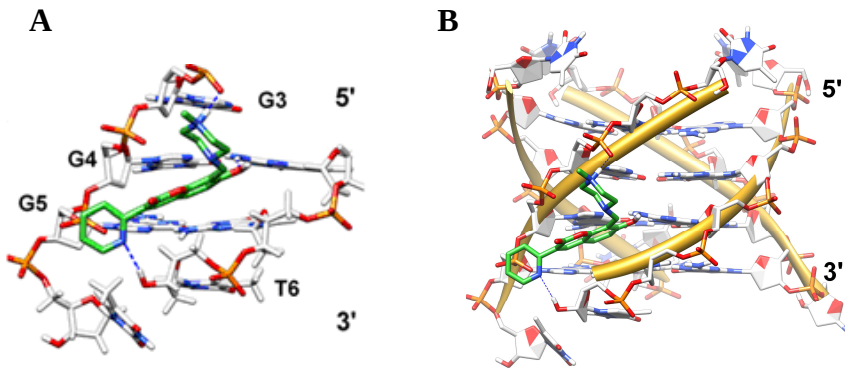
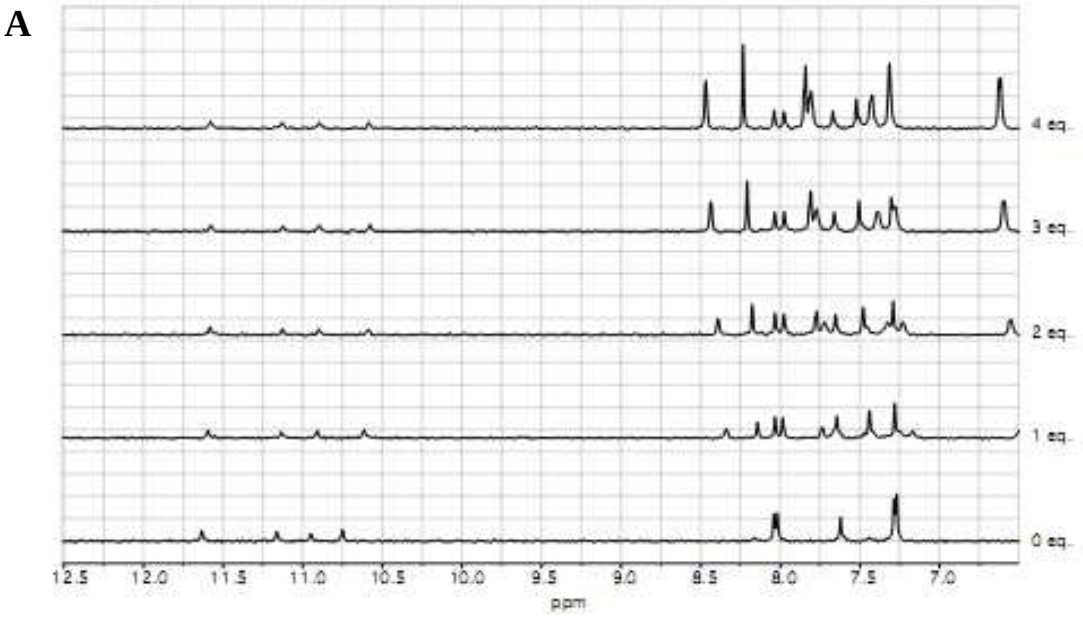


Figure 13: Binding poses calculated by AD4 for compounds 3 in the quadruplex structure. DNA backbone is represented as **A)** white stick bonds and **B)** in smooth gold ribbons and stick bonds, special filled rings representations are used for the sugars and bases.. Ligands are depicted as green sticks. H-bonds are represented as dashed blue lines.

Hydrophobic contacts between the pyridine and the T6 rings were also detected together with a charge reinforced H-bond established by the protonated nitrogen of the piperazine ring with the phosphate backbone. The same results are confirmed by NMR titration (figure 14).



B

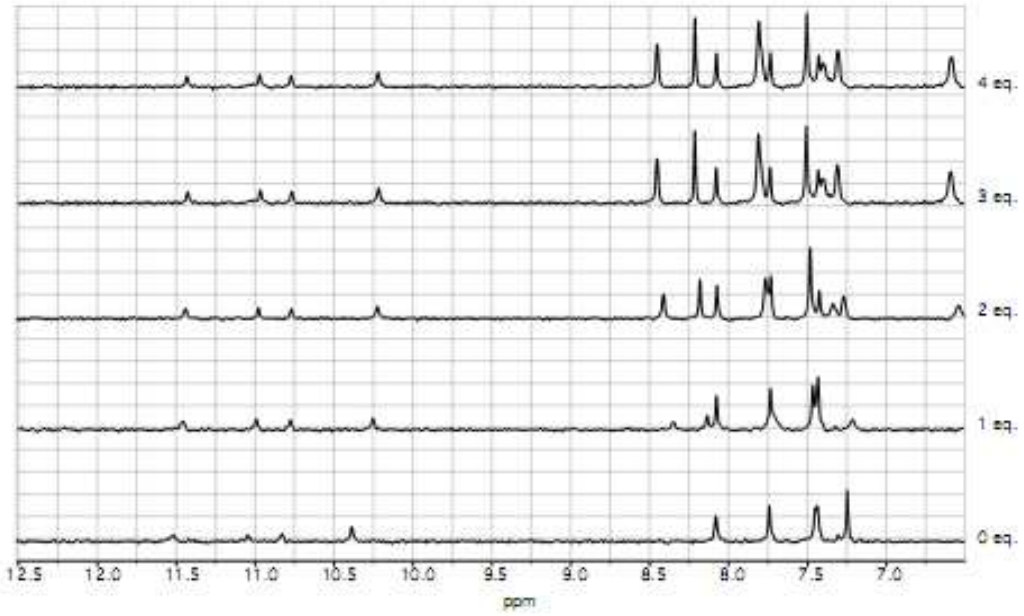


Figure 14: NMR titrations of compounds 3 with the modified quadruplex **A)** $[d(TGG^{Br}GGT)]_4$ and **B)** $[d(TGGGG^{Br}T)]_4$.

Interestingly, compound 4 displayed different behaviours for the three quadruplexes tested. With the quadruplex $[d(TGG^{Br}GGT)]_4$ (light-grey bars, figure 15), the titration provided higher Δδ values for the residues at the edges of the quadruplex with respect to unmodified $[d(TGGGGT)]_4$ (white bars, figure 15). On the contrary, titration of $[d(TGGGG^{Br}T)]_4$ caused major shift for the residue sitting in the very centre of the groove (dark-grey bars).

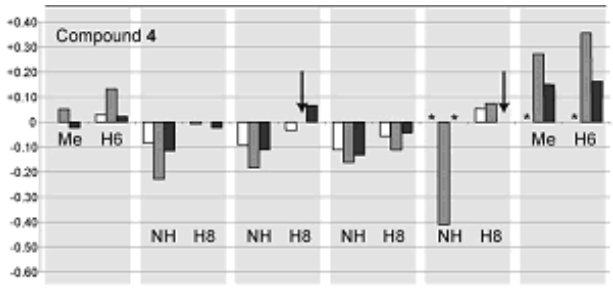


Figure 15: Δδ values of aromatic (H6/H8), methyl (Me) and imino protons (NH) for $[d(TGGGGT)]_4$ (white bars), $[d(TGG^{Br}GGT)]_4$ (light-grey bars) and $[d(TGGGG^{Br}T)]_4$ (dark-grey bars). Asterisks indicate a severe line broadening of the monitored signals. Arrows indicate the lack of bars due to the presence of bromine atoms.

This means that compound 4 can interact with the grooves, and, accordingly, when the very centre of the groove is hindered, it binds the end sides of the grooves, but it is able to bind the very centre of the grooves when their 3' edges are unavailable. Docking calculations suggest two binding poses especially in line with NMR data (figure 16 a, b and c), where the molecule can either interact with 5' residues (binding mode A, ΔG = -8.2) or with the 3' end (binding mode B, ΔG = -7.9).

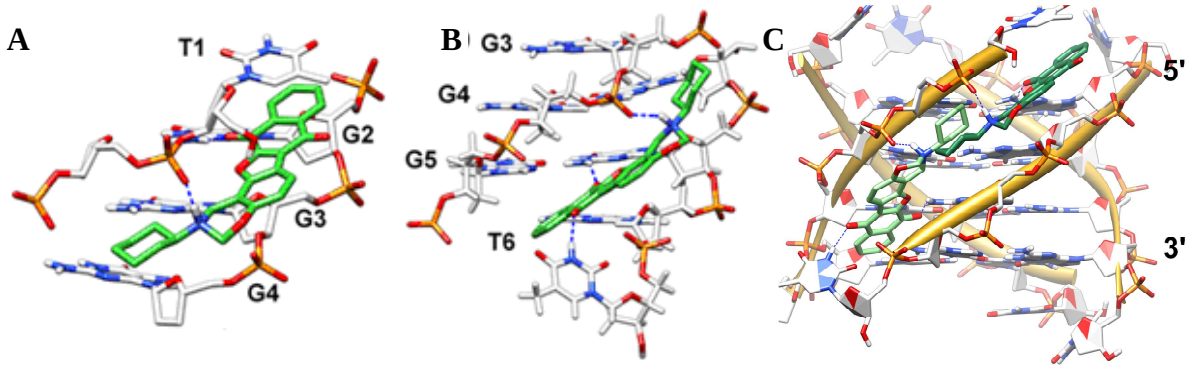


Figure 16: Two binding poses calculated by AD4 for compound 4 **A)** and **B)** in the quadruplex structure. DNA backbone is represented as white stick bonds, **C)** in smooth gold ribbons and stick bonds, special filled rings representations are used for the sugars and bases Ligands are depicted as green sticks. H-bonds are represented as dashed blue lines.. Ligands are depicted as green sticks. H-bonds are represented as dashed blue lines.

Specifically, in A, the benzoquinone ring establishes a π-π interaction with the T1 ring, while an H-bond between the G2 NH2 group and the carbonyl moiety is detected. In B, the two carbonyl moieties of the ligand are engaged in H-bonds with G4 NH2 and T6 NH respectively. In both binding modes a charge reinforced H-bond is observed between the protonated nitrogen of compound 4 and the phosphate backbone. Based on this finding and on the shift of NMR signals observed for all residues forming the groove (figure 17), a sliding motion of compound 4 inside the groove can be proposed.

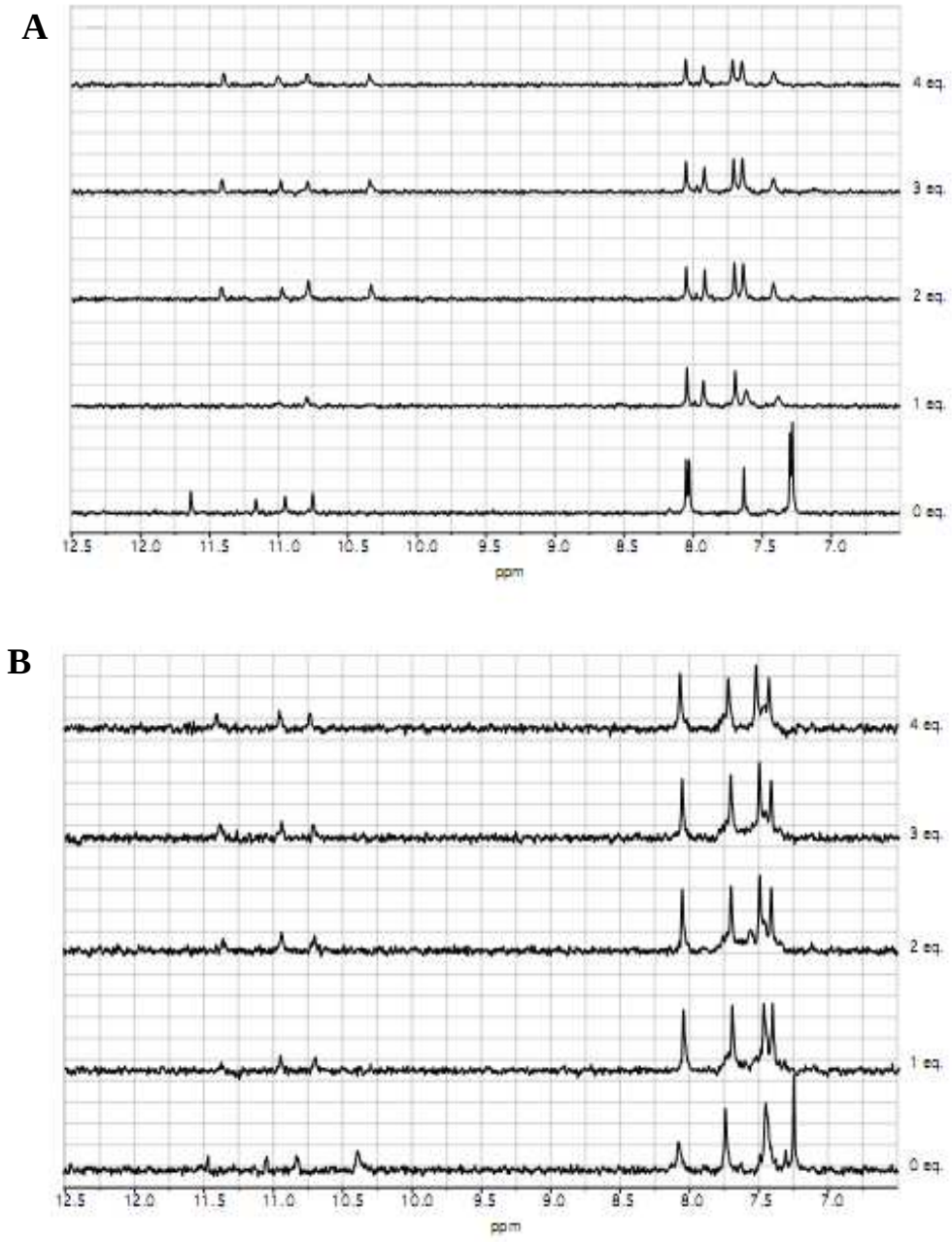


Figure 17: NMR titrations of compounds 4 with the modified quadruplex **A)** $[d(TGG^{Br}GGT)]_4$ and **B)** $[d(TGGGG^{Br}T)]_4$.

As far as compound 5 is concerned, the titration of $[d(TGGGGT)]_4$ (figure 18, white bars) clearly indicates that compound 5, like the other selected

compounds, prefers to bind the 3' edge of the groove. The titration of $[d(TGG^{Br}GGT)]_4$ displays significant lower Δδ values for the imino protons (figure 18, light-grey bars), while the shifts of the signal of G4, G5 and T6 remain substantially unchanged.

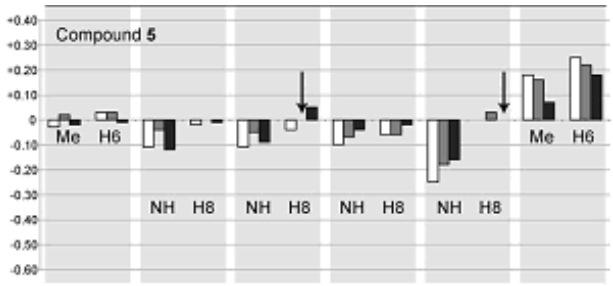


Figure 18: Δδ values of aromatic (H6/H8), methyl (Me) and imino protons (NH) for $[d(TGGGGT)]_4$ (white bars), $[d(TGG^{Br}GGT)]_4$ (light-grey bars) and $[d(TGGGG^{Br}T)]_4$ (dark-grey bars). Arrows indicate the lack of bars due to the presence of bromine atoms.

On the other hand, the titration of $[d(TGGGG^{Br}T)]_4$ indicates that compound 5 is still able to interact with the 3' edge of the quadruplex probably via end-stacking interaction. A mixed binding mode is therefore expected for compound 5. In line with these data, docking calculations suggested that in compound 5 the quinolone ring is adapted on the top of the 3'-end sandwiched between the T6 and G5 bases (figure 19), with the imidazo-pyrimidine moiety extending towards the groove incipience and H-bonding with the G5 NH2 group (ΔG = -6.7).

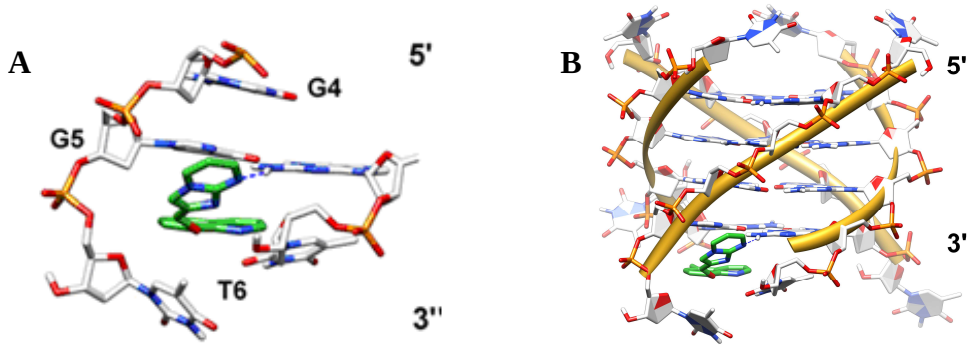
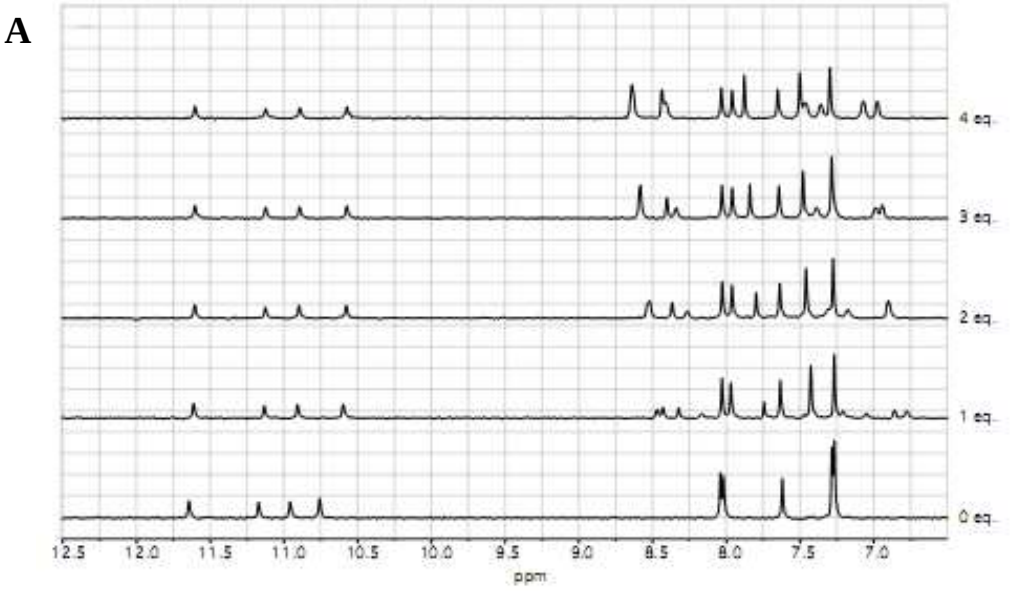


Figure 19: Binding poses calculated by AD4 for compounds 5 in the quadruplex structure. DNA backbone is represented as **A)** white stick bonds and **B)** in smooth gold ribbons and stick bonds, special filled rings representations are used for the sugars and bases.. Ligands are depicted as green sticks. H-bonds are represented as dashed blue lines.

The NMR titration is reported in figure 20.



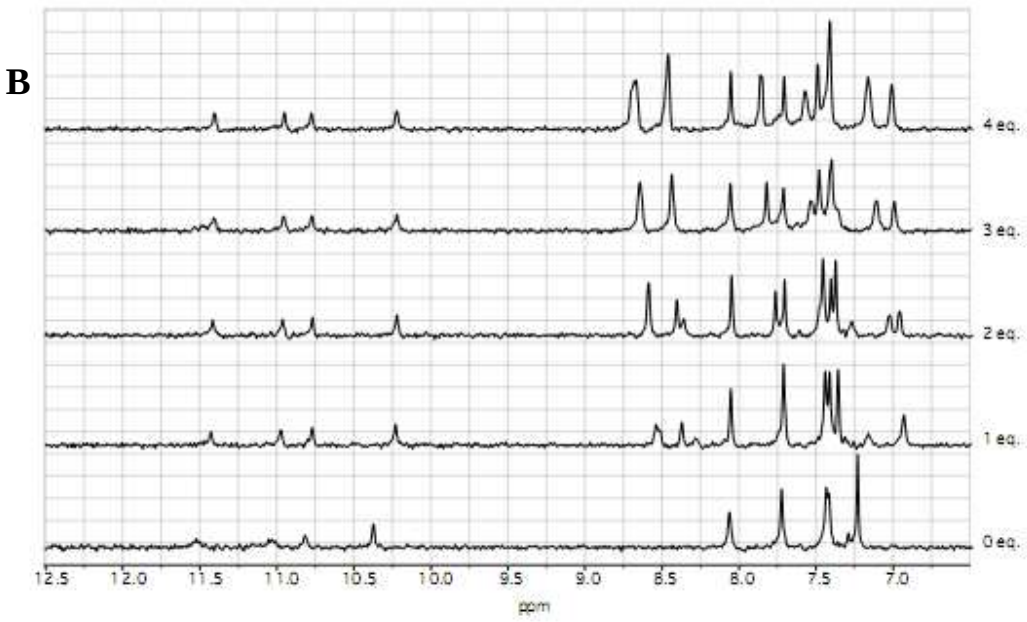


Figure 20: NMR titrations of compounds 5 with the modified quadruplex **A)** $[d(TGG^{Br}GGT)]_4$ and **B)** $[d(TGGGG^{Br}T)]_4$.

Finally, compound 6 displays peculiar titration behaviours. While the titration of the unmodified $[d(TGGGGT)]_4$ showed marked shifts of the G2, G3, G4 and G5 residues (figure 21, white bars), the titration of $[d(TGG^{Br}GGT)]_4$ is instead characterised by very low Δδ values for all residues (light-grey bars). Furthermore, the titration of $[d(TGGGG^{Br}T)]_4$ mainly caused shifts of the residues G2, G3 and T6, along with a severe line broadening of the imino proton signals (dark-grey bars, figure 21).

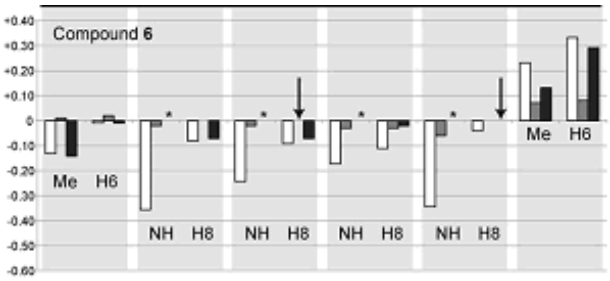


Figure 21: Δδ values of aromatic (H6/H8), methyl (Me) and imino protons (NH) for $[d(TGGGGT)]_4$ (white bars), $[d(TGG^{Br}GGT)]_4$ (light-grey bars) and $[d(TGGGG^{Br}T)]_4$ (dark-grey bars). Asterisks indicate a severe line broadening of the monitored signals. Arrows indicate the lack of bars due to the presence of bromine atoms

This means that compound 6 does bind with the grooves of the quadruplex, and, as in compound 4, the uniform perturbation of all groove residues of the quadruplex $[d(TGGGGT)]_4$, with respect to the limited ligand size, suggests that a fast sliding motion of compound 6 inside the groove is also possible. Molecular docking calculations also revealed that two molecules of compound 6 could in principle simultaneously anchor the two opposite ends of the groove although the 5' end is preferred (42 times out of 100) (figure 22 a, b and c).

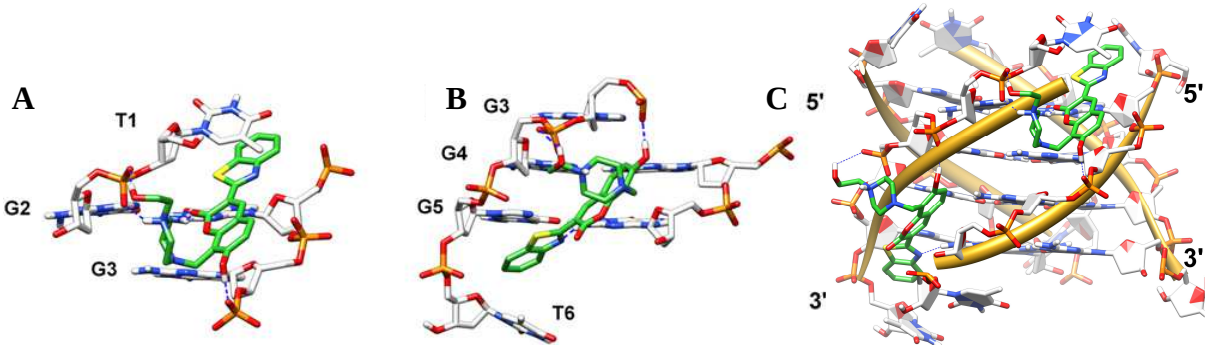
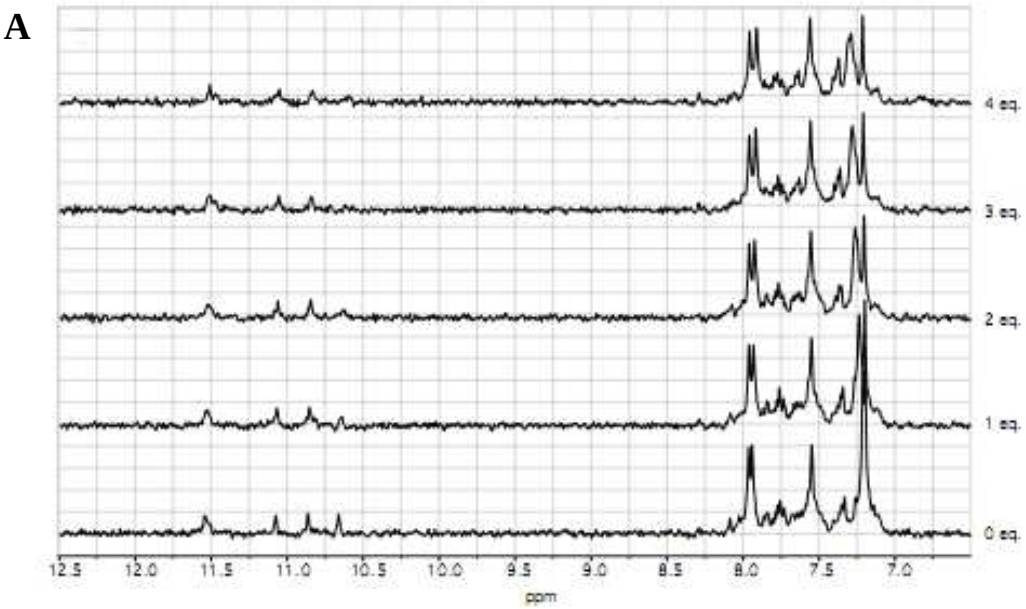


Figure 22: Two binding poses calculated by AD4 for compound 6 **A)** and **B)** in the quadruplex structure. DNA backbone is represented as white stick bonds, **C)** in smooth gold ribbons and stick bonds, special filled rings representations are used for the sugars and bases Ligands are depicted as green sticks. H-bonds are represented as dashed blue lines.. Ligands are depicted as green sticks. H-bonds are represented as dashed blue lines.

Notably, in perfect line with the changes experimentally observed for G2 and G5 aromatic protons signals, the ligand alternatively forms H-bonds with G2 (ΔG = -7.5) or G5 (ΔG = -6.1) bases in the two reported binding modes (figure 22). Finally, the two poses suggested by the docking program, in which a π-π interaction is formed with T1 or T6 ring, could somehow account for the T6-H6 proton and T1-methyl signals shift. The NMR titration is reported in figure 23.



B

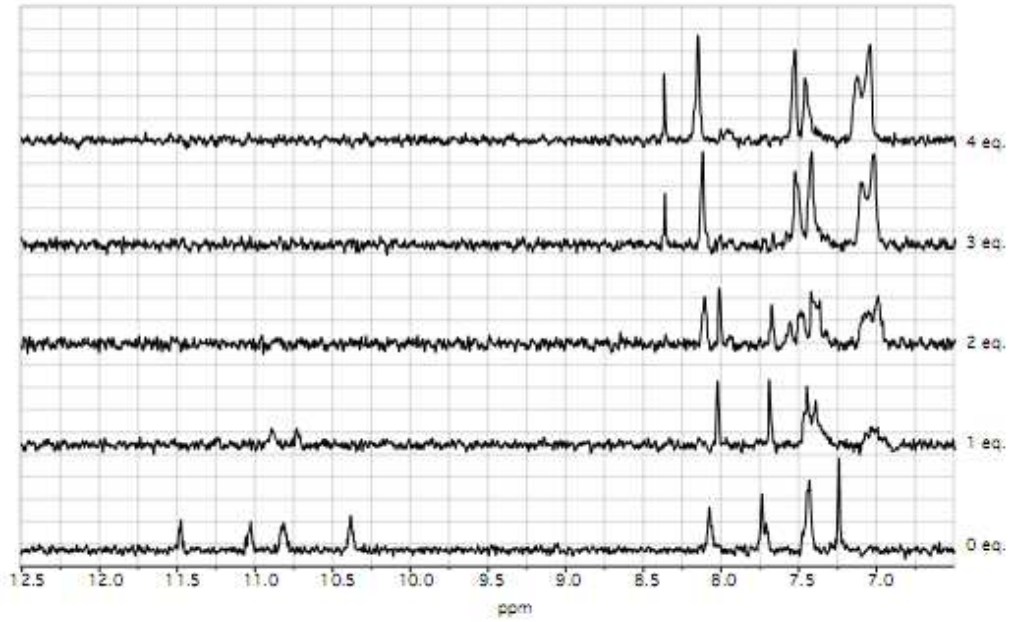


Figure 23: NMR titrations of compounds 6 with the modified quadruplex **A)** $[d(TGG^{Br}GGT)]_4$ and **B)** $[d(TGGGG^{Br}T)]_4$.

Isothermal Titration Calorimetry Measurements

In order to determine the binding affinity of the compounds 1-6 with the quadruplex $[d(TGGGGT)]_4$, ITC titration experiments were carried out.[12]

This kind of study was used in the past also to characterise the interaction of Distamycin A with the same quadruplex and it revealed that four molecules of the drug were accommodated in two opposite grooves of the quadruplex; moreover, the thermodynamic signature of the binding event suggested that the entropic contribution was the one driving the complex formation.[13] An analogous study however was not possible for the six new

ligands identified with the virtual screening, because their poor solubility in water prevented us from performing canonical ITC experiments,[12] in which a solution of the quadruplex is titrated with a concentrated drug solution in the identical aqueous buffer.

Nevertheless, an alternative approach to investigate the affinity of the new drugs for the DNA quadruplex was carried out by performing competition/displacement experiments, by analysing the ability of Distamycin A to bind to the quadruplex in the presence of another compound.14,15 Despite the solubility concerns, mixtures of the quadruplex for each of the six new drugs were successfully prepared by solubilising the molecules in DMSO and diluting them in a phosphate solution containing the quadruplex; these complexes were then titrated with Distamycin A, and the outcome followed by ITC. In figure 24 I report three representative ITC experiments.

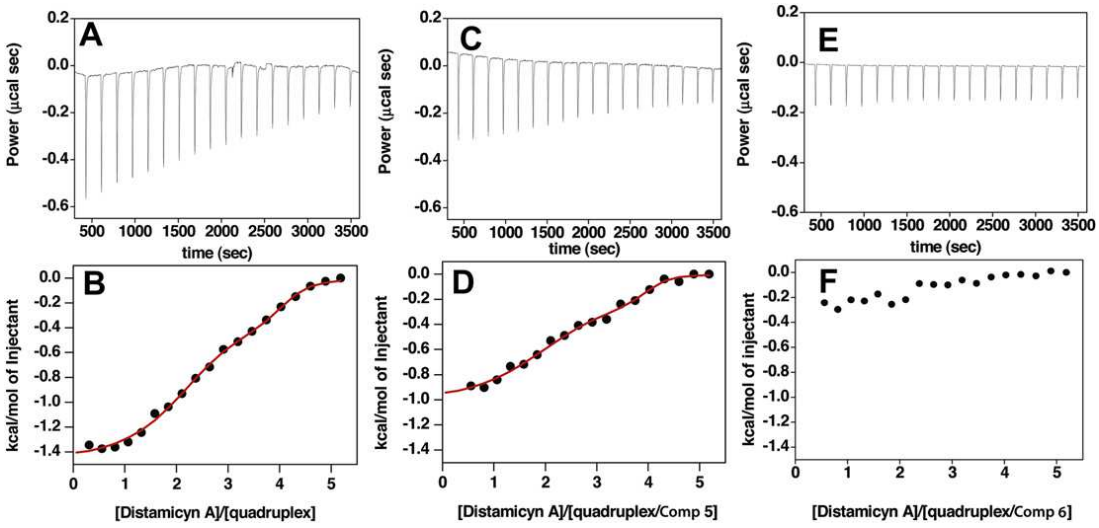


Figure 24: Calorimetric analysis of the interactions of the quadruplex [d(TGGGGT)$_4$] and its complexes with the drugs 5 and 6, with Distamycin A. Raw titration data showing the thermal effect of injecting Distamycin A into a calorimetric cell containing the quadruplex alone **A)**, quadruplex with compound 5 **C)** and quadruplex with compound 6 **E)**. The normalised heat for the titrations, shown in B, D and F respectively, was obtained by integrating the raw data and subtracting the heat of the ligand dilution. The heat effect reveals the typical double step binding curve of the Distamycin A interaction to the quadruplex in B and C, null interaction in F. The red lines in B and D represent the best fit derived by a non-linear least-squares procedure based on an independent binding sites model.

The titration of Distamycin A with [d(TGGGGT)$_4$] was repeated as control, and is clearly shows that in the new experimental conditions this binding event is identical to what was previously observed (figure 24, Panels A and B and Table 1).

Table 1: Thermodynamic parameters for the interaction of Distamycin A with [d(TGGGGT)$_4$] and its complexes formed with the 1, 3 and 5. 2, 4 and 6 are not reported since they abolish the binding of Distamycin A.

Compound in the calorimetric cell	n	Kb (M_1)	ΔH (kcal mol^{-1})	-TDS (kcal mol^{-1})	ΔG (kcal mol^{-1})
[d(TGGGGT)4]	1.8	$(6.0 \pm 0.2) \times 10^5$	-3.0	-7.4	-7.7
	4.0	$(7.0 \pm 0.1) \times 10^6$	-1.5	-7.7	-9.2
[d(TGGGGT)4]/**Compound 5**	1.9	$(3.0 \pm 0.1) \times 10^5$	-0.2	-7.1	-7.3
	4.0	$(7.0 \pm 0.1) \times 10^6$	-1.0	-8.2	-9.2
[d(TGGGGT)4]/**Compound 3**	1.7	$(2.0 \pm 0.1) \times 10^5$	-0.6	-6.5	-7.1
	3.8	$(7.0 \pm 0.1) \times 10^6$	-1.6	-7.6	-9.2
[d(TGGGGT)4]/**Compound 1**	1.9	$(3.0 \pm 0.1) \times 10^5$	-0.7	-6.6	-7.3
	3.8	$(7.0 \pm 0.1) \times 10^6$	-1.6	-7.6	-9.2

The experimental errors on the thermodynamic properties have been evaluated to be <5%

The binding occurs in two separate steps, the first one centred on a stoichiometry of 1:2 in which two molecules of Distamycin A interact with the quadruplex and then a second event in which two more molecules bind, leading to a final stoichiometry of 4 molecules of Distamycin A per quadruplex. Panels C and E report two displacement experiments carried out titrating with Distamycin A the complexes obtained with the drugs 5 and 6 respectively. The two experiments are quite different and indicative of two distinct phenomena.

In the first case, a binding curve similar to the control experiment is obtained (compare figure 24 D and B), indicating that the presence of the drug 5 in the mixture did not have an effect on the association of Distamycin A with the [d(TGGGGT)$_4$], this means that Distamycin A still binds in two distinct events, with stoichiometry and the binding is not significantly affected by the presence of compound 5 (Table 1).

Conversely, the ITC titration of Distamycin A into the quadruplex of the

compound 6 complex (figures 24 E and F) gives a completely different outcome. In this case, Distamycin A appears no longer able to interact with the DNA, suggesting that compound 6 binds to the quadruplex tightly than Distamycin A. It is important to highlight that the competition/displacement experiments do not give any information about the stoichiometry of the complex formed between the quadruplex and the compound competing with the Distamycin A.

A summary of the results for all six compounds is reported in Table 1: compounds 1 and 3 behave as compound 5, whereas the presence of compounds 2 or 4 or 6 abolishes the binding of Distamycin A.

In conclusion the competition ITC experiments represent a possible strategy to evaluate whether an insoluble groove binder is stronger or weaker than a soluble drug that could be used as reference point. In our case, using the Distamycin A as reference drug, we were able to group the six drugs identified with virtual screening in two extreme classes. The first class contains compounds that bind rather weakly to the quadruplex (compounds 1, 3 and 5), whereas in the second class the drugs that bind substantially stronger than Distamycin A (compounds 2, 4 and 6) are clustered.

This study is very interesting because it proved that the application of Virtual Screening calculations together with NMR experiments and Isothermal Titration Calorimetry is a successful strategy for the identification of new molecular chemo types able to bind the grooves of DNA quadruplex structures. The structural diversity of these inhibitors has provided valuable

alternative series for on-going lead optimization aimed at the identification of brand new pharmacological tools, endowed with better affinity and a pharmacokinetic profile and useful in the clarification of the mechanism targeting therapeutic potential use of G-quadruplexes.

Material and method

Oligonucleotide synthesis

The oligonucleotide d(TGGGGT) was synthesized on a Millipore Cyclone Plus DNA synthesizer using solid phase β-cyanoethyl phosphoramidite chemistry at 15 μmol scale. The oligomer was detached from the support and deprotected by treatment with concentrated aqueous ammonia at 55 °C for 12 h. The combined filtrates and washings were concentrated under reduced pressure, redissolved in H_2O, analyzed and purified by high-performance liquid chromatography (HPLC) on a Nucleogel SAX column (Macherey–Nagel, 1000-8/46); using buffer A: 20 mM KH_2PO_4/K_2HPO_4 aqueous solution (pH 7.0), containing 20% (v/v) CH_3CN; buffer B: 1 M KCl, 20 mM KH_2PO_4/K_2HPO_4 aqueous solution (pH 7.0), containing 20% (v/v) CH_3CN; a linear gradient from 0 to 100% B for 30 min and flow rate 1 ml/min were used. The fractions of the oligomer were collected and successively desalted by Sep-pak cartridges (C-18). The isolated oligomer proved to be >98% pure by NMR.

Virtual Screening Calculations

The AutoDock 4.0 (AD4)[3,16] software package, as implemented through the graphical user interface called AutoDockTools (ADT),[7] was used to dock small molecules to the G-Quadruplex structure. The DNA file was prepared using published coordinates (PDB 1S45).[17] Preparation of the DNA and ligand structures for docking calculations was attained according to parametrization suggested by Neidle and co-workers.[18] Thus, point charges were assigned to the DNA according to the AMBER94 force field[19] and all other atom values were generated automatically by ADT. The docking area was assigned visually and centred at the N2 nitrogen of the G4 residue in chain F. A grid of 22.5 Å x 15.0 Å x 26.25 Å with 0.375 Å spacing was calculated around the docking area for 13 ligand atom types using AutoGrid[17]. These atom types were sufficient to describe all atoms in the Life Chemicals diversity set database. For VS, compound structures of the database were prepared using the ZINC database server (http://zinc.docking.org/upload.shtml)[20] to take into account the different protomeric and tautomeric states of each compound. All the ligands were then converted in the AutoDock format file (.pdbqt). For each ligand, 100 separate docking calculations were performed. Each docking calculation consisted of 10 million energy evaluations using the Lamarckian genetic algorithm local search (GALS) method. The GALS method evaluates a population of possible docking solutions and propagates the most successful

individuals from each generation into the subsequent generation of possible solutions. A low-frequency local search according to the method of Solis and Wets is applied to docking trials to ensure that the final solution represents a local minimum. All dockings described in this research were performed with a population size of 150, and 300 rounds of Solis and Wets local search were applied with a probability of 0.06.

A mutation rate of 0.02 and a crossover rate of 0.8 were used to generate new docking trials for subsequent generations, and the best individual from each generation was propagated to the next generation. The docking results from each of the eight calculations were clustered on the basis of root-meansquare deviation (rmsd) between the Cartesian coordinates of the atoms and were ranked on the basis of free energy of binding. The top-ranked compounds were visually inspected for good chemical geometry. Pictures of the modelled ligand/enzyme complexes were rendered with UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco.[21]

Nuclear magnetic resonance experiments

30 NMR samples were prepared at a concentration of 1.7 mM, in 0.2 ml (H_2O/D_2O 9:1) buffer solution having 10 mM KH_2PO_4, 70 mM KCl, 0.2 mM EDTA, pH 7.0. NMR spectra were recorded with Varian ^{Unity}INOVA 700 MHz spectrometer. ^{1}H chemical shifts were referenced relative to external sodium

2,2-dimethyl-2-silapentane-5-sulfonate (DSS). 1D proton spectra were recorded using pulsed-field gradient DPFGSE[10,11] for H_2O suppression.

All 30 tested compounds were checked before titration by NMR and mass spectrometry. All samples turned out to be >95% pure.

The stability of $[d(TGGGGT)]_4$ in DMSO was tested by adding progressively deuterated DMSO in 2.0 mM concentration of quadruplex (0.5 ml H_2O/D_2O 9:1 buffer solution having 10 mM KH_2PO_4, 70 mM KCl, 0.2 mM EDTA, at pH 7.0). After each addition of DMSO-d6, the quadruplex was checked by 1D ^1H NMR experiments.

DMSO was added up to an overall percentage of 30%. At this stage we were able to still observe the four imino resonances, that are very diagnostic of the presence of the quadruplex.

Phasesensitive NOESY spectra[22] were recorded with mixing times of 100 and 200 ms (T = 25 °C). Pulsed-field gradient DPFGSE[10,11] sequence was used for NOESY experiments in H_2O.

NOESY experiments were recorded using STATES-TPPI[23] procedure for quadrature detection. The time domain data consisted of 2048 complex points in t2 and 400 fids in t1 dimension. A relaxation delay of 1.2 s was used. The NMR data were processed on iMAC running iNMR software (www.inmr.net).

Molecular modelling

The binding modes of compounds 1-6 were studied by means of docking experiments with the aid of Autodock4 (AD4)[3,24] and using the $[d(TGGGGT)]_4$ G-quadruplex DNA structure as deposited in the Protein Data Bank (PDB code 1S45) as macromolecules. The 3D structures of all the compounds were generated with the Maestro Build Panel.[25] For the purpose of docking each molecule has been constructed in the protonation state suggested by the MarvinSketch 5.2.5.1 package (http://www.chemaxon.com) using a pH 7.0 accordingly with the NMR titrations. The target DNA structures were prepared through the Protein Preparation Wizard of the graphical user interface Maestro 9.0.2116 and the OPLS-2001 force field. Water molecules were removed, hydrogen atoms were added and minimisation was performed until the RMSD of all heavy atoms was within 0.3 Å of the crystallographically determined positions. Then the constructed compounds and DNA structures were converted to AD4 format files using ADT generating automatically all other atom values. In order to allow the ligands to explore all the possible search space, the docking area has been centred on the mass centre of the quadruplex structure and defined by a box large enough to comprise the entire macromolecule.

Accordingly, grids points of 84× 84×84 with 0.375 Å spacing were calculated around the docking area for all the ligand atom types using AutoGrid4. 100 separate docking calculations were performed for each binder. Each

docking calculation consisted of 25×106 energy evaluations using the Lamarckian genetic algorithm local search (GALS) method. A low-frequency local search according to the method of Solis and Wets was applied to docking trials to ensure that the final solution represents a local minimum. Each docking run was performed with a population size of 150, and 300 rounds of Solis and Wets local search were applied with a probability of 0.06. A mutation rate of 0.02 and a crossover rate of 0.8 were used to generate new docking trials for subsequent generations. The docking results from each of the 100 calculations were clustered on the basis of root-mean square deviation (RMSD 2.0 Å) between the Cartesian coordinates of the ligand atoms and were ranked on the basis of the free energy of binding. All docking solutions were analysed for the coherency with NMR data and for each compound, the lowest energy solution more in line with the experimental data was further considered and subjected to energy minimisation through the OPLS-2001 force field. All figures were rendered using Chimera software package.[26]

Isothermal titration calorimetry

The [d(TGGGGT)$_4$] stock solution was prepared by dissolving the lyophilised compound in 10 mM phosphate buffer with 70 mM KCl, 0.2 mM EDTA, pH 7. The solution was annealed by heating at 95 °C for 5 min and slowly cooling to room temperature. The concentration of the dissolved oligonucleotide was

evaluated by UV measurement at 95 °C, using as molar extinction coefficient the value calculated by the nearest-neighbour model[27] for the sequence d(TGGGGT). Stock solutions of the six drugs were prepared by solubilising weighted amounts in DMSO to a final concentration of 8 mM. The complexes between the quadruplex and the drugs were prepared diluting the drug stock solution into the quadruplex solution to get a final DNA:drug molar ratio of 1:4.1 and a final DMSO concentration of 7%. Distamycin A was solubilised in the same buffer used for the oligonucleotide containing 7% of DMSO.

The titrations were carried out in 10 mM phosphate buffer, 70 mM, KCl, 0.2 mM EDTA, 7% DMSO, pH 7, at 293 K, using a high sensitivity ITC-200 microcalorimeter from Microcal (GE Healthcare). In each experiment, volumes of 2 mL of a 1.2mM Distamycin A solution were added into a 50 mM solution of quadruplex-DNA complex, using a computer-controlled 40-mL microsyringe, with a spacing of 180 s between each injection. Each titration was corrected for heat of dilution by subtracting the measured enthalpies of the injections following saturation. Integrated heat data obtained for the titrations were fitted using a non-linear least-squares minimisation algorithm to a theoretical titration curve, using the MicroCal-Origin 7.0 software package from which the binding parameters $\Delta H°$ (reaction enthalpy change in kcal mol^{-1}), K_b (binding constant in M^{-1}), and n (stoichiometry) were derived. The entropic contribution was calculated using the relationships $\Delta G° = -RT \cdot \ln K_b$ (R 1.987 cal mol^{-1} K^{-1}, T 293 K) and $\Delta G° = \Delta H° - T\Delta S°$.

References

1 Dash J., Shirude P. S., Hsu S. D., Balasubramanian S. **J. Am. Chem. Soc.** 2008, 130, 16048–16056.

2 Li Q., Xiang J., Li X., Chen L., Xu X., Tang Y., Zhou Q., Li L., Zhang H., Sun H., Guan A., Yang Q., Yang S., Xu G. **Biochimie** 2009, 91, 811–819.

3 Huey R., Morris G. M., Olson A. J., Goodsell D. S. **J. Comput. Chem.** 2007, 28, 1145–1152.

4 Dailey M. M., Hait C., Holt P. A., Maguire J. M., Meier J. B., Miller M. C., Petraccone L., Trent J. **O. Exp. Mol. Pathol.** 2009, 86, 141–150.

5 Caceres C., Wright G., Gouyette C., Parkinson G., Subirana J. A. **Nucleic Acids Res.** 2004, 32, 1097–1102.

6 Pellecchia M., Bertini I., Cowburn D., Dalvit C., Giralt E., Jahnke W., James T. L., Homans S. W., Kesler H., Luchinat C., Meyer B., Oschkinat H., Peng J., Schwalbe H., Siegal G. **Nat. ReV. Drug DiscoVery** 2008, 7, 738–745.

7 Sanner M.F. **J Mol Graph Model.** 1999, 17, 57-61.

8 Cosconati S., Marinelli L., Trotta R., Virno A., Mayol L., Novellino E., Olson A.J., Randazzo A. **J. Am. Chem. Soc.** 131 (2009) 16336e16337.

9 a) Virgilio A., EspositoV., Randazzo A., Mayol L., Galeone A. **Nucleic Acids Res.** 33 (19) (2005) 6188-6195; b) Petraccone L., Erra E., Esposito V., Randazzo A., Galeone A., Barone G., Giancola C., **Biopolymers** 77 (2) (2005) 75-85.

10 Hwang T.L., Shaka A.J., **J. Magn. Reson.** A112 (1995) 275-279.

11 Dalvit C. **J. Biomol. NMR** 11 (1998) 437-444.

12 Pagano B., Mattia C.A., Giancola C. **Int. J. Mol. Sci.** 10 (2009) 2935-2957.

13 a) Randazzo A., Galeone A., Mayol L. **Chem. Commun.** 11 (2001) 1030-1031; b) Randazzo A., Galeone A., Esposito V., Varra M., Mayol L. **Nucleosides Nucleotides Nucleic Acids** 21 (2002) 535-545; c) Martino L., Virno A., Pagano B., Virgilio A., Di Micco S., Galeone A., Giancola C., Bifulco G., Mayol L., Randazzo A., **J. Am. Chem. Soc.** 129 (2007) 15950-15956; d) Cosconati S., Marinelli L., Trotta R., Virno A., De Tito S., Romagnoli R., Pagano B., Limongelli V., Giancola C., Baraldi P.G., Mayol L., Novellino E., Randazzo A., **J. Am. Chem. Soc.** 132 (18) (2010) 6425-6433; e) Pagano B., Virno A., Mattia C.A., Mayol L., Randazzo A., Giancola C., **Biochimie** 90 (8) (2008) 1224-1232.

14 Sigurskjold S.W. **Anal. Biochem.** 277 (2000) 260-266.

15 Zhang Y.L., Zhang Z.Y., **Anal. Biochem.** 261 (1998) 139-148.

16 Morris G. M., Goodsell D. S., Halliday R.S., Huey R., Hart W. E., Belew R. K., Olson A. J. **J. Comput.Chemistry**, 1998, 19,1639-1662.

17 Caceres C., Wright G., Gouyette C., Parkinson G., Subirana J.A. **Nucleic Acids Res.** 2004, 32, 1097-1102.

18 Evans D. A., Neidle S. **J. Med. Chem.** 2006, 49, 4232-4238

19 Cornell W. D., Cieplak P., Bayly C. I., Gould I. R., Merz K. W. Jr., Ferguson D. M., Spellmeyer D. C., Fox T., Caldwell J. W., Kollman P. A. A. **J. Am.**

Chem.Soc. 1995, 117, 5179-5197.

20 Irwin J. J., Shoichet B. K. **J. Chem. Inf. Model.** 2005, 45, 177-182.

21 Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E. **J. Comput. Chem.** 2004, 25, 1605-1612.

22 Jeener J., Meier B., Bachmann H.P., Ernst R.R., **J. Chem. Phys.** 71 (1979) 4546-4553.

23 Marion D., Ikura M., Tschudin R., Bax A. **J. Magn. Reson.** 85 (1989) 393-399.

24 Cosconati S., Hong J.A., Novellino E., Carroll K.S., Goodsell D.S., Olson A.J., **J. Med. Chem.** 51 (2008) 6627-6630.

25 Maestro, Version 9.0.211. Schrodinger, L.L.C., New York, NY, 2009.

26 Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E. **J. Comput. Chem.** 25 (13) (2004) 1605-1612.

27 Cantor C.R., Warshaw R.R., Shapiro H. **Biopolymers** 9 (1970) 1059-1077.

“Fine tuning” of the selected compounds

In order to improve the binding affinity of the selected compounds we have also performed Similarity Search. Herein, I report some preliminary results about the structural optimization performed on compound 6, who showed the more promising results.

The rationale of this new approach is that we calculate the degree of similarity between the control compound and all the molecules present in a given databases; this value of similarity is measured with the *Tanimoto index* (T index) who can assume values between 0 (for molecules without a degree of similarity) and 1 (for molecules structurally identical to the reference compound).

In our case, the entire database ZINC8, having more than thirteen million compounds, and a T index = 0.8 (that means to consider 80% of similarity) have been used. In this way, we have found 61 new compounds that have in common with compound 6 the scaffold 2H-cromen-2-ones. Also in this case we have performed NMR titration for all compounds to evaluate the binding mode with the DNA quadruplex and we used the NMR data to draw a structure-activity relationship (SAR), in order to understand which part of the drug is important for a strong interaction with the quadruplex and which part do not permit the interaction. We have found 16 (figure 1) molecules that present a good affinity for the quadruplex, and for all was done a biological study to evaluate if these compounds cause a DNA damage.

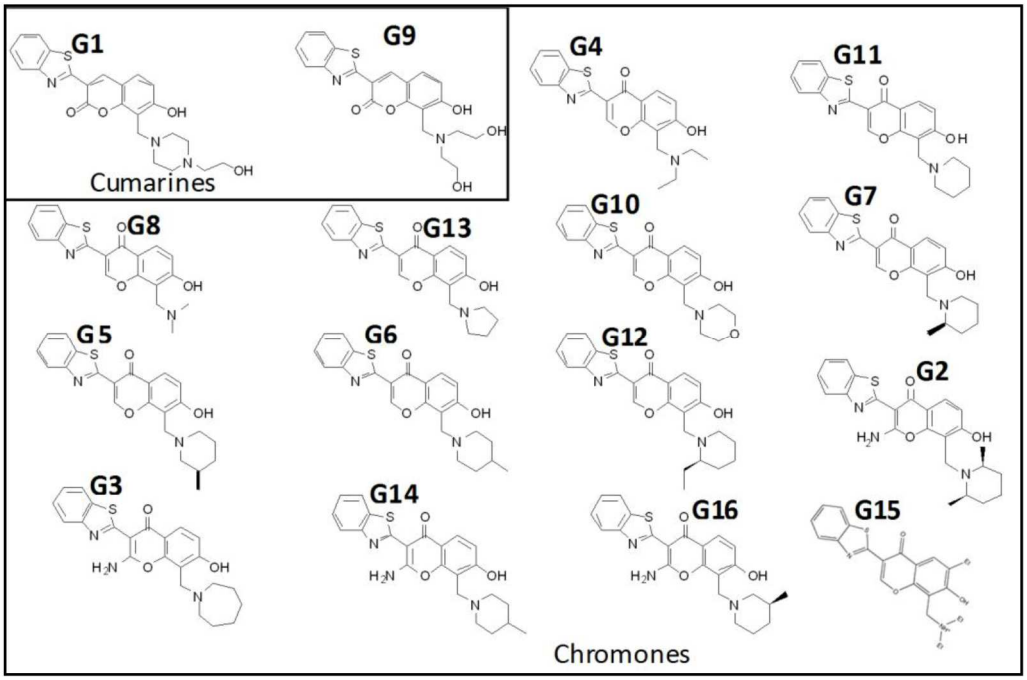


Figure 1: The 16 compounds founded, they have cumarine and chromone structures.

All the selected compounds were also tested from a biological point of view. In particular, human transformed fibroblasts (BJ-EHLT) were exposed to different concentrations (doses 0.1 and 0.5 μM) of each compounds for 24 hours and activation of damage response was measured by immunofluorescence (IF) (figure 2) using antibodies against γ-H2AX and TRF1 to mark DNA damage and telomeres respectively.

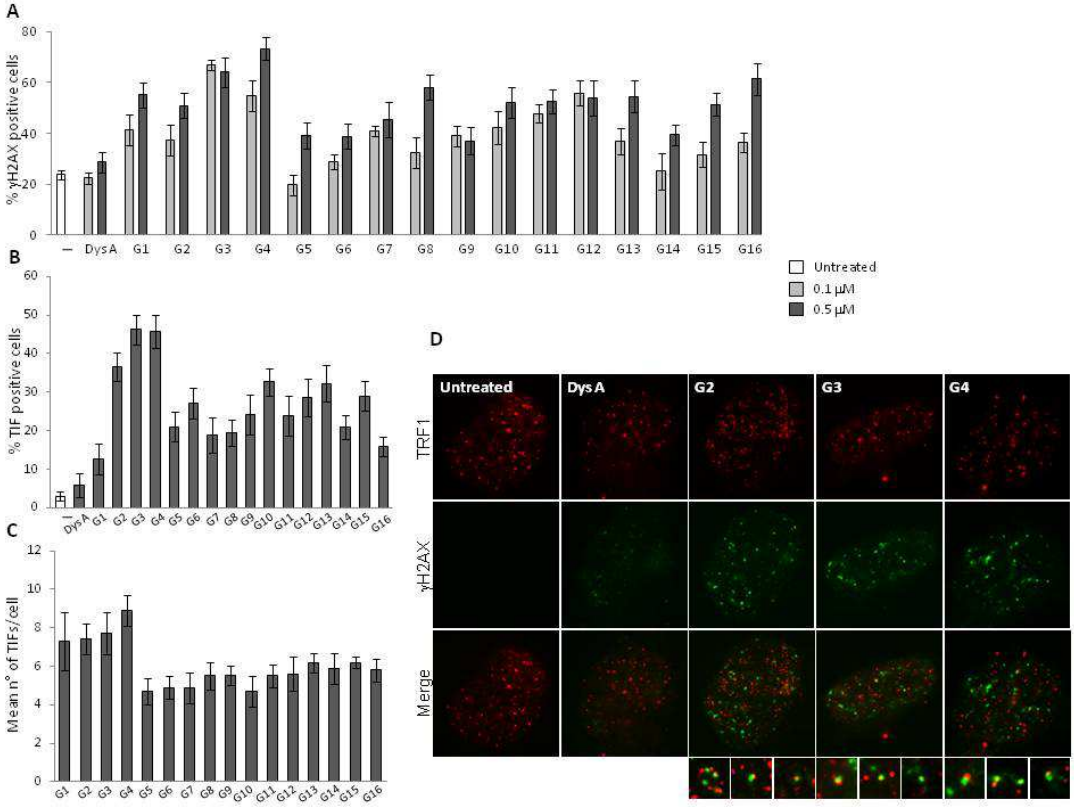


Figure 2: Analysis of DNA damage response activation at telomeres by different ligands. BJ-EHLT fibroblasts were treated for 24 hrs with the doses 0.1 and 0.5 μM of Dystamicin A (Dys) and ligands from G1 to G16. Cells were processed for immunofluorescence (IF) using antibodies against γH2AX and TRF1 to mark DNA and telomeres damage, respectively. Percentage of γH2AX- **A)** and TIF-positive cells **B)** in untreated and cells exposed to the indicated treatments. Cells with four or more γH2AX/TRF1 foci were scored as TIF positive. Mean number of γH2AX/TRF1 **C)** foci in the indicated samples. Three independent experiments were evaluated, and error bars indicate the standard deviation. **D)** Representative images of IF from γH2AX/TRF1 staining of untreated and Dys A, G2, G3, G4-treated BJ-EHLT fibroblasts are shown. Enlarged views are reported below the merged images from the samples treated with the G-ligands which have induced the highest percentage of TIF positive cells among all the compounds analyzed in (B). The images were acquired with a Leica Deconvolution microscope (magnification 100x).

It is noteworthy that the cells treated with the compounds from G1 to G4 (G1 was formerly named compound 6) induced strong phosphorylation of γ-H2AX, and this confirms that these compounds are DNA damage inducers (figure 2A). The four more active compounds (G1-G4) were also tested in order to clarify whether the damage could be localized at telomere level. Very interestingly, they were able to induce only telomeric damage (TIFs - telomere dysfunction-induced foci) (see figure 2B). The average number of TIFs per nucleus are visible in figure 2C. Cell with four or more γH2AX/TRF1 foci were scored as TIF positive.

In order to verify whether γ-H2AX was phosphorylated in response to dysfunctional telomeres, a deconvolution microscopy has been performed, that revealed that some of the damage foci induced by the compounds G2,G3,G4 colocalized with TRF1, an effective marker for interphase telomeres (figure 2D). [1,2]

We analyzed also the biological effect of treatment with G2 (the compound displaying better results) on HeLa cells. This drug is clearly capable to induce the tumor cell death in HeLa cells to different doses (figure 3A) and it is also capable to block the cell cycle in G0-G1 phase (figure 3B), that means we have a cellular death for apoptosis. The SA-b-gal activity (Senescence-associated-β-Galactosidase Staining) was also tested, and it showed (figure 3C) the presence of some cells in senescence.

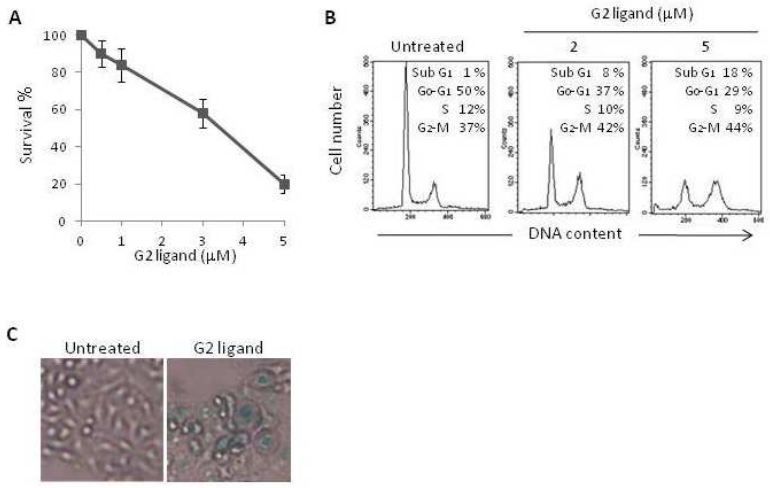
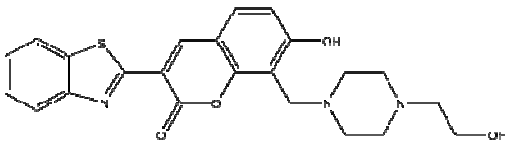
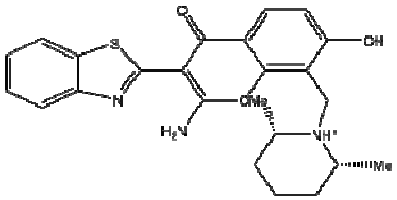


Figure 3: Biological effects of treatment with G2 ligand. **A)** Survival curve of the HeLa cells exposed to different doses of G2 ligand ranging from 0.5 to 5 μM. Surviving fractions were calculated as the ratio of absolute survival of the treated sample/absolute survival of the control sample. The data represent the mean of four independent experiments with standard deviations (SD). **B)** Cell cycle analysis of HeLa cells untreated or treated with 2 and 5 μM G2 ligand. Cells were processed 5 days after G2 exposure. y axis, cell numbers; x axis, relative DNA content on the basis of staining with propidium iodide. The percentages of the different cell cycle phases reported in the bar graph were calculated by the ModFit LT software. **C)** The induction of senescence was evaluated by the SA-β-gal staining of Hela cells untreated and treated with 5 μM G2 ligand for 5 days. A representative of three independent experiments with comparable results is shown. Original magnification, 40x.

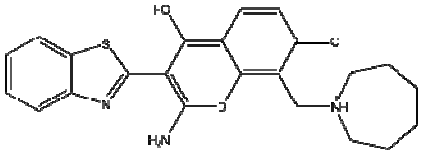
These interesting results need to be validated by in vivo experimental models of human cancers. In particular, the evaluation of toxicological profile and anticancer activity of this promising new groove binders (figure 4) will allow us to assess its therapeutic index and to proceed along the development process to future clinical applications.



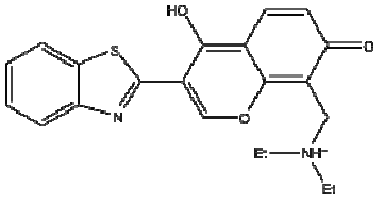
Composto 6 (G1)



Composto G2



Composto G3



Composto G4

Figure 4: The new groove binders.

References

1 van Steensel B., de Lange T. **Nature** 1997 385,740-743

2 Takai H., Smogorzewska A., de Lange T. **Curr. Biol.** 2003 13, 1549-1556

Conclusions

A successful structure-based virtual screening (VS) campaign was recently undertaken by our research group and it resulted in the identification of six small molecules able to interact with the groove of the quadruplex $[d(TGGGGT)]_4$.

In this dissertation, I have reported NMR spectroscopy experiments which, combined with extensive molecular docking studies, allow for a more detailed picture of the interaction between each VS-derived binder and the quadruplex DNA. Noteworthy, Isothermal Titration Calorimetry (ITC) measurements revealed that compounds 2, 4, and 6, despite their relatively small dimensions, bind substantially stronger than Distamycin A, which is, to the best of our knowledge, the most potent groove binder identified so far. As it is widely accepted that specificity among the various DNA G-quadruplexes, that might be simultaneously present in the human genome, is a fundamental requirement for the quadruplex binder to become a drug, extensive binding tests towards DNA duplex and different DNA quadruplex topologies will be the next step in our research program. The future results from the combination of a different approaches with structural studies will provide a source of inspiration for the design of next generation of potent and selective quadruplexDNAdrug-like binders.

CHAPTER III

La Protein

The La protein is a monomeric phosphorprotein that can be found in the nucleus of eukaryotic cells. The La protein was first discovered as an autoantigen found in patients suffering from the rheumatic diseases systemic lupus erythematosus and Sjogren's syndrome.[1, 2]

Although La was first characterized as a human protein, homologs have been identified in a wide variety of eukaryotes.[3–12]

A lot of experiments revealed that the La protein is associated with a very large number of nascent small RNAs for example: pre-tRNAs, pre-5S rRNA, U6 small nuclear RNA (snRNA), RN_{ase} P RNA, MRP RNA, 7SL RNA, Y RNAs, rodent $4.5S_I$ and $4.5S_{II}$ RNAs, and transcripts of Alu sequences.[13–20]

Others experiments revealed that La also binds a number of viral-encoded RNAs, including the adenovirus-encoded VA RNA_I and VA RNA_{II},[13, 21] the Epstein-Barr virus–encoded EBER 1 and EBER 2 RNAs,[22] and the leader RNAs of several negative-strand viruses.[23–26]

For the majority of cellular RNAs, La binds the precursors, rather than the mature RNAs. For instance the U6 RNAs, bound by the human La protein, are extended at the 3'-end and contain fewer modifications than the mature RNA.20 Similarly, the La protein binds precursors to tRNAs and 5S rRNA, but does not bind the corresponding mature RNAs.[15]

The reason why the La protein associates primarily with nascent RNA polymerase III transcripts became clear when it was shown that at least part of the recognition site for the protein is UUU_{OH}, which is at 3' end of virtually all newly synthesized RNA polymerase III transcripts.[27–29] These terminal uridylates are usually removed during RNA maturation, thus eliminating the binding site for the La protein. Consistent with this, the yeast La protein also binds certain nascent RNA polymerase II–transcribed small RNAs that end in UUU_{OH}.[30, 31]

In most cases, the 3'-end binding of the La protein is only transient, the protein being displaced upon excision or modification of the 3'-UUU_{OH} motif.

So far, the best-established role of the La protein is to provide protection against 3'-exonucleolytic degradation with various consequences depending on the nature of the transcript.

For pre-tRNAs and pre-U3 snRNA, La protein binding allows normal processing of the RNA.[31, 32] The La protein was also shown to exhibit an RNA chaperone activity,[33] promoting correct folding of certain pre-tRNAs[34] or facilitating the assembly of U snRNAs into functional ribonucleoprotein particles.[30, 35] As a consequence of these activities the La protein most probably takes part in the quality-control mechanism of newly synthesized noncoding RNAs.[36-38] In addition genuine La proteins are versatile factors that were reported in yeasts, flies, and mammals to bind to certain mRNAs and to influence their translation efficiency,[39-44] their stability, or subcellular localization;[45-47] many reports established that the HsLa (human genuine La) protein directly binds to several viral RNAs, often enhancing their translation and replication

efficiency[48-52] and, at least for some viruses, shielding them from the host's defence systems.[52]

In general, most mRNAs for which La positively effects translation have lengthy and complex 5′ UTRs, whereas studies of specific mRNAs whose outcomes revealed a negative effect on translation were on mRNAs with short presumably nonstructured 5′ UTRs such as the very short 5′ UTR containing 5′TOP mRNAs.[53,54-56] In this regard, it is relevated that the nonphosphorylated (cytoplasmic) isoform of the human La protein can recognize the 5′ GpppN cap (and related structures) present on cellular mRNAs.[57]

Binding to the caps of 5′ TOP mRNAs by nonphospho-La may inhibit cap-mediated translation initiation. The same activity on complex, e.g., IRES mRNAs can conceivably shift the balance toward internal initiation. This would not necessarily exclude a more direct role for La at the IRES element (HCV mRNA) itself or just upstream of the AUG, as for MDM2 mRNA.[42]

Structural analysis of La proteins

Alignment of La proteins from species ranging from humans to trypanosomes reveals that these proteins can be divided into at least three regions (figure 1). The N-terminus of all known La proteins contains the ~60 amino acid and this is called **La motif**, a highly conserved sequence that is also present in a number of otherwise unrelated proteins.[7,58,59,60] Following the La motif, there is a less well conserved RNA recognition motif, called **RRM** (RNA Recognition Motif),[61,62] and a highly charged, weakly conserved C-terminus. Consistent with the idea that the C-terminus represents a distinct domain of La, this portion of the human protein can be severed from the N-terminal half of the protein by a variety of proteases.[4,63] The molecular size of the La protein varies from ~50 kilodaltons in vertebrates to 32 kilodaltons in the yeast S. cerevisiae (figure 1). Almost all of the additional mass of the larger La proteins is due to an expanded C-terminal domain (figure 1).

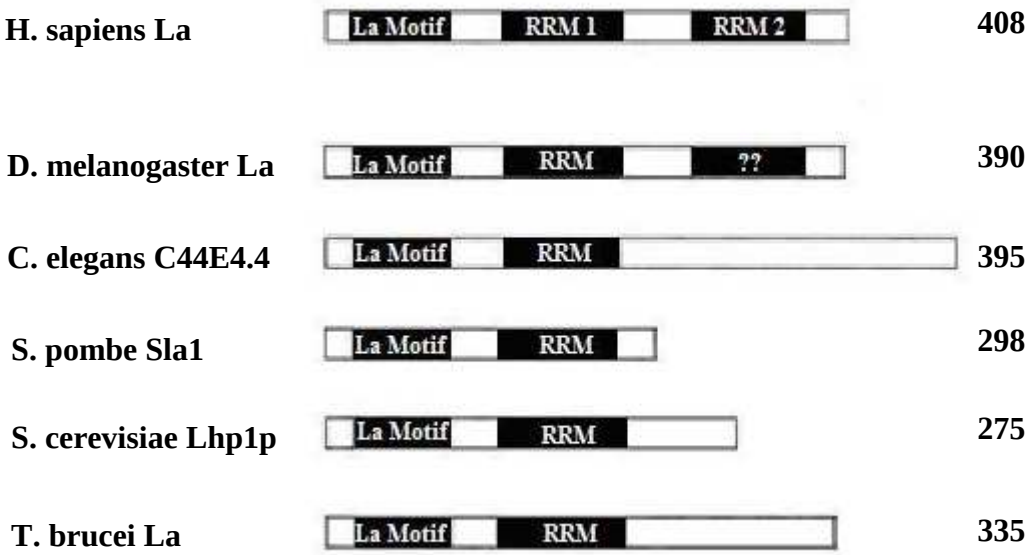


Figure 1: Structure of La proteins. The human, Drosophila, C. elegans, S. pombe, S. cerevisiae, and T. brucei proteins are aligned. For each protein, the length in amino acids is indicated on the right.

The C-terminal domain is the least conserved part of the La protein, varying in both size and sequence between species. With the exception of genuine La homologs from a few species,[64,65] the COOH-terminal domain (CTD) of genuine La proteins contains the atypical RRM (RRM2) followed by an unstructured terminal region that contain several phosphorylation sites, a short basic motif (SBM), a nuclear localization signal (NLS), and a nuclear retention element (NRE) [10,66-69] (figure 1). Disruption of these trafficking elements causes mislocalization of La and malfunction in tRNA maturation, and in some cases leads to disorder of some of the pre-tRNA processing steps.[70,71]

Metabolic labeling experiments have revealed that La proteins from humans

to yeast are phosphorylated in vivo.[10,72,73] The human La protein is phosphorylated at multiple sites, all of which appear to be located within the C-terminal domain.[63,71,72] To date, four phosphorylation sites have been mapped in the human protein. The major site of phosphorylation is a casein kinase II site at Ser 366.[73,74] In addition, the protein is phosphorylated at Thr 302, Ser 325, and Thr 362.[74] On the basis of sequence comparison, none of these sites appear conserved beyond vertebrate La proteins. Consistent with a lack of conservation, the major sites of phosphorylation were recently mapped in the *S. cerevisiae* La protein. Two of the three major sites of phosphorylation map to the N-terminus of the protein, and the major C-terminal site does not correspond to a casein kinase site.[75]

In the past the structure of the La motif has been debated, because some predicted that it folded as a canonical RRM, others contended that it adopted a predominantly helical configuration;[64,68,76] but all doubts were resolved with a work of Alfano et al.[77] where an NMR study resolved the structure of both domains: La Motif and RRM1 (figure 2).

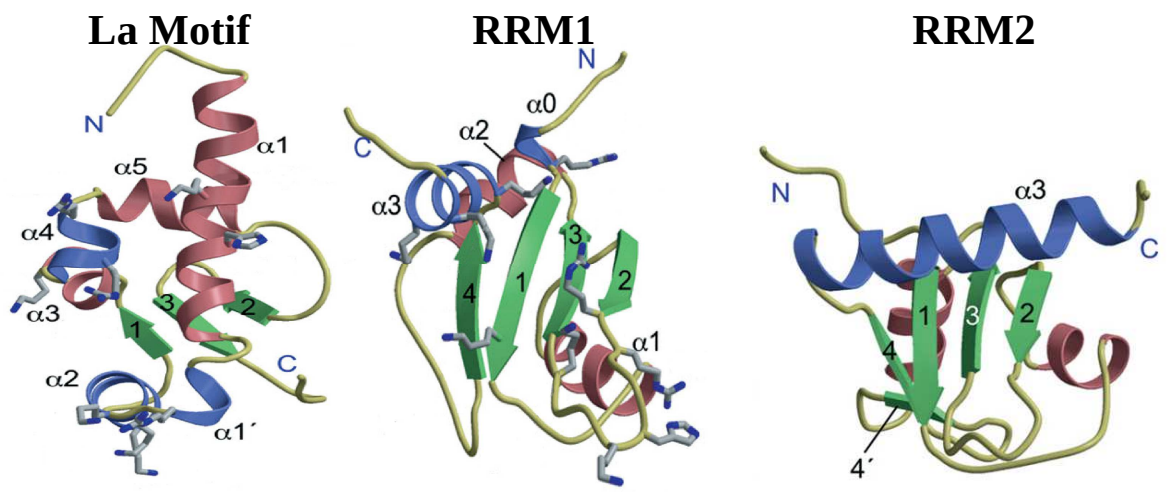


Figure 2: Secondary structure of La Motif, central RRM (RRM1) and C-terminal RRM2 of La genuine protein. (Taken from Alfano et al.[77]).

They have demonstrated that the La Motif does not adopt an RRM-like fold, as was proposed in the past, but its compact structure comprises six a-helices and a three-stranded antiparallel β-sheet[78-80] (figure 2). The structure of the La motif is dominated by a very long amphipathic a-helix (a1), which is followed by a shorter helix labeled a1'. The polar flank of helix a1 is largely solvent exposed but the opposite apolar face makes hydrophobic contacts with every other secondary structure element in the La Motif, forming the hydrophobic core of the domain. The twisted three-strand β-sheet contributes apolar residues to the domain core but also presents a solvent-exposed surface on the opposite side of the domain to helix a1. The C-terminal extension of strand β3 contains a well-defined loop followed by an extended arm that folds back onto the molecule, contacting helix a1' and the loops connecting it to a1 and a2 (figure 2).

RRM1 (figure 2) has a classical RRM structure consisting of a four-strand β-sheet backed by two a-helices but has an unusual C-terminal a-helix that projects up from the β-sheet RNA-binding surface. Helix a3 is predominantly hydrophilic and protrudes away from the body of the domain and toward the solvent. This configuration is also in marked contrast to the long, amphipathic helix a3 of the RRM2 of hLa, which forms an extensive network of hydrophobic interactions with an apolar patch on the β-sheet, thereby obscuring the putative RNA-binding surface[81] (figure 2). Thus the β-sheet surface of the RRM1 seems to be largely available for RNA interactions.

This NMR study has demonstrated that the La Motif and RRM1 are independently folded and maintain the same three-dimensional structure in the context of an NTD fragment. It was also confirmed that the two domains act synergistically in the RNA binding because when they are present in isolation, neither La Motif nor the RRM1 has affinity for oligo (U) RNA.

Whilst the conserved NTD of La binds to 3'-UUU_{OH} sequences common to pol III transcripts and other small RNAs, for the CTD the role in RNA binding is less well characterized, but it appears to have a distinct role in RNA recognition and is reported to be involved in binding 5'-triphosphates and internal sequences in structured RNA molecules.[57,64,68,82-85] Moreover, the CTD has been proposed to play a role in pol III reinitiation, although the involvement of La in transcription termination and the regulation of recycling of pol III transcription complexes is still controversial.[64,68]

In the 2003 Jacks et al.[81] resolved the structure of the RRM2 (figure 2).

This RRM2 has a different structure from canonical RRMs, in fact it possesses a β1-α1- β2- β3- α2- β4'- β4- α3 topology and folds to generate a five stranded, antiparallel β-sheet that is terminated by a long α helix which obscured the RNA binding surface. Until recently, it seemed that the RRM2 domain did not display the ability to bind RNA,[81] although several reports suggested that the C-terminal domain is involved in the regulation of the complex nuclear and cytoplasmic activities of genuine La proteins, at least in mammals.[86]

We know that the La protein binds specifically to the 3' polyU ends characteristic of most Pol III precursors, but in the 2011 Martino L. et al.[87] have demonstrated that La is capable to bind to several viral RNAs. They studied the binding to the Hepatitis C Virus (HCV) genome in the stem-loop IV of the IRES domain (Internal Ribosomal Entry Site). By systematic biophysical investigations, they have discovered that La binds to domain IV using an RNA recognition mechanism that is quite distinct from its mode of binding to RNAs with a 3' UUU_{OH} trailer: although the La Motif and first RNA recognition motif (RRM1) are sufficient for high-affinity binding to 3' oligoU, recognition of HCV domain IV requires the La Motif and RRM1 to work in concert with the atypical RRM2 which has not previously been shown to have a significant role in RNA binding. This new mode of binding does not appear sequence specific, but recognizes structural features of the RNA, in particular a double-stranded stem flanked by single stranded extensions.

Poly-U binding of La protein

Previous studies have established that both domains, the La Motif and RRM1, were required for recognition of 3′-UUUOH by La;[85,88] a co-crystal structure containing the La domain from human La (hLa) and a short RNA ending in UUU-3′OH revealed that this mode of binding surprisingly occurs through contacts other than to the expected surfaces,[89] but nonetheless confirming the requirement for both motifs for RNA binding (figure 3). Furthermore this co-crystal structure analysis has highlighted that the linker polypeptide connecting the two domains fold into an a helix when there is an RNA binds, this appears to be an indirect effect of RNA binding because there are no contacts between the RNA and this portion of the polypeptide.[89]

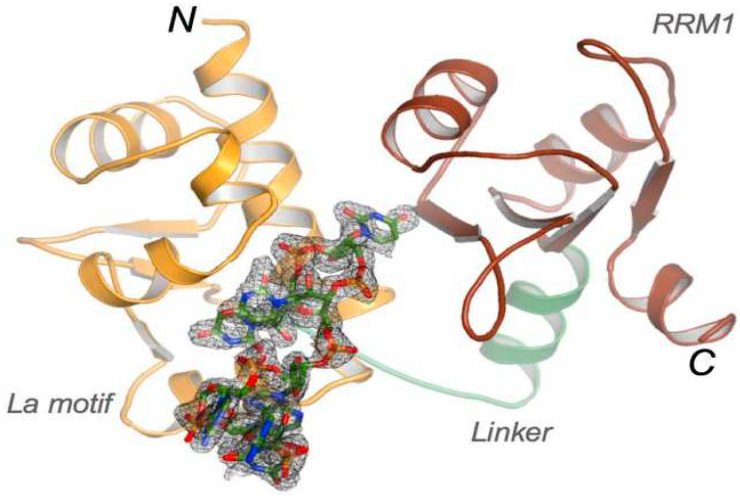


Figure 3: A co-crystal structure of binding modes of the La motif and RRM1 of the human La protein (NTD) with AUAAUUU RNA bound. La motif is in orange; interdomain linker in green and RRM1 in brown. (Taken from Kotik-Kogan et al. 2008.[89])

To explain the binding mode of La protein Kotik-Kogan et al.[89] have made NMR and crystal studies with different RNA oligomers ended with 3'-UUU_{OH}; in particular they studied the following oligos: AUUUU, AUAAUUU, AUAUUUU,UUUUUUUU for the crystal structures, whilst for the NMR study they used UU, UUU, UUUU, UCUU, AUUUU, UUUUUUUU RNA oligomers.

The crystal structure of the complex LaNTD: AUUUU revealed a number of important structural details of the recognition of the 3' end of the RNA oligomer that help to explain the specificity of the LaNTD for sequences terminating in UUUOH. The major interactions are made by the terminal pair of nucleotides (U_{-1}/U_{-2}), which consistently make the closest contacts with the protein and appear to provide the majority of the binding energy for the interaction (figure 4).

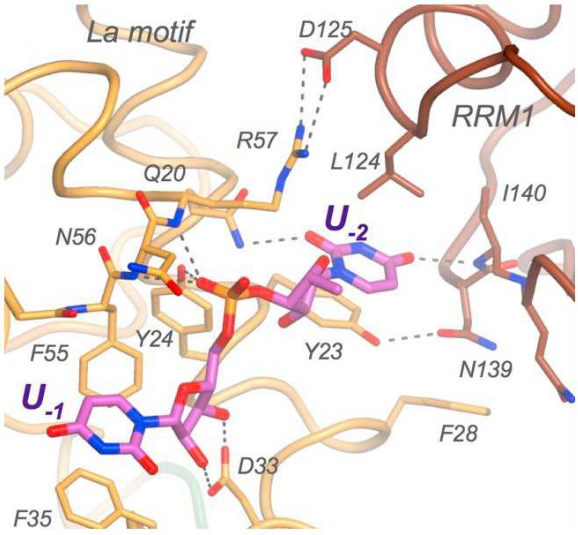


Figure 4: Structure of the LaNTD:AUUUU complex showing just the last two nucleotides, U_{-1} and U_{-2}. The RNA is shown as a stick model. The protein is colored by domain: La Motif orange and RRM1 brown; selected side chains are shown as sticks with their carbon atoms colored by domain. Hydrogen bonds are shown as dashed lines. (Taken from Kotik-Kogan et al. 2008.[89])

The NMR studies have outlined that very similar patterns of chemical shift perturbation were observed in all classes, consistent with a common mode of 3' end recognition that involves both the La Motif and RRM1 domains and appears largely dominated by the interactions made by the last two nucleotides in the sequence, U_{-1} and U_{-2}, as found in all crystal structures.

U_{-1} interacts exclusively with the La motif (figure 4), specific H-bond interactions between U_{-1} and the La motif only involve backbone features of the nucleotide: the 2' and 3' OH groups from the ribose ring of U_{-1} are specifically recognized by hydrogen bonds to the O atoms of the side-chain carboxylate moiety of D33 (figure 4), additionally, the O1 atom of the U_{-1} phosphate group makes hydrogen bonds with both the backbone NH groups of N56 and R57 and with the side-chain OH of Y24. Consistent with the structure, these interactions with the ribose-phosphate backbone have been shown by mutagenesis to be important for 3' end recognition.[79,90] The penultimate nucleotide, U_{-2}, makes intimate contacts with the protein that involve both the La motif and RRM1 (figure 4). In contrast to U_{-1}, it is the nucleotide base of U_{-2} and not its backbone that makes specific hydrogen bonds to the protein. Atoms O2 and O4 from the pyrimidine ring of U_{-2} hydrogen bond to the side-chain amide of Q20 (La motif) and the main-chain amide of I140 (RRM1), respectively, a pair of interactions that clearly helps to draw the two domains together. The pyrimidine ring of U_{-2} is also stacked directly on top of Y23, a residue from the La motif that is positioned by a hydrogen bond from the tyrosine side chain to the side chain of N139 of

RRM1, which in turn stacks on top of F28 from the La motif (figure 4). The U_{-2} pyrimidine ring is also packed directly underneath L124 from RRM1, a residue that is at least partly secured in place by the salt bridge made by its neighbor (D125) with R57 from the La motif.

Thus, the recognition of U_{-2} in the complex engages a concerted set of protein:RNA and protein:protein interactions involving residues from both domains to form a tightly defined pocket that is specific in size, shape, and hydrogen-bonding capacity for a uridylate base; NMR analyses confirm that these interdomain interactions only form upon RNA binding. This induced fit of the binding pocket around U_{-2} accounts well for the cooperative nature of RNA binding by both domains of LaNTD.[76,77,79,85]

References

1 Mattioli M., Reichlin M. **Arthritis Rheum.** 197417:421–29

2 Alspaugh M.A., Tan E.M. **J. Clin.Invest.** 1975.55:1067–73

3 Chambers J.C., Keene J.D. **Proc. Natl. Acad. Sci. USA** 1985.82:2115–19

4 Chambers J.C., Kenan D., Martin B.J., Keene J.D. **J. Biol. Chem.** 1988.263: 18043–51

5 Topfer F., Gordon T., McCluskey J. **J. Immunol.** 1993.150:3091–100

6 Scherly D., Stutz F., Lin-Marq N., Clarkson S.G. **J. Mol. Biol.** 1993.231:196–204

7 Yoo C.J., Wolin S.L. **Mol. Cell. Biol.** 1994.14:5412–24

8 Bai C.Y., Li Z.H., Tolias P.P. **Mol. Cell. Biol.** 1994. 14:5123–29

9 Lin-Marq N., Clarkson S.G. **J. Mol. Biol.** 1995.245:81–85

10 Van Horn D.J., Yoo C.J., Xue D., Shi H., Wolin S.L. **RNA** 1997.3:1434–43

11 Marchetti M.A., Tschudi C., Kwon H., Wolin S.L., Ullu E. **J. Cell Sci.** 2000. 113:899–906

12 Westermann S., Weber K. **Biochim. Biophys. Acta** 2000.1492:483–87

13 Lerner M.R., Boyle J.A., Hardin J.A., Steitz J.A. **Science** 1981.211:400–2

14 Hendrick J.P., Wolin S.L., Rinke J., Lerner M.R., Steitz J.A. **Mol. Cell. Biol.** 1981.1:1138–49

15 Rinke J., Steitz J.A. **Cell** 1982.29:149–59

16 Hashimoto C., Steitz J.A. **J. Biol. Chem.** 1983.258:1379–82

17 Chambers J.C., Kurilla M.G., Keene J.D. **J. Biol. Chem.** 1983.258:11438–

41

18 Reddy R., Tan E.M., Henning D., Nohga K., Busch H. **J. Biol. Chem.** 1983, 258: 1383–86

19 Shen C.K., Maniatis T. **J. Mol. Appl. Genet.** 1982.1:343–60

20 Rinke J., Steitz J.A. **Nucleic Acids Res.** 1985.13:2617–29

21 Rosa M.D., Gottlieb E., Lerner M.R., Steitz J.A. **Mol. Cell. Biol.** 1981.1:785–96

22 Lerner M.R., Andrews N.C., Miller G., Steitz J.A. **Proc. Natl. Acad. Sci. USA** 1981.78:805–9

23 Kurilla M.G., Cabradilla C.D., Holloway B.P., Keene J.D. **J. Virol.** 1984.50:773–78

24 Kurilla M.G., Keene J.D. **Cell** 1983.34:837–45

25 Wilusz J., Kurilla M.G., Keene J.D. **Proc. Natl. Acad. Sci. USA** 1983. 80:5827–31 26.

26 Wilusz J., Keene J.D. **Virology** 1984.135: 65–73

27 Francoeur A.M., Mathews M.B. **Proc. Natl. Acad. Sci. USA** 1982.79:6772–76

28 Reddy R., Henning D., Tan E., Busch H. **J. Biol. Chem.** 1983. 258:8352–56

29 Stefano J.E. Cell 1984.36:145–54

30 Xue D., Rubinson D.A., Pannone B.K., Yoo C.J., Wolin S.L. **EMBO** J. 2000.19: 1650–60

31 Kufel J., Allmang C., Chanfreau G., Petfalski E., Lafontaine D.L., Tollervey D. **Mol. Cell. Biol.** 2000.20:5415–24

32 Yoo C.J., Wolin S.L. **Cell.** 199789:393-402.

33 Belisova A., Semrad K., Mayer O., Kocian G., Waigmann E., Schroeder R., Steiner G. **RNA** 2005. 11: 1084–1094.

34 Chakshusmathi G., Kim S.D., Rubinson D.A., Wolin S.L. **EMBO** J. 2003.22: 6562–6572.

35 Pannone B.K., Xue D., and Wolin S.L. **EMBO** J. 1998.17: 7442–7453.

36 Copela L.A., Chakshusmathi G., Sherrer R.L., Wolin S.L. **RNA** 2006.12: 644–654.

37 Huang Y., Bayfield M.A., Intine R.V., Maraia R.J. **Nat. Struct. Mol. Biol.** 2006. 13: 611–618.

38 Kadaba S., Wang X., Anderson, J.T. **RNA** 2006.12: 508–521.

39 Holcik and Korneluk; **Mol Cell Biol.** 200020:4648-57.

40 Kim Y.K., Back S.H., Rho J., Lee S.H., Jang S.K. **Nucleic Acids Res.** 2001.29: 5009–5016.

41 Cardinali B., Carissimi C., Gravina P., Pierandrei-Amaldi P. **J. Biol. Chem.** 2003.278: 35145–35151.

42 Trotta R., Vignudelli T., Candini O., Intine R.V., Pecorari L., Guerzoni C., Santilli G., Byrom M.W., Goldoni S., Ford L.P., et al. **Cancer Cell** 2003.3: 145–160.

43 Inada and Guthrie **Proc. Natl. Acad. Sci. U. S. A.** 2004;101: 434–439

44 Vazquez-Pianzola P., Urlaub H., Rivera-Pomar R. **Proteomics** 2005.5: 1645–1655.

45 McLaren R.S., Caruccio N., Ross J. **Mol. Cell. Biol.** 1997.17: 3028–3036.

46 Adilakshmi T. Laine R.O. **J Biol Chem.** 2002;277:4147-51.

47 Brenet F., Dussault N., Borch J., Ferracci G., Delfino C., Roepstorff P., Miquelis R., L. **Mol. Cell. Biol.** 2005.25: 7505–7521

48 Costa-Mattioli M., Svitkin Y., Sonenberg N. **Mol. Cell. Biol.** 2004;24: 6861–6870

49 Raha T., Pudi R., Das S., Shaila M.S. **Virus Res.** 2004;104:191-200.

50 Domitrovich A.M., Diebel K.W., Ali N., Sarker S., Siddiqui A. **Virology.** 2005; 25;335:72-86

51 Xue Q., Ding H., Liu M., Zhao P., Gao J., Ren H., Liu Y., Qi Z.T. **Arch. Virol.** 2007; 152: 955–962.

52 Bitko V., Musiyenko A., Bayfield M.A., Maraia R.J., Barik S. **J. Virol.** 2008.82: 7977–7987.

53 Schwartz E., Intine R.V., Maraia R.J., **Mol. Cell. Biol.** 2004 24 9580–9591.

54 Svitkin Y.V., Evdokimova V.M., Brasey A., Pestova T.V., Fantus D., Yanagiya A., Imataka H., Skabkin M.A., Ovchinnikov L.P., Merrick W.C., Sonenberg N., **EMBO J.** 2009 28 58–68.

55 Svitkin Y.V., Ovchinnikov L.P., Dreyfuss G., Sonenberg N., **EMBO J.** 1996 15 7147–7155.

56 Zhu J., Hayakawa A., Kakegawa T., Kaspar R.L., **Biochim. Biophys. Acta** 2001 1521 19–29.

57 Bhattacharya R., Perumal K., Sinha K., Maraia R., Reddy R., **Gene Exp.** 2002 10 243–253.

58 Sobel S.G., Wolin S.L. **Mol. Biol. Cell** 1999 10:3849–62

59 Chauvet S., Maurel-Zaffran C., Miassod R., Jullien N., Pradel J., Aragnol D. **Dev. Dynam.** 2000. 218:401–13

60 Tan Q., Li X., Sadhale P.P., Miyao T., Woychik N.A. **Mol. Cell. Biol.** 2000. 20:8124–33

61 Bandziulis R.J., Swanson M.S., Dreyfuss G. **Genes Dev.** 1989. 3:431–37

62 Query C.C., Bentley R.C., Keene J.D. **Cell** 1989. 57:89–101

63 Chan E.K.L., Francoeur A.M., Tan E.M. **J. Immunol.** 1986. 136:3744–49

64 Wolin S.L., Cedervall T. **Annu. Rev. Biochem.** 2002. 71: 375–403.

65 Fleurdepine S., Deragon J.M., Devic M., Guilleminot J., Bousquet-Antonelli C. **Nucleic Acids Res.** 2007. 35: 3306–3321.

66 Simons F.H., Broers F.J., Van Venrooij W.J., Pruijn G. J. **Exp. Cell Res.** 1996. 224: 224–236

67 Rosenblum J.S., Pemberton L.F., Bonifaci N., Blobel G. **J. Cell Biol.** 1998. 143: 887–899.

68 Maraia R.J., Intine R.V. **Mol. Cell. Biol.** 2001. 21: 367–379.

69 Intine R.V., Dundr M., Misteli T., Maraia R.J., **Mol. Cell** 2002 9 1113–1123.

70 Bayfield M.A., Kaiser T.E., Intine R.V., Maraia R.J. **Mol. Cell. Biol.** 2007 27 3303–3312.

71 Pizer L.I., Deng J-S, Stenberg R.M., Tan E.M. **Mol. Cell. Biol.** 1983. 3:1235–45

72 Francoeur A.M., Chan E.K.L., Garrels J.I., Mathews M.B. **Mol. Cell. Biol.** 1985. 5:586–90

73 Fan H., Sakulich A.L., Goodier J.L., Zhang X., Qin J., Maraia R.J. **Cell**

1997. 88:707–15

74 Broekhuis C.H., Neubauer G., van der Heijden A., Mann M., Proud C.G., et al. **Biochemistry** 2000. 39:3023–33

75 Long K.S., Cedervall T., Walch-Solimena C., Noe D.A., Huddleston M.J., et al. **RNA** 2001. 7:1589–602

76 Ohndorf U.M., Steegborn C., Knijff R. & Sondermann P. **J. Biol. Chem.** 2001 276, 27188–27196.

77 Alfano C., Sanfelice D., Babon J., Kelly G., Jacks A., Curry S., Conte M.R. **Nat. Struct. Mol. Biol.** 2004 7–14

78 Kenan D.J., Keene J.D. **Nat. Struct. Mol. Biol.** 200411 303–305

79 Dong G., Chakshusmathi G., Wolin S.L., Reinisch K.M., **EMBO J.** 200423 1000–1007.

80 Clark K.L., Halay E.D., Lai E., Burley S.K., **Nature** 1993 364 412–420.

81 Jacks A., Babon J., Kelly G., Manolaridis I., Cary P.D., Curry S., Conte Maria R. **Structure** 2003 11, 833–843.

82 Maraia R.J., Intine R.V. **Gene Expr.** 2002.10, 41–57.

83 Ali N., Pruijn G.J., Kenan D.J., Keene J.D., Siddiqui A. **J. Biol. Chem.** 2000. 275, 27531– 27540.

84 Fan H., Goodier J.L., Chamberlain J., Engelke D.R., Maraia R.J. **Mol. Cell. Biol.** 1998 18, 3201– 3211.

85 Goodier J.L., Fan H., Maraia R.J., **Mol. Cell. Biol.** 1997 17 5823–5832.

86 Park J.M., Intine R.V., Maraia R.J. **Gene Expr.** 2007.14: 71–81.

87 Martino L., Pennell S., Kelly G., Bui Tam T. T., Kotik-Kogan O., Smerdon

Sthephen J., Drake Alex F., Curry S., Conte Maria R. **Nucleic Acids Research** 2011 1–14.

88 Horke S., Reumann K., Schulze C., Grosse F., Heise T., **J. Biol. Chem.** 2004 279 50302–50309.

89 Kotik-Kogan, O., Valentine E.R., Sanfelice D., Conte M.R., Curry S. **Structure** 2008.16: 852-862.

90 Teplova M., Yuan Y.R., Phan A.T., Malinina L., Ilin S., Teplov A., Patel D.J., **Mol. Cell** 2006 21 75–85.

CHAPTER IV

La-related proteins (LARPs)

The La Module (La Motif plus RRM1) is also present on proteins otherwise unrelated to genuine La, referred to as La-related proteins (LARPs).[1-11] Contrary to the well characterized genuine La protein, functional and structural data on LARPs are very scarce, and to date only a few LARPs have been characterized.

Phylogenetic analyses, together with structural motif searches allowed us to propose that besides the genuine La proteins family, LARPs can be divided into four clearly distinct families (LARP1, 4, 6, and 7). The greatest degree of homology between these various LARP classes of proteins occurs in the La Motif and also the RRM1 (figure 1).

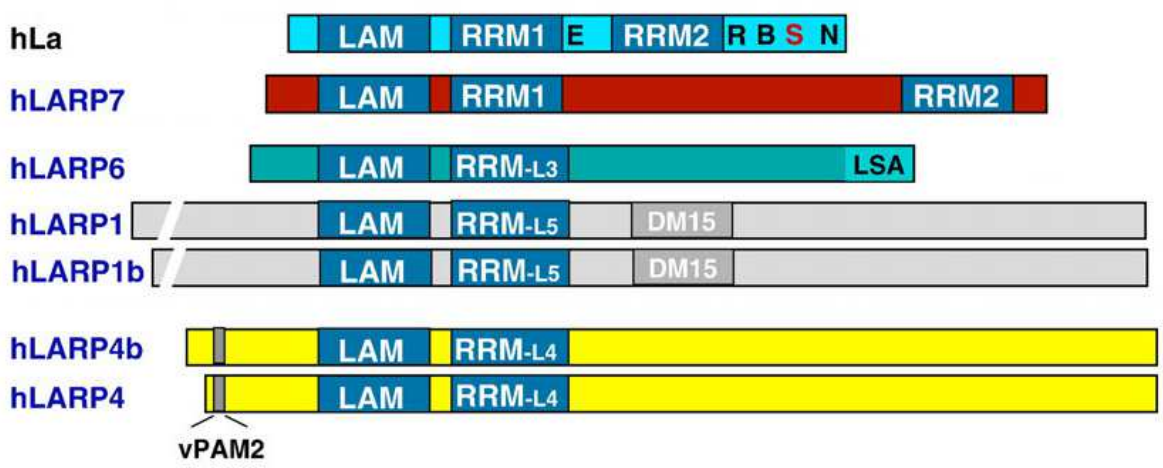


Figure 1: Schematic alignment indicating the conserved La Motif–RRM domain architecture of the genuine La proteins followed by LARPs (Taken from Bayfield Mark A. et al.[13]).

Initially, because of their similarity to La genuine protein, the researchers thought that LARPs also had same features of binding, in other words that they were capable to bind ligand with UUU-3'OH ended, but this has not always proved true. In fact, whereas some LARPs are indeed capable to recognize the RNA ending with the UUU-3'OH like the genuine La protein, LARP4 shows a different mechanism of recognition in that instead of binding poly (U) RNA it binds poly (A) RNA.

In humans two LARP4 family members are found (hLARP4 and hLARP4b, previously called hLARP5),[12] which share 37% identity and 53% similarity throughout their amino acid sequence. Unlike other LARPs, LARP4 and LARP4b diverge in several of the amino acids critical for canonical UUU-3'OH recognition, suggesting that, as already outlined, that the LARP4 family is the most diverged from genuine La proteins and other LARPs in the RNA binding pockets of their La Module.[13]

Role of LARP4

In 2010 the Conte's group at King's College in collaboration with others, have discovered the role and functions of LARP4.[14]

Using several techniques as Nuclear Magnetic resonance (NMR), crystallography, Isothermal Titration Calorimetry (ITC), researchers have found that LARP4 can bind poly(A) RNA via the La Module and interacts with the MLLE domain of poly (A) binding protein (PABP) via a non-canonical PAM2

sequence.

This discovery is very important because, contrary to what it was previously believed (that the two domains, La Motif and RRM, of LARPs protein such as in the La genuine protein, work together to bind the RNA ligands ending in UUU-3'OH) it is now clear that instead LARP4 binds poly-A RNA, and this different interaction involves in a different function. In fact the protein LARP4 binding the poly (A) RNA present at the 3' untranslated region (UTR) of most mRNAs, stabilising it and therefore stimulating translation.

Biophysical and nuclear magnetic resonance (NMR) studies have shown that a peptide representing LARP4 PAM2w interacts with the MLLE of PABP with a K_d within the range found for other PAM2 motifs. This is significant, because it suggests that LARP4 (and probably LARP5/4b) functions as parts of a network of proteins that compete for the PABP MLLE domain.

This PAM2w differs from all other known PAM2 sequences in that it contains a Trp in place of Phe, the latter being the most important contributor to MLLE binding[15] and this study showed that this mutation did not affect the overall binding affinity to MLLE PABP.

The interaction of LARP4 with the RNA has been investigated by EMSA assays, where the binding properties of the La modules from of human La(1-235) and LARP4(1-286) have been compared. Among the homopolymers tested, A(20) exhibited the best affinity for LARP4-NTD, as reflected by the ratio of free to bound RNA, whereas U(20) showed less binding and C(20) and G(20) showed no binding (figures 2 A to D).

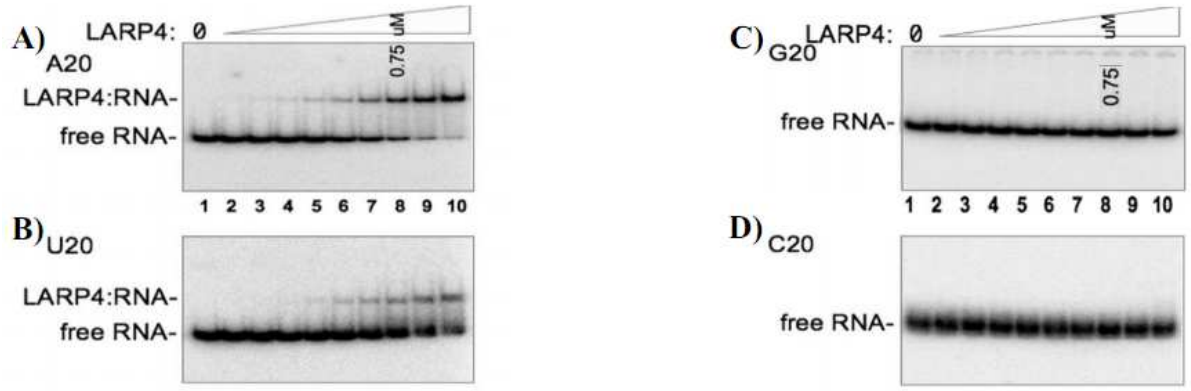


Figure 2: EMSA of homopolymeric 20-mer RNAs for binding to the RNA binding domain of LARP4 protein. (Taken from Yang R. et al.[14])

LARP4 binding to RNA was not affected by addition of 3' phosphate; this is a difference with La protein that is capable to bind only ssRNA ending with 3' hydroxyl (figure 3).

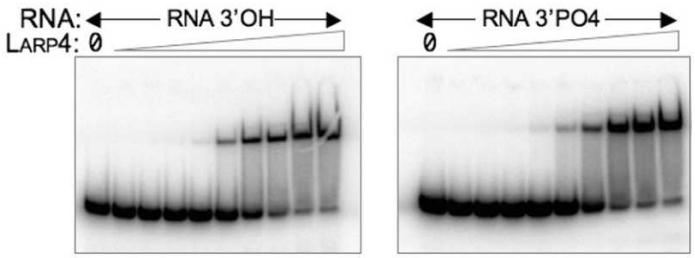
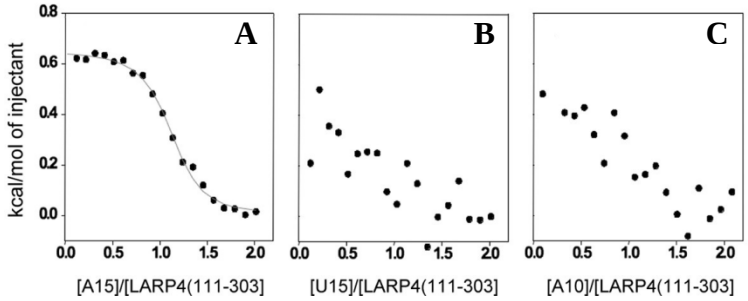


Figure 3: Comparison of LARP4(1-286) binding to otherwise identical 36-mer RNAs that end with a 3'-OH or a 3'-phosphate. (Taken from Yang R. et al.[14])

ITC (figure 4) was employed to study in more detail the RNA binding properties of La Motif-RRM of LARP4. The interaction of LARP4 (111-303) with

three different RNAs was tested. This study confirmed that LARP4 is not capable to bind to short RNAs, and that it only binds to poly (A) and not poly (U) RNAs. In fact as shown in panels below, it binds to poly A15 (see panel A) but not poly A10 (see panel C) indicating a requirement for a specific length of its RNA target. This is a specific feature that differs from the binding profile of genuine La protein.



Figures 4: (A to C) Isothermal titration calorimetric analysis of LARP4(111-303) interactions with A(15), U(15), and A(10). The K_d and other thermodynamic parameters derived from this analysis are reported in Table 1. (Taken from Yang R. et al.[14])

TABLE 1. Thermodynamic parameters of the interaction of La(1-194) and LARP4(111-303) with single-stranded RNA sequence[a]

Protein	n	K_d ($1/K_b$) (nM)	$\Delta H°$ (kcal mol^{-1})	$-T\Delta S°$ (kcal mol^{-1})	$\Delta G°_{298K}$ (kcal mol^{-1})
La(1-194)/U(15)	0.8	752	−39.0 ± 0.2	30.7 ± 0.2	−8.3 ± 0.4
La(1-194)/A(10)					
La(1-194)/A(15)					
LARP4(111-303)/U(15)					
LARP4(111-303)/A(10)					
LARP4(111-303) A(15)	1.0	714	0.7 ± 0.1	−9.1 ± 0.3	−8.4 ± 0.4

[a] The values represent the averages and the standard deviations over three independent measurements.

The titration analysis of LARP4(111-303) with A(15), has shown that binding occurs as one event centred on a molar ratio of 1. LARP4(111-303) interacts

with A(15) with a K_d of 714 nM, with entropically driven binding (Table 1). In contrast, LARP4(111-303) displayed very weak association with U(15) and A(10) with a K_d beyond the threshold that could be rigorously measured by ITC, i.e., ≥0.1 mM. Therefore, LARP4 binds to A(15) at least 200-fold more tightly than U(15) and A(10).

References

1 Yoo, C.J. and Wolin, S.L. **Mol. Cell. Biol.** 1994.14: 5412–5424.

2 Saget, O., Forquignon, F., Santamaria, P., and Randsholt, N.B. **Genetics** 1998.149: 1823–1838.

3 Sobel, S.G. and Wolin, S.L. **Mol. Biol. Cell** 1999.10: 3849–3862.

4 Remillieux-Leschelle, N., Santamaria, P., and Randsholt, N.B. **Genetics** 2002. 162: 1259–1274.

5 Aigner, S., Postberg, J., Lipps, H.J., and Cech, T.R. **Biochemistry** 2003. 42: 5736–5747.

6 Witkin, K.L. and Collins, K. **Genes & Dev.** 2004. 18: 1107–1118.

7 Valavanis, C., Wang, Z., Sun, D., Vaine, M., and Schwartz, L.M. **Gene** 2007. 393: 101–109.

8 He, N., Jahchan, N.S., Hong, E., Li, Q., Bayfield, M.A., Maraia, R.J., Luo, K., and Zhou, Q. **Mol Cell.** 2008. 29: 588–99.

9 Krueger, B.J., Jeronimo, C., Roy, B.B., Bouchard, A., Barrandon, C., Byers, S.A., Searcey, C.E., Cooper, J.J., Bensaude, O., Cohen, E.A., et al. **Nucleic Acids Res.** 2008. 36: 2219–2229.

10 Markert, A., Grimm, M., Martinez, J., Wiesner, J., Meyerhans, A., Meyuhas, O., Sickmann, A., and Fischer, U. **EMBO Rep.** 2008. 9: 569–575.

11 Nykamp, K., Lee, M.H., and Kimble J. **RNA** 2008. 14: 1378–1389.

12 Bousquet-Antonelli, C., and Deragon J. M. **RNA** 2009. 15:750–764.

13 Bayfield Mark A., Yang R. and Maraia Richard J. **Biochimica et**

Biophysica Acta 2010 1799 365–378

14 Yang R., Gaidamakov Sergei A., Xie J., Lee J., Martino L., Kozlov G., Crawford Amanda K., Russo Amy N., Conte Maria R., Gehring K., and Maraia Richard **J. Mol-Cell. Biol.** 2011.31: 542-555

15 Kozlov G., De Crescenzo G., Lim Nadia S, Siddiqui N., Fantus D., Kahvejian A., Trempe Jean-François, Elias D., Ekiel I., Sonenberg N., O'Connor-McCourt M. and Gehring K. **EMBO J.** 2004. 23:272–281.

16 Kozlov, G., and Gehring K. **PLoS One** 2010. 5:e10169.

CHAPTER V

Materials and Methods

My project concentrated on the study of the structure of the two domains of LARP4: La Motif 111-196 and RRM 196-287.

Before studying the NMR spectra, I expressed protein in bacteria system, so I have learnt how to purify the proteins.

In this chapter, I will illustrate the different protocols that I have applied to express and to purify the protein, in order to obtain a good NMR sample.

Plasmid and bacterial cell strains used in this study

The two domain of LARP4: La Motif(111-196) and RRM(196-287) were cloned in Conte's laboratory in a pET-Duet-1 expression vector and expressed in E.Coli Rosetta II cells.

Transformation of competent cells

The competent cell (BL21 Rosetta from Novagen) was taken and 1 µl of required plasmid DNA was added in a sterile Eppendorf tube. The mixture was then left on ice for 30 minutes. After that the cells were "heat shocked"

at 42°C for 30 seconds and put in the ice for further 2 minutes. After that 250 µl of LB were added followed by incubation at 37°C for 1 hour. An aliquot (100 µl) of transformed cells was spread onto an LB agar plate containing the proper antibiotic (0.100 mg/ml Amplicillin and 34 mg/ml Chloramphenicol), and incubated at 37°C.

Expression studies

A single colony was inoculated in 10 ml of LB media supplemented with antibiotics, like Ampicillin and Chloramphenicol, as appropriate, and grown to an OD600 of 0.6. Next, isopropyl-β-D-thiogalactopyronoside (IPTG) was added to a final concentration of 1 mM so as to induce production of the desired protein. The cultures were grown for a further 3 hours at 37°C before harvesting by centrifugation (3000 x g/10 minutes at 4°C).

Large scale expression of unlabelled protein

Cells previously tested for protein expression were then grown at 37°C overnight in 100 ml of LB medium and on the day after diluted in 2L of LB medium containing antibiotics before being incubated at 37°C at 220 rpm. When the OD_{600} reached values around 0.6, the protein expression was induced for 3 hours with 100 µg/ml of IPTG. In the next stage the cells were harvested at 7000 rpm for 20 minutes and frozen until further use.

Large scale expression of isotope-labelled proteins

The expression of isotopically enriched proteins was carried out in M9 medium (per litre: 6.0 g Na_2HPO_4, 3.0 g KH_2PO_4, 0.5 g NaCl, 1.0 g NH_4Cl, sterilised in autoclave and supplemented with 2 g Glucose, 1 mL 1 M Thiamine, 1 ml 0.1M $CaCl_2$, 2 ml 1 M $MgSO_4$, antibiotics and 10 ml M9 trace elements see Table 2). For the expression of protein [^{15}N]-labelled, $^{15}NH_4Cl$ from Cambridge Isotope Laboratories was used at 1.0 g/L. In the case of [^{13}C]-labelled protein, uniformly labelled [^{13}C]-Glucose (Cambridge Isotope Laboratories) was used at 2.0 g/L. As described above, for unlabelled medium the cells were grown until OD_{600} of 0.6 and then the protein expression was induced using IPTG. After 3 hours of expression time the cells were harvested and frozen as above.

Purification of proteins

All the cell pellets were resuspended in 30 mL of Nickel buffer A (Table 2) with the addition of Lysozime and phenylmethylsulfonyl fluoride (PMSF) in final concentration 2.0 mg/ml, used as protease inhibitor. The cells were then cracked by sonication (3 minutes with a pulser 5 sec on and 5 sec off at amplitude of 30%). The cells lysate was harvested at 1700 rpm for 1 hour. The supernatant was then separated from the cells particles and ready for

purification.

Below I report the sequential steps for each purification and the details of the columns and buffers used.

Ni-affinity column → TEV Cleavage → Ni-affinity column → Heparin column

Chromatography procedures

Ni-affinity column

A Ni-NTA (nichel-nitilotriacetic acid) resin from Qiagen was used in this project for affinity purification of both domains of LARP4 recombinant protein expressed with a $(His)_n$ tag.

Superflow resin is capable of binding up to 30 mg of protein. The column was equilibrated with 2 column volume (CV) of Ni-column Buffer A (see Table 2), then the protein sample was loaded with flow rate of 2 mL/min. Elution of bound proteins was obtained using a linear gradient of 20 CV of Ni-column Buffer B and was collected in fractions of 3 mL.

TEV Cleavage

The TEV protease is a highly site-specific cysteine protease that is found in the Tobacco Etch Virus (TEV). The optimum recognition site for this enzyme is the

sequence Glu-Asn-Leu-Tyr-Phe-Gln-Ser and cleavage occurs between the Gln and Ser residues. Some of the advantages of this enzyme are its high specificity and its high activity rate. One of the main uses of this protein is for removing affinity tags from purified proteins.

To remove the Tag present at the N-terminal of the protein, a TEV Cleavage was used. The protein after a Ni-NTA column was dialysed overnight against TEV buffer (see Table 2). When the reaction was complete, the solution was loaded on a benchtop Ni-affinity column to separate the cleaved protein from Tag-TEV (which has His-Tag) and uncleaved product, at the end the fractions collected were put overnight against Heparin buffer dialysis.

Heparin column

5 ml HiTrap™ Heparin HP Columns prepacked with Heparin Sepharose™ High Performance (Pharmacia) was used for purification of both LARP4s.

Before the use, the column was equilibrated with 2 CV of Heparin-Buffer A, at 3 mL/min flow rate. Once the protein was loaded, the elution occurred using a linear gradient of a 30 CV Heparin-Buffer B.

TABLE 2: List of buffers used during the purification steps performed in this project

Ni-column A: 50 mM Tris, 10 mM Imidazol, 300 mM NaCl, 5% Glycerol, pH 8.0
Ni-column B: 50 mM Tris, 500 mM Imidazol, 300 mM NaCl, 5% Glycerol, pH 8.0
TEV cleavage: 50 mM Tris, 1 mM EDTA, 1 mM DTT, 50 mM NaCl, pH 8.0
Heparin buffer dialysis: 50 mM Tris, 100 mM KCl, 0,2 mM EDTA, 1 mM DTT, pH 7.25
Heparin A: 50 mM Tris, 10% Glycerol, pH 7.25
HeparinB: 50 mM Tris, 10% Glycerol, 2M KCl, pH 7.25
NMR buffer dialysis: 20 mM Tris, 100 mM KCl, 0,2 mM EDTA, 1 mM DTT, pH 7.25
M9 trace elements: 0.5 g $FeCl_3$, 0.05 g $ZnCl_2$, 0.01 g $CuCl_2$, 0.01 g $CoCl_2$, 0.01 g H_3BO_3, 5 g EDTA per 1L

NMR samples preparation

All the NMR samples were prepared at the required concentration by centrifugation at 4000 rpm using Vivaspin with a filter cut off 3 kDa, from a diluted solution coming from overnight dialysis in NMR buffer. A maximum of 10% D_2O was added to all samples.

NMR measurements

All spectra were acquired at 293K on Bruker NMR spectrometers operating at a 500 and 700 MHz proton frequency.

Spectral width of 10 ppm for the ^{1}H and 32ppm for ^{15}N dimensions were used. The spectral width of ^{13}C were chosen according to the spectrum 28 ppm for the HNCA,[1] 70 ppm for the CBCA(CO)NH.[2,3]

The mixing time for the heteronuclear 3D NOESY experiments was 100 or 120 ms. Squared shifted sine-bell windows functions were used in all dimensions before Fourier transformation. Proton chemical shifts are referred to DSS at 0 ppm and were calibrated on the water resonance at 4.65 ppm; nitrogen and carbon chemical shifts were calculated by indirect referencing.[4]

References

1 Kay L.E., Ikura M., Bax A. **J. Am. Chem. Soc.**, 1990. 112: 888-889.

2 Grzesiek S., Bax A. **J. Am. Chem. Soc.**, 1992. 114: 6291–6293.

3 Grzesiek S., Bax A. **J. Magn. Reson.**, 1992. 99: 201–207.

4 Wishart David S., Bigam Colin G., Yao J., Abildgaard F., Dyson H. J., Oldfield E., Markley John L., Sykes Brian D. **Journal of Biomolecular NMR**, 1995. 6: 135-140.

Results

In my work I carried out structural studies of the two domains of LARP4, namely LARP4 (111-196) (corresponding to the La Motif) and LARP4 (196-287) (the RRM1). This entails sample preparation followed by acquisition and analysis of the NMR experiments.

For LARP4 (111-196) a previous student in the Conte's lab found the best condition to study this domain by NMR, but as part of my work I carried out the purification for this domain for NMR studies. I also carried out the NMR assignment, the side chain assignment and structural calculation.

For LARP4 (196-287) domain I found the best condition to obtained a good NMR sample and I have done the purification and completed the backbone assignment.

For the La module I have carried out the purification and backbone assignment (based on the data found for individual domains).

In this part of the thesis I describe in detail the application of the NMR techniques such as multidimensional heteronuclear experiments, applied in combination with routine isotopic enrichment of the protein, which permitted the detailed characterisation of the two domains.

Purification of LARP4 (111-196)

The LARP4 (111-196) clone used tag containing six histidines at the N-terminal end of the protein that allow an easy purification on immobilised metal affinity columns. The supernatant, containing tagged LARP4 (111-196) obtained after centrifugation at 17000 rpm for 40 minutes, was loaded onto a Ni-NTA column from Qiagen equilibrated with the Ni-affinity column buffer A (see Table 2 Chapter V). Elution was performed using 250 mM Imidazol with linear gradient (figure 5).

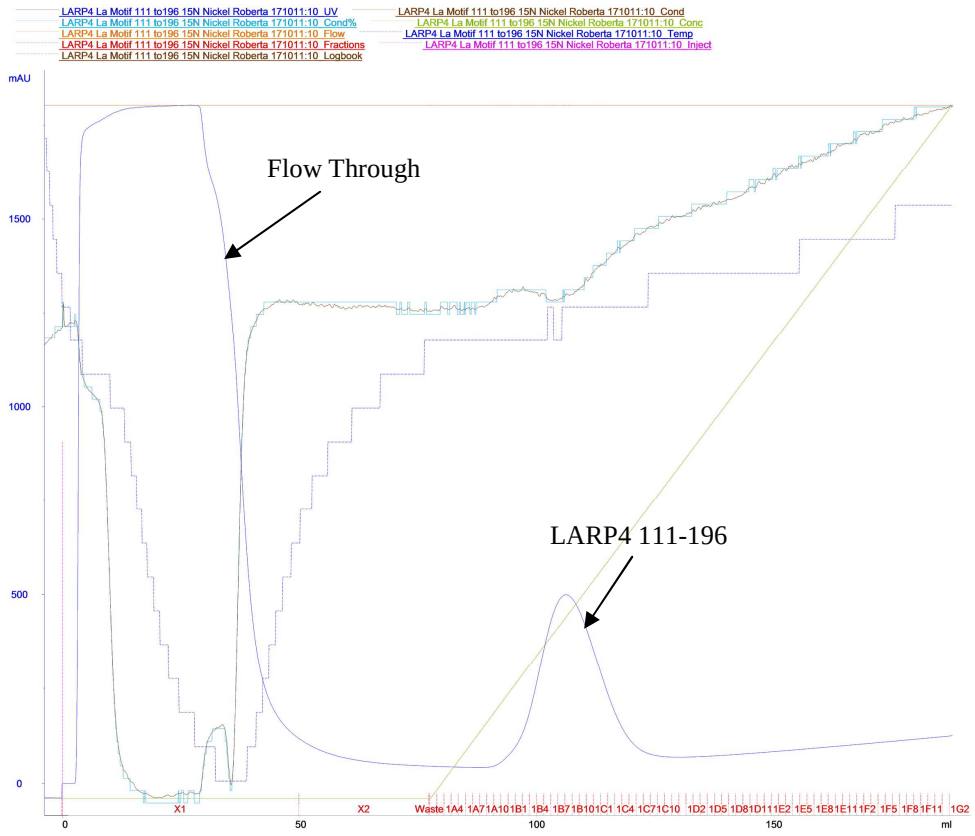


Figure 5: Ni-NTA column purification profile for LARP4 (111-196). The protein was loaded on the column at 3 mL/min flow rate. After washing out the unbound contaminants (when the absorbance went back to the baseline) Imidazol gradient (elution buffer) was started. The wide peak eluted is mainly LARP4 (111-196) together with nuclei acids.

From the chromatogram is clear that the protein is present in the fractions between A10 and C6, as confirmed from the previous study made by another student; so I have collected and dialysed these fractions against the TEV overnight, in order to remove the N-terminal Histidine tag.

After TEV cleavage, to obtain the protein free from the TEV uncleaved products, additional Ni-NTA column was required: LARP4 (111-196) from dialysis was loaded onto a bench top Ni-NTA column.

In figure 6 we can see the cleaved LARP4 (111-196) after purification.

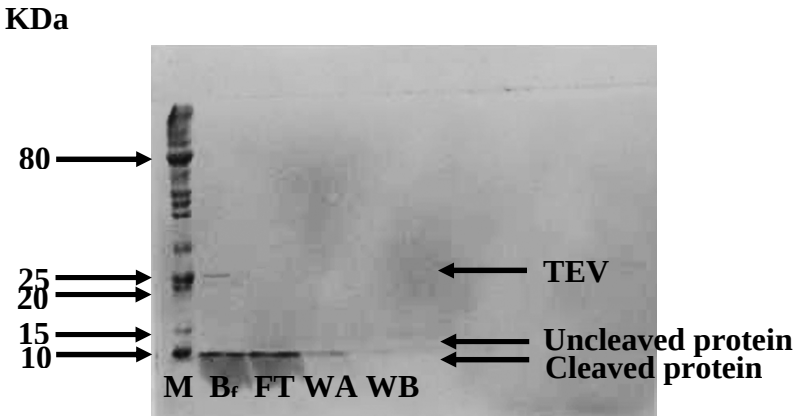


Figure 6: SDS PAGE analysis of Ni-NTA column of LARP4 (111-196) following TEV cleavage. Bf is the fraction before the TEV cleavage, FT (Flow Through) is the protein that not binds the column, WA is the washing of the column with Ni-buffer A and WB is the histidine tag that binds the column and is eluted using 100% Ni-buffer B.

By looking at the SDS PAGE it is clear that the protein free from His Tag is in the fractions FT and WA, these fractions were pooled and dialysed against Heparin buffer dialysis (see Table 2) overnight.

LARP4 (111-196) from dialysis was loaded on a 5 ml Hi-Trap DEAE (diethilaminethil) FF column prepacked with DEAE-Sepharose Fast Flow

(Pharmacia). Before loading, the column was equilibrated using 2 CV of Heparin buffer A at 3 mL/min flow rate. The protein was collected in the FT and the column washed using 5 CV 100% Heparin buffer B (figure 7).We used the DEAE column in order to remove nucleic acids.

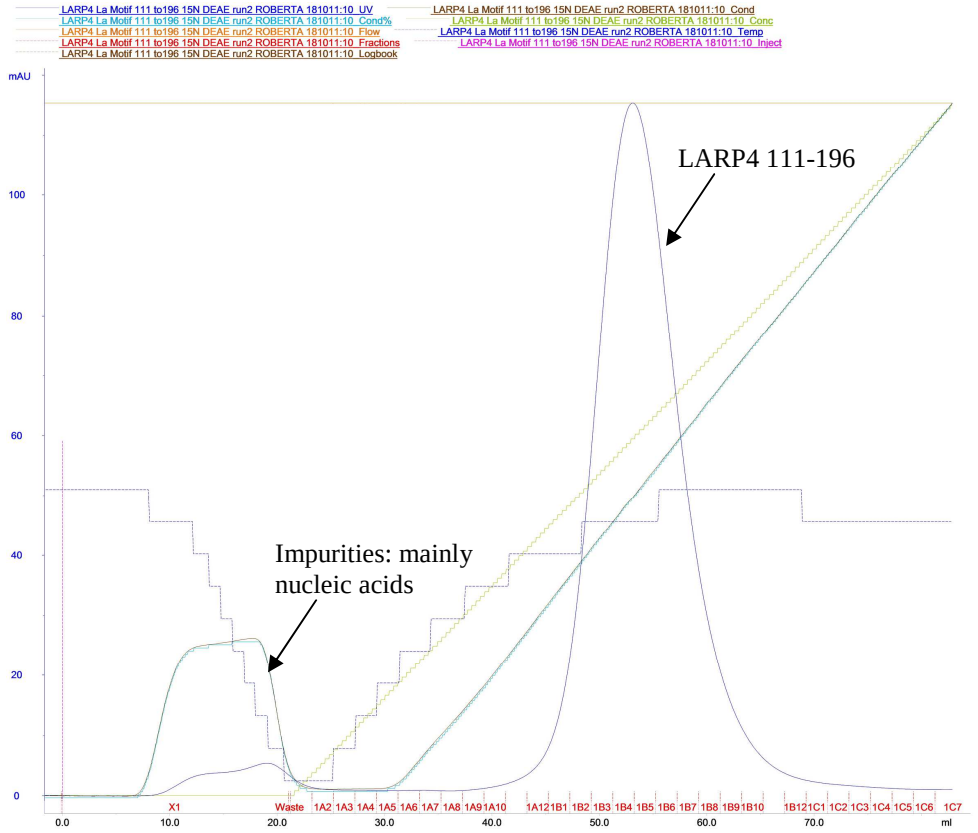


Figure 7: Hi-Trap DEAE purification profile for LARP4 (111-196). This purification step is essential to separate the protein from nucleic acids. It is possible to appreciate the separation of nucleic acids mainly present in the flow through.

The chromatogram shows that the pure protein is present in the fractions between A12 and B10, therefore I collected all the purest fractions and dialysed them overnight in NMR buffer pH 7.25; the day after the protein was concentrated on Vivaspin with a cut off of 3 kDa. The final concentration was then evaluated by UV measurements using an empirical extinction

coefficient (http://www.expasy.org/tools/protparam.html).

Purification of LARP4 (196-287)

The LARP4 (196-287) clone used tag containing six histidines at the N-terminal end of the protein that allows an easy purification on immobilised metal affinity columns. The supernatant, containing tagged LARP4 (196-287) obtained after centrifugation at 17000 rpm for 40 minutes, was loaded onto a Ni-NTA column from Qiagen equilibrated with the Ni-column buffer A (see Table 2 Chapter V). Elution was performed using 250 mM Imidazol with linear gradient (figure 8).

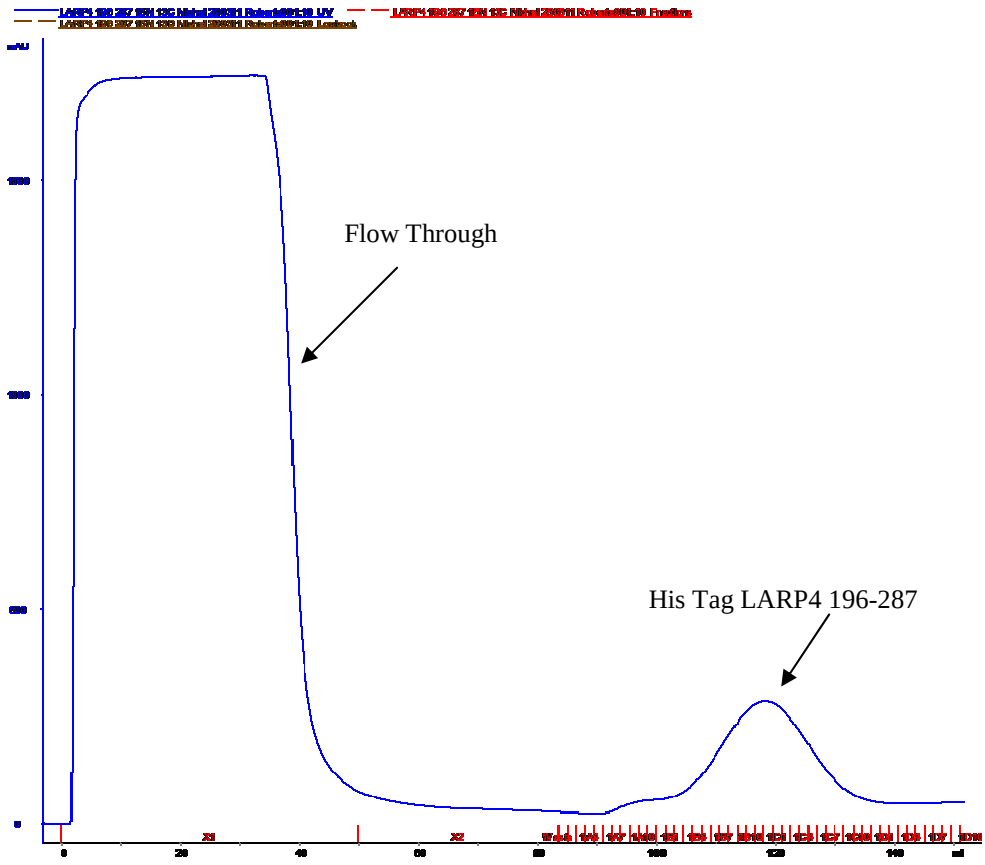


Figure 8: Ni-NTA column purification profile for LARP4 (196-287). The protein was loaded on the column at 3 ml/min flow rate. After washing out the unbound contaminants (when the absorbance went back to the baseline) Imidazol gradient (elution buffer) was started. The wide peak eluted is mainly LARP4 (196-287) together with some impurities and nuclei acids.

Fractions were monitored by SDS PAGE (figure 9) and as shown below my protein is present in the fractions between B5 and C9, so I have collected and dialysed these fractions against the TEV overnight, in order to remove the N-terminal Histidine tag.

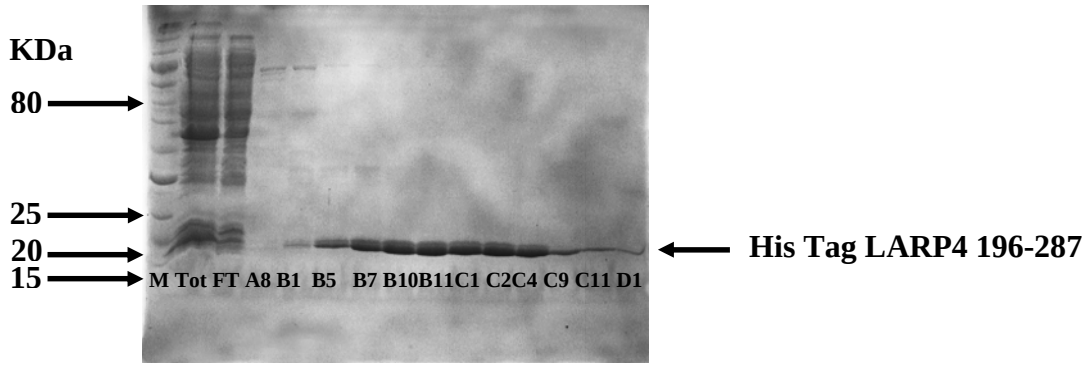


Figure 9: SDS PAGE analysis of Ni-NTA column elution of LARP4 (196-287). The presence of the N-teminal Histidine Tag allows an easy purification of the protein. Protein ladder is loaded on the left (M), and each subsequent lane is numbered according to the corresponding fraction eluted from the Ni-NTA column. His-tagged LARP4 (196-287) migrates at an approximate MW of 10 kDa.

After TEV cleavage, to obtain the protein free from the TEV uncleaved products, additional Ni-NTA column was required: LARP4 (196-287) from dialysis was loaded onto a Ni-NTA column, in figure 10 we can see the cleaved LARP4 (196-287) after purification.

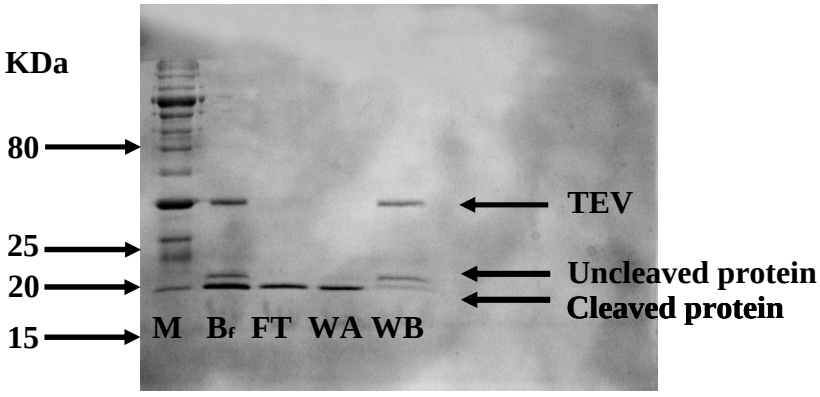


Figure 10: SDS PAGE analysis of Ni-NTA column of LARP4 (196-287) following TEV cleavage. B_f is the fraction before the TEV cleavage, FT (Flow Through) is the protein that not binds the column, WA is the washing of the column with Ni-buffer A and WB is the histidine tag that binds the column and is eluted using 100% Ni-buffer B.

The SDS PAGE (figure 10) indicated the presence of the free protein from His Tag in the fractions FT and WA. These fractions were pooled and dialysed against Heparin buffer dialysis (see Table 2) overnight.

LARP4 (196-287) from dialysis was loaded on a 5 ml Hi-Trap Heparin column (Amersham-Pharmacia Biotech) equilibrated in Heparin-column Buffer A and eluted using linear gradient of 2 M KCl over 15 CV (figure 11), in order to remove nucleic acids.

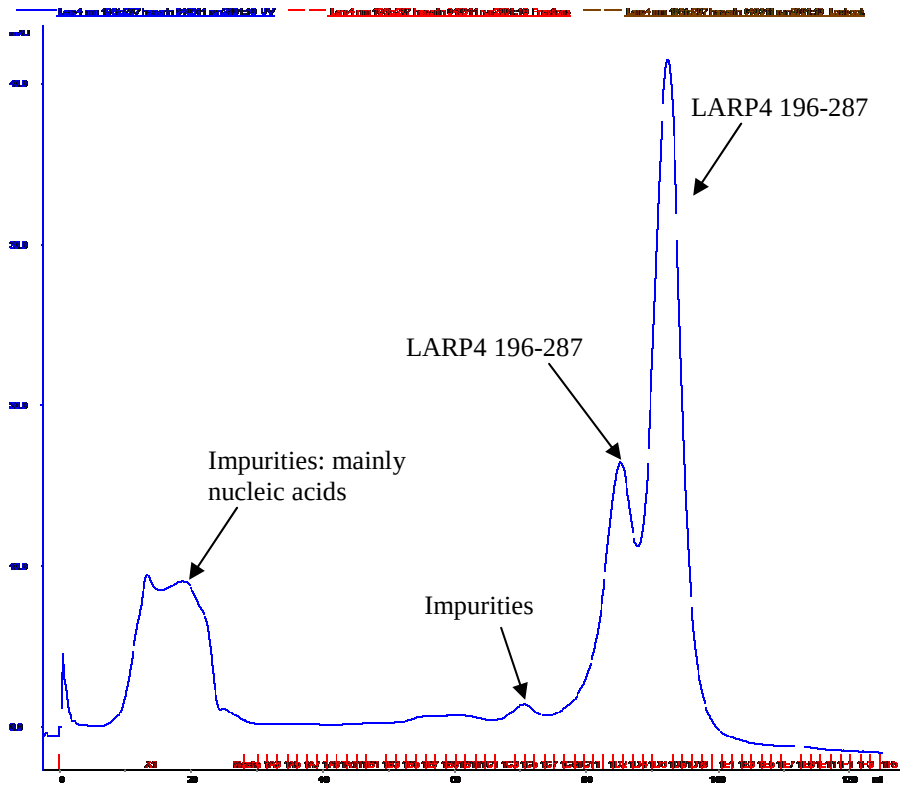


Figure 11: Hi-Trap Heparin purification profile for LARP4 (196-287). This purification step is essential to separate the protein from nucleic acids. It is possible to appreciate the separation of nucleic acids mainly present in the flow through.

The figure 12 shows the pure LARP4 (196-287) protein obtained after the heparin column step: the purest fractions were pooled together and dialysed

overnight in NMR buffer pH 7.25; the day after the protein has been concentrate on Vivaspin with a cut off of 3 kDa. The final concentration was then evaluated by UV measurements using an empirical extinction coefficient (http://www.expasy.org/tools/protparam.html).

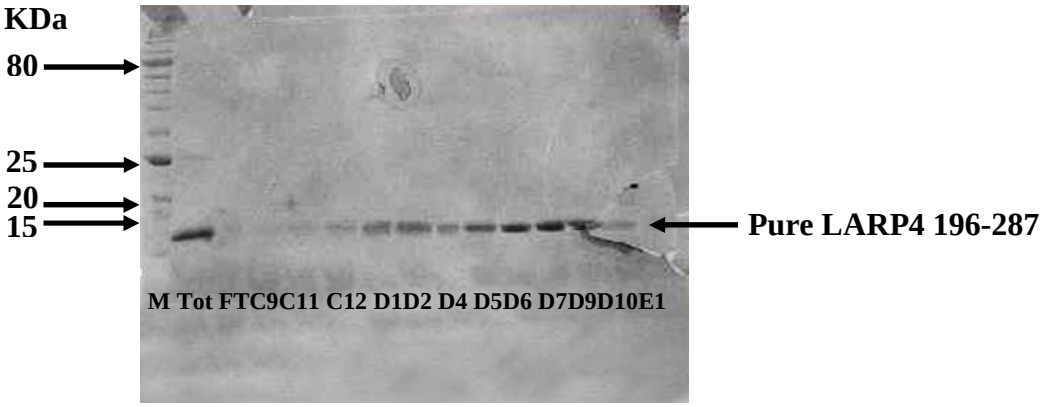


Figure 12: SDS PAGE analysis of Heparin column elution of LARP4 (196-287): pure protein is obtained after this purification step. Fraction from C12 to D10 were pooled together and dialysed overnight in NMR buffer.

While for LARP4 (111-196) a previous student from Conte lab found the best condition to prepare a good NMR sample, for LARP4 (196-287) I carried out a careful analysis to find the best NMR conditions for structural studies.

The project started with the analysing NMR data of the longer LARP4 (196-303) domain at pH 7.25 (previously prepared). However, I found difficulties in obtaining complete backbone assignment, in particular the resonance corresponding to last part of the domain appeared to be missing. As a result I decided to use a shorter clone where the last part, seemingly not folding into a secondary structure, was cut. I then started studying the LARP4 (196-287) domain obtaining a complete backbone assignment.

By comparing the ^{1}H NMR of two RRMs it can be seen that there is not a big difference between the two domains, therefore I started to study the LARP4(196-287) (figure 13).

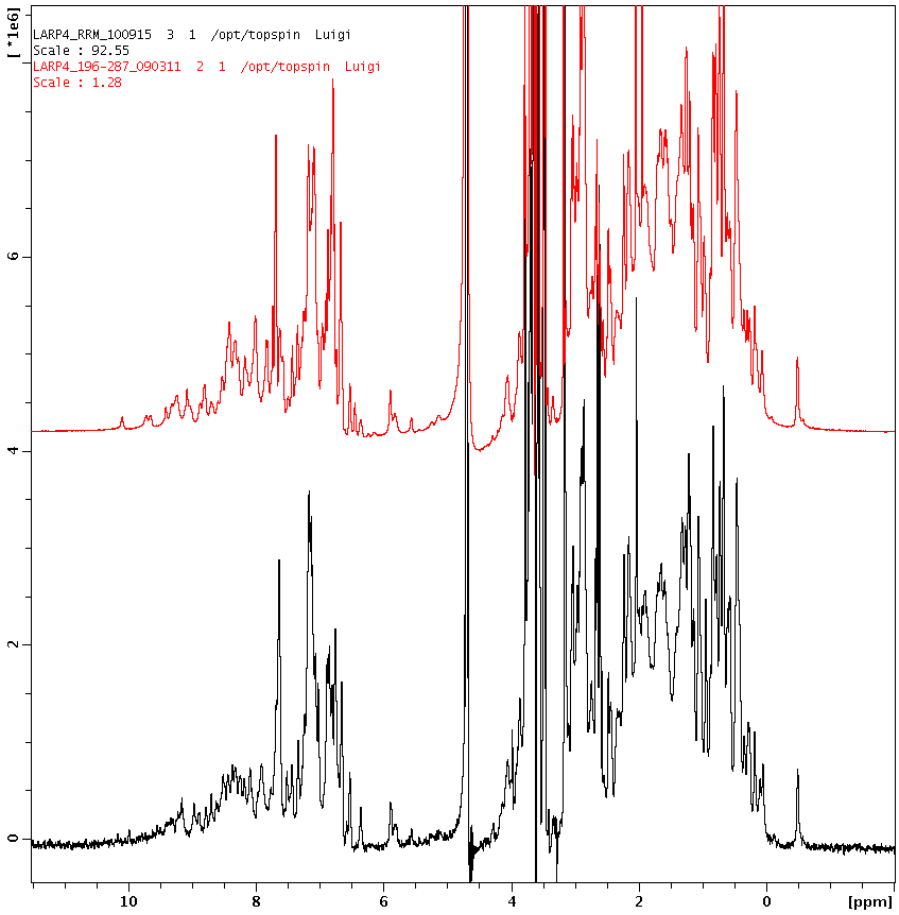


Figure 13: Comparison of ^{1}H NMR spectra of LARP4 (196-303) (on the top) with LARP4 (196-287) (on the bottom) at the same condition pH 7.25.

Among the tests carried out to optimise the sample for NMR structural determination, I checked the stability of the domain over time and in different buffer, temperature and concentration conditions.

To start with, I took the NMR sample of LARP4 (196-287) at pH 7.25 shown in figure 13 and I checked its stability after one week investigating for any

changes, but no change was detected as the sample maintained the same profile shown before.

Next, I decided to change the pH, in particular I tested the protein at pH 5. The buffer 20 mM Sodium Acetate, 100 mM Sodium Chloride, 0.2 mM EDTA and 1 mM DTT, pH 5 was used.

^{1}H NMR spectra at pH 5 was compared with the same sample at pH 7.25, and no great differences could be seen in the 1D experiment (figure not reported). I further checked the stability of the protein at pH 5 after several weeks and the sample looked stable.

Further analysis was then conducted with an HSQC experiment of LARP4(196-287) at pH 5 overlapping it with the HSQC of LARP4 (196-303) at pH 7.25. The advantage of the ^{1}H-^{15}N HSQC was much better resolution compared to the 1D proton spectra, enabling us to see individual resonances.

Figure 14 shows that some shifts occurred, and this was a predictable outcome for having changed the condition of the experiment, but interestingly some new peaks seemed to appear.

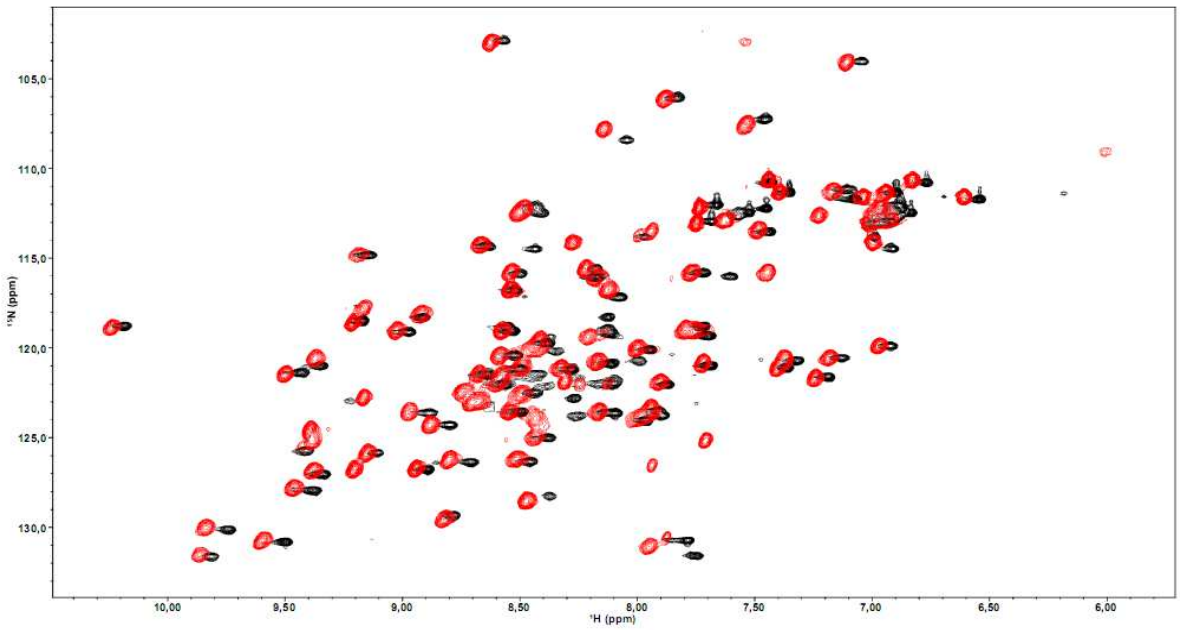


Figure 14: Overlapped of HSQCs spectra of LARP4 (196-303) pH 7.25 (black) and LARP4 (196-287) pH 5 (red).

It is well known that reducing the concentration of salts in the NMR sample will provide a better sensitivity hence allowing the detection of more peaks, because this is a better condition for the NMR CryoProbe. For this reason I decided to use an NMR buffer at pH 5 but without salts and also in this case no changes were found even after some weeks (figure 15).

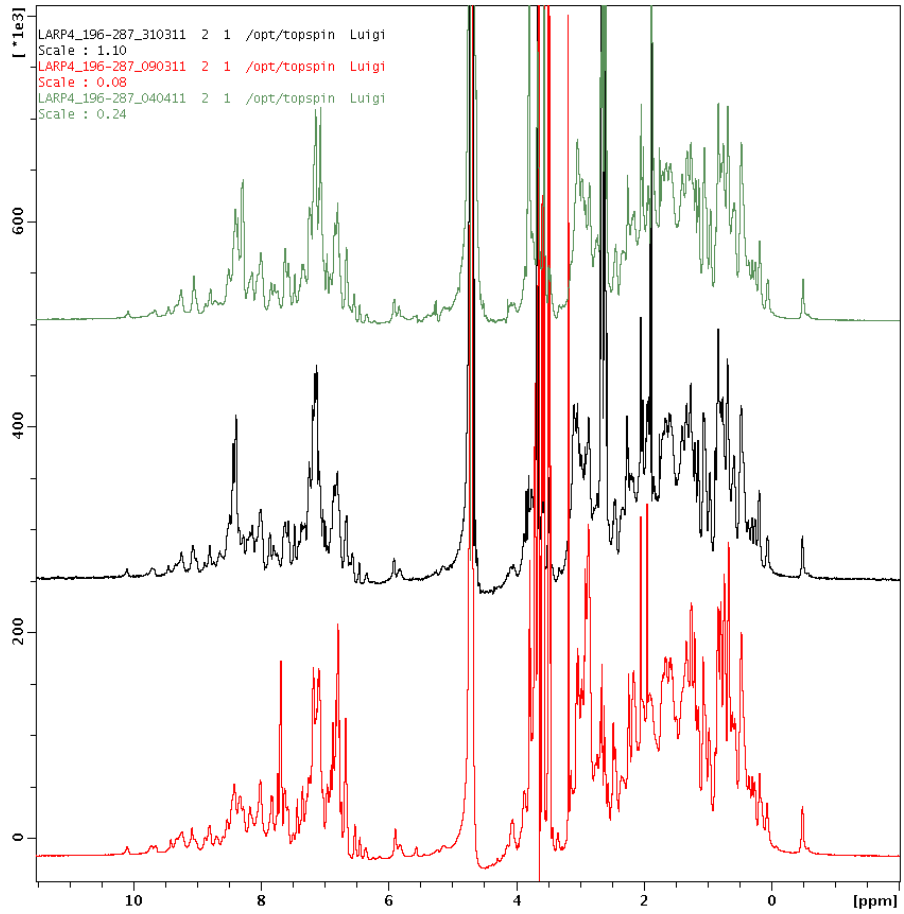


Figure 15: Comparison of ^{1}H NMR spectra of LARP4 (196-287) at pH 5 without salts (on the top) and with salts (in the middle) and LARP4 (196-287) at pH 7 (on the bottom).

Nonetheless, when I actually acquired all the NMR spectra of LARP4(196-287) at pH 5 without salts I detected a stability problem that was not detected in the early preliminary tests. In fact by comparing the previously acquired HSQC with the one obtained in the latest experiments, I noticed some differences: few peaks moved and others disappeared (figure 16).

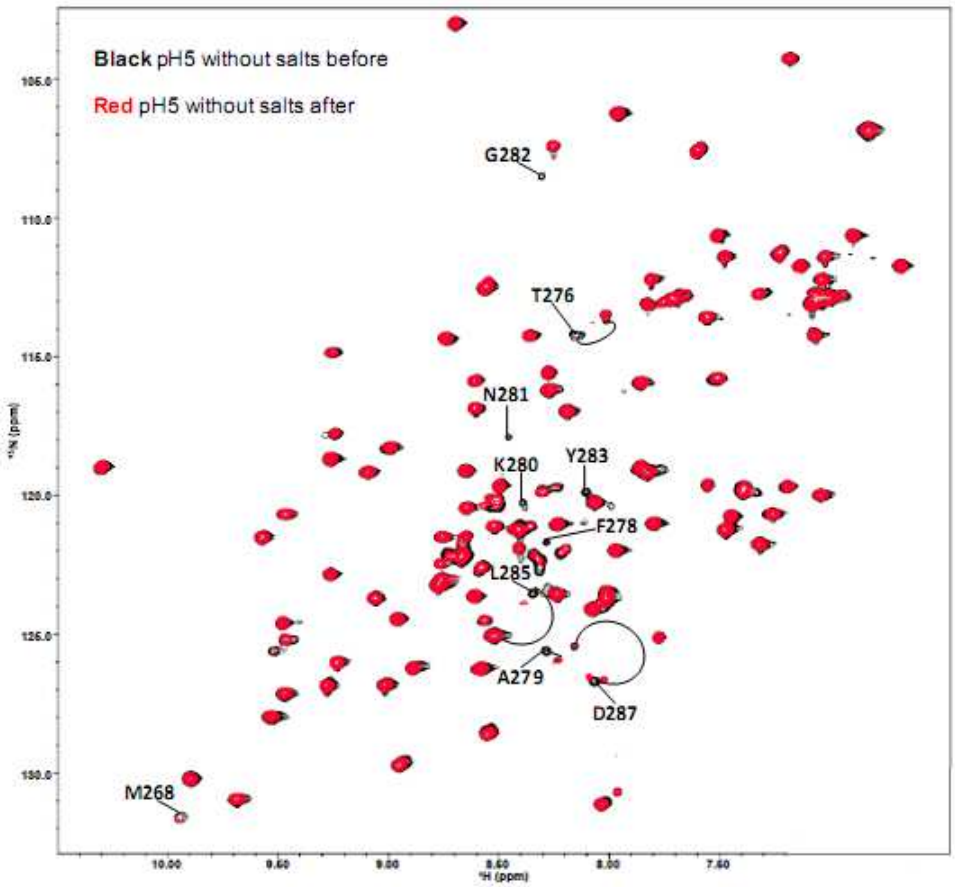


Figure 16: Overlap of HSQC spectra of LARP4 (196-287) pH 5 without salts.

Following this evidence I decided to compare those ^{1}H-^{15}N HSQC of LARP4 (196-287) at pH 5 without salts with the HSQC of LARP4(196-287) at pH 5 with salts (figure 17); in particular I have overlapped the spectra of a fresh sample without salts with the sample at pH 5 with salts, then I have overlapped a sample, at pH 5 without salts, acquired after several experiments run at room temperature, with the sample at pH 5 with salts.

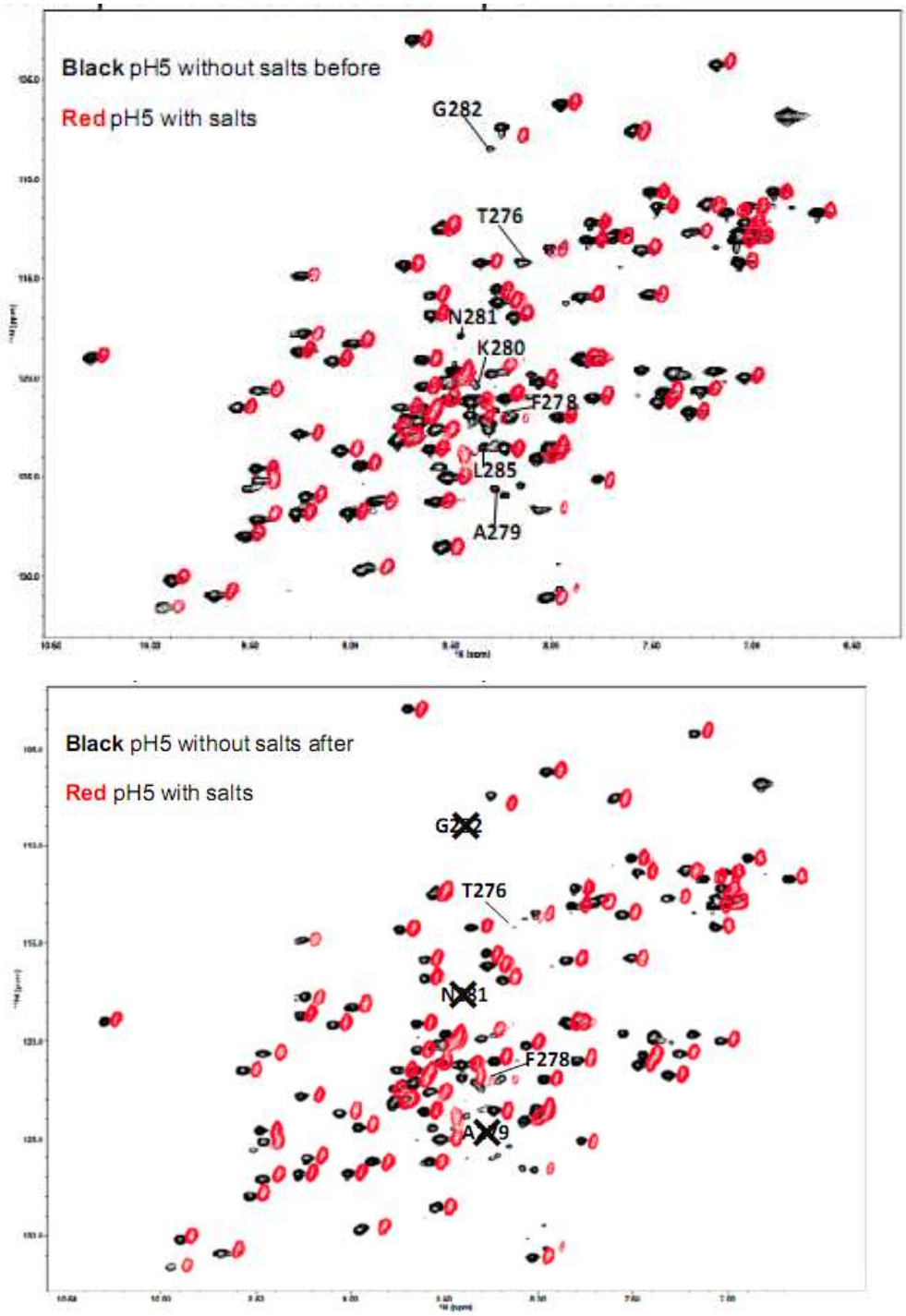


Figure 17: On the top I reported the overlap of HSQC spectra of LARP4 (196-287) pH 5 without salts before (black) with LARP4 (196-287) pH 5 with salts (red), whilst on the bottom I reported the overlap of HSQC spectra of LARP4 (196-287) pH 5 without salts after (black) with LARP4 (196-287) pH 5 with salts (red).

In this analysis I noticed that some peaks had disappeared in the spectra at pH 5 with salts, these are in particular amino acids situated at the end of the protein's primary sequence.

Having performed all these analysis I drew the following conclusions:

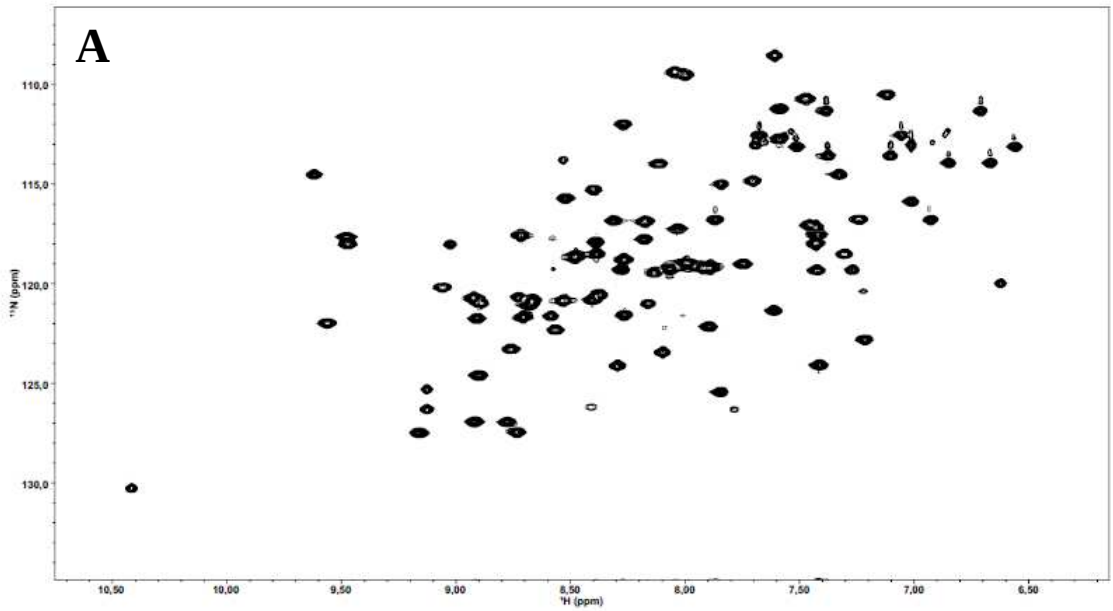
- to study the LARP4 (196-287) because, in contrast with the longer protein LARP4 (196-303), it allows for a complete assignment.
- to acquire all the spectra needed for the complete NMR domain's assignment , at pH 7.25 which proved to produce very stable results. Although at pH 5 we have been able to get a better NMR spectra's resolution whether with salts or without but as a downside we have detected a stability problem in the NMR spectra acquisition lasting several days during the experiments.

The decision to work at pH 7.25 was also justified by the fact that while performing the analysis for different pH conditions, I also prepared a sample of LARP4 (111-287), La Module (including both La Motif and RRM domains), for crystallographic studies and as is known the best condition for them is to work at pH 7.25.

NMR analysis of LARP4 (111-196) and (196-287)

To test the suitability of LARP4 (111-196) and (196-287) samples for structural determination by NMR spectroscopy an initial [^{1}H-^{15}N] Heteronuclear Single Quantum Coherence (HSQC)[1] 2D experiment was recorded. Figure 18 reports the HSQC of the two domains of LARP4; the [^{1}H-^{15}N] HSQC displays cross-peaks for each proton bound to nitrogen, in particular backbone or side chain amide protons: this spectrum is considered the "finger-print" of the protein. The dispersion of the cross-peaks indicates that both proteins are folded and feasible for NMR analysis.

The purified samples of [^{15}N] and [^{15}N, ^{13}C] labelled LARP4 (111-196) and LARP4 (196-287) were used to record NMR spectra on Bruker spectrometers.



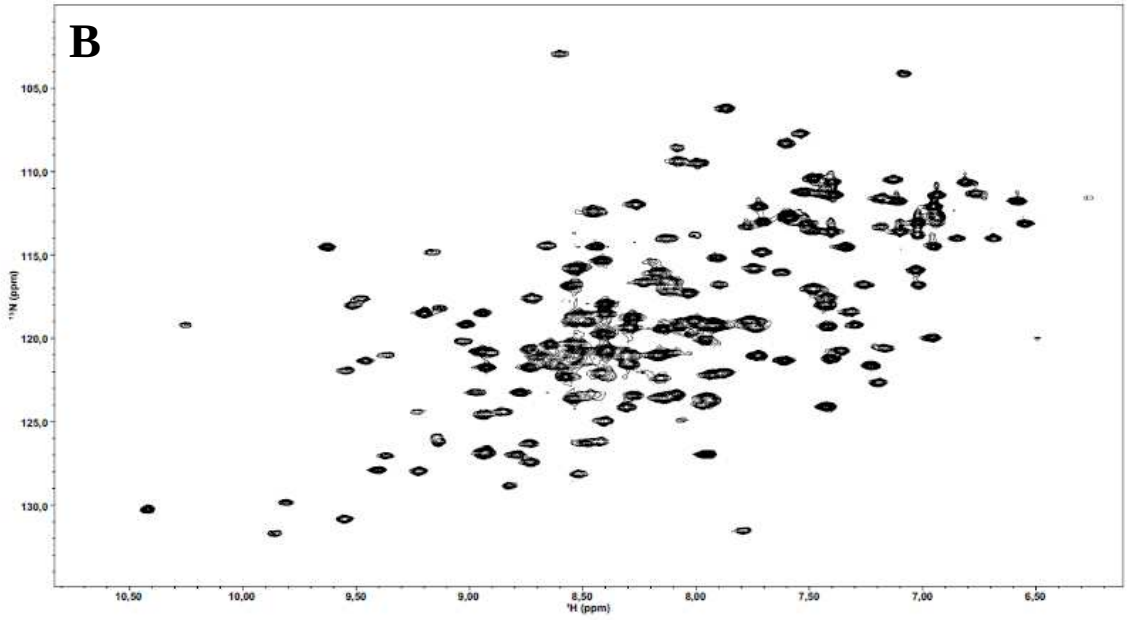


Figure 18: [^{1}H-^{15}N] HSQC of LARP4 (111-196) $^{(A)}$ and LARP4 (196-287) **(B)**.

Assignment of NMR spectra

In order to solve the structure of a protein by NMR spectroscopy, the first stage of the analysis is spectral assignment, in which each signal in the spectra has to be associated with a specific nucleus in the molecule. In general this process can be divided in two parts, the sequential assignment for the nuclei in the protein backbone and the assignment of the nuclei in the amino acid side chains.

The assignment of LARP4 (111-196) and LARP4 (196-287) was achieved by employing triple resonance NMR experiments. These experiments are so called because three different nuclei (^{1}H, ^{15}N and ^{13}C) are correlated. All these experiments rely on the ability to transfer magnetisation between

covalently attached nuclei through-bond scalar couplings. The most common way to transfer magnetisation between ^{1}H and ^{15}N nuclei or between ^{1}H and ^{13}C nuclei is an INEPT (insensitive nuclei enhanced by polarisation transfer) sequence.

The INEPT transfers the magnetisation from the ^{1}H nucleus (I spin) to the attached ^{13}C or ^{15}N nucleus (the S spin). After the transfer the S spin nucleus can be frequency labelled. At this point the magnetisation can then be transferred either to another nucleus or back to the original ^{1}H nucleus for detection. All of the triple resonance experiments used for assigning both LARP4 domains work on the same basic principles. The first step is to transfer the magnetisation from ^{1}H to ^{15}N and then to the other nuclei of interest.

Figure 19 shows a schematic representation of the magnetisation transfer through bounds that occurs in some of the used experiments: the nomenclature directly refers to the atoms detected during the acquisition.[2]

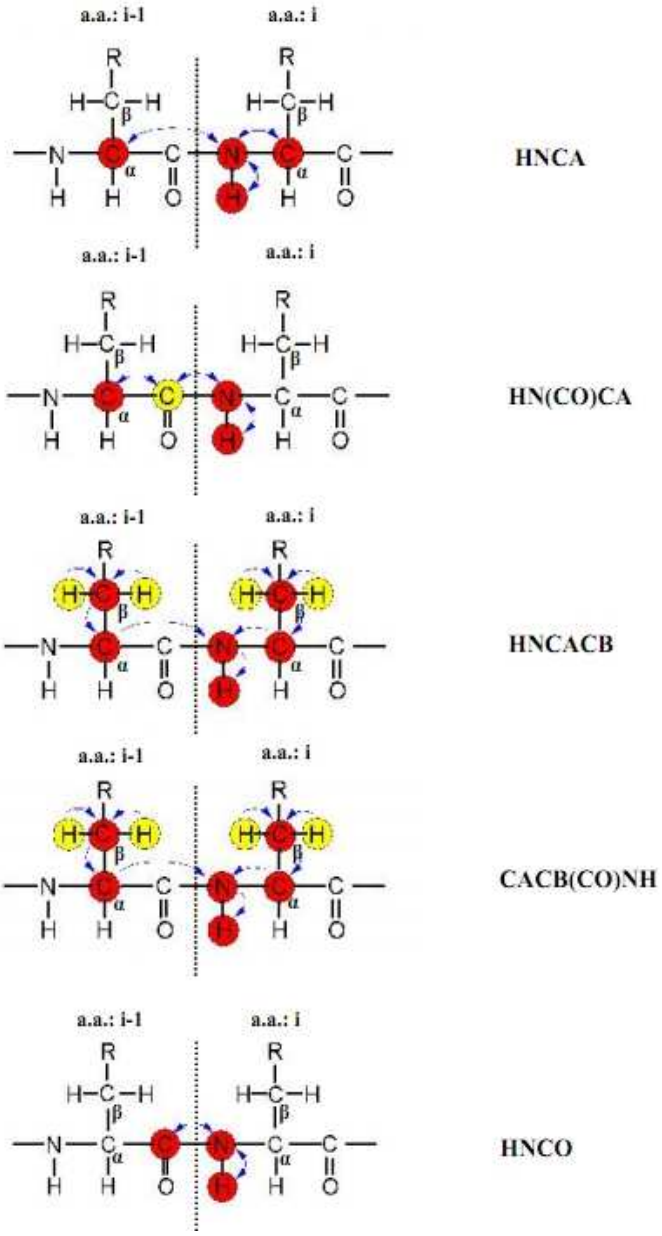


Figure 19: Schematic representation of magnetisation transfer occurring in the 3D NMR experiments used for the sequential assignment of protein backbone. The nomenclature on the right[2] indicates the magnetisation transfer pathway and spin that are frequency labelled. The magnetisation originates from the amide proton and the arrows indicates the transfer direction.

^{1}H, ^{13}C and ^{15}N backbone assignment of LARP4 (111-196) and LARP4 (196-287)

In the HNCA experiment[3] the magnetisation is transferred from amide proton (^{1}HN) to the directly attached nitrogen nuclei (^{15}N) through an INEPT sequence, in a so called "out-and-back" coherence transfer pathway, where the magnetisation is detected on the same spin from which it originated. The normal out-and-back type contribution can be written in a short notation as:

$$^{1}H^{N}\rightarrow{}^{15}N\rightarrow{}^{13}Ca(t1)\rightarrow{}^{15}N(t2)\rightarrow{}^{1}HN(t3)$$

where $^{1}H^{N}$, ^{15}N and ^{13}Ca stand for the polypeptide backbone amide proton, nitrogen and a-carbon nuclei, respectively, and t1, t2 and t3 indicate the chemical shift evolution times. In each step, the magnetisation is transferred via strong ^{1}J coupling between the nuclei (see Table 3). The nitrogen atom of a given amino acid is correlated with both ^{13}Ca, its own and the one of the preceding amino acid (i-1). Because the coupling which connects the nitrogen atom with the ^{13}Ca carbon of the preceding amino acid ($^{13}Ca_{i-1}$) (^{2}J = 4-7Hz),[3] it is sometimes possible to discern the cross signal between intra- and inter-residual carbons, due to a different peak intensity. However, this is not always clear because geminal coupling constants can vary with the dihedral angles; furthermore, spectral overlap often complicates the analysis

requiring a second experiments, namely HN(CO)CA, to be used in conjunction with HNCA. The HN(CO)CA experiment correlates the backbone amide proton and nitrogen (H^{N}_{i} and N_i) to the alpha-carbon of the preceding residue ($^{13}Ca_{i-1}$). The magnetisation transfer between the amide nitrogen and the alpha carbon takes place in two steps, *via* the carbonyl carbon (see figure 19): the first step is the transfer from $^{1}H^{N}_{i}$ to the CO of the preceding amino acid, then the magnetisation travel to the $^{13}Ca_{i-1}$ and is detected. The respective constants coupling are J_{NCO}= 15Hz and J_{CaCO}= 55Hz as reported in Table 3.

1 bond couplings		**2 bonds couplings**	
$^{1}J_{CH}$	140Hz		
$^{1}J_{NH}$	90-100 Hz		
$^{1}J_{NCO}$	15 Hz	$^{2}J_{NC\beta}$	9-11 Hz
$^{1}J_{NCa}$	11Hz	$^{2}J_{NCa}$	4-7 Hz
$^{1}J_{CaCo}$	55 Hz		
$^{1}J_{CaCa}$	35 Hz		

Table 3: The basic coupling constants used to optimise the pulse sequence. Since the ^{1}H-^{15}N and ^{1}H-^{13}C one bound coupling constants are comparably large, the 3D heteronuclear experiments are highly efficient.

To facilitate the analysis of 3D experiments such as HNCA and HN(CO)CA, strips are extracted in correspondence of peaks in the [^{1}H-^{15}N] HSQC spectrum and strip plots are made to trace the resonance down the backbone of the protein as illustrated in figure 19. Here the strips are placed

in a row to show the sequential connectivities from each amino acid to the preceding one: the blue line connects the ^{13}Ca of the preceding amino acid (i-1) to the subsequent ^{13}Ca (i). As previously explained it is possible to appreciate a difference in the cross signal intensities due to the different J-coupling constants.

In the CBCA(CO)NH,[4,5] the resonances of Ca and Cβ of the preceding i-1 amino acid are correlated to the HN of the own i amino acid. In the HNCACB, both the i-1 and the i Ca and Cβ are correlated to the NH_i. In figure 20 I report the example of HNCA of LARP4 (111-196).

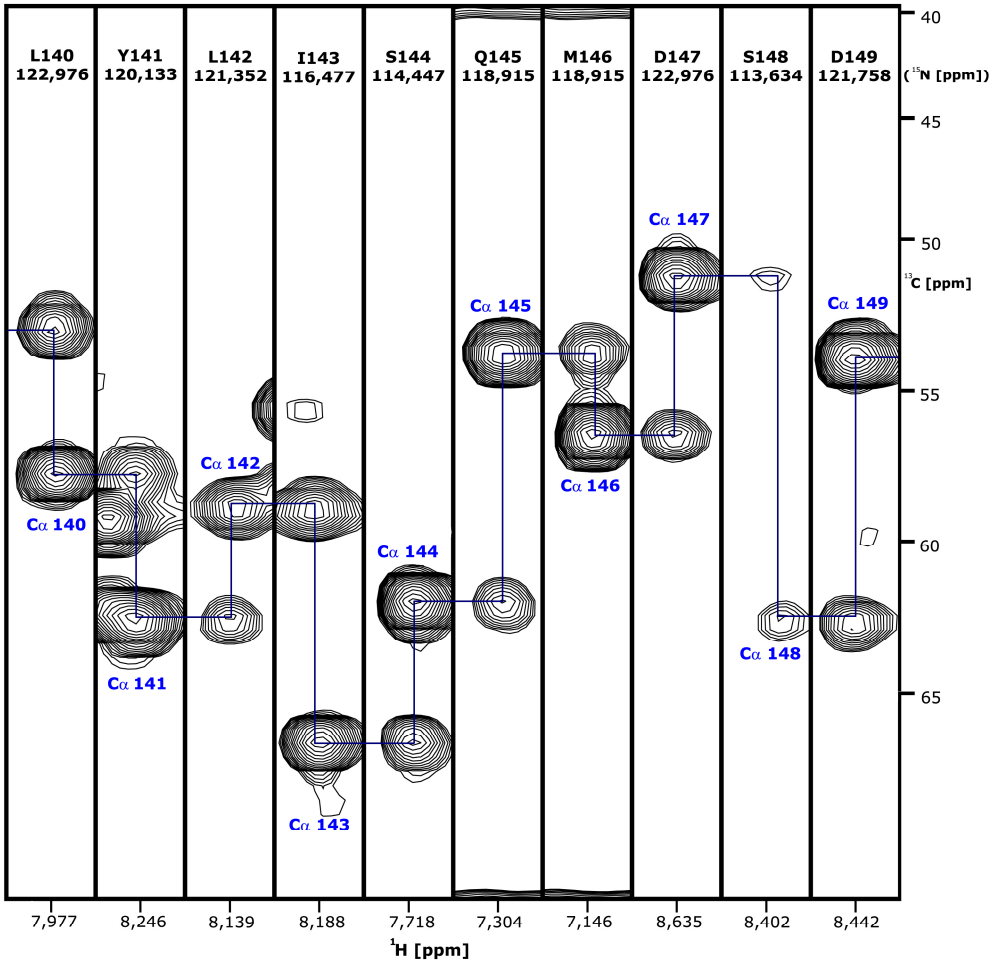


Figure 20: Sequentially arranged strips extracted from the HNCA spectrum of [^{15}N-^{13}C]-LARP4 (111-196). Each ^{1}HN correlates with two peaks, the $^{13}Ca_i$ and the $^{13}Ca_{i-1}$. these strips were placed in arrow to show the sequential connectivities from each amino acid to the preceding one: the i-1 residues are connected to the i residues with a blue line.

Thus for every strip extracted in correspondence of a cross-peak in the [^{1}H-^{15}N]-HSQC, a spin system consisting of two pairs of carbon resonances are known, one relative to the i amino acid the other to the preceding one: this allows the determination of a sequence of spin systems in a similar way to

that for the analysis of the HNCA/HN(CO)CA experiments.

The HNCACB/CBCA(CO)NH set provide an additional advantage over the HNCA/HN(CO)CA, in that the Cβ chemical shift dispersion is wider than that observed for Ca signals. The additional Cβ chemical shift information can therefore help to resolve ambiguities in the sequential assignment due to overlap of Ca resonances. Furthermore, the Cβ chemical shift are more residue-specific than the Ca ones and often can be used to identify the amino acid type that corresponds to the peaks in the spectrum. For example, characteristic chemical shifts such as the Cβ of alanine, serine and threonine can give fundamental clues in the identification of the amino acid type.

In the HNCO experiment,[4,5] the magnetisation is transferred from the amide proton to the directly attached $^{13}CO_{(i-1)}$ carbon atom (figure 20). The HNCO also is very useful for resolving accidental signal degeneracy in the HSQC projection. In proteins, every amide proton is covalently bonded to a single CO group, and therefore we expect to find only one cross signal per frequency in the HNCO. Hence, if two cross signals are observed at the frequency of an amide proton, this indicate the presence of two amino acids with accidentally degenerate amide protons. One limitation of the use of this experiment is for the amino acid preceding proline residues: there is no magnetisation transfer because the $^{1}H^{N}$ is not present.

By analysing all of these spectra described above, an almost complete backbone assignment was obtained for the two LARP4's domains.

Secondary structure prediction

The secondary structure prediction of LARP4 (111-196) and (196-287) was performed using TALOS software (figure 22).[6]

All the collected chemical shift data were used to calculate φ and ψ dihedral angles using TALOS, which combined the chemical shift information of ^{1}Ha, ^{13}Ca, $^{13}C\beta$, ^{13}CO, ^{15}N together with sequence information derived from a database of high-resolution NMR and X-ray structures.

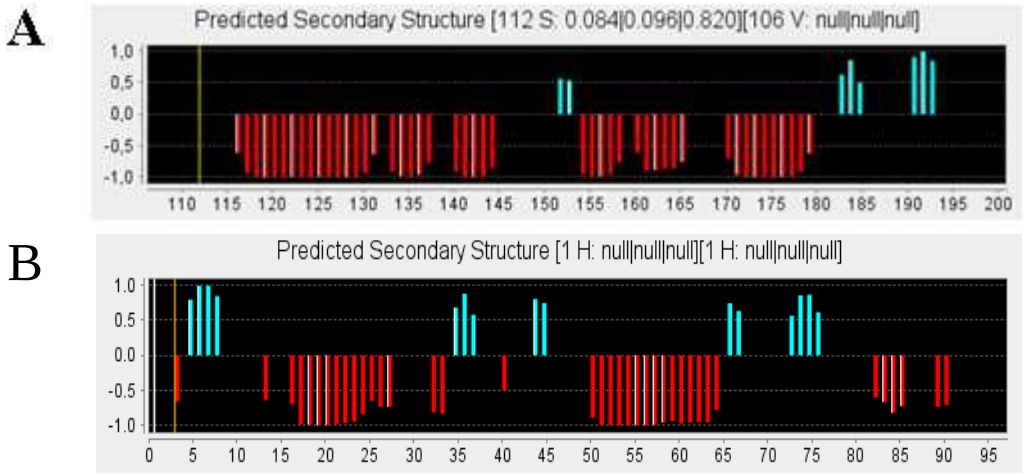


Figure 22: Secondary structure prediction of LARP4 (111-196) **(A)** and LARP4(196-287) **(B)** obtained by TALOS.

TALOS evaluates the similarity in amino acid sequence and secondary structure for a string of three sequential amino acids relative to all triplets of sequential residues contained in the database. The TALOS output for the φ and ψ backbone angles of the central residue in each string consists of the average of the corresponding angles in the 10 strings in the database with the highest degree of similarity.

The program classifies only those predictions for which at least nine out of ten predictions fall in the same populated region of the Ramachandran plot.

TALOS analysis shows that for the domain LARP4 (111-196) we have the same organization such as in the La Motif 1-103 of the genuine La protein; whilst for LARP4 (196-287) domain we have a difference from the RRM 105-202 of La genuine protein, perhaps a small additional strand before strand β3 (figure 23).

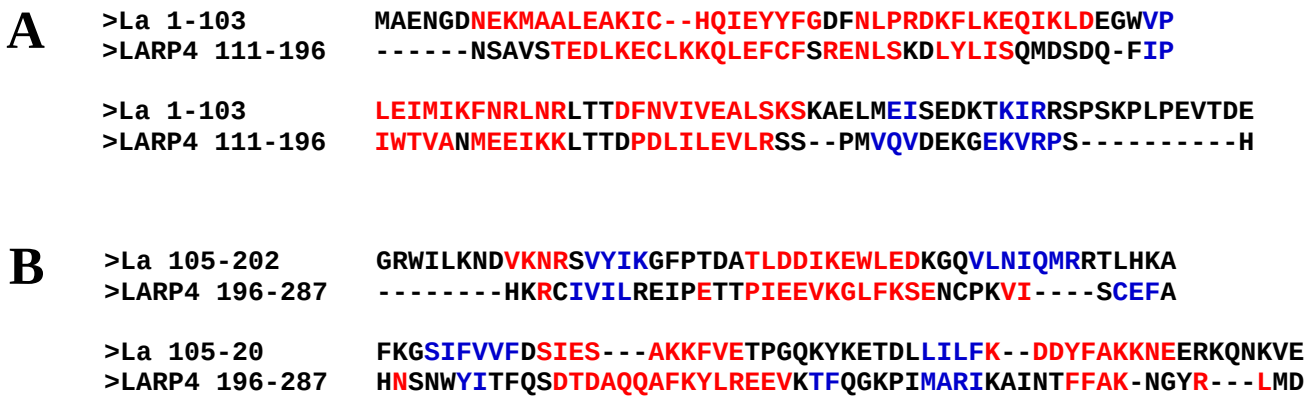


Figure 23: Comparison of the TALOS data of the two domains of LARP4 with the two domains of La genuine protein. **A)** Comparison of the La Motif domain of the two different proteins. **B)** comparison of the RRM domain of the two proteins. (In red the a-helix and in blue the β-strand). This is the alignment amino acid sequences obtained from TCOFFE.[7]

^{1}H and ^{13}C side chain assignment of LARP4 (111-196) and LARP4 (196-287).

The sequential assignment by triple resonance experiments yields the chemical shifts for the backbone atoms, whereas the amino acid side chain chemical shifts can be assigned using ^{13}C- edited NOESY in H_2O, ^{13}C- edited NOESY in D_2O and ^{15}N- edited NOESY.

The combination of these experiments permitted completion and confirmation of the assignment of the side chains: for almost all residues a complete assignment was obtained, some protons from the aromatic side chains (δ, ε and ζ) as well as H_ε of some Lysines were not assigned. The percentage of assigned carbons is also relatively high, at 85%.

In figure 24 I report, an example, of the ^{13}C- edited NOESY analysis of Ile154 of LARP4 (111-196) side chain. The C_α plane is the starting point: the selected diagonal peak (asterisk) is the H_α; in correspondence with this chemical shift it is possible to observe the complete pattern of NOE signals to the aminoacid side chain. Lines were scanned in order to find similar patterns and the simultaneous analysis of the ^{15}N- edited NOESY facilitated the assignment of the side chains of the two domains. With this methodology most of the side chains were assigned, creating the basis for the collection of NOE-based restraints which was performed next.

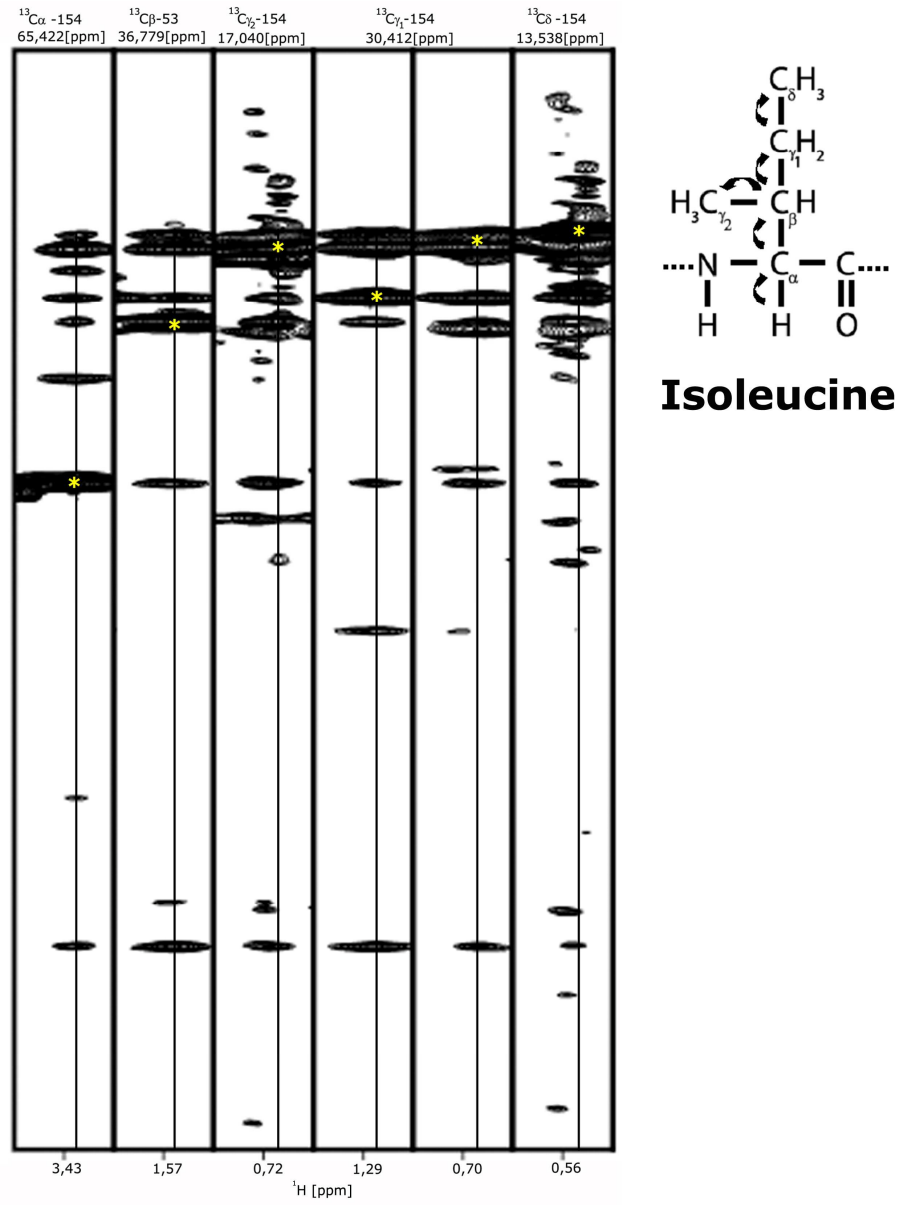


Figure 24: Analysis of the ^{13}C- edited NOESY of LARP4 (111-196) for Ile 154. Strips are extracted from the spectrum and placed in a row. Starting from the $^{13}C_{\alpha}$, the side chain pattern was recognised and selected. ^{13}C planes were subsequently scanned to find comparable patterns that defined the ^{13}C chemical shift of the side chain. The diagonal peak (indicated by asterisk) is the directly attached proton.

STRUCTURE DETERMINATION

Obtaining a chemical shift assignment is not a goal in itself, and is usually used in conjunction with other experiments to obtain a protein structure. In principle, all observable NMR, such as NOE, chemical shifts, hydrogen bonds and dihedral angles provide information on the protein conformation.

The three-dimensional structures of the LARP4 La Motif (111-196) and RRM (196-287) were determined using the new software *UNIO10* with a procedure where we have combined automated NOE assignment (CANDID)[8], automated NOESY peak picking (ATNOS)[9] and structure's determination module.

The input for ATNOS/CANDID consists of:

- the protein's amino acid sequence
- the chemical shift lists from the specific sequence's resonance assignment
- and one or several 3D NOESY spectra.

ATNOS performs multiple cycles of NOE's peak identification, it was used in concert with the software CANDID to obtain the automated NOE assignment while the protein's structure calculation was identified by the CNC software.

In the second and subsequent cycles, the intermediate protein's structures are used as an additional guide for the interpretation of the NOESY spectra.

CANDID includes two new elements making it robust with respect to the

presence of artefacts in the input data: ***network-anchoring*** and ***constraint-combination***, which have a key role in the de novo protein's structure determinations for the successful generation of the correct polypeptide fold by the first CANDID cycle.

Network-anchoring is based on the evidence that any network of correct NOE cross-peak assignments forms a self-consistent set.

The initial chemical shift-based assignments for each individual NOE cross-peak are therefore weighted by the extent to which they can be embedded into the network formed by all other NOE cross-peak assignments.

Constraint-combination reduces the deleterious impact of artefact NOE, assigning an upper distance constraints to the input for the protein's structure calculation by combining the assignments for two or several peaks into a single upper limit distance constraint, which lowers the probability that the presence of an artefact peak will influence the outcome of the structure's calculation.

The protocol of ATNOS/CANDID comprised seven iterative cycles of NOESY peak picking and NOE cross-peak identification with ATNOS[9], automated NOE assignment with CANDID[8], 3D structure calculation with CNS.

The second and subsequent cycles differ from the first cycle for the further addition of selection criteria targeting the cross-peaks and NOE assignments which are based on the assessments of the protein's structure-guided search of the experimental NOESY spectra.

The increase in the number of NOE cross-peaks identified and of the

conformationally meaningful NOE upper distance constraints between the first and second cycles, reflects the fact that the regions near the diagonal and the solvent line were added for the spectral analysis in cycle 2, and that additional protein's structure-based information was available in cycle 2 to guide the analysis of the NOESY spectra.

These numbers are nearly constant after the second cycle, reflecting that the correct protein fold had already been found after the first cycle of calculation, whereby the slight decrease of the number of NOE distance constraints during the cycle 7 is caused by an additional filtering step ensuring that all NOEs have either unique assignments to a single pair of hydrogen atoms, or are eliminated from the input for the structure's calculation (figure 25).

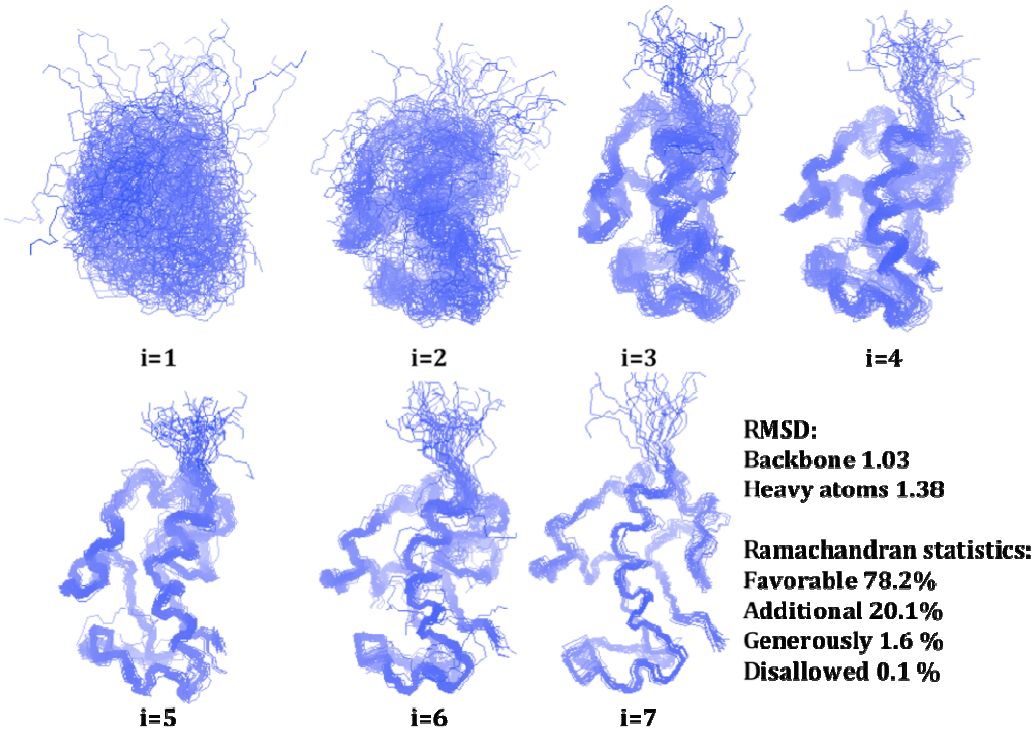


Figure 25: Comparison of the structures obtained from the first cycle until the 7th.

The output from a ATNOS/CANDID cycle includes:

- list of NOESY cross-peak assignments
- list of comments about individual assignment decisions that can help to recognize potential artefacts in the input data
- 3D protein's structure in the form of a bundle of conformers.

The results obtained from the Automated Structure Calculation are very promising because the comparison of the La Motif structures of genuine La protein (obtained with manual method), and the La Motif of LARP4 protein (obtained with automatic method), shows that they are very similar.

The main difference between this two methods is the considerable reduction in computation and the related working time.

References

1 Bodenhausen G., Ruben D.J. **Chem. Phys. Lett.** 1980 69: 185–189.

2 Ikura M., Kay L. E., Bax A. **Biochemistry** 1990 29: 4659-4667

3 Kay L.E., Ikura M., Bax, A. **J. Am. Chem. Soc.**, 1990. 112: 888-889.

4 Grzesiek S., Bax, A. **J. Am. Chem. Soc.**, 1992. 114: 6291–6293.

5 Grzesiek S., Bax, A. **J. Magn. Reson.**, 1992. 99: 201–207.

6 Cornilescu G., Delaglio F., Bax A.. **Journal of Biomolecular NMR**, 1999 13: 289–302, 1999.

7 T-Coffee: A novel method for multiple sequence alignments. Notredame,Higgins,Heringa,**JMB**,302 (205-217) 2000.

8 Herrmann T., Guüntert P., Wuü¨thrich K., **J. Mol. Biol.** 2002a 319, 209–227.

9 Herrmann T., Guüntert P., Wuüthrich K., **J. Biomol. NMR** 2002b. 24, 171–189.

CHAPTER VI

Discussion

Through this research I have started the study of a new component of the La related protein (LARPs) big family.

My protein, LARP4, has the same domain as the Human La genuine protein, La Motif and RRM. In contrast with the latter, LARP4 has a new domain PAM2, that is involved in the interaction with poly (A) RNA, but it hasn't got the RRM2 domain.

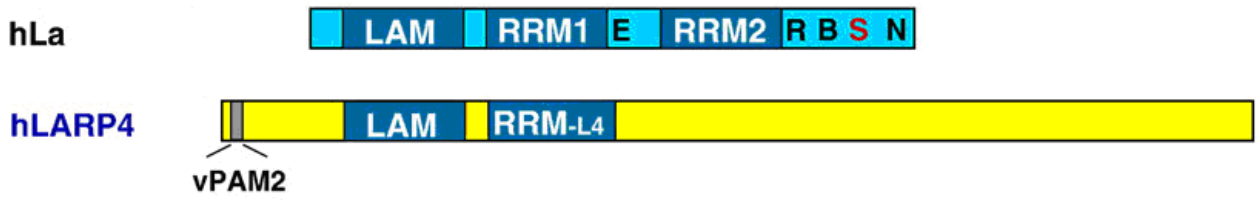


Figure 1: Allignment of hLa protein and hLARP4 protein.

LARP4 La motif

From a careful NMR study of LARP4 (111-196) I obtained the structure using a new software UNIO-ATNOS/CANDID, with a new approach of automated calculation and I obtained a good superimposition of 20 best structures at lower energy. In figure 2 I report the superimposition of these structures and in Table 1 I reported the statistics of protein NMR structure determination.

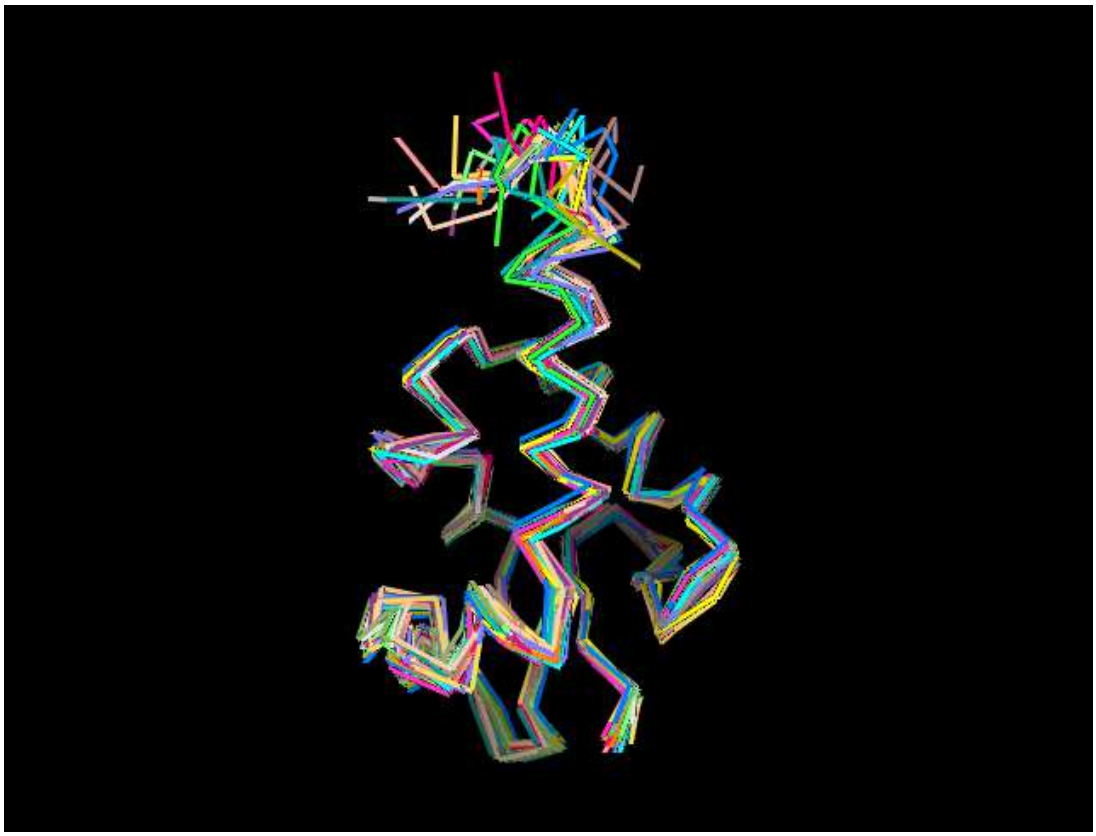


Figure 2: Superimposition of 20 best structures at lower energy.

Table 1: Statistics of protein structure.

Unio- ATNOS/CANDID – Statistics of protein NMR structure determination

Number of residues	86 (111-196)
Molecular weight [Da]	9925.22
Number of models	20
Target function [A^2]	10.96 +/- 0.55 (9.75..11.72)
Setup-given RMSD range	111-196
- Backbone RMSD[A]	0.93 +/- 0.17 (0.65..0.93)
- Heavy atom RMSD [A]	1.29 +/- 0.15 (0-97..1.58)
Optimal RMSD range	119-196
- Backbone RMSD[A]	0.41 +/- 0.07 (0.32..0.57)
- Heavy atom RMSD [A]	0.93 +/- 0.11 (0-70..1.11)
NOE restraints [#]	908
- intraresidual (| i-j | =0)	358 (41.15%)
- sequential (| i-j | =1)	239 (27.47%)
- medium range (1< | i-j | <5)	137 (15.75%)
- long-range (| i-j | >4)	136 (15.63%)
NOE restraints per residue	10.12
RMS NOE restraint violation [A]	0.0407
Dihedral restraints [#]	157
RMS dihedral restraint violation [^{0}]	0
Ramachandran statistics	
- most favoured [%]	77.34
- additionally allowed [%]	22.34
- generously allowed [%]	0.19
- disallowed [%]	0.13

By overlapping the structure obtained from an automatic calculation method LARP4 (111-196) and the structure obtained from a manually method La Motif of La genuine protein, it can be shown that we obtained a very similar 3D structure (figure 3).

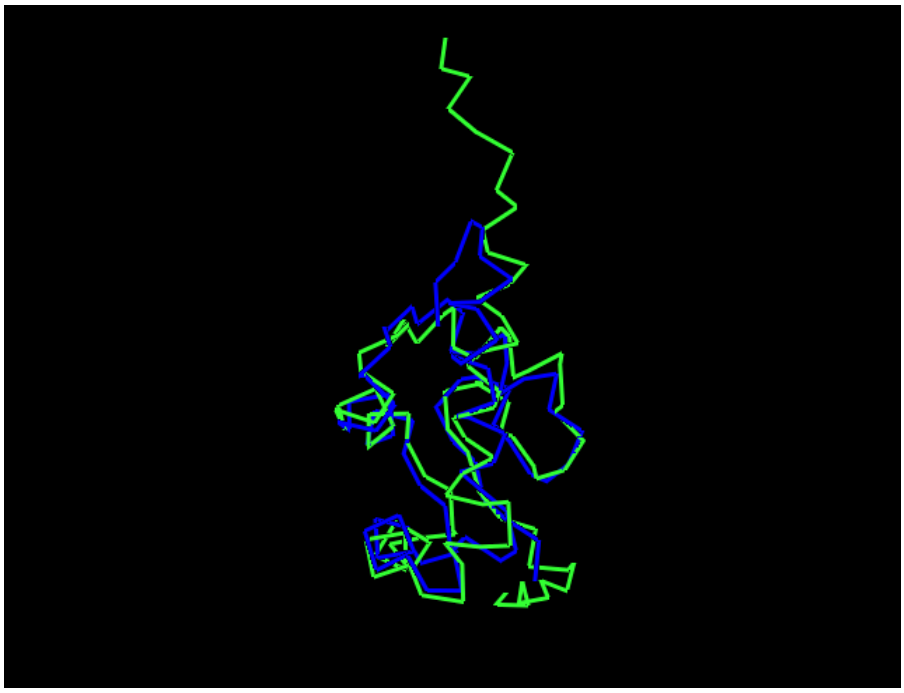


Figure 3: Overlap of the domain La Motif of La genuine protein (green) and LARP4 (blue).

At a first analysis we can see that LARP4 (represented in blue) has some regions that are positioned in a different way from La genuine protein, in other words some regions of LARP4 are not perfectly aligned with La protein.
This is not a mistake of assignment because we can still see NOE correlations between the amino acid residues involved, and by overlapping the HSQC of LARP4 (111-287) La Module, LARP4 (111-196) and LARP4 (196-287) it can clearly be seen that all the residues shifted represent the region not aligned with La protein (figure 4).

Relative orientation of La motif and RRM

As a next analysis I anted to investigate the relative mobility and orientation of the La motif and RRM1 in the La Module of LARP4.

As I describe in chapter three, La genuine protein bind poly (U) RNA with both domain: La Motif and RRM1 and only during the interaction the linker that connects the two domain fold into an a helix, this folding appears to be an indirect effect of RNA binding because there are not contacts between the RNA and this part of the polypeptide. Whether this was the case for LARP4 needed to be investigate. For this I analysed the HSQC of the two individual domains compared to the La Module. If the two domains are mainly tumbling independently, I would expect very little differences in chemical shift compared to the double domain. If on the contrary shifts are seen, this might indicate regions of the two domains that interact with each other.

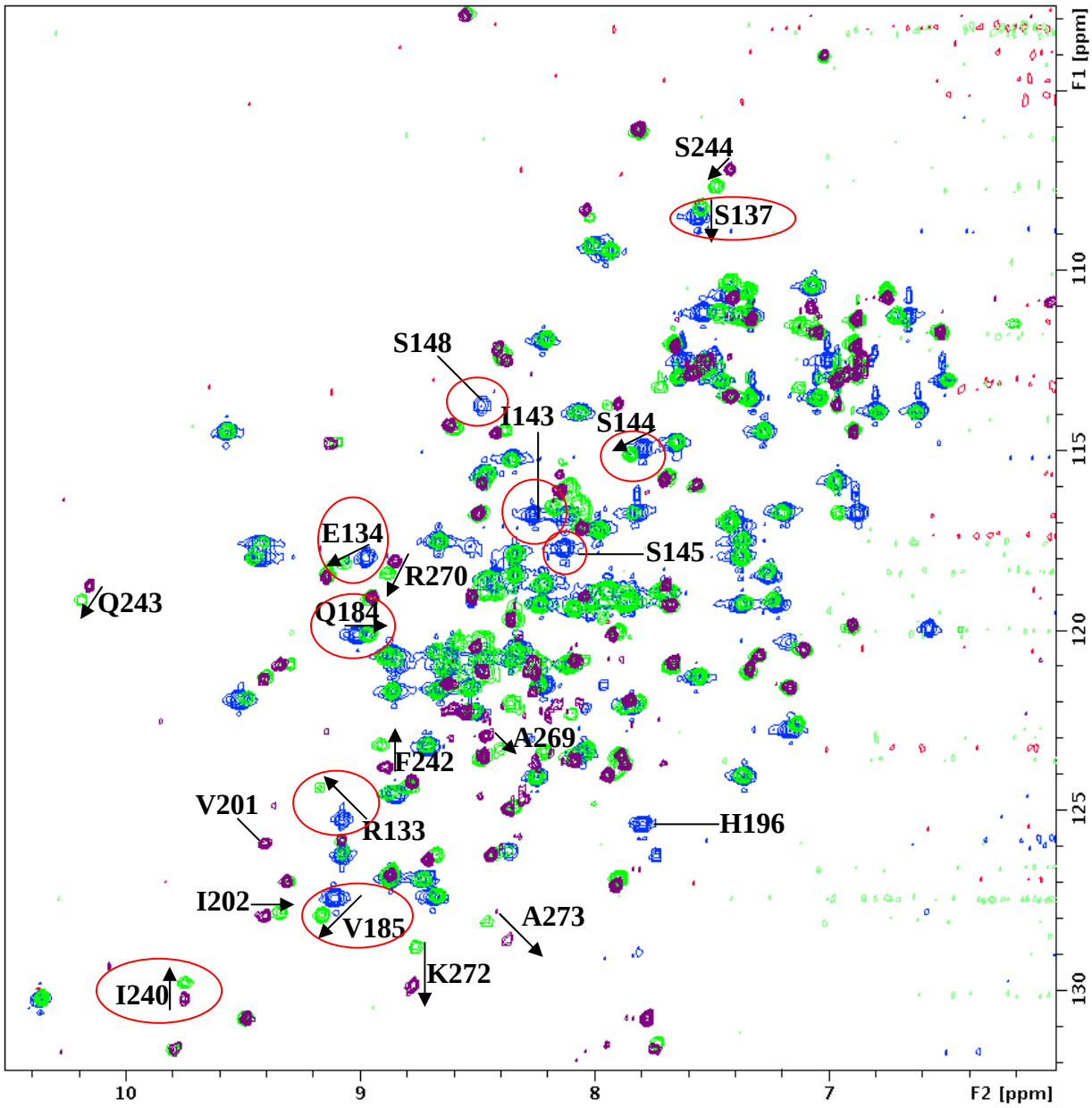


Figure 3: Overlap of HSQC of LARP4 (111-287) in green, LARP4 (111-196) in blue and LARP4 (196-287) in purple. I drew a circle for the residues of LARP4 (111-196) that moved and they are involved in a different arrangement structure.

In figure 3 I drew a circle for the residue of LARP4 (111-196) moved with respect to the entire domain La Module: the residues involved in the shift are the amino acids present on the bottom of the structure. In figure 4 I report the

overlap of La motif's structure of La genuine protein and LARP4 (111-196); in red are represented the residues of the La motif that undergo chemical shift perturbations in the tandem La Module and therefore probably involved in the interaction with the RRM1. By analysing this figure, we can assume that the RRM1 could interact with La Motif placed on the bottom of that domain.

These hypothesis need to be confirmed by further study of the relaxation time of the two LARP4's domains together and separately; such a study could clarify whether, like the La genuine protein, the two domains are interdependent or not in the La module.

In my research I have also developed the assignment for LARP4 (111-287), La Module, but it proved quite difficult to assign the linker that connect the two domain.

Few explanations can be given which account for the different arrangement of LARP4 (111-196) compared with the La genuine protein:

- probably because in LARP4 we did not have the linker as in La genuine protein that fold in a helix
- or because it is very short
- or it has a different folding

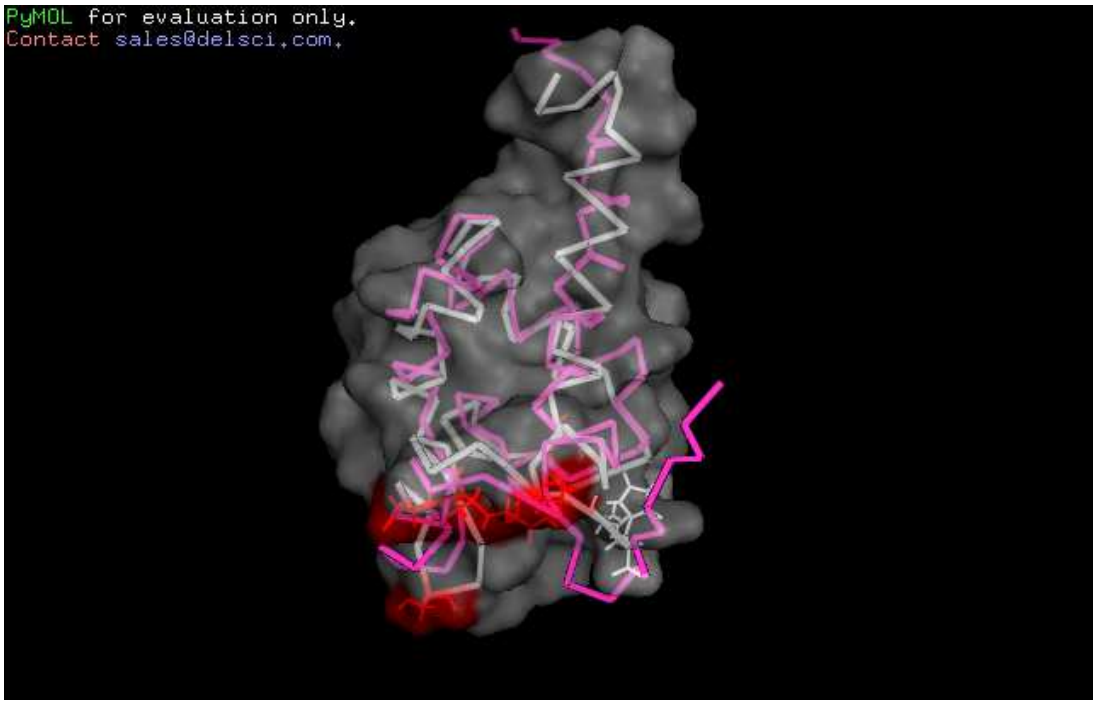


Figure 4: Overlap of the La motif's structure of La genuine (magenta) LARP4 (grey)

LARP4 RRM1

For LARP4 (196-287) I find the best condition to study the protein and the TALOS files show the different prediction at pH 5 or pH 7 (figure 5).

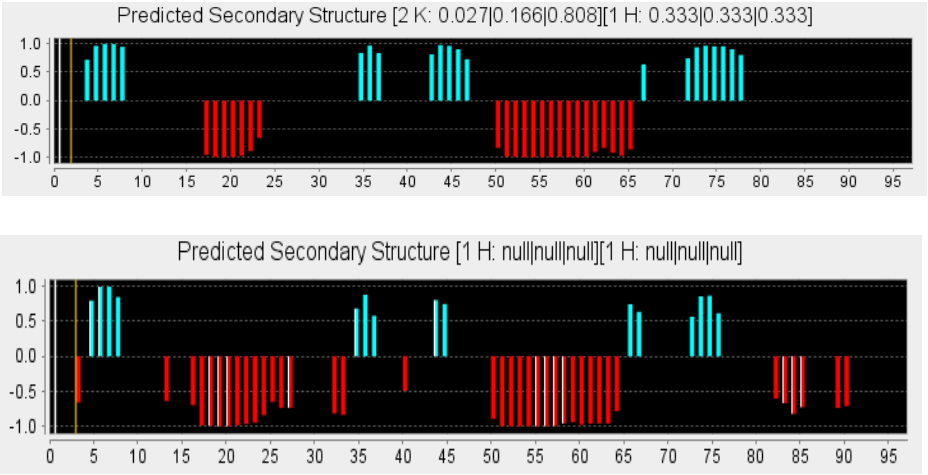


Figure 5: Comparison of TALOS prediction at different condition: pH5 on the top and pH7 on the bottom.

At pH 7 the last a helix reappears, like the one found in the RRM1 of La genuine protein. This analysis also indicated that in LARP4 we have either a longer β3 strand or a small additional strand preceding β3 (figure 6).

```
>La 105-202       GRWILKNDVKNRSVYIKGFPTDATLDDIKEWLEDKGQVLNIQMRRTLHKA
>LARP4 196-287    --------HKRCIVILREIPETTPIEEVKGLFKSENCPKVI----SCEFA

>La 105-20        FKGSIFVVFDSIES---AKKFVETPGQKYKETDLLILFK--DDYFAKKNEERKQNKVE
>LARP4 196-287    HNSNWYITFQSDTDAQQAFKYLREEVKTFQGKPIMARIKAINTFFAK-NGYR---LMD
```

Figure 6: Allignment of RRM domain of La genuine protein and LARP4

The next step is to pursue structural determination of LARP4 and LARP4-polyA RNA complex by NMR and X-ray.

Additional studies on the NMR data are in progress, in particular the side chains assignment of LARP4 (196-287) is almost completed in order to have a structure of this domain and to compare it with La genuine protein RRM1.

Also in progress are the acquisition of the relaxation time experiments for both domains of LARP4 separately (La Motif and RRM) and together (La Module) and then with the poly-A RNA. This experiments are very important because will allow us to understand which domain and in particular which residues are involved in the interaction with the RNA, so that it will be possible to find the difference with the La genuine protein.

We are also planning an NMR and ITC titration of LARP4 (111-287) with poly-A RNA to compare all the results against each other.

CONCLUSION

At this stage my research can only report the structure of LARP4 (111-196) with the important finding that it does not show differences with the same domain in La protein, if we exclude the region that is very liked involved in the interaction with the RRM1, but this hypothesis need to be confirmed by further experiments.

The most interesting outcome of my research is that this structure has been obtained with automatic method which represent an important step forward in the calculations domain, having achieved the same results that in the past Alfano et al. had reached applying a manual method.