



TESI DOCTORAL

Títol	<i>In silico</i> strategies for the design of RNA binders: focus on nucleotide repeat expansion disorders and HIV-1
Realitzada per	Alejandro López González
en el Centre	IQS School of Engineering
i en el Departament	Química Orgànica
Dirigida per	Dr. Jordi Teixidó Closa i Dr. Roger Estrada Tejedor

A la meva família.

The important thing is not to stop questioning. Curiosity has its own reason for existing. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries merely to comprehend a little of this mystery every day. Never lose a holy curiosity.

- Albert Einstein

AGRAÏMENTS

Vull agrair la col·laboració de totes aquelles persones i institucions que han fet possible la realització d'aquesta tesi.

En primer lloc, vull agrair als meus directors de tesi. Al Dr. Jordi Teixidó i al Dr. Roger Estrada, que van confiar en mi per dur a terme aquest projecte i van introduir-me en el món computacional, i també per ajudar a posar en ordre les idees d'aquest treball. També vull agrair el seu suport i interès al Dr. José Ignacio Borrell que ha conduït una part important d'aquest projecte. Gràcies també a la Fundació La Marató de TV3 per la beca predoctoral que m'ha permès continuar amb aquesta tasca, i als grups col·laboradors en aquest projecte IUCT i Universitat de València. Especialment al grup del Dr. Rubén Artero per tot el suport tècnic i científic durant aquest projecte de recerca.

Als computacionals, i als que ho van ser, a en Marc, l'Agustí, en Javi, la Pili i la Júlia; perquè vam patir junts el pes de la teòrica però ho vam ensucrar amb molts somriures. I també als nouvinguts, Eli, Pere, Regi, Elsa, Quico, Jaume i Anna per fer que la tornada al departament fos més càlida. Als fotoquímics amb els quals he conviscut durant tota aquesta etapa, la Bea, l'Uri, en Rubén, l'Ester, en Roger i en Joaquim. I també amb les que no hi són present al departament però em van animar durant les etapes més fosques, gràcies Júlia i Alicia. Als companys sintètics, especialment a l'Albert i a la Marta que van fer que els dinars fossin el moment especial. A l'Annie, que la trobem a faltar des de la distància i a l'Héctor, que m'ha acompanyat des dels meus inicis a l'IQS.

I am very grateful to my research advisor in The Scripps Research Institute, Dr. Matthew D. Disney for his constant support and guidance during the course of my time in his lab and for introducing me to the RNA chemical biology world. I would also like to thank all the members of Disney's lab for their support and encouragement. Thanks to Jessica, Slava, Vala, Sai, Yong and Yoshio, and I want to express my special thanks to Suzanne for showing me about how a translational research is done.

Vull agrair també el suport d'en Fredi, en Claudio, l'Andrea i la Rebecca per ser sempre amb mi quan més ho he necessitat. También quiero agradecer esta tesis a los españoles perdidos por Florida, a Tamara, Alicia, Akaitz y Mónica por hacer que aquella estancia fuera inolvidable. Así cualquiera se siente como en casa.

Finalmente, mi más sincero agradecimiento a mi familia, que ha sido un apoyo constante e incondicional durante toda esta etapa. Gracias a mi madre, mi padre, mis hermanos y mi tía. Gràcies a tots per haver-me acompanyat en aquesta etapa y per ajudar-me a tancar aquest capítol personal tan important.

SUMARI

Tradicionalment, el disseny de fàrmacs s'ha basat en utilitzar molècules per tractar dianes biològiques clau o en una via metabòlica concreta, típicament proteïnes. No obstant, aquest paradigma està evolucionant de forma que el problema a resoldre no es troba només en entendre l'activació o inactivació de proteïnes sinó en controlar la maquinària interna que utilitza la cèl·lula per produir aquestes proteïnes. De la mateixa manera, les malalties conegudes com a minoritàries tenen els seus orígens en la genètica, i la utilització de proteïnes com a diana terapèutica ha esdevingut insuficient.

L'ARN presenta una importància fonamental en els éssers vius degut al seu paper en multitud d'esdeveniments de regulació i control. La unió de molècules a l'ARN pot permetre el silenciament gènic o sobreexpressió d'un gen en concret. Fins la data, pocs fàrmacs amb la capacitat d'unir-se a l'ARN han tingut un impacte clínic significatiu pel que és imperatiu identificar nous esquelets químics que superin les dificultats associades al reconeixement d'aquesta macromolècula. Tanmateix, s'ha observat que el tractament per inhibir l'expressió gènica de determinats patògens és una estratègia prometedora ja que universalitza la seva activitat virucida.

El factor limitant actual es troba en la manca de metodologies que permetin identificar de forma ràpida i efectiva potencials fàrmacs per tractar aquest tipus de patologies. Per aquest motiu, la present tesi es basa en la recerca de noves metodologies computacionals per al reconeixement de potencials fàrmacs per al tractament de dues malalties ARN-dependents com són la distròfia miotònica de tipus 1 (DM1) i el virus de la immunodeficiència humana (VIH-1). Gran part d'aquestes metodologies es basen en l'estructura del receptor pel que és fonamental una millor comprensió de les particularitats estructurals de l'ARN en un entorn dinàmic. Per aquest motiu, en el present treball s'han aplicat protocols de dinàmica molecular i models de xarxes elàstiques a fi d'investigar la promiscuïtat estructural d'aquests ARNs. Aquestes tècniques han permès obtenir informació per a refinar la modelització de noves famílies de compostos amb activitat biològica en models *in cellulo*. De la mateixa manera, s'ha realitzat una racionalització de l'activitat biològica de pèptids i peptoids actius en front de DM1 i VIH-1 per a poder-ne millorar la seva activitat en futures modificacions.

Finalment, s'ha estudiat la dinàmica d'una proteïna essencial per al desenvolupament muscular, la desregulació de la qual està involucrada en múltiples malalties minoritàries, i especialment en DM1. Aquesta proteïna (MBNL1) està involucrada en múltiples processos de regulació i l'estudi realitzat en aquesta tesi proposa que el seu comportament dinàmic condiciona les seves propietats d'unió a certs transcrits d'ARN per a expressar o silenciar determinats gens.

SUMARIO

Tradicionalmente, el diseño de fármacos se ha basado en usar moléculas para tratar dianas biológicas clave o una vía metabólica concreta, típicamente proteínas. No obstante, este paradigma está evolucionando de forma que el problema a resolver no se encuentra sólo en la activación o inactivación de proteínas sino en controlar la maquinaria interna que usa la célula para producir estas proteínas. Del mismo modo, las enfermedades conocidas como minoritarias tienen sus orígenes en la genética, y la utilización de proteínas como diana terapéutica es, a día de hoy, insuficiente.

El ARN presenta una importancia fundamental en los seres vivos debido a su papel en multitud de procesos de regulación y control. La unión de moléculas al ARN potencialmente permite el silenciamiento o sobreexpresión de un gen en concreto. Hasta la fecha, pocos fármacos con capacidad de unirse al ARN han tenido un impacto clínico significativo por lo que es imperativo identificar nuevos esqueletos químicos que superen las dificultades asociadas al reconocimiento de esta macromolécula. Igualmente, se ha observado que el tratamiento para inhibir la expresión génica de determinados patógenos es una estrategia prometedora ya que universaliza su actividad virucida.

El factor limitante actual se encuentra en la falta de nuevas metodologías que permitan identificar de forma rápida y efectiva potenciales fármacos para tratar este tipo de patologías. Por este motivo, la presente tesis se basa en la prospección de nuevas metodologías computacionales para el reconocimiento de potenciales fármacos para el tratamiento de dos enfermedades ARN-dependientes como son la distrofia miotónica de tipo 1 (DM1) y el virus de la inmunodeficiencia humana (VIH-1). La mayor parte de estas metodologías se basan en la estructura del receptor por lo que es fundamental una mejor comprensión de las particularidades estructurales del ARN en un entorno dinámico. Por este motivo, en el presente trabajo se han aplicado protocolos de dinámica molecular y modelos de redes elásticas con el fin de investigar la promiscuidad estructural de estas estructuras. Estas técnicas han permitido obtener información para refinar la modelización de nuevas familias de compuestos con actividad biológica en modelos *in cellulo*. Del mismo modo, se ha realizado una racionalización de la actividad biológica de péptidos y peptodios activos frente a DM1 y VIH-1 para poder mejorar su actividad en futuras modificaciones.

Finalmente, se ha estudiado la dinámica de una proteína esencial para el desarrollo muscular, cuya desregulación está involucrada en múltiples enfermedades minoritarias, especialmente en DM1. Esta proteína (MBNL1) está involucrada en múltiples procesos de regulación y el estudio realizado en esta tesis propone que su comportamiento dinámico condiciona las propiedades de unión a ciertos transcritos de ARN para expresar o silenciar determinados genes.

SUMMARY

Traditionally, drug design has focused on using molecules for treating biological targets or a specific metabolic pathway, typically proteins. However, this model is changing in such a way that the problem to be resolved is not only to be found in the activation or inactivation of proteins but also in controlling the internal cellular machinery that regulates these proteins. Similarly, the new generation of diseases known as genetic or 'rare' diseases is rooted in genetics, and the use of proteins as a therapeutic target is, today, insufficient.

RNA has a fundamental importance in life and it plays a role in multiple processes of regulation and control. Small molecules that bind RNA can potentially silence or promote gene expression. To date, few drugs that potentially bind RNA had a significant clinical impact so it is essential to identify new chemical scaffolds that are able to overcome the inherent difficulties of this macromolecule recognition. Equally important, it the observation that inhibition of gene expression of certain pathogens is a promising strategy due to their universal virucidal activity.

The current limiting factor is the lack of alternative methodologies to quickly and effectively identify potential drugs to treat such diseases. Therefore, this thesis is based on the exploration of new computational methodologies for identifying potential drugs for the treatment of two RNA-dependent diseases such as myotonic dystrophy type 1 (DM1) and the human immunodeficiency virus (HIV-1). Most of these methods are based on the receptor's structure which is why a better understanding of the structural characteristics of RNA in a dynamic environment is essential. Therefore, in this thesis in order to investigate the structural promiscuity of these targets molecular dynamics and elastic network model protocols have been applied. These techniques have provided important information to refine the modeling of new families of compounds with biological activity on *in cellulo* models of DM1. Similarly, a rationalization of the biological activity of peptides and peptoids active against DM1 and HIV-1 based on structure has been conducted to improve their potency in future modifications.

Finally, the dynamics of an essential protein for muscle development, whose deregulation is involved in many rare diseases, especially in DM1, has been studied. This protein (MBNL1) is involved in many regulatory processes and the study in this thesis proposes that its dynamic behavior determines its binding properties in order to bind specific RNA transcripts to enhance or silence gene expression.

INDEX

1. Introduction	
1.1. The central dogma of molecular biology	3
1.1.1. Biological information and inheritance	4
1.1.2. Transfer of biological sequential information	5
1.1.3. Functional structure of a gene	10
1.2. Structural biology of RNA	11
1.2.1. Chemical structure and stability of nucleic acids	12
1.2.2. General nomenclature of nucleotides	15
1.2.3. Geometric classification of RNA base pairs	18
1.2.4. Movement of bases into the helical space	19
1.2.5. Secondary and tertiary structural elements in RNA	20
1.3. Nucleotide repeat expansion disorders	22
1.3.1. Brief characteristics of nucleotide repeat expansion disorders	23
1.3.2. Structural classes of isolated repeats	24
1.3.3. Clinical features of myotonic dystrophies	25
1.3.4. Molecular pathomechanism of myotonic dystrophies	26
1.3.5. From mechanism to therapeutics of DM1	29
1.3.6. Pathogenesis of spinocerebellar ataxia type 10	30
1.4. Targeting of HIV-1 RNA	32
1.4.1. Structure and replication cycle	33
1.4.2. Trans-acting response element as a therapeutic target	34
1.5. Objectives	37
1.6. References	39
2. Methods	
2.1. Consilience of a multidisciplinary enterprise	47
2.1.1. Definition of terms	47
2.1.2. Components of a model	48
2.2. Molecular mechanics	50
2.2.1. Force field definition	50

2.2.2.	Statistical mechanics	52
2.2.3.	Molecular dynamics	53
2.2.4.	Enhanced sampling techniques	55
2.2.5.	Simplifying the models: elastic networks	57
2.3.	Computer-aided drug design	59
2.3.1.	Drug design in the discovery pipeline	59
2.3.2.	Structure-based drug design	60
2.3.3.	Ligand-based drug design	63
2.4.	References	67
3.	Structural complexity of the RNA	
3.1.	Background of the study	73
3.2.	Unveiling the dynamics of CNG repeats	74
3.2.1.	Conformational sampling through molecular dynamics	75
3.2.2.	Validation of the structural parameters	76
3.2.3.	Simulation of rCUG hairpin models	77
3.2.4.	U-U conformational sampling of r(CUG) _n structures	80
3.2.5.	Stiffness analysis of r(CUG) _n models	81
3.2.6.	Cluster analysis and hydration effect	82
3.2.7.	Significance of dynamics for drug design	85
3.3.	Simplifying the model: elastic network models as RNA conformational sampling tools	87
3.3.1.	Anisotropic network models of the rCUG ensemble	87
3.3.2.	Conventional MD of the rCUG system model	90
3.3.3.	Accelerated MD of the rCUG system model	91
3.3.4.	Comparison of ANM, PCA and EDA modes	93
3.3.5.	The simpler the better (or not)	95
3.4.	The context matters: structured rAUUCU	98
3.4.1.	Dynamics of the ⁵ UCU ³ / ³ UCU ⁵ model system	99
3.4.2.	Stabilization of C-C pairs by water-mediated hydrogen bonds	100
3.4.3.	Inherent instability of the ⁵ UCU ³ / ³ UCU ⁵ internal loop	103
3.5.	Protocols	105
3.5.1.	Equipment	105
3.5.2.	Protocols	106

3.6. References	112
4. RNA drug design strategies for DM1	
4.1. Background of the study	117
4.2. Novel scoring function for nucleic acids – small molecule complexes	119
4.2.1. Artificial neural networks function	120
4.2.2. Benchmarking several docking protocols for RNA	123
4.3. Ligand-based selection of compounds for DM1	127
4.3.1. Description of the chemical library	128
4.3.2. Activity-binding correlation	129
4.3.3. <i>De novo</i> drug design: theophylline dimers	130
4.3.4. Pharmacophore screening: pentamidine-like compounds	131
4.3.5. Chemoinformatic analyses for scaffolds identification	132
4.3.6. Pyrido[2,3- <i>d</i>]pyrimidines as an active scaffold	134
4.4. Druggability study of rCUG structures	137
4.4.1. Druggability analysis	137
4.4.2. Molecular recognition depends on essential dynamics	139
4.5. Quantitative structure-activity relationship (QSAR)	142
4.5.1. Description of the dataset	143
4.5.2. ANN-QSAR model performance	146
4.6. Protocols	150
4.6.1. Equipment	150
4.6.2. Protocols	150
4.7. References	152
5. Protein and peptide strategies	
5.1. Background of the study	159
5.2. Investigating the MBNL1 protein	161
5.2.1. Sequence coevolution of the CCCH domain	163
5.2.2. MBNL domains from a structural perspective	164
5.2.3. MBNL2 experimental fluctuations correlate with predicted normal modes	165
5.2.4. ZnF1/2 from MBNL1 and MBNL2 exhibit equivalent large-scale motions	167
5.2.5. Global dynamics of ZnF3/4 differ from those of ZnF1/2	168

5.2.6.	Effect of RNA binding over local fluctuations	169
5.2.7.	Both ZnFs of MBNL1 have differentiated affinity for RNA	170
5.2.8.	Significance of the structural study	172
5.3.	D-hexapeptides as DM1 drugs	176
5.3.1.	Peptide structural sampling	177
5.3.2.	Peptide modifications	181
5.4.	Peptide strategies for HIV-1 targeting	183
5.4.1.	TAR molecular recognition of penta-C- α -PAA Ic and IIIb	184
5.4.2.	Interfacial waters play an essential role in recognition	186
5.5.	Protocols	188
5.5.1.	Equipment	188
5.5.2.	Protocols	188
5.6.	References	192
6.	Conclusions	197
7.	Annexes	199
8.	Publications	213

INDEX OF FIGURES

Figure 1.1. Chemical structure of a three base pair fragment of a DNA double helix	5
Figure 1.2. Schematic representation of DNA replication.	6
Figure 1.3. Schematic process of RNA transcription.	8
Figure 1.4. Simplified representation of the alternative splicing mechanism.	9
Figure 1.5. Regulatory sequence controls when and where expression occurs for the protein coding region.	10
Figure 1.6. Primary, secondary and tertiary nucleic acids representation.	14
Figure 1.7. Representation of the dipole-dipole interaction caused by London dispersion forces.	15
Figure 1.8. Nucleotide nomenclature.	17
Figure 1.9. Definition of the interacting edges and glycosidic bond orientations.	18
Figure 1.10. Base pair parameters, step parameters and helical parameters.	20
Figure 1.11. RNA secondary and tertiary structural elements.	21
Figure 1.12. Structural classes of isolated triplet repeats in the human genome.	25
Figure 1.13. Postulated pathological mechanism underlying DM1 and DM2.	28
Figure 1.14. Diagram of HIV-1 virion.	34
Figure 2.1. Bonded and non-bonded interactions representation.	51
Figure 2.2. Simplified flowchart of a typical MD simulation.	54
Figure 2.3. Periodic boundary conditions.	54
Figure 2.4. Replica-exchange MD, accelerated MD and steered MD methods.	57
Figure 2.5. Schematic illustration of a docking process.	61
Figure 2.6. Representation of a set of pharmacophoric keys.	63
Figure 2.7. Representation of an artificial neural network.	65
Figure 3.1. DMPK mRNA repeated expansion folds into a metastable hairpin and causes DM1 by sequestering MBNL1.	75
Figure 3.2. Schematic representation and three-dimensional model of ds[(CUG) ₆] ₂ .	76
Figure 3.3. Three-dimensional representation of ds[(CUG) ₆] ₂ , r(CUG) ₁₆ and r(CUG) ₈ .	79
Figure 3.4. Representation of the fraying and U2 conformational states from r(CUG) ₁₆ .	80
Figure 3.5. Stiffness constants along the sequence of the different model fragments.	82

Figure 3.6. Density plot of η/θ states observed along the trajectory.	83
Figure 3.7. Isodensity contour plots of cluster pairs.	84
Figure 3.8. Classification of the average CUG repeats obtained with MD.	86
Figure 3.9. Design and validation of the distance dependent force constant for ANM.	89
Figure 3.10. Maximum overlapping achieved by using different cutoffs.	89
Figure 3.11. Cumulative overlap of ANM soft modes of PC1 to PC3.	90
Figure 3.12. Representation of the all-atom and coarse-grained models.	91
Figure 3.13. Structural analyses of the cMD and aMD simulations of rCUG.	92
Figure 3.14. Projection over the PC1-3 subspace of cMD, aMD and PDB.	93
Figure 3.15. Atomic global fluctuations extracted from the cMD, aMD and PDB.	95
Figure 3.16. Sequence, RMSD and clusters of the rAUUCU simulation.	100
Figure 3.17. Stiffness force constant of the helical space of rAUUCU.	101
Figure 3.18. 2D-PMF surface for the C-C conformations along the MD simulation.	102
Figure 3.19. Stick model and charge distribution of the potential hydrogen bond of the rAUUCU internal loop and a water molecule.	102
Figure 3.20. Pulling force and cumulative work distribution profiles vs distance from the center of mass.	104
Figure 4.1. Distribution of main physicochemical properties of the training and test sets.	121
Figure 4.2. Schematic representation of the parameterization of the ANN scoring function.	121
Figure 4.3. Recognition vs RMSE and experimental vs predicted correlation of the ANN models.	122
Figure 4.4. Score vs RMSD representation of poses scored with Vina and ANNScore.	123
Figure 4.5. Small molecule – RNA docking benchmark summary.	124
Figure 4.6. Superposition of native conformations and docking generated poses.	125
Figure 4.7. Percentage of average recovery of the <i>Drosophila</i> population.	127
Figure 4.8. Physicochemical properties and fragment frequency of in-house molecules.	128
Figure 4.9. ROC plot of docking results vs <i>Drosophila</i> model data.	129
Figure 4.10. Theophylline's derivatives and binding through the major groove.	130
Figure 4.11. Relative fluorescence polarization for caffeine and derivatives.	131
Figure 4.12. Scaffold tree map of the qHTS chemoinformatic analysis.	133
Figure 4.13. Projection of active, inactive and inconclusive molecules over the PC1 and PC2 subspace and physicochemical properties.	134
Figure 4.14. Projection of pyrido[2,3- <i>d</i>]pyrimidines over the PC1 and PC2 subspace.	135

Figure 4.15. Druggability analysis of an r(CUG) ₃ model.	138
Figure 4.16. ROC curve obtained from the rigid and flexible docking.	140
Figure 4.17. Superposition of monomer fragment of LR08 and the binding hotspots.	141
Figure 4.18. Physicochemical properties of the selected molecules from the qHTS.	145
Figure 4.19. Explained variance vs number of PCs and descriptor weights.	146
Figure 4.20. Prediction, recognition, accuracy and RMSE of QSAR-ANN models.	147
Figure 4.21. False positive molecules obtained in the training and test sets.	148
Figure 5.1. Logo and sequence coevolution analysis of CCCH domains.	164
Figure 5.2. Three-dimensional structure, sequence alignment and electrostatic potential of MBNL1 domains.	166
Figure 5.3. Correlation matrix between ANM and PCA and normalized squared fluctuations for MBNL2.	167
Figure 5.4. PC1 histogram for ZnF1/2 MD and squared fluctuations.	168
Figure 5.5. PC1 histogram for ZnF3/4 MD and squared fluctuations.	169
Figure 5.6. Local fluctuation of ZnFs with and without RNA.	171
Figure 5.7. Work profiles for each SMD pulling process.	173
Figure 5.8. Local fluctuations extracted from the SMD simulations and pocket volume distribution.	174
Figure 5.9. PPX Ramachandran plots extracted from the MD simulations.	178
Figure 5.10. General Ramachandran plot representations of the non-PPX peptides.	179
Figure 5.11. Normalized contribution of each amino acid to the overall secondary structure.	179
Figure 5.12. Correlation between PPII content and activity in animal model.	180
Figure 5.13. ABP1 modifications: peptoid analog and modified peptide structures.	181
Figure 5.14. Structural sampling of representative conformations of the modifications.	182
Figure 5.15. Molecular structure of four PAAs active against HIV-1.	184
Figure 5.16. Frame representations for compounds Ic and IIIb complexes with TAR.	186
Figure 5.17. Correlation between experimental and calculated ΔG of binding.	187
Figure 5.18. Interfacial density maps for Ic , RFFR, IIIb and KRFR.	187

ABBREVIATIONS

A	Adenine
aMD	Accelerated Molecular Dynamics
ANM	Anisotropic Network Models
ANN	Artificial Neural Networks
C	Cytosine
CADD	Computer-aided Drug Design
cMD	Conventional Molecular Dynamics
DM1	Myotonic Dystrophy type 1
DM2	Myotonic Dystrophy type 2
DNA	Deoxyribonucleic Acid
EDA	Essential Dynamics Analysis
ENM	Elastic Network Models
G	Guanine
GNM	Gaussian Network Models
HIV	Human Immunodeficiency Virus
LBDD	Ligand-based Drug Design
LMO	Leave-many-out
LOO	Leave-one-out
MBNL	Muscleblind-like (protein)
MLR	Multivariate Linear Regression
NA	Nucleic Acids
MD	Molecular Dynamics
OMIM	Online Mendelian Inheritance in Man
PC	Principal Component
PCA	Principal Components Analysis
PDB	Protein Data Bank
PLS	Partial Least Squares
PPI	Protein – Protein Interactions
QSAR	Quantitative Structure-Activity Relationship

RAN	Repeat Associated Non-ATG (translation)
REMD	Replica-Exchange Molecular Dynamics
RMSD	Root-mean-squared Deviation
RMSE	Root-mean-squared Error
RNA	Ribonucleic Acid
SBDD	Structured-based Drug Design
SCA10	Spinocerebellar Ataxia type 10
STR	Short Tandem Repeats
T	Thymine
TNR	Trinucleotide Repeats
TRED	Triplet Repeat Expansion Disorder
U	Uracil

CHAPTER I. INTRODUCTION

1.1. THE CENTRAL DOGMA OF MOLECULAR BIOLOGY

I called this idea the central dogma, for two reasons, I suspect. I had already used the obvious word hypothesis in the sequence hypothesis, and in addition I wanted to suggest that this new assumption was more central and more powerful. [...] As it turned out, the use of the word dogma caused almost more trouble than it was worth. Many years later Jacques Monod pointed out to me that I did not appear to understand the correct use of the word dogma, which is a belief that cannot be doubted. I did apprehend this in a vague sort of way but since I thought that all religious beliefs were without foundation, I used the word the way I myself thought about it, not as most of the world does, and simply applied it to a grand hypothesis that, however plausible, had little direct experimental support.

- Francis Crick

What Mad Pursuit: A Personal View of Scientific Discovery. 1988.

The classical view of the central dogma of molecular biology describes the two-step process, **transcription** and **translation**, by which the information in the genetic code flows into proteins: DNA → RNA → protein. The central dogma states that "the coded genetic information hard-wired into deoxyribonucleic acid (DNA) is transcribed into individual transportable cassettes, composed of messenger ribonucleic acids (RNA, mRNA); each mRNA cassette contains the program for synthesis of a particular protein (or small number of proteins)."¹

In 1958, Francis Crick described all possible directions of information flow between DNA, RNA and protein.¹ He concluded that once information was transferred from nucleic acid (DNA or RNA) to protein it could not flow back to nucleic acids. Thus, the transfer of information to proteins is irreversible. However, the central dogma was restated by Crick in 1970 at a time when Howard Temin and David Baltimore both independently discovered the enzyme responsible for reverse transcription. According to Crick, the correct, concise version of the central dogma is:²

... The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

The central dogma is the backbone of an entire discipline. However, announcement of its demise has come every so often because of recent discoveries, such as non-coding RNAs (ncRNA), alternative splicing, reverse transcriptase, introns, junk DNA, epigenetics, RNA viruses, trans-splicing, transposons, prions, epigenetics, and gene rearrangements among others (further details about some of

these events will be forthcoming in the next sections). Although Crick's perspective has never been challenged, Benjamin Lewin adapted its definition as follows:³

... The central dogma states that information in nucleic acid can be perpetuated or transferred but the transfer of information into protein is irreversible.

This dissertation is based on pathogenic RNA targets which are directly related to the deregulation of many processes that occur during the classical "DNA makes RNA makes protein" flow of biological information. For this reason, a brief introduction of the most common biological events will be presented.

1.1.1. BIOLOGICAL INFORMATION AND INHERITANCE

As stated by the central dogma, all living organisms are regulated by the sequential transfer of biological information. Moreover, this information must be discrete and inheritable. Thus, how a 'biological information unit' is described? The necessary information to build and maintain an organism's cells is contained into the **genes**. The word 'gene' is used to define stretches of DNA (and RNA in some cases) that code for a polypeptide or for a functional RNA chain. Then, nucleic acids are established as a molecular repository of genetic information.

Genes are inherited from two parents which divide out copies of their genes to their offspring. Sometimes this process is compared with mixing two hands of cards, shuffling them, and then dealing them out again. Humans have a pair of copies of each of the parental genes which come from copies that are found in eggs or sperm - but they only include one copy of each type of gene. The join of the biological information from an egg and sperm form a complete set of genes. The eventually resulting child has the same number of genes as their parents, but for any gene, one of their two copies comes from their father and the other from their mother. The effects of this mixing depend on the types - the **alleles** - of the gene.

The vast majority of living organism encode their genes in long strands of DNA. From a biochemical perspective, DNA consists of a chain made from four types of nucleotide units, each of which contains: a five-carbon sugar (**2'-deoxyribose**), a **phosphate** group, and one of the four bases named **adenine (A)**, **cytosine (C)**, **guanine (G)** and **thymine (T)** (refer to figure 1.1, page 5).

Two chains of DNA twist around each other to form a DNA double helix with the phosphate-sugar backbone spiraling around the outside, and the bases pointing inwards with adenine base pairing to thymine and guanine to cytosine. The specificity of base pairing occurs because when adenine and thymine align form two hydrogen bonds, whereas cytosine and guanine form three hydrogen bonds. The two strands in a double helix must therefore be complementary.

Moreover, DNA strands have directionality due to the chemical composition of the pentose residues of the bases. One end of a DNA polymer contains an exposed hydroxyl group on the 3' position of the deoxyribose; this is known as the **3' end** of the molecule. The other end contains an exposed phosphate group and is called the **5' end**. The two strands of a double-helix run in antiparallel directions.

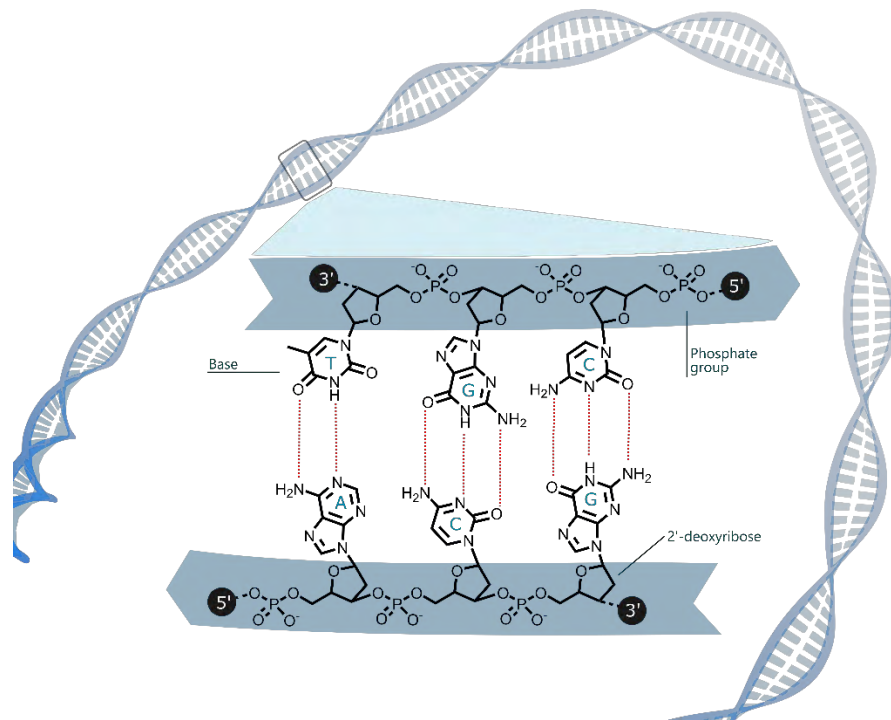


Figure 1.1. Chemical structure of a three base pair fragment of a DNA double helix. The sugar-phosphate backbone chains run in opposite direction with the bases pointing inwards, base-pairing A (cyan) to T (dark blue) and C (dark green) to G (green) with hydrogen bonds (red dashed lines).

1.1.2. TRANSFER OF BIOLOGICAL SEQUENTIAL INFORMATION

The central dogma definition suggests several classes of biological information transfer processes (table 1.1). These processes can be roughly separated into general classes (observed in most of the living organisms) and special classes (mainly observed in viruses and some eukaryotes). Herein, a brief description of the most relevant events during DNA replication, transcription and translation will be introduced.

Table 1.1. General and special classes of information transfer suggested by the dogma.			
General		Special	
DNA → DNA:	DNA replication	RNA → DNA:	Reverse transcription
DNA → RNA:	Transcription	RNA → RNA:	RNA replication
RNA → protein:	Translation	*DNA → protein:	Direct translation

*DNA → protein process has only been demonstrated in a cell-free system, but not in intact cells.

DNA replication is one of the fundamental steps in the central dogma due to its implications for the progeny of any cell (figure 1.2, page 6). The first step in DNA replication is the separation of the two DNA strands (parent strands) that make up the helix. A helicase unwinds the superhelix (a coil of DNA helices) and the double-stranded DNA helix itself to create a replication fork. This step involves the breaking of hydrogen bonds between bases of the two antiparallel strands. The splitting happens in A•T rich regions, which contains two hydrogen bonds per base pair (there are three hydrogen bonds between C•G pairs). At this point, single-stranded binding (SSB) proteins bind to the open double-stranded DNA to prevent its re-association.

When the two parent strands of DNA are separated, one strand is oriented in the 5'→3' direction while the other strand is oriented in the 3'→5' direction. However, the enzyme that carries out the

replication, DNA polymerase, only functions in the $5' \rightarrow 3'$ direction. Thus, the daughter strands must be synthesized through different methods: one adding nucleotides one by one in the direction of the replication fork (henceforth named *leading strand*), the other able to add nucleotides only in small fragments (*lagging strand*).

Since DNA replication moves along the parent strand in the $5' \rightarrow 3'$ direction, replication can occur very easily on the **leading strand**. Triggered by RNA primase, which adds the first nucleotides to the nascent chain (primers), the DNA polymerase simply sits near the replication fork, moving as the fork does, adding nucleotides one after the other, preserving the proper anti-parallel orientation. This sort of replication is called continuous replication.

On the **lagging strand** the enzyme must move away from the fork because DNA polymerase is only able to add nucleotides in the $5' \rightarrow 3'$ direction. Thus, each time the helicase unwinds the parent strands, DNA polymerase must operate in the opposite direction. For this reason, the lagging strand replicates in small segments, called Okazaki fragments. These fragments are stretches of 100 to 200 nucleotides in humans that are synthesized in the $5' \rightarrow 3'$ direction away from the replication fork, and initiated in the primer position generated by the RNA primase. Yet while each individual segment is replicated away from the replication fork, each subsequent Okazaki fragment is replicated more closely to the receding replication fork than the fragment before. These fragments are then stitched together by DNA ligase, creating a continuous strand. This type of replication is called discontinuous replication.

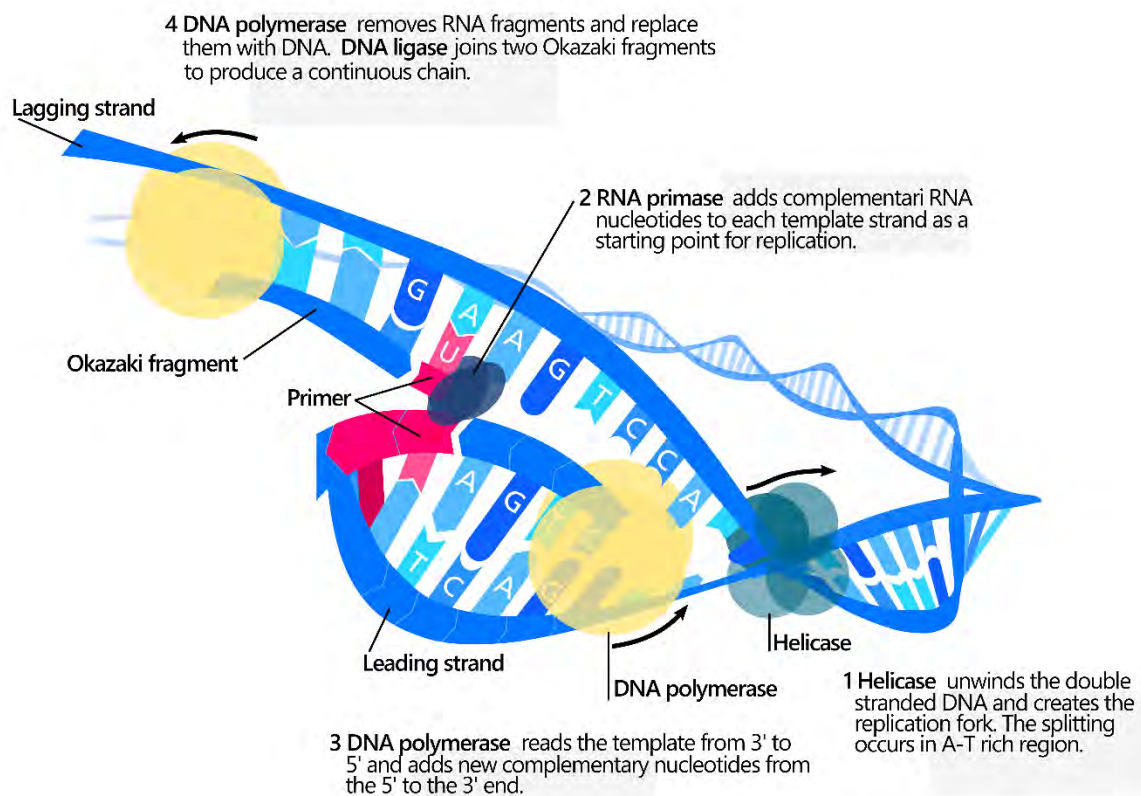


Figure 1.2. The helicase unwinds the double-stranded DNA for replication, making a forked structure. RNA primase generates short strands of RNA that bind to the single-stranded DNA to initiate DNA synthesis by the DNA polymerase. This enzyme can work only in the $5' \rightarrow 3'$ direction, so it replicates the leading strand continuously. Lagging-strand replication is discontinuous, with short Okazaki fragments being formed and later linked together by DNA ligase.

A DNA **transcription** unit encoding for a protein may contain both a coding sequence, which will be translated into the protein, and regulatory sequences, which direct and regulate the synthesis of that protein. The regulatory sequence '**upstream**' (before from) the coding sequence is called the five prime untranslated region (**5'UTR**); the sequence '**downstream**' (after from) the coding sequence is called the three prime untranslated region (**3'UTR**).

As opposed to DNA replication, transcription results in a complementary RNA chain that substitutes the nucleotide uracil (U) in all instances where thymine (T) would have occurred in a DNA. In transcription, only one of the two DNA strands serve as a template for transcription. The non-template DNA strand is called the **coding strand**, because its sequence is the same as the newly created RNA transcript (except for the substitution of uracil for thymine). The DNA is read by RNA polymerase in the 3'→5' direction. Thus, the complementary RNA is synthesized in the opposite direction (5'→3'). This directionality occurs because RNA polymerase can only add nucleotides to the 3' end of the growing RNA chain. This removes the need for an RNA primer to initiate RNA synthesis, as is the case in DNA replication.

Transcription is divided into pre-initiation, initiation, promoter clearance, elongation and termination (figure 1.3, page 8, refer to steps 1 to 3). During the pre-initiation, RNA polymerase requires the presence of a core promoter sequence in the DNA which comprise a sequence found at -30, -75, and -90 base pairs upstream from the transcription start site (TSS). Transcription factors are proteins that bind to these promoter sequences and facilitate the binding of RNA polymerase.

RNA polymerase does not directly recognize core promoter sequences but instead the transcription factors mediate the binding of RNA polymerase (initiation phase). The completed assembly of transcription factors and RNA polymerase bind to the promoter, forming a transcription initiation complex. After initiation, the RNA polymerase must clear the promoter. During this process there is a tendency to release the RNA transcript and produce truncated transcripts, which results in an abortive initiation. After several rounds of abortive initiation, promoter clearance coincides with a TFIIF's phosphorylation of serine 5 on the carboxy terminal domain of RNAP II, leading to the recruitment of capping enzyme (CE).

As transcription proceeds, RNA polymerase traverses the template strand and uses base pairing complementarity with the DNA template to create an RNA copy. This produces an exact copy of the coding strand, except for the replacement of U for T, and the new nucleotides are composed of a ribose sugar with an addition hydroxyl group at 2' position.

Finally, the termination involves cleavage of the new transcript followed by template-independent addition of adenines [poly(A)] at its new 3' end, in a process called polyadenylation. Cleavage and polyadenylation is directed by a poly(A) signal in the RNA. The core poly(A) signal consists of an almost invariant AAUAAA hexamer that lies 20-50 nucleotides upstream of a more variable element rich in U or GU residues. Cleavage can be mediated by a minimal protein complex that can be separated into five factors. In particular, two of these factors are the cleavage and polyadenylation specificity factor (CPSF) which binds the AAUAAA motif and the cleavage stimulation factor (CstF) which binds the downstream U-rich element.

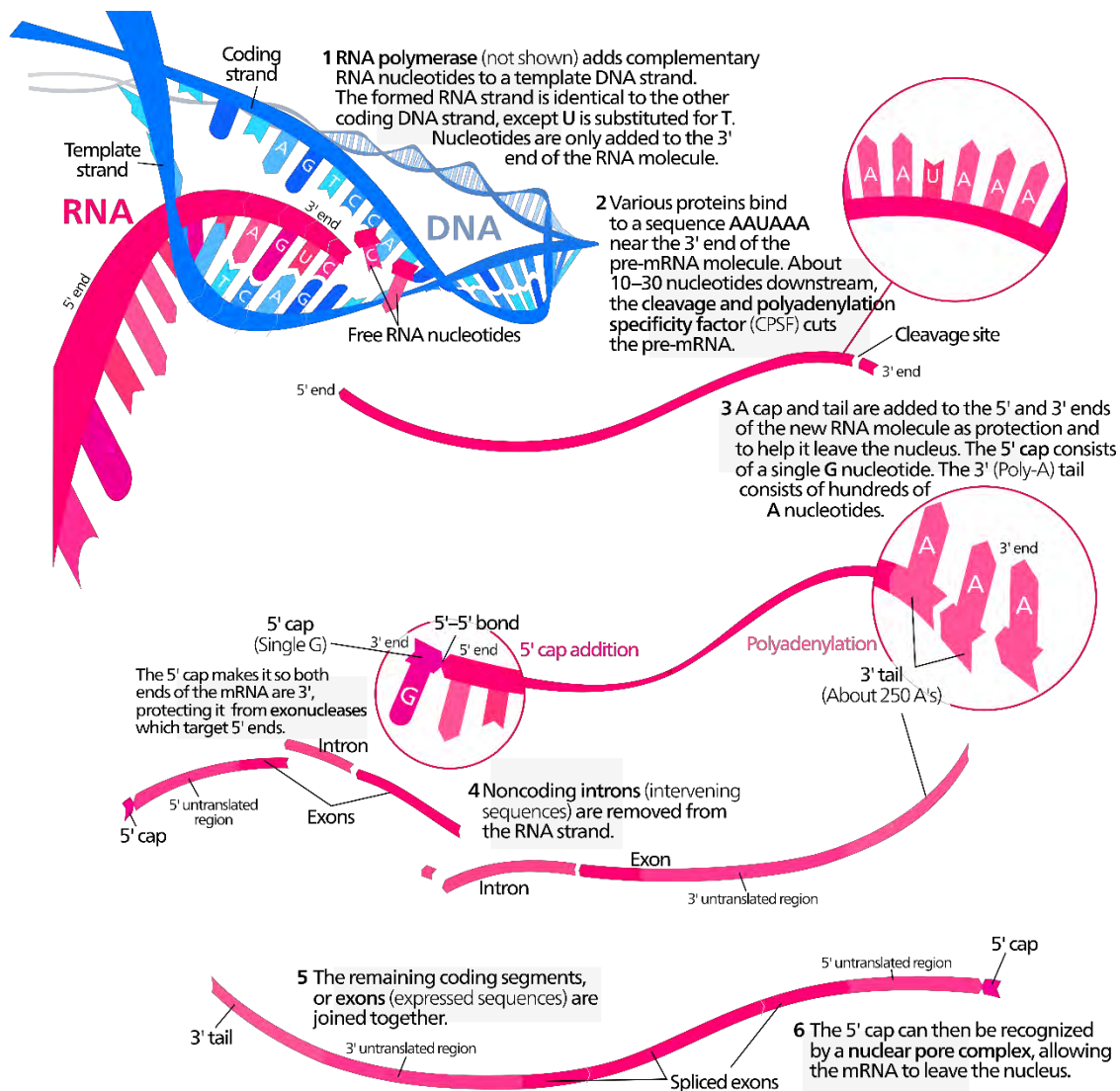


Figure 1.3. Schematic process of RNA transcription. RNA polymerase uses a template DNA strand to form a pre-mRNA transcript. Several post-transcriptional modifications occur, which include polyadenylation, capping of the transcript and RNA splicing. Image by Kelvin Song, distributed under a CC-BY 2.0 license.

After the transcription process, the resulting RNA chain is called pre-messenger RNA (**pre-mRNA**). At this point, the pre-mRNA molecule undergoes two general modifications: 5' capping, and RNA splicing, which occur in the cell nucleus before the RNA is translated. These processes are called **post-transcriptional modifications** (figure 1.3, steps 3 to 6) and make a messenger RNA (**mRNA**) as a result.

Capping of the pre-mRNA involves the addition of 7-methylguanosine (m7G) to the 5' end. To achieve this, the terminal 5' phosphate requires removal by a phosphatase enzyme. The enzyme guanosyl transferase then catalyses the reaction, which produces the diphosphate 5' end. The diphosphate 5' end then attacks the gamma phosphorus atom of a GTP molecule in order to add the guanine residue in a 5'-5' triphosphate bond. The enzyme (guanine-N7)-methyltransferase (cap MTase) transfers a methyl group from S-adenosyl methionine to the guanine ring. This type of cap, with just the (m7G) in position is called a cap 0 structure. The ribose of the adjacent nucleotide may also be methylated to give a cap 1. Methylation of nucleotides downstream of the RNA molecule

produce cap 2, cap 3 structures and so on. In these cases the methyl groups are added to the 2' OH groups of the ribose sugar. The cap protects the 5' end of the primary RNA transcript from attack by ribonucleases that have specificity to the 3'-5' phosphodiester bonds.

RNA splicing is the process by which introns - regions of RNA that do not code for protein - are removed from the pre-mRNA and the remaining exons connected to re-form a single continuous molecule. Although most RNA splicing occurs after the complete synthesis and end-capping of the pre-mRNA, transcripts with many exons can be spliced co-transcriptionally. The splicing reaction is catalyzed by a large protein complex called the spliceosome assembled from proteins and small nuclear RNA molecules that recognize splice sites in the pre-mRNA sequence.

Many pre-mRNAs, including those encoding antibodies, can be spliced in multiple ways to produce different mature mRNAs that encode different protein sequences. This process is known as **alternative splicing**, and allows production of a large variety of proteins from a limited amount of DNA (see figure 1.4). This mechanism is regulated by alternative splicing proteins which inhibit the spliceosome formation at certain regions of the pre-mRNA.

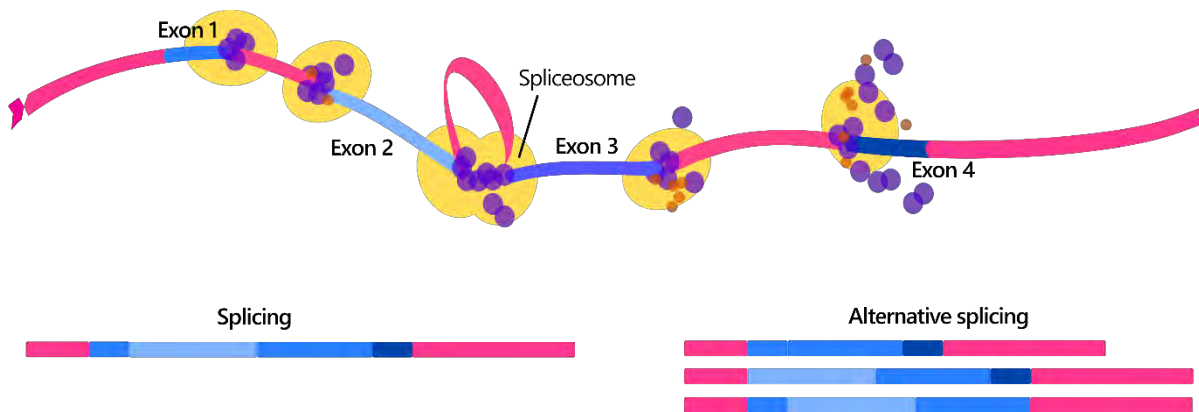


Figure 1.4. Simplified representation of the alternative splicing mechanism. When one or several spliceosomes are inhibited by specific splicing proteins which inhibit the spliceosome formation.

Finally, **translation** is the process by which a protein is synthesized from the information contained in a mature mRNA. During translation, an mRNA sequence is read using the genetic code, which is a set of rules that defines how an mRNA sequence is to be translated into the 20-letter code of amino acids. Each amino acid is matched to a three nucleotide subsequence of the mRNA (**codon**). All this process occurs in a structural complex called the ribosome, which is the factory of protein synthesis, and is composed of several ribosomal RNA molecules and proteins.

First, a small ribosomal subunit binds to the start of the mRNA sequence. Then, a transfer RNA (**tRNA**) molecule carrying the amino acid methionine binds to what is called the start codon of the mRNA sequence. The start codon contains the sequence AUG in all mRNA molecules and codes for methionine. Next, the large ribosomal subunit binds to form the complete initiation complex.

The ribosome continues to translate each codon one by one. Each corresponding amino acid is added to the growing chain and linked via a peptide bond. Elongation continues until all of the codons have been read.

Termination occurs when the ribosome reaches a 'stop' codon (UAA, UAG, and UGA). Since there are no tRNA molecules that can recognize these codons, the ribosome recognizes that translation is complete. The new protein is then released, and the translation complex comes apart.

1.1.3. FUNCTIONAL STRUCTURE OF A GENE

All genes have regulatory regions in addition to regions that explicitly code for a protein or RNA product. A regulatory region shared by almost all genes - or **promoter** - provides a position that is recognized by the transcription machinery when a gene is about to be transcribed and expressed. A gene can have more than one promoter, resulting in RNAs that differ in how far they extend in the 5' end. Although promoter regions have a consensus sequence consisting of the most common nucleotide at each position, some genes have "strong" promoters that bind the transcription machinery well, and others have "weak" promoters that bind poorly. These weak promoters usually permit a lower rate of transcription than the strong promoters, because the transcription machinery binds to them and initiates transcription less frequently. Other possible regulatory regions include enhancers, which can compensate for a weak promoter.⁴

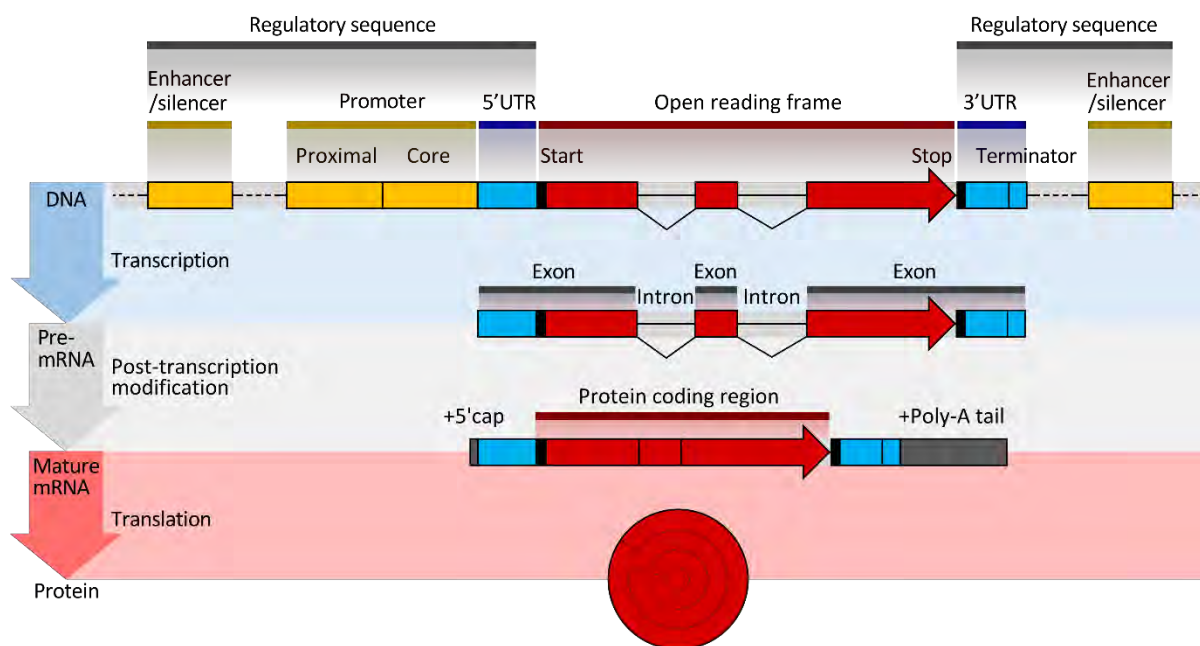


Figure 1.5. Regulatory sequence controls when and where expression occurs for the protein coding region (red). Promoter and enhancer regions (yellow) regulate the transcription of the gene into a pre-mRNA which is modified to add a 5' cap and poly(A) tail (grey) and remove introns. The mRNA 5' and 3' untranslated regions (blue) regulate translation into the final protein product. Image by Thomas Shafee, distributed under a CC-BY 2.0 license.

1.2. STRUCTURAL BIOLOGY OF RNA

The major credit I think Jim and I deserve [...] is for selecting the right problem and sticking to it. It's true that by blundering about we stumbled on gold, but the fact remains that we were looking for gold. Both of us had decided, quite independently of each other, that the central problem in molecular biology was the chemical structure of the gene. [...] We could not see what the answer was, but we considered it so important that we were determined to think about it long and hard, from any relevant point of view.

- Francis Crick

What Mad Pursuit: A Personal View of Scientific Discovery. 1988.

[Molecular biology] is concerned particularly with the forms of biological molecules and with the evolution, exploitation and ramification of these forms in the ascent to higher and higher levels of organization. Molecular biology is predominantly three-dimensional and structural - which does not mean, however, that it is merely a refinement of morphology. It must at the same time inquire into genesis and function.

- William Thomas Astbury

Perspectives on the Emergence of Scientific Disciplines. 1974.

Nowadays it is clear that in reality, the established conception of the central dogma of molecular biology is not entirely accurate insofar as it puts emphasis on proteins as the mediator of biological function. About 80% of the human genome is transcribed even though only 1% codes for proteins.⁵ Although it would be possible that this may correspond to simple transcriptional noise, it seems to be an unlikely waste of cellular energy resources, and considering the major role played by RNA in regulation of gene expression, it may as well have a role. This chapter builds upon nucleic acids introduced in the prior chapter to include a description of RNA structure and folding. Table 1.2 (page 12) contains a list of the most relevant (but not all) known RNA types and its function.

Table 1.2. List of RNA types and function.			
Type	Abbreviation	Function	Distribution
RNAs involved in protein synthesis			
Messenger RNA	mRNA	Codes for protein	All organisms
Ribosomal RNA	rRNA	Translation	All organisms
Signal recognition particle RNA	SRP RNA	Membrane integration	All organisms
Transfer RNA	tRNA	Translation	All organisms
RNAs involved in post-transcriptional modifications			
Small nuclear RNA	snRNA	Splicing and other functions	Eukaryotes and archaea
Small nucleolar RNA	snoRNA	Nucleotide modification of RNAs	Eukaryotes and archaea
SmY RNA	SmY	mRNA trans-splicing	Nematodes
Small Cajal body-specific RNA	scaRNA	Type of snoRNA; Nucleotide modification of RNAs	
Guide RNA	gRNA	mRNA nucleotide modification	Kinetoplastid mitochondria
Ribonuclease P	RNase P	tRNA maturation	All organisms
Ribonuclease MRP	RNase MRP	rRNA maturation, DNA replication	Eukaryotes
Y RNA		RNA processing, DNA replication	Animals
Telomerase RNA Component	TERC	Telomere synthesis	Most eukaryotes
Regulatory RNAs			
Antisense RNA	aRNA, asRNA	Transcriptional attenuation / mRNA degradation / mRNA stabilisation / Translation block	All organisms
Cis-natural antisense transcript	cis-NAT	Gene regulation	
CRISPR RNA	crRNA	Resistance to parasites, probably by targeting their DNA	Bacteria and archaea
Long noncoding RNA	lncRNA	Regulation of gene transcription, epigenetic regulation	Eukaryotes
MicroRNA	miRNA	Gene regulation	Most eukaryotes
Piwi-interacting RNA	piRNA	Transposon defense, maybe other functions	Most animals
Small interfering RNA	siRNA	Gene regulation	Most eukaryotes
Trans-acting siRNA	tasiRNA	Gene regulation	Land plants
Repeat associated siRNA	rasiRNA	Type of piRNA; transposon defense	Drosophila
7SK RNA	7SK	negatively regulating CDK9/cyclin T complex	
Parasitic RNAs			
Retrotransposon		Self-propagating	Eukaryotes and some bacteria
Viral genome		Information carrier	Double-stranded RNA viruses, positive-sense RNA viruses, negative-sense RNA viruses, many satellite viruses and reverse transcribing viruses
Viroid		Self-propagating	Infected plants
Satellite RNA		Self-propagating	Infected cells

1.2.1. CHEMICAL STRUCTURE AND STABILITY OF NUCLEIC ACIDS

Chemically speaking, RNA is very similar to DNA and their structure can be defined at four different levels: primary, secondary, tertiary and quaternary. The **primary** structure consist of a linear succession of nucleotides linked together by phosphodiester bonds. Figure 1.1 contained the primary structure representation of a DNA sequence (TGC) making hydrogen bonds with its complementary sequence (ACG). In a similar way, figure 1.6A (page 14) shows the primary structure of the four RNA nucleotides (C, G, A, U). The basic structure of the bases is the same

for both nucleic acids; on the one hand, A and G are considered **purines (R)**, which consist of a six membered and a five membered ring containing nitrogen. On the other hand, C and T (or U in the case of RNA) are classified as **pyrimidines (Y)**. The main difference between DNA and RNA nucleotides lies in the hydroxyl in the 2' position of the ribose. Notice that the phosphodiester bond is made from the 5' position from the phosphate group of the first nucleotide to the 3' hydroxyl of second nucleotide (5'→3' direction, sense strand). The sum of all the phosphate and ribose units constitutes the **backbone** of the structure.

The next level of structural characterization is the **secondary** structure which is defined as the set of interactions between bases (figure 1.6B, page 14). The most general case is the DNA double helix where the two strands of DNA are held together by hydrogen bonding interactions. The secondary structure is responsible for the shape of the nucleic acids which is known to be a key factor for the activity of some RNA single-stranded polynucleotides.

Tertiary structure refers to the location of the atoms into the three-dimensional space. It takes into consideration steric and geometrical constraints when large scale folding occurs into a specific three-dimensional shape. For instance, the tertiary structure of the DNA double helix includes A-DNA, B-DNA and Z-DNA configurations (figure 1.7C, table 1.3). Some single stranded RNAs can fold and make double stranded fragments, which are usually in A-RNA (structurally analogous to A-DNA) or A'-RNA forms.^{6,7} The main differences between all those types are the handedness (right or left), the length of the helix turn, the number of base pairs per turn and the size difference between major and minor grooves.

Table 1.3. Comparison of the structural properties of A, B and Z DNAs as derived from single-crystal X-ray analysis.⁸

	A	B	Z
Overall proportions	Short and broad	Longer and thinner	Elongated and slim
Rise per base pair	2.3 Å	3.32 Å	3.8 Å
Helix-packing diameter	25.5 Å	23.7 Å	18.4 Å
Helix rotation sense	Right-handed	Right-handed	Left-handed
Base pairs per helix repeat	1	1	2
Base pairs per turn of helix	~11	~10	12
Rotation per base pair	33.6°	35.9°	-60° per 2 bp
Pitch per turn of helix	24.6 Å	33.2 Å	45.6 Å
Tilt of base normal to helix axis	+19°	-1.2°	-9°
Base pair mean propeller twist	+18°	+16°	~0°
Helix axis location	Major groove	Through base pairs	Minor groove
Major groove proportions	Extremely narrow but very deep	Wide and of intermediate depth	Flattened out on helix surface
Minor groove proportions	Very broad but shallow	Narrow and of intermediate depth	Extremely narrow but very deep
Glycosyl bond conformation	anti	anti	anti at C, syn at G

Finally, the **quaternary** structure refers to a higher level of organization of nucleic acids and its interaction with other macromolecules. The most commonly seen form of quaternary structure is the formation of chromatin, which leads to its interaction with the small protein histones. Another example would be the interaction between separate RNA units in the ribosome or the spliceosome.

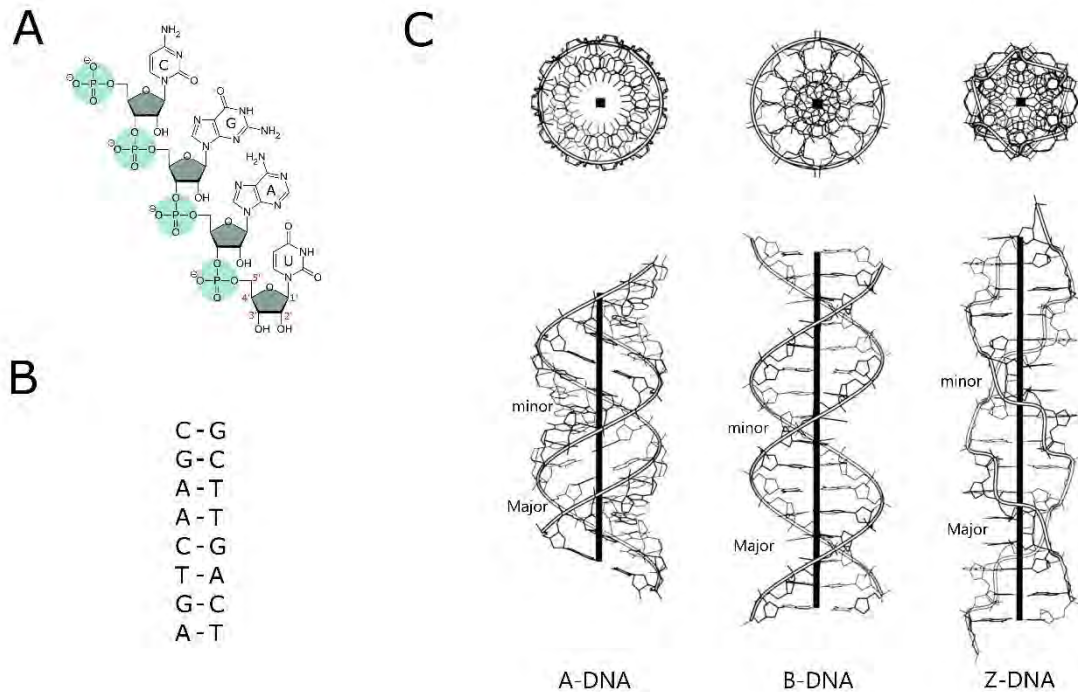


Figure 1.6. (A) RNA fragment containing the chemical structure of a CGAU fragment. The primary structure corresponds to the sequence alone. (B) Secondary structure of a double helix DNA. All the nucleotides are forming a base pair with its complementary nucleotide. (C) Most common tertiary structures of DNA (A, B and Z types) with its corresponding minor and major grooves.

Before more detailed structural features of the polynucleotides are described, a few remarks about the forces that govern base-base interactions are required. All these structural levels of organization are achieved by two main interactions: those in the plane of the bases (horizontal) which are the **hydrogen bonds**; those perpendicular to the base planes or **base stacking**.

Hydrogen bonds have been previously mentioned and correspond to the most common nucleic acid stabilization force. They are electrostatic in character and, in general, are formed if a hydrogen atom connects two atoms of higher electronegativity (nitrogen, oxygen or fluorine). Thus, the strength of this bond depends on the partial charges located on the component atoms of the bond. The name *bond* is something of a misnomer, as it is a particular strong dipole-dipole attraction and should not be confused with a covalent bond. The hydrogen bond has a wide energy range (0.2–40 kcal/mol) and is stronger than van der Waals interaction forces, but weaker than covalent or ionic bonds. However, it also has some features of covalent bonding because it is directional, produces interatomic distances shorter than the van der Waals radii and usually involves a limited number of interaction partners (it can be interpreted as a type of valence). A hydrogen atom attached to a relatively electronegative atom will play the role of the hydrogen bond **donor**. The other electronegative atom will be the hydrogen bond **acceptor**.

In aqueous solution, the bases in a single stranded oligonucleotide are stacked such that the base planes are separated by their van der Waals distance (3.4 Å) and parallel one to another. This effect of **base stacking** is the least understood but most important stabilizing force. The associated entropy (ΔH) is highly unfavorable while the enthalpies ($T\Delta S$) are strongly favorable. However,

the overall base stacking interaction is favorable since the solvent enthalpy and entropy are both strongly favorable. This interaction can be explained by two separate forces:

- (i) **Hydrophobic interactions.** When a hydrophobic base is dissolved into water, the water molecules cluster around it in an ordered fashion due to the effect of hydrogen bonding. The hydrophobic unit cannot form hydrogen bonds with water (or can only perform a very limited amount of them) hence the water molecules adopt an ordered structure around the non-polar molecule to maximize the hydrogen bonding. The 'reordering effect' results in an overall entropy gain. In nucleic acids, the effect of burying the non-polar bases results in a dramatic increase of both enthalpy and entropy.
- (ii) **London dispersion forces.** The bases stack one to another at their van der Waals distance which is the distance where the two molecules have an attraction effect. However, at too close distances the electrons of the two approaching molecules overlap, causing a repulsion effect (figure 1.7). At that instant, the electronic charge distribution within the atomic groups is asymmetric due to the electronic fluctuation. Therefore, the induced dipole in that group of atoms is able to polarize the electronic system of the neighboring molecules, hence inducing parallel dipoles that attract each other. These forces are additive and extremely distance dependent, falling off with the sixth power of distance. Moreover, stacking requires aromaticity of the bases due to its polarizable p electron cloud.

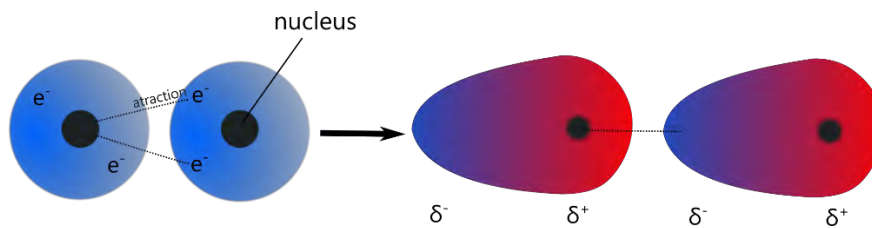


Figure 1.7. Representation of the dipole-dipole interaction caused by London dispersion forces.

1.2.2. GENERAL NOMENCLATURE OF NUCLEOTIDES

Each strand of the nucleic acid chain has a terminal C5'-OH group on one end and a terminal C3'-OH group on the other. Thus, if the structure is a double helix, the two intertwined strands run in anti-parallel directions. To number the $2n$ bases of a fully paired structure (hence it has n base pairs), **strand I** is specified as the sense strand and **strand II** as the antiparallel strand. The residue numbering of strand II is specified so that they coincide with the base pairs of strand I: the first base pair involves base 1 in strand I with base $2n$ of strand II; the second base pair involves the pairing of strand I's base 2 and strand's II base $2n-1$; and so on.⁶

Usually, nucleotides are abbreviated by a pair of letters: a lower-case Roman letter denotes the sugar type ('r' or 'd', for ribose and deoxyribose respectively) and an upper-case Roman letter represents the base type (C, G, A, T, U). Nucleotides can be also abbreviated by adding a lower-case 'p' (for phosphate) to the nucleoside symbol. Thus, the sequence rGpCprA is trimer in a RNA where G is at the 5' end and A at the 3' end. For brevity, the lower-case letters will be omitted along the text for RNA, and a 'd' will be used for denoting DNA.

Standard atomic labeling schemes have been recommended for nucleic acids as follows (see figure 1.9, page 17):⁶

- Base atoms are numbered systematically (as shown in figure 1.8A). On nitrogen of the base (N1 for **Y**, pyrimidines, N9 for **R**, purines) is always connected to the C1' sugar of the sugar by a glycosyl bond (N1/9-C1').
- Sugar atoms are distinguished from the base atoms by a prime suffix, and within the sugar numbering sequence is counted clockwise from the ring oxygen in the direction of the carbon attached to the base nitrogen (see figure 1.8B).
- In the polynucleotide backbone, the counting direction for torsion angles (α , β , γ , δ , ϵ and ζ) is specified by the sequence: P→O ζ '→C ζ '→C4'→C3'→O3'→P (see figure 1.8B, page 17). The full list of torsion angle definitions is attached in table 1.4.⁹

The 5-membered sugar ring is generally nonplanar in nucleic acids. One or two atoms may *pucker* out of the plane defined by the remaining ring atoms. The **sugar pucker** type is described by the atoms that deviate from that ring plane. Atoms displaced on the same side of C5' are designated as *endo*, and atoms displaced on the opposite site are called *exo*.⁶ A convenient description of the sugar conformations is achieved by using the pseudorotation path (see figure 1.8C). Following the Altona and Sundaralingam description,¹⁰ the five endocyclic torsion angles (τ) are restricted to the values:

$$\tau_j = \tau_{max} \cos \left[P + \frac{4\pi}{5} (j - 2) \right], \quad j = 0,1,2,3,4 \quad (1.1)$$

where τ_{max} and P correspond to the amplitude and phase shift respectively of a sinusoidal motion between conformations of equal energy with respect to the mean plane. The wave-like pseudorotation path described by eq. (1.1) is often divided in N, S, E and W sugar-pucker regions, and positive (+) and negative (-) torsions (refer to figure 1.8C). The two major types of pucker modes are C3'-*endo* ($0^\circ < P < 36^\circ$, usually found in RNA) and C2'-*endo* ($144^\circ < P < 188^\circ$, common in DNA).

A final torsion is found relative to the sugar moiety; the base can assume two major orientations about the glycosyl C1'-N1/9 bond: *syn* (0°) and *anti* (180°) conformations. Roughly speaking, four major conformations can be found which correspond to the combinations of C3'-*endo* and C2'-*endo* sugar pucker with *syn* and *anti* values for the glycosyl rotation (χ). This combination of $\{P, \chi\}$ pairs vary for the different nucleotides depending on the chemical structure of the sugar, the size of the base, and the nature of the nucleoside substituents (chemical derivatives, for instance).

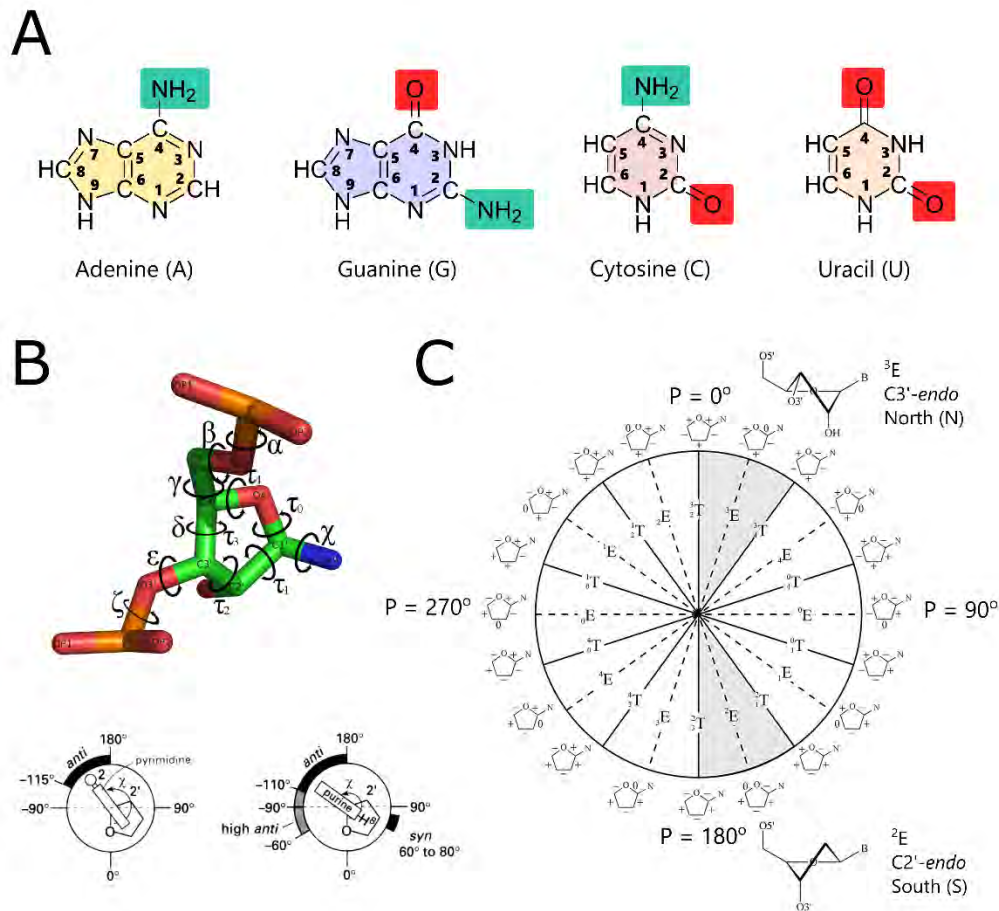


Figure 1.8. (A) Bases atom structure and numbering. Amino and carbonyl groups are highlighted. (B) RNA backbone representation containing atom names and torsions. The most common χ torsions for pyrimidines (Y) and purines (R) are represented in the radial plots below. (C) Pseudorotation wheel and preferred conformation of ribose (top) and deoxyribose (bottom). Each point on the circle represents a specific value of the pseudorotation angle P . On the periphery of the cycle, riboses with signs of the endocyclic torsion angles (τ_0 to τ_4) are indicated (+) positive, (-) negative and (0) torsion angle at 0° .

Nucleic acids present a large number of backbone conformational parameters in striking contrast to the two torsion angles (φ, ψ) found in protein structures. However, a simplified representation of the backbone conformation using the pseudo-torsion angles (η, θ) has been proposed (see definitions in table 1.4). Pyle et al.¹¹ developed a modified version of the pseudo-torsions (η', θ') which led to a better representation of the crystallographic density.

Table 1.4. Nucleic acid torsion angle definitions. Torsion may involve atoms from consecutive nucleotides.

Angle	Sequence	Angle	Sequence
α	O3'-P-O5'-C5'	τ_0	C4'-O4'-C1'-C2'
β	P-O5'-C5'-C4'	τ_1	O4'-C1'-C2'-C3'
γ	O5'-C5'-C4'-C3'	τ_2	C1'-C2'-C3'-C4'
δ	C5'-C4'-C3'-O3'	τ_3	C2'-C3'-C4'-O4'
ϵ	C4'-C3'-O3'-P	τ_4	C3'-C4'-O4'-C1'
ζ	C3'-O3'-P-O5'	χ	O4'-C1'-N1-C2 (Y) O4'-C1'-N9-C4 (R)
Pseudo-torsion angle definitions			
η	C4'-P-C4'-P	η'	C1'-P-C1'-P
θ	P-C4'-P-C4'	θ'	P-C1'-P-C1'

1.2.3. GEOMETRIC CLASSIFICATION OF RNA BASE PAIRS

Watson and Crick dictated the intermolecular interactions between canonical base pairs [C•G, and A•T(U)] which permitted the formation of short double stranded DNA helices. However, the growing literature on RNA structural biology is hampered by the lack of a systematic nomenclature for ‘non-standard’ base pairing interactions. For this reason, ambiguous and confusing terms are often used to describe base pairing. In this work the classification proposed by Leontis and Westhof will be used.¹² This classification is based on the observation that, while only about 60% of bases in structured RNAs participate in canonical Watson-Crick base pairs (WC), the great majority participate in another type of interactions such as edge-to-edge.

According to their classification there are 12 basic families of base pairs. RNA purine (**R**) and pyrimidine (**Y**) bases present three edges that can make hydrogen bond interactions, named **Watson-Crick**, **Hoogsteen edge** (for **R**) or **C-H edge** (for **Y**) and **Sugar edge**. For clarity purposes and following the same criteria as the original authors, the C-H edge will also be named Hoogsteen. A given edge can interact in a plane with any one of the other edge types of a second base, and can do so either in a *cis* or *trans* orientations of the glycosidic bonds. Thus, a total 12 distinct edge-to-edge interactions are possible, each of which is designated by stating the interacting edges of each of the two bases and the relative glycosidic bond orientation (figure 1.9, table 1.5).

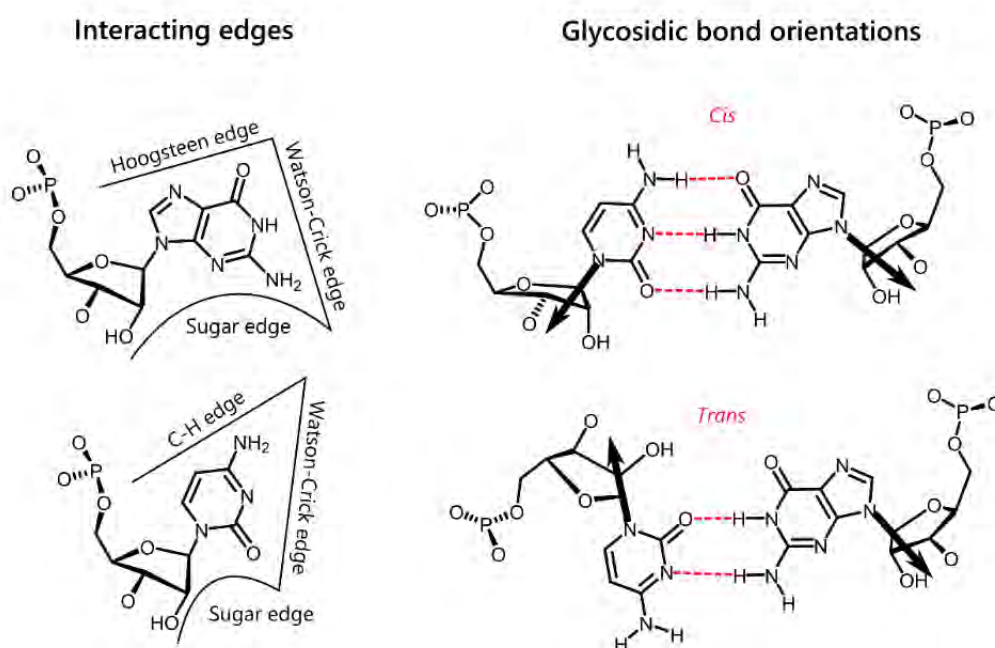


Figure 1.9. Definition of the interacting edges (Watson-Crick, Hoogsteen or CH-edge and sugar-edge) and glycosidic bond orientations.

Table 1.5. Leontis and Westhof geometric classification of RNA base pairs.¹²

Type	Nomenclature	Annotation
Cis Watson-Crick/Watson-Crick	G•A <i>cis</i> W.C./W.C.	
	C•C <i>cis</i> W.C./W.C. (wobble)	
	G•U <i>cis</i> W.C./W.C. (wobble)	
	U•C <i>cis</i> W.C./W.C.	
	U•U <i>cis</i> W.C./W.C. (wobble)	
Trans Watson-Crick/Watson-Crick	A•U <i>trans</i> W.C./W.C.	
	A•A <i>trans</i> W.C./W.C.	
	G•G <i>trans</i> W.C./W.C.	
	G•C <i>trans</i> W.C./W.C.	
	A•C <i>trans</i> W.C./W.C.	
	G•U <i>trans</i> W.C./W.C.	
	U•C <i>trans</i> W.C./W.C.	
	C•C <i>trans</i> W.C./W.C.	
U•U <i>trans</i> W.C./W.C.		
Cis Watson-Crick/Hoogsteen	G•G <i>cis</i> W.C./Hoogsteen	
	U•A <i>cis</i> W.C./Hoogsteen	
	G•A <i>cis</i> W.C./Hoogsteen	
	A+•G <i>cis</i> W.C./Hoogsteen	
Trans Watson-Crick/Hoogsteen	A•A <i>trans</i> W.C./Hoogsteen	
	G•G <i>trans</i> W.C./Hoogsteen	
	U•A <i>trans</i> W.C./Hoogsteen	
	C•A <i>trans</i> W.C./Hoogsteen	
Cis Watson-Crick/Sugar-edge	A•G <i>cis</i> W.C./Sugar-edge	
	A•U <i>cis</i> W.C./Sugar-edge	
Trans Watson-Crick/Sugar-edge	A•G <i>trans</i> W.C./Sugar-edge	
	C•G <i>trans</i> W.C./Sugar-edge	
Cis Hoogsteen/Hoogsteen	-	
Trans Hoogsteen/Hoogsteen	A•A <i>trans</i> Hoogsteen/Hoogsteen	
Cis Hoogsteen/Sugar-edge	-	
Trans Hoogsteen/Sugar-edge	A•G <i>trans</i> Hoogsteen/Sugar-edge	
	A•A <i>trans</i> Hoogsteen/Sugar-edge	
	C•U <i>trans</i> Hoogsteen/Sugar-edge	
Cis Sugar-edge/Sugar-edge	-	
Trans Sugar-edge/Sugar-edge	G•G <i>trans</i> Sugar-edge/Sugar-edge	

1.2.4. MOVEMENT OF BASES INTO THE HELICAL SPACE

The architecture of nucleic acids double helix can be described in terms of helical parameters, which are derived from spatial location of bases without taking into account the phosphate backbone. Depending on the translational or rotational relative movement of the bases around the {x,y,z} axes, there exist six possible definitions of **base pair parameters** (*shear, stretch, stagger, buckle, propeller, opening*; figure 1.10). Moreover, if a base pair is taken as a rigid block, six parameters are necessary to rigorously describe the position and orientation of one base pair relative to another. There are two sets of local parameters used for this conformational analysis: **step parameters** (*shift, slide, rise, tilt, roll, twist*; figure 1.10) describe the stacking geometry conformation of a double helix at every base pair step; and the **helical parameters** (*x-displacement, y-displacement, inclination, tip*; figure 1.10) which demonstrate the position and orientation the each base pair relative to the helical axis. Usually, some of these parameters can deviate significantly from each macromolecule type, and therefore they can be used to perform a general structural classification.¹³

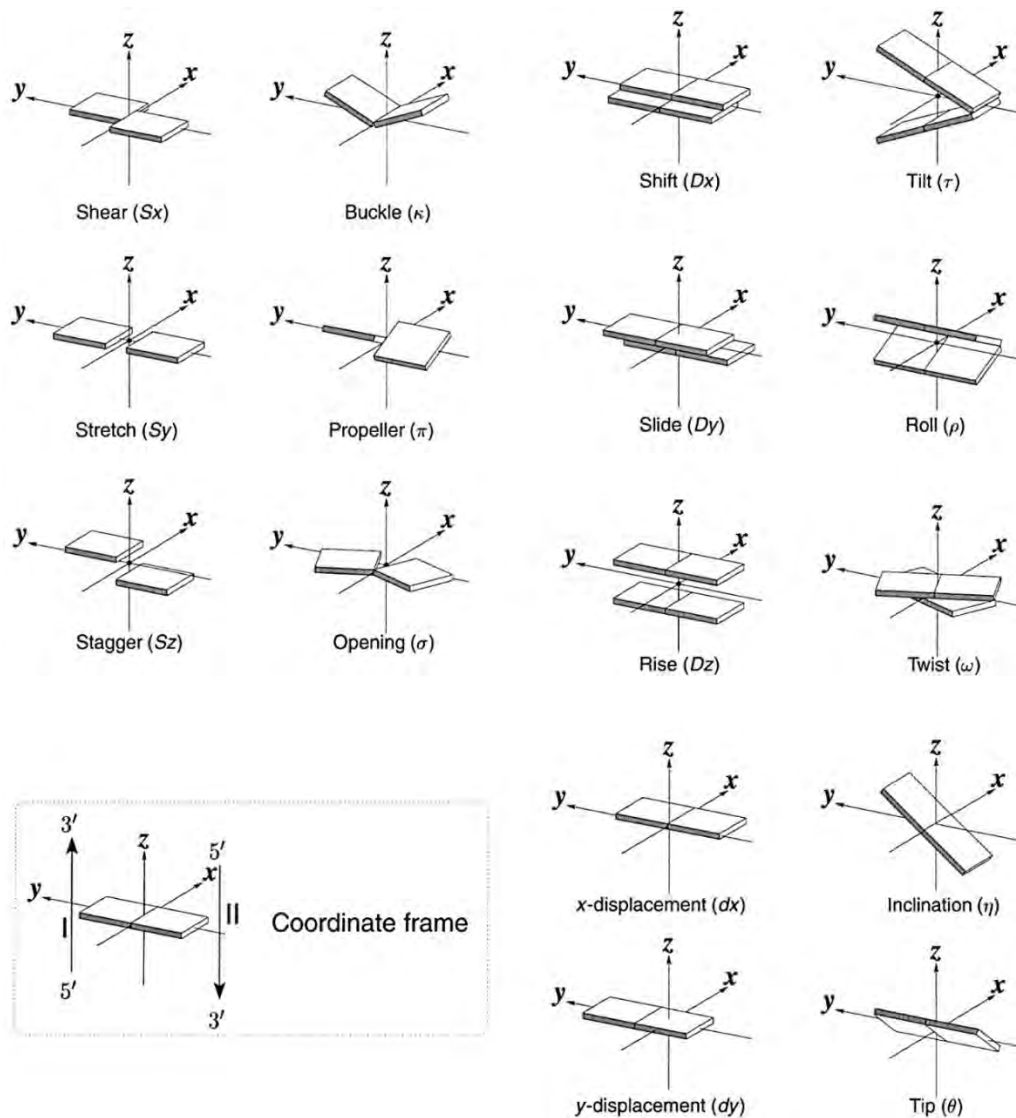


Figure 1.10. Base pair parameters (*shear, stretch, stagger, buckle, propeller, opening*), step parameters (*shift, slide, rise, tilt, roll, twist*) and helical parameters (*x-displacement, y-displacement, inclination, tip*) definitions using a block schematic representation. Image taken from Lu et al.¹³

1.2.5. SECONDARY AND TERTIARY STRUCTURAL ELEMENTS IN RNA

The tendency of DNA to form double stranded structures is well known since the work of Watson & Crick.¹⁴ However, single stranded nucleic acids sequences will generally contain many complementary regions that have the potential to form double helices when the molecules fold back onto itself. Characteristic double helical stretches that define the secondary structure are usually overserved in RNA strands. Secondary structure elements may in turn be arranged in space to form a three-dimensional tertiary structure. In energetic terms, these tertiary element are formed by weaker non-covalent bond interactions than in secondary structure; thus RNA folding can be regarded as a hierarchical process in which secondary structure forms before tertiary structure. Figure 1.12 shows a schematic representation of the most common secondary and tertiary structural elements found in RNA structures.

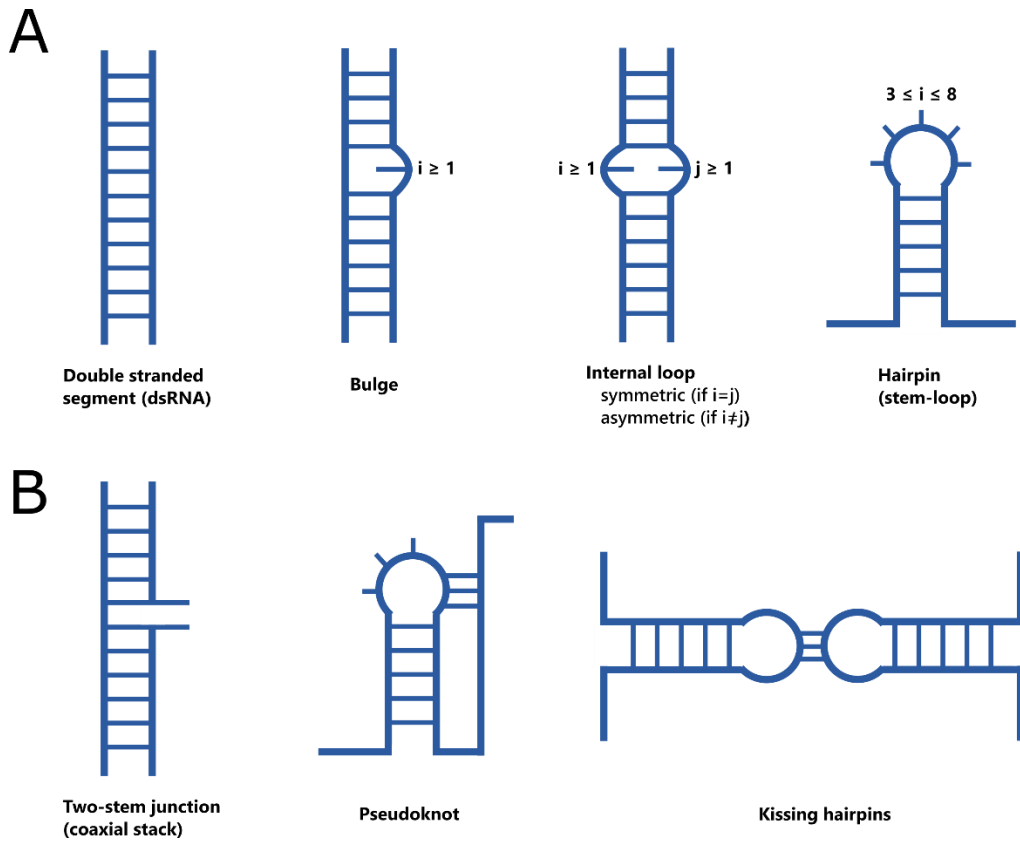


Figure 1.11. (A) RNA secondary structural elements and (B) tertiary structural element

1.3. NUCLEOTIDE REPEAT EXPANSION DISORDERS

In fact, the human genome is littered with pseudogenes, gene fragments, 'orphaned' genes, 'junk' DNA, and so many repeated copies of pointless DNA sequences that it cannot be attributed to anything that resembles intelligent design.

- **Kenneth R. Miller**

Life's Grand Design. Technology Review. **1994**, 97(2):24-32.

Inheritance can sometimes become a problem. Some diseases are hereditary hence run in families; others, are caused by the environment, such as infectious diseases. Even some diseases come from a combination of genes and the environment. Genetic disorders are diseases that are caused by a single allele of a gene and are inherited in families. These include Huntington's disease, cystic fibrosis, Duchenne muscular dystrophy or myotonic dystrophy, among others. Other diseases are influenced by genetics, but the genes a person gets from their parents only change their risk of getting a disease. Most of these diseases are inherited in a complex way, with either multiple genes involved, or coming from both genes and the environment.

Several endogenous or viral diseases are RNA related, thus the relevance of this target has been growing interest among researchers. In particular, short tandem repeats (STRs) or microsatellites – composed of 1-6 nucleotide motifs repeated <30 times - make a huge contribution to the entire sequence of the human genome. Trinucleotide repeats (TNRs) are a particular group of STRs that may have several regulatory roles in gene expression or be the source of rapid evolutionary changes, depending on their type, length and localization in the genome.¹⁵ Abnormal expansion of certain TNR tracts induce human neurological diseases known as triplet repeat expansion disorders (TREDs).^{16,17} In the past two decades several pathomechanisms have been posed serious challenges for researchers and still remain poorly recognized.¹⁸⁻²²

The discovery of disease-causing RNAs is yielding a wealth of new therapeutic targets, and the growing understanding of RNA biology and chemistry is providing new RNA-based tools for developing therapeutics.²³⁻²⁵

1.3.1. BRIEF CHARACTERISTICS OF NUCLEOTIDE REPEAT EXPANSION DISORDERS

Repetitive sequences constitute around 30% of the human genome and, in most species, alterations in lengths of repetitive DNA sequences during evolution create diversity.²⁶ However when longer than a threshold length, these simple repeats expand during most parent-child transmissions and during development of progeny. The changes in length can be substantial. For instance, repeats in coding sequences become unstable at ~29-35 units in length and the changes in tract size are typically <10 repeats per generation. By contrast, expansion repeats in noncoding regions initiate from ~55-200 units and increase by 100-10,000 units per generation.²⁶

Nucleotide repeat expansion disorders comprise a heterogeneous group of diseases that result from expansion of these specific repetitive DNA microsatellite sequences, in both coding and noncoding regions of the genes (table 1.6).^{16,20} The disease becomes more severe and presents an earlier age of onset with each successive generation, a phenomenon known as anticipation. This is explained by the addition of further repeats in the gene in the progeny of affected individuals.²⁷

Among the nucleotide repeat expansions, the most common is the trinucleotide repeat expansion which originate TREDs. Triplet expansions are likely caused by *slippage* during DNA replication. These sequences are instable into these tandem regions, hence 'loop out' structures may form during DNA replication while maintaining complementary base pairing between the parent strand and the daughter strand being synthesized. Then, the repetitive triplet is extended and sealed by DNA polymerase and DNA ligase, respectively. If the 'loop out' structure is formed from the sequence on the daughter strand this will result in an increase in the number of repeats. However, many questions and puzzling features remain and many alternative ideas turned out to be an ongoing issue of mechanistic studies ever since.¹⁵

The biological function of nucleotide repeats might be carried out at the transcript or protein level. In mRNA, their functionality may depend on its type, structure and localization in the different functional regions of mature transcripts. The highest frequency is found in the open reading frame (ORF, 59%), followed by the five prime (5'UTR, 28%) and three prime untranslated region (3'UTR, 13%).²⁰

The general types of pathogenic mechanisms are as follows: toxic RNA gain-of-function; toxic protein gain-of-function; and mutant transcript and mutant protein loss-of-function. The latter includes Friedreich's ataxia (FRDA), in which the expression of a product from the mutant gene is inhibited by the expanded GAA repeats located in the intron of the *FXN* gene.¹⁶ Mutant RNA toxicity (or RNA gain-of-function) was first shown to be involved in DM1 and FXTAS.²⁸ But the largest group of TREDs is caused by expanded CAG repeats present in the ORF of various genes.^{16,19,29,30} These repeats are translated into abnormally elongated polyglutamine [poly(Q)] tracts in proteins which derives into toxicity (protein gain-of-function). Moreover, these poly(Q) tracts derive from CAG transcripts that may be also toxic.^{19,21,28,30-33}

In addition, the perspective of TREDs pathogenesis have been complicated by the discovery of repeat associated non-ATG translation (RAN translation).³³⁻³⁶ It suggests that toxic proteins may also be derived from repeats previously thought to be noncoding. RAN translation was first discovered in SCA8 by Zu et al.³⁶ when control experiment to block ATG-initiated ATXN8 poly(Q) translation did not prevent expression of the protein. These results demonstrated that

CAG and CUG expansion mutations can express proteins without the canonical ATG initiation codon.

Table 1.6. Nucleotide repeat disorders including trinucleotide repeat expansion disorders (TREDs).							
Disease	Sequence	Location	Parent of origin of expansion	Repeats (normal)	Repeats (pre-mutation)	Repeats (disease)	Somatic instability
Diseases with coding nucleotide repeats							
DRPLA	CAG	<i>ATN1</i> (exon 5)	P	6-35	35-48	49-88	Yes
HD	CAG	<i>HTT</i> (exon 1)	P	6-29	29-37	38-180	Yes
OPMD	GCN	<i>PABPN1</i> (exon 1)	P and M	10	12-17	>11	None found
SCA1	CAG	<i>ATXN1</i> (exon 8)	P	6-39	40	41-83	Yes
SCA2	CAG	<i>ATXN2</i> (exon 1)	P	<31	31-32	32-200	Unknown
SCA3	CAG	<i>ATXN3</i> (exon 8)	P	12-40	41-85	52-86	Unknown
SCA6	CAG	<i>CACNA1A</i> (exon 47)	P	<18	19	20-33	None found
SCA7	CAG	<i>ATXN7</i> (exon 3)	P	4-17	28-33	>36 to >460	Yes
SCA17	CAG	<i>TBP</i> (exon 3)	P > M	25-42	43-48	45-66	Yes
SMBA	CAG	<i>AR</i> (exon 1)	P	13-31	32-39	40	None found
Diseases with noncoding nucleotide repeats							
DM1	CTG	<i>DMPK</i> (3' UTR)	M	5-37	37-50	>50	Yes
DM2	CCTG	<i>CNBP</i> (intron 1)	Uncertain	<30	31-74	75-11,000	Yes
FXTAS	GCC	<i>AFF2</i> (5' UTR)	M	4-39	40-200	>200	Unknown
FRDA	GAA	<i>FXN</i> (intron 1)	Recessive	5-30	31-100	70-1,000	Yes
FXS	CGG	<i>FMR1</i> (5' UTR)	M	6-50	55-200	200-4,000	Yes
HDL2	CTG	<i>JPH3</i> (exon 2A)	M	6-27	29-35	36-57	Unknown
SCA8	CTG	<i>ATXN8OS</i> (3' UTR)	M	15-34	34-89	89-250	Unknown
SCA10	ATTCT	<i>ATXN10</i> (intron 9)	M and P (smaller)	10-29	29-400	400-4,500	Yes
SCA12	CAG	<i>PPP2R2B</i> (5' UTR)	M and P (more unstable)	7-28	28-66	66-78	None found
ALS/FTD	GGGGCC	<i>C9ORF72</i>	P and M	<30	30-1,000	1,000-2,000	Yes

AFF2, AF4/FMR2 family, member 2; ALS, amyotrophic lateral sclerosis; AR, androgen receptor; *ATN1*, atrophin 1; *ATXN*, ataxin; *ATXN8OS*, *ATXN8* opposite strand (non-protein coding); *CACNA1A*, calcium channel, voltage-dependent, P/Q type, alpha 1A subunit; *CNBP*, CCHC-type zinc finger nucleic acid binding protein; DM, myotonic dystrophy; *DMPK*, dystrophia myotonica-protein kinase; DRPLA, dentatorubral-pallidoluysian atrophy; *FMR1*, fragile X mental retardation 1; FXTAS, mental retardation, X-linked, associated with FRAXE; FRDA, Friedreich's ataxia; FTD, frontotemporal dementia; *FXN*, frataxin; FXS, fragile X syndrome; FXTAS, fragile X-associated tremor/ataxia syndrome; HD, Huntington's disease; HDL2, Huntington's disease-like 2; *HTT*, huntingtin; *JPH3*, junctophilin 3; M, maternal; N, arbitrary nucleotide; OPMD, oculopharyngeal muscular dystrophy; P, paternal; *PABPN1*, poly(A) binding protein nuclear 1; *PPP2R2B*, protein phosphatase 2, regulatory subunit B; SCA, spinocerebellar ataxia; SMBA, spinomuscular bulbar atrophy; *TBP*, TATA-box binding protein.

1.3.2. STRUCTURAL CLASSES OF ISOLATED REPEATS

Repeated RNA sequences may differ from the non-pathogenic transcript. It was therefore necessary to demonstrate the architecture of these particular nucleotide expansion-related transcripts. Nucleotide repeats have been divided into four classes: (I) ultra-stable G-quadruplex structures, (II) semi-stable hairpins; (III) unstable hairpins; and (IV) not forming any higher order structures (figure 1.12). The most common group of semi-stable hairpins is composed of

CGA, CGU and CNG repeat hairpins (where *N* is an arbitrary nucleotide). The stem portion of CNG hairpins contains C•G and G•C pairs separated by periodic *N•N* interactions. The terminal loop present in the CNG hairpins typically contains four nucleotides. Moreover, CUG, CAG and CCG show a tendency to align ‘in register’ conformations like a ‘slippery hairpin’. The slippery effect can be reduced by capping the stem with canonical C•G and G•C pairs.²⁰ This result proves that the sequence context of the repeats may strongly influence their structural features.

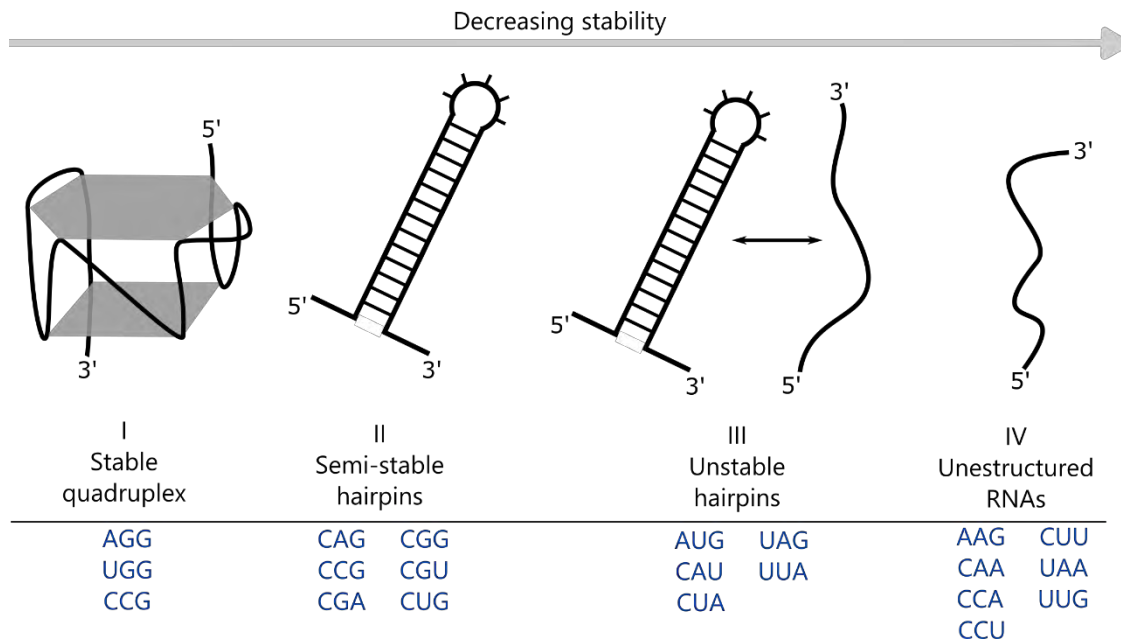


Figure 1.12. Structural classes of isolated triplet repeats in the human genome based on the results of biochemical and biophysical studies.

1.3.3. CLINICAL FEATURES OF MYOTONIC DYSTROPHIES

Myotonic dystrophies (DMs) are autosomal dominant, multisystemic diseases with a core pattern of clinical presentation including myotonia, muscular dystrophy, cardiac conduction defects, posterior iridescent cataracts, and endocrine disorders.³⁷ Myotonic dystrophy type 1 (DM1) was described by Steinert in 1909 as the ‘classic’ type of myotonic dystrophy which was called Steinert’s disease (OMIM 160900, Online Mendelian Inheritance in Man).³⁷ It was in 1992 that the gene defect responsible for DM1 was discovered and found to be caused by an expansion of an unstable CTG trinucleotide repeat in the 3’UTR of the myotonic dystrophy protein kinase (*DMPK*; OMIM 605377).³⁸ This gene is located on chromosome 19q13.3.³⁹ In 1994, a different multisystemic disorder was described with dominantly inherited myotonia, proximal greater than distal weakness, and cataracts but lacking the gene responsible for DM1. In Europe, the disease was named proximal myotonic myopathy (PROMM, OMIM* 160900) or proximal myotonic dystrophy (PDM) while in the United States it was termed myotonic dystrophy type 2 (DM2).^{38,40} Later studies demonstrated that DM2 was caused by an unstable tetranucleotide repeat expansion, CCTG, in intron 1 of the nucleic acid-binding protein (*CNBP*) gene on

chromosome 3q21.^{22,41,42} Although DM1 and DM2 have similar symptoms, they also present a number of very dissimilar features making them clearly separate diseases (table 1.7).

Table 1.7. Comparison of clinical manifestations between DM1 and DM2.

Clinical features	DM1	DM2
General features		
Epidemiology	Widespread	European
Age of onset (years)	0 to adult	8 to 60
Anticipation	Always present	Exceptional
Congenital form	Present	Absent
Life expectancy	Reduced	Normal range
Core features		
Clinical myotonia	Evident in adult-onset	Present in <50%
EMG myotonia	Always present	Absent or variably in many
Muscle weakness	Disabling at age 50	Onset after age 50-70
Cataracts	Always present	Present in minority
Muscle symptoms		
Facial and jaw weakness	Always present	Usually absent
Bulbar weakness-dysphagia	Always later	Absent
Respiratory muscles weakness	Always later	Exceptional
Distal limb muscle weakness	Always prominent	Only <i>flexor digitorum profundus</i> , rare
Proximal limb muscle weakness	Absent or mild	Most disabling symptoms in many
Sternocleidomastoid weakness	Face, temporal, distal hands and legs	Usually absent
Myalgic pain	Absent	Present in ≥50%
Systemic features		
Tremors	Absent	Prominent in many
Behavioral change	Early in most	Not apparent
Cognitive disorders	Prominent	Not apparent
Hypersomnia	Prominent	Not apparent
Cardiac arrhythmias	Always present	From absent to severe
Male hypogonadism	Manifest	Subclinical in most
Manifest diabetes	Frequent	Infrequent

DM1 is the second most common cause of muscular dystrophy after Duchenne muscular dystrophy and the most common cause of adult onset muscular dystrophy with a worldwide incidence of 1 in 8,500. Similar to other TREDs, disease severity and age of onset approximately correlates with the size of the expansion. The congenital form is much more severe and is characterized by general muscle hypotonia and respiratory distress at birth, as well as delayed motor and cognitive development.²⁷

1.3.4. MOLECULAR PATHOMECHANISM OF MYOTONIC DYSTROPHIES

As introduced above, pathology of DM1 and DM2 stems from the corresponding CUG (rCUG^{exp}) and CCUG repeats (rCCUG^{exp}) which form stable hairpin loops. The most studied myotonic dystrophy is DM1, and its full molecular pathomechanism is still a matter of debate. Several hypotheses have been proposed over the years such as a CUG-induced inhibition of DMPK protein expression (which induces its haploinsufficiency)^{40,43} or inhibition of the expression of genes adjacent to the *DMPK* gene (e.g. *SIX5*).^{15,27} However, knockout mice models proved that both theories were insufficient to reproduce all the symptoms observed in DM1.^{44,45} Furthermore, rCUG^{exp} hairpins were discovered to aggregate into the ribonuclear foci (the central site in which the disease localizes and develops) and evidenced the toxicity of expanded

RNA.⁴⁶ These discoveries provided the foundations for the paradigm based on RNA gain-of-function (figure 1.13). The molecular mechanism by which rCUG^{exp} results in disease include:

- (i) Accumulation of rCUG^{exp} in a hairpin-like structure which is able to bind and sequester MBNL proteins with their subsequent loss-of-function in the nucleoplasm.
- (ii) Activation of the protein kinase C (PKC) pathway and suppression of the expression of specific miRNAs culminating in up-regulation of the CELF₁ protein, resulting in its gain-of-function.

Both MBNL₁ and CELF₁ are required for splicing regulation during development. Disruption of their functions lead to missplicing of many genes such as *CLCN1*, *BIN1*, *IR*, *PKM* and *TNNT2*, which explains the prominent features of the disease.⁴⁷ However, some of the observed transcriptional and posttranscriptional perturbations and widespread signaling caused by rCUG^{exp} cannot be wholly explained by MBNL₁ and/or CELF₁.

Loss-of-function of MBNL proteins

MBNL proteins were first discovered as factors involved in DM₁ pathogenesis but were subsequently shown to be direct regulators of alternative splicing.^{42,48} There are three MBNL paralogues in mammals: MBNL₁ and MBNL₂ are expressed in many tissues including brain, heart, muscle, and liver, whereas MBNL₃ is expressed mainly in placenta.⁴⁹ Thus, the observation that MBNL₁ is highly expressed in hearts and skeletal muscle corroborated that these tissues manifest the most severe DM₁ phenotypes.

MBNL proteins contain four zinc finger domains each composed of three cysteines and one histidine that coordinate zinc atoms to bind single-stranded in a sequence specific manner.⁵⁰ Genome-wide analyses, together with crystal structure and biochemical studies revealed that MBNL₁ recognizes YGCY motifs, hence it strongly binds CUG, CCUG and CAG repeats.⁴² Mouse model studies showed that loss of MBNL₁ accounts for greater than 80% of the splicing pathology due to rCUG^{exp} RNA.⁵¹ Intriguingly, MBNL₁ affinity for CCUG repeats is either higher or within the same range as for CUG, and given typically larger expansions in DM₂, its milder phenotype remains elusive.

Recent studies showed that RNA helicase p68 (DDX5) enhances MBNL₁ binding to the stem-loop of rCUG^{exp}.⁵² As a possible scenario, the mismatches in the RNA structure provide an anchoring site for the helicase to initiate the strand separation, facilitating the interaction of MBNL₁ with consensus binding motifs. More interesting, another dead box helicase DDX6 was recently found to exert the opposite effect. DDX6 induced relocalization of rCUG^{exp} to the cytoplasm, dissociation of MBNL₁ from mutant transcripts and partial correction of splicing defects.⁵³

Gain-of-function of CELF₁ protein

CELF₁, also named CUG binding protein 1 (CUGBP1), was discovered in a screen from HeLa cell proteins that bound the CUG motif.⁵⁴ It was later found that CELF₁ binds preferentially the base of the CUG hairpin structure and its affinity for the *DMPK* transcript is not proportional to the CUG repeat size. More importantly, the localization of CELF₁ in healthy and DM₁ cells is identical, indicating that unlike MBNL proteins rCUG^{exp} does not sequester CELF₁. However,

CELF1 levels are increased 2- to 4-fold in DM1 cells and tissues. This up-regulation process occurs via two independent mechanisms: CELF1 stabilization mediated by hyperphosphorylation by PKC; and through reduced levels of miR-23a/b, which should suppress CELF1 protein translation in heart tissue.^{40,54-56} Increased CELF1 protein levels are thought to significantly contribute to the overall DM1 pathogenesis.

Developmental reprogramming of the transcriptome

Alternative pre-mRNA splicing is a primary source of transcriptome complexity in humans that allows multiple transcripts with potentially different function to be produced from a single gene. Misregulation of several splicing processes is an important feature in DM1. The misregulated events are normally developmentally regulated, which in DM1 undergo an adult-to-embryonic switch in splicing patterns. Many of the embryonic isoforms are unable to fulfill the proper adult tissue requirements resulting in specific defects that contribute to the overall disease pathogenesis. Splicing misregulations of some pre-mRNA whose splicing is disrupted, CELF1 and MBNL1 regulate them antagonistically.^{57,58}

In addition, alternative poly(A) also contributes in expanding the RNA transcript diversity for tissue development. Alternative 3'UTR from through alternative poly(A) alter interactions of transcripts with specific factors to modulate their translation, localization and turnover.⁵⁹

Beyond splicing

Although sequestration of MBNL1 accounts for the majority of splicing perturbations, many aspects of DM1 cannot be explained by the splicing defects alone. For instance, rCUG^{exp} in DM1 undergo RAN translation producing a DM1-poly(Q) expansion protein in DM1 patients and mice that aggregates both in nucleus and plasma.⁴¹ This effect is produced because *DMPK* is transcribed bidirectionally hence it may transcribe CAG and CUG transcripts. How and whether RNA proteins affect directly will further improve the understanding of the molecular mechanism focused on RNA toxicity.

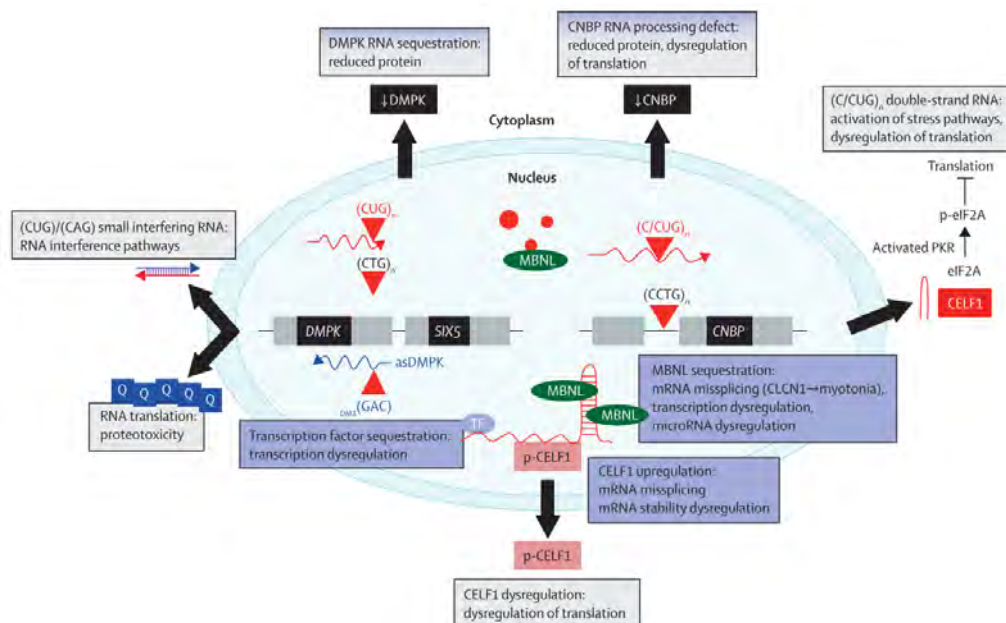


Figure 1.13. Postulated pathological mechanism underlying myotonic dystrophy types 1 and 2. Image taken from Udd et al.⁴¹

Moreover, tissue-specific changes in miRNA expression may also play a major role in DM1 pathology. For instance, nuclear accumulation of rCUG^{exp} in mouse heart decrease expression of a group of miRNAs including *miR-1*. The same group is also downregulated in DM1 patients whose depletion is associated with cardiac arrhythmias and fibrosis. Comparison of miRNA expression in *Drosophila* model and muscle biopsies of DM1 patients showed that multiple conserved miRNAs including *miR-1*, *miR-7*, and *miR-10* are downregulated.⁶⁰⁻⁶²

1.3.5. FROM MECHANISM TO THERAPEUTICS OF DM1

There are no treatments for DMs at this time, and drugs currently in clinical trials are only for symptom management. Several strategies have been proposed. For instance, overexpression of MBNL1 in DM1 ameliorates splicing abnormalities, myotonia and myopathology.⁵⁸ Reducing CELF1 steady state levels through inhibition of the PKC pathway increases survival rate in mice.⁶³ Thus, combination of MBNL1 and CELF1 levels modulation could address DM1 symptoms, but direct targeting of RNA-binding proteins could have deleterious effects.

Since most of these diseases result from the formation of toxic gain-of-function RNA, targeting the toxic RNA should provide the most comprehensive results. Targeting can be approached in two ways: either targeting the RNA for degradation, or stabilizing and covering the RNA in order to render its binding sites unavailable for interaction with RNA-binding proteins. One proposed approach to target toxic repeat RNA is antisense oligonucleotides (ASO). ASO studies thus far have produced promising results in DMs, but each study revealed issues that limited the efficacy of the treatment.^{47,64}

Despite the promising results from ASO development, small molecules continue to be the most attractive therapeutic approach due to their efficiency and low cost production, as well as simplicity of administration (table 1.8, page 31). Small molecule strategies involve binding of the RNA to sterically inhibit interaction with RNA-binding proteins. The first small molecule that proved its efficacy was discovered in a screen of nucleic acid binding compounds for disruption of MBNL1 binding to rCUG^{exp} in DM1. This compound, pentamidine (see table 1.8, *n*-amidine with *n*=5), rescued splicing defects in DM1 cell culture model.⁶⁵ Although its ability to rescue missplicing were modest, the study demonstrated that small molecules are a viable therapeutic approach for RNA-related diseases. Höchst 33258 and derivatives were discovered to improve pentamidine *in vitro* potency by selective binding to rCUG^{exp} RNAs.⁶⁶

Significant improvements were achieved in search for small molecule therapeutic with the evidence that modular assembly methods and chemical similarity searching could be applied to improve specificity and affinity of previously discovered RNA-binding compounds. By using an azide handle to modularly assemble a Höchst 33258 derivative on a peptoid backbone an improvement in potency, selectivity, cell permeability, and localization was achieved if compared to the parent compound.⁶⁷ Several molecules were identified from the National Cancer Institute and eMolecules databases by similarity screening, which provided compounds with increases efficacy for improving DM1 splicing defects.²⁴ Taken together, not only the screening of hit compounds capable of binding RNA is a viable therapeutic approach, but modularly assembled compounds can drastically improve its efficiency.

Further improvements around pentamidine were achieved by the development of heptamidine, which demonstrated for the first time that small molecules can revert DM1 phenotype in an animal model.⁶⁸ In addition to the mentioned small molecules, a triaminotriazine-acridine conjugate has been discovered to inhibit MBNL1-CUG repeat complex formation.⁶⁹ Unfortunately, the compound was proved to be ineffective in cell culture due to its low solubility and permeability. However, the attachment of a cationic polyamine side chain to the compound took advantage of the polyamine transporting system of cells.⁷⁰

Upon realization of the fact that rCUG^{exp} hairpin was structurally similar to the HIV-1 frameshift RNA stem loop a new compound was synthesized based on DB213, replacing its dimethylammonium groups with triaminotriazine units.⁷¹ This compound improved the glossy and rough eye phenotype in a *Drosophila* DM1 model.

In addition to small molecules and ASOs, Garcia-Lopez et al. discovered ABP1 peptide through a *D*-amino acid hexapeptide screening using a *Drosophila* model.⁷² The peptide induced relaxation of the RNA secondary structure and prevented MBNL1 binding.

Much of the molecular basis of nuclear repeat expansion disorders pathology has been successfully uncovered in the past decade. Although the pathogenic mechanism has become increasingly more complex as more details are uncovered, the available information and models serve as an excellent tools for further development of potential therapeutics, which has been accelerating since the discovery of Höchst 33258 in 2000.

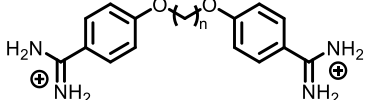
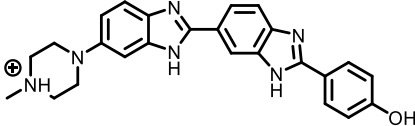
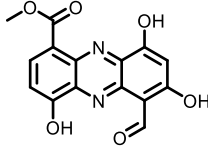
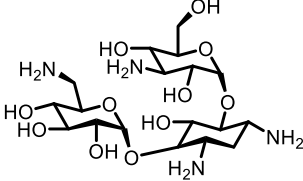
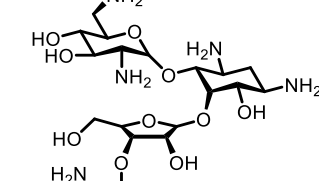
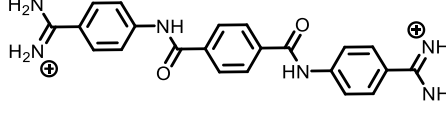
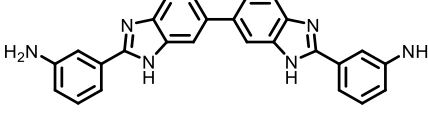
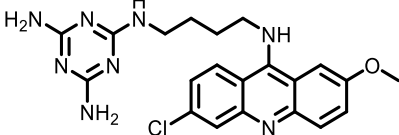
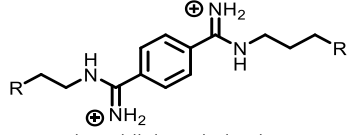
1.3.6. PATHOGENESIS OF SPINOCEREBELLAR ATAXIA TYPE 10

Spinocerebellar ataxia type 10 (SCA10) is characterized by slowly progressive cerebellar ataxia (dysfunction of the cerebellum) that usually starts as poor balance and unsteady gait, followed by upper-limb ataxia, scanning dysarthria (spoken words are broken up into syllables), and dysphagia (difficulty in swallowing). The disease is exclusively found in Latin American populations, particularly those with Amerindian admixture. SCA10 is inherited in an autosomal dominant disease caused by the expansion of the non-coding ATTCT pentanucleotide repeat in the *ATAXIN10* gene chromosome 22q13.31. SCA10, DM2 and ALS/FTD are the only human diseases caused by non-trinucleotide microsatellite expansion mutations. Studies have shown that the pathology of repeat expansion disorders is predominantly caused by two modes of the RNA transcript toxicity (repeated AUUCU, r(AUUCU)^{exp}): repeats bind and sequester proteins involved in RNA biogenesis (RNA gain-of-function);⁷³ and production of toxic homopolymeric proteins that accumulate as inclusion bodies through RNA translation. Expanded r(AUUCU)^{exp} sequester heterogeneous ribonucleoprotein K (hnRNP K), which induces a translocation of protein kinase Cδ to mitochondria and caspase-3 mediated apoptosis of neuronal cells.⁷⁴

Repeat length scales with disease severity. For example, in SCA10, healthy individuals typically have <50 repeats while those afflicted with disease have up to ~5000 repeats. Structural studies have been reported for various triplet repeats and revealed common structural features. For example, they adopt an overall A-form geometry, with variations in base pair and helical parameters. A biophysical study by Handa et. al. suggests r(AUUCU)₉ forms a structured A-form helix via CD and NMR analysis.⁵ Their NMR studies revealed evidence of A-U and U-U base

pairing, suggesting that $r(\text{AUUCU})$ repeats harbor 3×3 nucleotide $5' \text{UCU}_3 / 3' \text{UCU}_5'$ internal loops with two U-U and one C-C non-canonical pairs.

Table 1.8. Subset of representative small molecules, peptides and peptidomimics used for targeting CUG repeats involved in DM1.

Minor groove binders	 <p>n-amidine</p>	 <p>Höchst 33258</p>
Obtained from HTS campaigns	 <p>Lomofungin</p>	
Aminoglycosides	 <p>Kanamycin A</p>	 <p>Neomycin B</p>
Elucidated from <i>in silico</i> screening	 <p>P1</p>	 <p>H1</p>
<i>De novo</i> design	 <p>Triaminotriazine-acridine conjugates</p>	 <p>Bisamidinium derivatives</p>
Peptide and peptidomimic	<p>[Pip-Pro-Cys-Lys]₂</p>	<p>ppyawe (<i>D</i>-aminoacids)</p>

14. TARGETING OF HIV-1 RNA

A lot of people ask, 'Do you think humans are parasites?' It's an interesting idea and one worth thinking about. People casually refer to humanity as a virus spreading across the earth. In fact, we do look like some strange kind of bio-film spreading across the landscape. A good metaphor? If the biosphere is our host, we do use it up for our own benefit. We do manipulate it. We alter the flows and fluxes of elements like carbon and nitrogen to benefit ourselves—often at the expense of the biosphere as a whole. If you look at how coral reefs or tropical forests are faring these days, you'll notice that our host is not doing that well right now. Parasites are very sophisticated; parasites are highly evolved; parasites are very successful, as reflected in their diversity. Humans are not very good parasites. Successful parasites do a very good job of balancing—using up their hosts and keeping them alive. It's all a question of tuning the adaptation to your particular host. In our case, we have only one host, so we have to be particularly careful.

- Carl Zimmer

Talk at Columbia University, 'The Power of Parasites'. 2000

The targeting of RNA with small molecules has the potential to offer a complementary approach to the targeting of proteins. As mentioned in previous sections, many newly discovered functions of RNA are regulatory mechanism in which proteins do not participate. Small molecule ligands for RNA have been developed with three major classes of targets in mind: antibacterial, antiviral and mRNA.

Replication of RNA viruses, such as the human immunodeficiency virus (HIV), is dependent upon multiple specific interactions between viral RNAs and viral and cellular proteins. Antiretroviral therapy to treat AIDS uses molecules that target the reverse transcriptase and protease enzymes of human immunodeficiency virus, type 1 (HIV-1). A major problem associated with these treatments, however, is the emergence of drug-resistant strains. Thus, there is a compelling need to find drugs against other viral targets such as viral RNA targets.

1.4.1. STRUCTURE AND REPLICATION CYCLE

HIV is different in structure from other retroviruses (figure 1.14). It is composed of two copies of positive single-stranded RNA that codes for the virus' nine genes enclosed by a conical capsid composed of 2,000 copies of the viral protein p24. The single-stranded RNA is tightly bound to nucleocapsid proteins, p7, and enzymes needed for the development of the virion such as reverse transcriptase, proteases, ribonuclease and integrase. A matrix composed of the viral protein p17 surrounds the capsid ensuring the integrity of the virion particle.⁷⁵

This is, in turn, surrounded by the viral envelope that is composed of two layers of phospholipids taken from the membrane of a human cell when a newly formed virus particle buds from the cell. Embedded in the viral envelope are proteins from the host cell and about 70 copies of a complex HIV protein that protrudes through the surface of the virus particle. This protein, known as Env, consists of a cap made of three molecules called glycoprotein (gp) 120, and a stem consisting of three gp41 molecules that anchor the structure into the viral envelope.⁷⁶ This glycoprotein complex enables the virus to attach to and fuse with target cells to initiate the infectious cycle. Both these surface proteins, especially gp120, have been considered as targets of future treatments or vaccines against HIV.⁷⁵

The RNA genome consists of at least seven structural landmarks (LTR, TAR, RRE, PE, SLIP, CRS, and INS), and nine genes (gag, pol, and env, tat, rev, nef, vif, vpr, vpu, and sometimes a tenth tev, which is a fusion of tat env and rev), encoding 19 proteins. Three of these genes, gag, pol, and env, contain information needed to make the structural proteins for new virus particles. For example, env codes for a protein called gp160 that is broken down by a cellular protease to form gp120 and gp41. The six remaining genes, tat, rev, nef, vif, vpr, and vpu (or vpx in the case of HIV-2), are regulatory genes for proteins that control the ability of HIV to infect cells, produce new copies of virus (replicate), or cause disease.⁷⁵

The two Tat proteins (p16 and p14) are transcriptional trans-activators for the LTR promoter acting by binding the trans-acting response (TAR) RNA element. The TAR may also be processed into miRNAs that regulate the apoptosis genes ERCC1 and IER3. The Rev protein (p19) is involved in shuttling RNAs from the nucleus and the cytoplasm by binding to the RRE RNA element. The Vif protein (p23) prevents the action of APOBEC3G (a cellular protein that deaminates Cytidine to Uridine in the single stranded viral DNA and/or interferes with reverse transcription). The Vpr protein (p14) arrests cell division at G2/M. The Nef protein (p27) down-regulates CD4 (the major viral receptor), as well as the MHC class I and class II molecules.⁷⁷

Nef also interacts with SH3 domains. The Vpu protein (p16) influences the release of new virus particles from infected cells. The ends of each strand of HIV RNA contain an RNA sequence called the long terminal repeat (LTR). Regions in the LTR act as switches to control production of new viruses and can be triggered by proteins from either HIV or the host cell. The Psi element is involved in viral genome packaging and recognized by Gag and Rev proteins. The SLIP element (TTTTTT) is involved in the frameshift in the Gag-Pol reading frame required to make functional Pol.⁷⁵

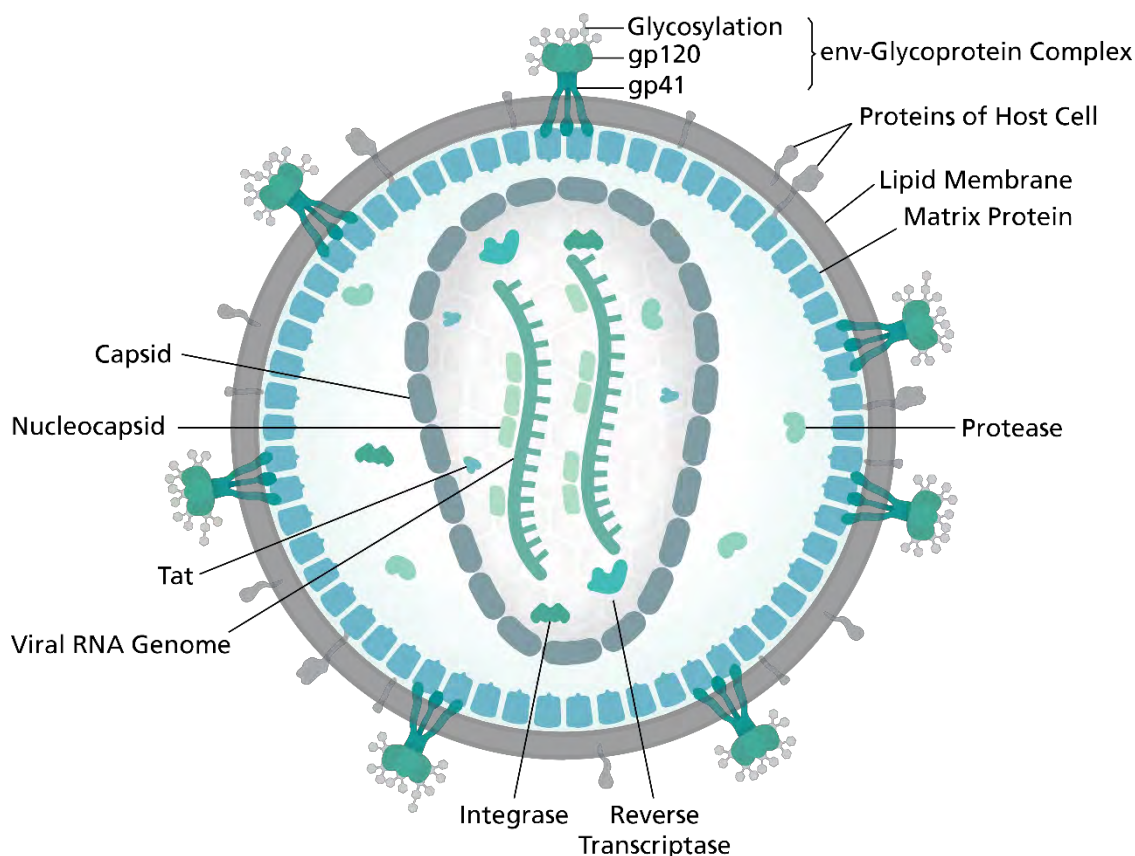


Figure 1.14. Diagram of HIV-1 virion. Image by Thomas Spletstoesser, distributed under a CC-BY-SA 4.0 license.

1.4.2. TRANS-ACTING RESPONSE ELEMENT AS A THERAPEUTIC TARGET

After the HIV genomic RNA is integrated in the host cell's genome, transcription of the viral RNA is initiated by HIV Tat protein. Tat recognizes and binds to an asymmetric bulge located in the TAR element, which in turn is located at the beginning of viral transcripts. Binding of endogenous cyclin Tr with the Tat-TAR complex enhances viral RNA transcription. Thus, inhibition of this step of viral replication is considered a promising alternative to current antiviral approaches.^{17,78}

Several small molecule structures targeting viral RNA have been identified using standard and virtual screening approaches (see table 1.9, page 35). However, a deeper understanding of RNA recognition processes by both cognate partners and synthetic ligands is essential for improving the current drug design.

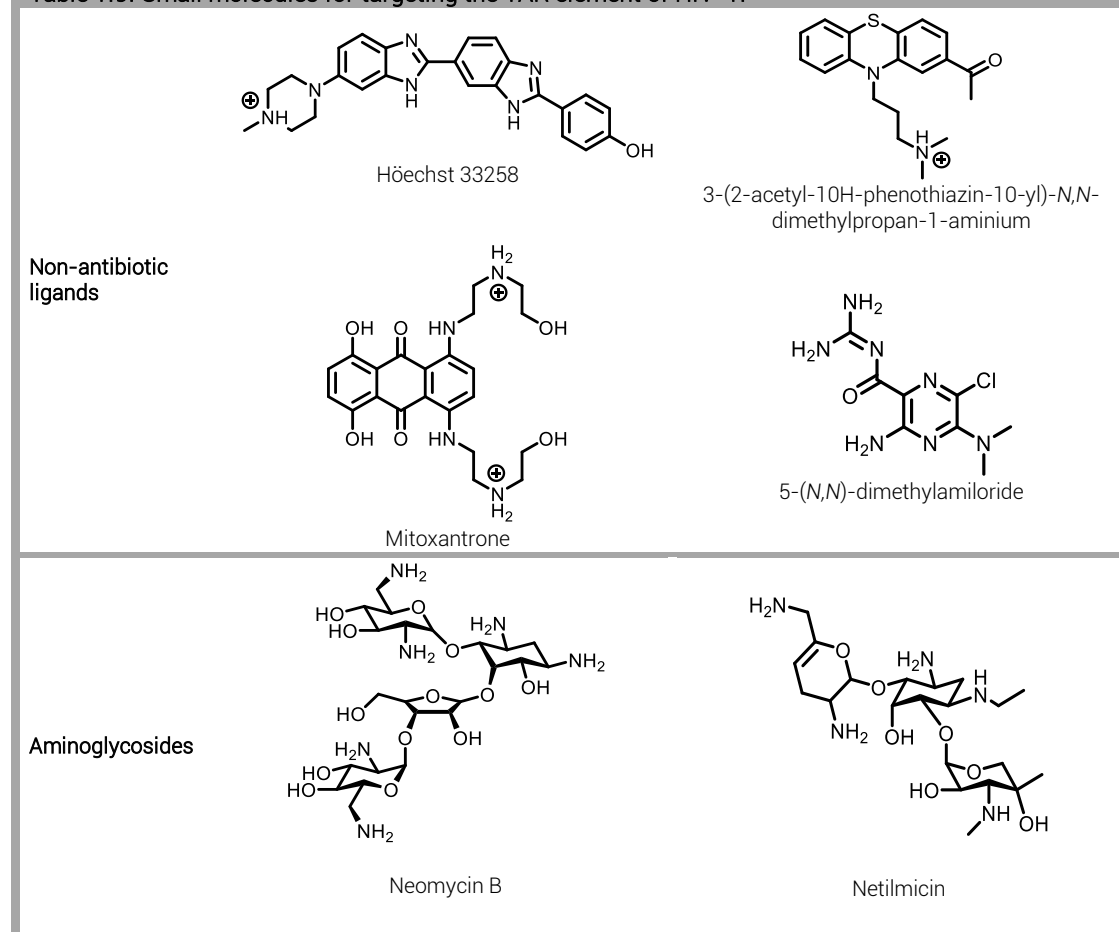
The great debate is whether to use peptides or small molecules. Peptides are more selective being derived by the linear protein sequences, and they should mimic the catalytic or the regulatory subunits of the cell cycle controller complexes, but on the other side they usually present poorer pharmacokinetic characteristics. In contrast, small molecules have better pharmacokinetic features but lower specificity because many RNA targets show high sequence similarity within the active site. Aminoglycosides are a class of antibiotic small molecules known to inhibit the Tat-TAR complex. For instance, neomycin is one of the most potent inhibitors, and dimerization of the aminoglycoside with different linkers proved to be a successful strategy.²⁵

Computational techniques have been also employed to identify small molecules that disrupt this complex. Al-Hashimi and coworkers used a combined molecular dynamics (MD) and NMR residual decoupling approach to identify RNA folds into a dynamic ensemble of structures. The group identified a total of 57 compounds, including aminoglycosides and mitoxantrone among others, which binded TAR with high affinity values.²⁵

RNA-protein complexes such as Tat-TAR have proven difficult targets for conventional small molecule inhibitor design. This may be due to the large surface interfaces buried in the complexes, as well as the inherent flexibility of both RNA and protein in the free states, which seem to favor mutually induced fit mechanisms of binding. Conformationally restrained peptidomimetics having well-defined folded structures offered the prospect of developing structure-activity relationships useful in designing and optimizing ligand affinity to RNA. Athanassiou et al.⁷⁹ demonstrated that these TAR-Tat peptidomimetics fold into stable hairpin conformations in free solution and that formation of this structure favors binding.

Cyclic peptidomimics of Tat also can induce changes in the apical loop of TAR which may interfere in cyclin T1 binding.⁸⁰ Pascale et al. developed a series of α -polyamide amino acids (PAA) that presented promising selectivity towards the TAR RNA.⁸¹ These molecules are constituted by a poly-(2-aminoethylglycyl) backbone onto which are condensed L- α -amino acids.

Table 1.9. Small molecules for targeting the TAR element of HIV-1.



1.5. OBJECTIVES

This thesis is built upon the assumption that conventional drug-design techniques can be applied and modified in order to elucidate new chemical entities for targeting pathogenic endogenous RNA and viral RNA. According to what has been exposed during this introduction the objectives of this thesis are:

1. Construct a computational study of pathogenic RNA structures involved in myotonic dystrophy type 1 (DM1) and spinocerebellar ataxia type 10 (SCA10) using molecular modeling techniques.
2. Use the structure-based and ligand-based information in DM1 to elucidate novel small molecules using conventional and improved chemoinformatic techniques.
3. Combine the molecular modeling and chemoinformatic techniques from objectives 1 and 2 to study macromolecular complexes such as RNA-protein, which may be essential for the study of novel peptides for targeting DM1 pathogenic RNA and the TAR element involved in HIV-1.

1.6. REFERENCES

1. Crick, F. On protein synthesis. *Symp. Soc. Exp. Biol.* **12**, 138–163 (1958).
2. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).
3. Lewin, B. *Genes VIII. Chemistry and Biology* (Pearson Prentice Hall, 2004).
4. Rao, V. S. *Transgenic Herbicide Resistance in Plants*. (CRC Press, 2014).
5. Ball, P. DNA: Celebrate the unknowns. *Nature* **496**, 419–20 (2013).
6. Schlick, T. *Molecular Modeling and Simulation: An Interdisciplinary Guide*. **21**, (Springer New York, 2010).
7. Walker, J. M. *RNA Folding*. (2014).
8. Dickerson, R. E. *et al.* The Anatomy of A-, B-, and Z-DNA. *Science (80-.)*. **216**, 475–485 (1982).
9. Richardson, J. S. *et al.* RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* **14**, 465–481 (2008).
10. Altona, C. & Sundaralingam, M. Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. *J. Am. Chem. Soc.* **94**, 8205–8212 (1972).
11. Humphris-Narayanan, E. & Pyle, A. M. Discrete RNA libraries from pseudo-torsional space. *J. Mol. Biol.* **421**, 6–26 (2012).
12. Leontis, N. B. & Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA* **7**, 499–512 (2001).
13. Lu, X. J. & Olson, W. K. 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31**, 5108–5121 (2003).
14. Watson, J. D. & Crick, F. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Clin. Orthop. Relat. Res.* **462**, 2 (2007).
15. Pearson, C. E., Nichol Edamura, K. & Cleary, J. D. Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.* **6**, 729–742 (2005).
16. Broda, M., Kierzek, E., Gdaniec, Z., Kulinski, T. & Kierzek, R. Thermodynamic stability of RNA structures formed by CNG trinucleotide repeats. Implication for prediction of RNA structure. *Biochemistry* **44**, 10873–10882 (2005).

17. Tor, Y. Targeting RNA with small molecules. *ChemBioChem* **4**, 998–1007 (2003).
18. Meola, G. & Cardani, R. Myotonic dystrophies: An update on clinical aspects, genetic, pathology, and molecular pathomechanisms. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1852**, 594–606 (2015).
19. Nalavade, R., Griesche, N., Ryan, D. P., Hildebrand, S. & Krauss, S. Mechanisms of RNA-induced toxicity in CAG repeat disorders. *Cell Death Dis.* **4**, e752 (2013).
20. Galka-Marciniak, P., Urbanek, M. O. & Krzyzosiak, W. J. Triplet repeats in transcripts: Structural insights into RNA toxicity. *Biol. Chem.* **393**, 1299–1315 (2012).
21. Krzyzosiak, W. J. *et al.* Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nucleic Acids Res.* **40**, 11–26 (2012).
22. Udd, B. *et al.* 140th ENMC International Workshop: Myotonic Dystrophy DM2/PROMM and other myotonic dystrophies with guidelines on management. *Neuromuscul. Disord.* **16**, 403–413 (2006).
23. Foff, E. P. & Mahadevan, M. S. Therapeutics development in myotonic dystrophy type 1. *Muscle Nerve* **44**, 160–169 (2011).
24. Parkesh, R. *et al.* Design of a bioactive small molecule that targets the myotonic dystrophy type 1 RNA via an RNA motif-ligand database and chemical similarity searching. *J. Am. Chem. Soc.* **134**, 4731–4742 (2012).
25. Disney, M. D., Yildirim, I. & Childs-Disney, J. L. Methods to enable the design of bioactive small molecules targeting RNA. *Org. Biomol. Chem.* **12**, 1029–39 (2014).
26. McMurray, C. Mechanisms of trinucleotide repeat instability during human development. *Nat. Rev. Genet.* **11**, 786–799 (2010).
27. Longman, C. Myotonic Dystrophy. *J. R. Coll. Physicians Edinb.* **36**, 51–55 (2006).
28. Todd, P. K. & Paulson, H. L. RNA-mediated neurodegeneration in repeat expansion disorders. *Ann. Neurol.* **67**, 291–300 (2010).
29. Yildirim, I., Park, H., Disney, M. D. & Schatz, G. C. A dynamic structural model of expanded RNA CAG repeats: A refined X-ray structure and computational investigations using molecular dynamics and umbrella sampling simulations. *J. Am. Chem. Soc.* **135**, 3528–3538 (2013).
30. Fiszer, A. & Krzyzosiak, W. J. RNA toxicity in polyglutamine disorders: Concepts, models, and progress of research. *J. Mol. Med.* **91**, 683–691 (2013).
31. Zhang, X. *et al.* A potent small molecule inhibits polyglutamine aggregation in Huntington's disease neurons and suppresses neurodegeneration in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 892–897 (2005).

32. Li, L.-B. & Bonini, N. M. Roles of trinucleotide-repeat RNA in neurological disease and degeneration. *Trends Neurosci.* **33**, 292–298 (2010).
33. Cleary, J. D. & Ranum, L. P. W. Repeat associated non-ATG (RAN) translation: New starts in microsatellite expansion disorders. *Current Opinion in Genetics and Development* **26**, 6–15 (2014).
34. Cleary, J. D. & Ranum, L. P. W. Repeat-associated non-ATG (RAN) translation in neurological disease. *Hum. Mol. Genet.* **22**, 45–51 (2013).
35. Wojciechowska, M., Olejniczak, M., Galka-Marciniak, P., Jazurek, M. & Krzyzosiak, W. J. RAN translation and frameshifting as translational challenges at simple repeats of human neurodegenerative disorders. *Nucleic Acids Res.* **42**, 11849–11864 (2014).
36. Zu, T. *et al.* RAN proteins and RNA foci from antisense transcripts in C9ORF72 ALS and frontotemporal dementia. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E4968–77 (2013).
37. Harper, P. S. Congenital myotonic dystrophy in Britain. I. Clinical aspects. *Arch. Dis. Child.* **50**, 505–513 (1975).
38. Ranum, L. P. W. & Day, J. W. Myotonic dystrophy: RNA pathogenesis comes into focus. *Am. J. Hum. Genet.* **74**, 793–804 (2004).
39. Dansithong, W. *et al.* Cytoplasmic CUG RNA foci are insufficient to elicit key DMI features. *PLoS One* **3**, e3968 (2008).
40. Lee, J. E. & Cooper, T. A. Pathogenic mechanisms of myotonic dystrophy. *Biochem. Soc. Trans.* **37**, 1281–1286 (2009).
41. Udd, B. & Krahe, R. The myotonic dystrophies: Molecular, clinical, and therapeutic challenges. *Lancet Neurol.* **11**, 891–905 (2012).
42. Kino, Y. *et al.* Muscleblind protein, MBNL1/EXP, binds specifically to CHHG repeats. *Hum. Mol. Genet.* **13**, 495–507 (2004).
43. Jiang, H., Mankodi, A., Swanson, M. S., Moxley, R. T. & Thornton, C. a. Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons. *Hum. Mol. Genet.* **13**, 3079–3088 (2004).
44. Suenaga, K. *et al.* Muscleblind-like 1 knockout mice reveal novel splicing defects in the myotonic dystrophy brain. *PLoS One* **7**, (2012).
45. Kanadia, R. N. *et al.* A muscleblind knockout model for myotonic dystrophy. *Science* **302**, 1978–1980 (2003).
46. Pettersson, O. J., Aagaard, L., Jensen, T. G. & Damgaard, C. K. Molecular mechanisms in DMI -- a focus on foci. *Nucleic Acids Res.* 1–9 (2015). doi:10.1093/nar/gkvo29

47. Singh, R. K. & Cooper, T. A. Pre-mRNA splicing in disease and therapeutics. *Trends Mol. Med.* **18**, 472–482 (2012).
48. Warf, M. B., Diegel, J. V, von Hippel, P. H. & Berglund, J. A. The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9203–9208 (2009).
49. Kinter, J. & Sinnreich, M. Molecular targets to treat muscular dystrophies. *Swiss Med. Wkly.* **144**, w13916 (2014).
50. Teplova, M. & Patel, D. J. Structural insights into RNA recognition by the alternative-splicing regulator muscleblind-like MBNL1. *Nat. Struct. & Mol. Biol.* **15**, 1343–1351 (2008).
51. Piccirillo, R., Demontis, F., Perrimon, N. & Goldberg, A. Developmental insights into the pathology and therapeutic strategies for DMR: back to the basics. *Dev. Dyn.* 1–47 (2013). doi:10.1002/dvdy.
52. Laurent, F. X. *et al.* New function for the RNA helicase p68/DDX5 as a modifier of MBNL1 activity on expanded CUG repeats. *Nucleic Acids Res.* **40**, 3159–3171 (2012).
53. Pettersson, O. J. *et al.* DDX6 regulates sequestered nuclear CUG-expanded DMPK-mRNA in dystrophia myotonica type 1. *Nucleic Acids Res.* **42**, 7186–7200 (2014).
54. Kuyumcu-Martinez, N. M., Wang, G. S. & Cooper, T. a. Increased Steady-State Levels of CUGBP1 in Myotonic Dystrophy 1 Are Due to PKC-Mediated Hyperphosphorylation. *Mol. Cell* **28**, 68–78 (2007).
55. Masuda, A. *et al.* CUGBP1 and MBNL1 preferentially bind to 3' UTRs and facilitate mRNA decay. *Sci. Rep.* **2**, 209 (2012).
56. Teplova, M., Song, J., Gaw, H. Y., Teplov, A. & Patel, D. J. Structural Insights into RNA Recognition by the Alternate-Splicing Regulator CUG-Binding Protein 1. *Structure* **18**, 1364–1377 (2010).
57. Warf, M. B. & Berglund, J. A. MBNL binds similar RNA structures in the CUG repeats of myotonic dystrophy and its pre-mRNA substrate cardiac troponin T. *RNA* **13**, 2238–2251 (2007).
58. Kanadia, R. N. *et al.* Reversal of RNA missplicing and myotonia after muscleblind overexpression in a mouse poly(CUG) model for myotonic dystrophy. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 11748–11753 (2006).
59. Batra, R. *et al.* Loss of MBNL Leads to Disruption of Developmentally Regulated Alternative Polyadenylation in RNA-Mediated Disease. *Mol. Cell* **56**, 1–12 (2014).
60. Dardenne, E. *et al.* RNA Helicases DDX5 and DDX17 Dynamically Orchestrate Transcription, miRNA, and Splicing Programs in Cell Differentiation. *Cell Rep.* **7**, 1900–1913 (2014).

61. Perfetti, A. *et al.* Plasma microRNAs as biomarkers for myotonic dystrophy type 1. *Neuromuscul. Disord.* **24**, 509–515 (2014).
62. Erriquez, D., Perini, G. & Ferlini, A. Non-coding RNAs in muscle dystrophies. *Int. J. Mol. Sci.* **14**, 19681–19704 (2013).
63. Ketley, A. *et al.* High-content screening identifies small molecules that remove nuclear foci, affect MBNL distribution and CELF1 protein levels via a PKC-independent pathway in myotonic dystrophy cell lines. *Hum. Mol. Genet.* **23**, 1551–1562 (2014).
64. Wojtkowiak-Szlachcic, a. *et al.* Short antisense-locked nucleic acids (all-LNAs) correct alternative splicing abnormalities in myotonic dystrophy. *Nucleic Acids Res.* **43**, 3318–3331 (2015).
65. Warf, M. B., Nakamori, M., Matthys, C. M., Thornton, C. a & Berglund, J. A. Pentamidine reverses the splicing defects associated with myotonic dystrophy. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 18551–18556 (2009).
66. Pushechnikov, A. *et al.* Rational Design of Ligands Targeting Triplet Repeating Transcripts That Cause RNA Dominant Disease: Application to Myotonic Muscular Dystrophy Type 1 and Spinocerebellar Ataxia Type 3. *J. Am. Chem. Soc.* **131**, 9767–9779 (2009).
67. Lee, M. M. *et al.* Controlling the specificity of modularly assembled small molecules for RNA via ligand module spacing: Targeting the RNAs that cause myotonic muscular dystrophy. *J. Am. Chem. Soc.* **131**, 17464–17472 (2009).
68. Coonrod, L. a. *et al.* Reducing levels of toxic RNA with small molecules. *ACS Chem. Biol.* **8**, 2528–2537 (2013).
69. Jahromi, A. H. *et al.* Developing bivalent ligands to target CUG triplet repeats, the causative agent of myotonic dystrophy type 1. *J. Med. Chem.* **56**, 9471–9481 (2013).
70. Jahromi, A. H. *et al.* A novel CUGexp·MBNL1 inhibitor with therapeutic potential for myotonic dystrophy type 1. *ACS Chem. Biol.* **8**, 1037–1043 (2013).
71. Wong, C. H. *et al.* Targeting toxic RNAs that cause myotonic dystrophy type 1 (DM1) with a bisamidinium inhibitor. *J. Am. Chem. Soc.* **136**, 6355–6361 (2014).
72. García-López, A., Llamusi, B., Orzáez, M., Pérez-Payá, E. & Artero, R. D. In vivo discovery of a peptide that prevents CUG-RNA hairpin formation and reverses RNA toxicity in myotonic dystrophy models. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 11866–11871 (2011).
73. Handa, V., Yeh, H. J. C., McPhie, P. & Usdin, K. The AUUCU repeats responsible for spinocerebellar ataxia type 10 form unusual RNA hairpins. *J. Biol. Chem.* **280**, 29340–29345 (2005).
74. White, M. C. *et al.* Inactivation of hnRNP K by expanded intronic AUUCU repeat induces apoptosis via translocation of PKCδ to mitochondria in spinocerebellar ataxia

10. *PLoS Genet.* **6**, 1–12 (2010).
75. Leitner, T. *et al.* HIV Sequence Compendium 2008 Editors. (2008).
76. Chan, D. C., Fass, D., Berger, J. M. & Kim, P. S. Core structure of gp41 from the HIV envelope glycoprotein. *Cell* **89**, 263–273 (1997).
77. Vasudevan, A. A. J. *et al.* Structural features of antiviral DNA cytidine deaminases. *Biol. Chem.* **394**, 1357–1370 (2013).
78. Stelzer, A. C. *et al.* Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. *Nat. Chem. Biol.* **7**, 553–559 (2011).
79. Athanassiou, Z. *et al.* Structural Mimicry of Retroviral Tat Proteins by Constrained β -Hairpin Peptidomimetics: Ligands with High Affinity and Selectivity for Viral TAR RNA Regulatory Elements. *J. Am. Chem. Soc.* **126**, 6906–6913 (2004).
80. Davidson, A., Patora-Komisarska, K., Robinson, J. a. & Varani, G. Essential structural requirements for specific recognition of HIV TAR RNA by peptide mimetics of Tat protein. *Nucleic Acids Res.* **39**, 248–256 (2011).
81. Pascale, L. *et al.* Thermodynamic studies of a series of homologous HIV-1 TAR RNA ligands reveal that loose binders are stronger Tat competitors than tight ones. *Nucleic Acids Res.* **41**, 5851–5863 (2013).

CHAPTER II. METHODS

2.1. CONSILIENCE OF A MULTIDISCIPLINARY ENTERPRISE

Art is the lie that helps tell the truth.

- Pablo Picasso

Modeling molecular systems by computer has been drawing increasing attention from scientist in varied disciplines. In particular, modeling large biological systems – proteins, nucleic acids, and lipids – is a multidisciplinary enterprise. Biologists describe the cellular process; chemist fill in the atomic and molecular details; physicists extend these views to the electronic level and underlying forces; mathematician analyze and formulate numerical models and algorithms; and computer scientists and engineers provide the crucial implementational support for running large computer programs. Thus, the many names for the field remarks its cross-disciplinary nature: computational biology, computational chemistry, *in silico* biology, computational structural biology, computational biophysics, theoretical biophysics, theoretical chemistry, and the list goes on.¹

This multidisciplinary approach is important not only because the many aspects involved but also since the best computational approach is often linked to the biological problem. Close connection between theory and experiments are essential because computational models evolve as experimental data become available. In return, biological theories and new experiments are performed as a result of computational insights.

2.1.1. DEFINITION OF TERMS

Molecular modeling is the science and art of studying molecular structure and function through model building and computation. The continuing growth of computing power has made it possible to analyze, compare and characterize large and complex data sets obtained from experiments. Molecular modeling is a relatively young discipline that is developing with astonishing speed. However, a cautionary usage of molecular modeling tools as well as a critical perspective of the field's strength and limitations must be warranted.

When undertaking a molecular modeling study of a system, the level of modeling - spatial resolution, time scale, and degrees of freedom of interest - must be considered. The questions being addressed by computational approaches are as intriguing and complex as the biological or chemical systems themselves. As experimental results are being reported in structure determination, including new methodologies such as NMR, cryoelectron microscopy, and single-molecule biochemistry techniques, modeling approaches are needed to understand many fundamental questions concerning their biological motions and functions. Modeling techniques provides a way to systematically explore 'structural – dynamical - thermodynamic' patterns, test and develop hypotheses, interpret and extend experimental data, and extend basic laws that govern molecular structure, flexibility, physicochemical features and function.

Molecular modeling began with the notion that molecular geometry, energy and other molecular properties can be calculated from models subject to physical forces. A wide variety of different procedures or models have been developed to calculate molecular structure and energetics. These have generally been separated into two categories: quantum chemical models and molecular mechanics models. Which level of modeling is chosen to describe a particular molecular process depends on the type of process itself.

Quantum chemical (QM) models stem from the Schrödinger equation (Eq. 2.1). It treats molecules as collections of nuclei and electrons, without any reference to 'chemical bonds'. The solution to the Schrödinger equation is in terms of the motions of the electrons, which leads to molecular structure and energy, and information about bonding. Unfortunately, the Schrödinger equation cannot actually be solved for any but one-electron system (hydrogen atom), and several approximations have to be made. The existent quantum chemical models differ in the nature of these approximations, and span a wide range, both in terms of their capability and reliability and their 'computational cost'.²

$$\hat{H}\Psi = E\Psi \quad (2.1)$$

The typical alternative to quantum chemical models are called **molecular mechanics (MM) models**. These models do not start from an 'exact' theory as the Schrödinger equation, but rather from a simple but reasonable approximation of molecular structure. Molecular mechanics describes molecules in terms of bonded atoms, which have been distorted from some idealized geometry due to non-bonded van der Waals and Coulombic interactions. Further details about MM models will be given in the next sections.²

2.1.2. COMPONENTS OF A MODEL

A model can be composed as the sum of four basic subunits: (1) the **degrees of freedom**; (2) the description of the potential energy or **force field**; (3) the **sampling method**; (4) the simulation of external forces, the **boundaries** and the conditions of the system. Depending on the selection of these four parameters, different modeling 'levels' can be achieved including QM and MM (table 2.1, page 49).

Table 2.1. Examples of levels of modeling in computational biochemistry and molecular biology.³

Methods	Degrees of freedom	Properties, processes	Time scale
Quantum dynamics	Atoms, nuclei, electrons	Excited states, relaxation, reaction dynamics	Picoseconds
Quantum mechanics	Atoms, nucleic, electrons	Ground and excited states, reaction mechanisms	No time scale
Molecular mechanics	Atoms, solvent	Ensembles, averages, system properties, folding	Nanoseconds
Statistical methods (database analysis)	Groups of atoms, amino acids residues, bases	Structural homology and similarity	No time scale
Continuum methods (hydrodynamics and electrostatics)	Electrical continuum, velocity continuum, etc.	Rheological properties	Supramolecular
Kinetic equations	Populations of species	Populations dynamics, signal transduction	Macroscopic

2.2. MOLECULAR MECHANICS

Molecular modeling simulations occur over a wide range of time and space scales, and the approach to study them depends on the question asked. In many cases, the best alternative is an experimental technique. However, theoretical methods have made huge advances the last few decades, and simulations can either provide more detail or are more efficient to use compared to setting up a new experiment.

When the aim is to predict the structure and/or function of proteins, the best tool is normally bioinformatics that detect relation proteins from amino acid sequence similarity. For computational drug design often it is more productive to use less accurate but fast statistical methods like QSAR (Quantitative Structure-Activity Relationship, refer to section 2.3). In this section, one of the most used MM methods, molecular dynamics (MD), will be introduced. Traditionally, the role of simulations has been to test if simple theoretical models can explain experimental observations. However, this is changing rapidly and today simulations frequently make predictions about properties such as binding or folding dynamics that are later confirmed in the lab.

From an ideal physics point-of-view, the Schrödinger equation should be able to predict all states (wave functions) of any molecule *ab initio*. However, when many particles are involved it is necessary to introduce several approximations such as the aforementioned MM methods. The conceptual difference is that QM is excellent at describing the electronic structure and enthalpy (ΔH) of a small system, while MM instead excels at sampling billions of states of a particular macromolecule – in particular this means they properly include the entropy part of free energy ($T\Delta S$). Since MM methods have been parametrized from experiments they also perform better when it comes to reproducing observations on microsecond scale.

2.2.1. FORCE FIELD DEFINITION

Macroscopic properties measured in an experiment are not direct observations but averages over billions of molecules representing a statistical mechanics **ensemble**. From a practical point of view there are several limitations: (1) it is not sufficient to work with individual structures, but

systems have to be expanded to generate a representative ensemble of structures; (2) properties related to free energy (ΔG) cannot be calculated directly from individual simulations (Eq 2.2); (3) for equilibrium properties the aim is to examine the ensemble of structures and not necessarily the atomic trajectories.⁴

$$\Delta G = \Delta H - T\Delta S \quad (2.2)$$

All classical MM simulations rely on empirical sets of parameters called force fields to calculate interactions and evaluate the potential energy of a system as a function of atomic coordinates. A force field consist of a set of equations used to calculate the potential energy and forces from particle coordinates, and also contains a collection of parameters used in the equations. For most purposes this approximations work great, but they cannot reproduce quantum effects such as bond formation or breaking.

All classical force fields subdivide potential functions in two classes (figure 2.1): **bonded interactions** cover stretching of covalent bonds, angle-bending, torsion potentials when rotating around bonds, and out-of-plane ‘improper torsion’ potentials. **Non-bonded interactions** that are close in space consist of van der Waals interactions (for repulsion and dispersion) and Coulomb electrostatic interactions.

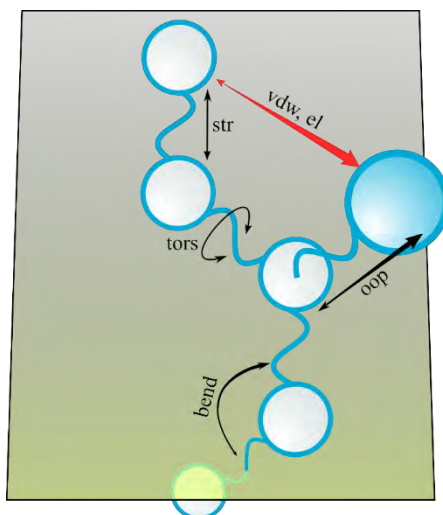


Figure 2.1. Bonded and non-bonded interactions representation: stretching, str; bending, bend; torsion, tors; out-of-plane, oop; van der Waals, vdw; electrostatic, el.

A particular set of force fields, the AMBER⁵ force fields, have been shown to accurately reproduce the structural and dynamics properties of a large variety of canonical and non-canonical nucleic acids. AMBER (Assisted Model Building with Energy Refinement) compose a family of force fields developed by Peter Kollman’s group at the University of California, San Francisco which refers to the following functional form (Eq. 2.3):

$$V(r^N) = E_{str} + E_{bend} + E_{tors} + E_{vdw} + E_{el} = \sum_{bonds} k_b(l - l_0)^2 + \sum_{angles} k_a(\theta - \theta_0)^2 + \sum_{torsions} \sum_n \frac{1}{2} V_n [1 + \cos(n\omega - \gamma)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^N f_{ij} \left\{ \epsilon_{ij} \left[\left(\frac{r_{oij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{oij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} \quad (2.3)$$

First, notice that despite the term ‘force field’, this equation defines the potential energy of the system [$V(r^N)$]; force is the derivative of this potential with respect to position. The summing over bonds (first term) represent the energy between covalently bonded atoms at a distance l .

This harmonic force (or ideal spring) is a good approximation near the equilibrium bond length (l_0). The term k_b represent the force constant for separating the pair of atoms.

The second term (summing over angles) represents the energy due to electron orbitals involved in covalent bonding. In this case, k_a is the force constant of bending and θ is the angle formed by three consecutive bonded atoms within a θ_0 equilibrium angle.

The torsions term represents the energy for twisting a bond due to bond order and neighboring bonds or lone pairs of electrons. The energy barrier need to succeed in the ‘twist’ is defined by V_n , ω is the torsion angle and γ represents the phase.

The last term represents the non-bonded energy between all atom pairs, which contains the van der Waals energies (first term of summation) and the electrostatic energies (second term). The van der Waals energy is calculated using the Lennard-Jones potential equation with an equilibrium distance r_{0ij} and well depth ϵ_{ij} . The factor of 2 ensures that the equilibrium distance is r_{0ij} . The electrostatic energy part assumes that the charges in every atom can be represented by a single point charge (q_i and q_j) using Coulomb’s equation (where $1/4\pi\epsilon_0$ represents the electrostatic constant).

2.2.2. STATISTICAL MECHANICS

The concept of ‘ensemble’ comes from statistical mechanics theories. Statistical mechanics is a branch of physical sciences that studies macroscopic systems from a molecular point of view. The goal is to understand and to predict macroscopic phenomena from the properties of individual molecules composing a system. In order to connect the macroscopic system to the microscopic system, time independent statistical averages are often introduced.

The thermodynamic state of a system is usually defined by a small set of parameters (e.g. temperature, pressure and number of particles). Other thermodynamic properties may be derived from fundamental thermodynamic equations. The microscopic state of a system is defined by atomic positions (\mathbf{q}), and momenta (\mathbf{p}) which can be considered as coordinates in a multidimensional space called **phase space**. For a system of N particles, this space has $6N$ dimensions, and a single point in this space describes the state of the system. An ensemble is a collection of points in phase space satisfying the conditions of a particular thermodynamic state. In other words, it is a collection of all possible systems which have different microscopic states but have an identical macroscopic or thermodynamic state. For instance, a molecular dynamics

Acronym	Description	Condition for equilibrium	Name of ensemble
NVT	System with constant number of particles, volume and temperature.	Minimum A (Helmholtz free energy)	Canonical
NVE	System with constant number of particles, volume and total energy.	Minimum S (entropy)	Microcanonical
NpT	System with constant number of particles, pressure and temperature.	Minimum G (Gibbs free energy)	Isothermal-Isobaric
μVT	System with constant chemical potential, volume and temperature.	Maximum PV	Grand canonical

simulation generates a sequence of points in phase space as a function of time; these states belong to the same ensemble, and they correspond to different conformations of the system and their respective momenta. There exist different ensembles with different characteristics (table 2.2, page 52).²

2.2.3. MOLECULAR DYNAMICS

The two most common ways to generate statistically faithful equilibrium ensembles are Monte Carlo and molecular dynamics (MD) simulations. MD is a computer simulation of physical movements of atoms and molecules acting under a potential energy (force field) that involves numerical solution of the differential system arising from Newtonian dynamics. The atoms and molecules are allowed to interact for a period of time, giving a view of the motion of the atoms. Unfortunately, these methods cannot handle structures that are very far from equilibrium (for instance, if two atoms are overlapping in space). Thus, prior to simulation, an energy minimization is required in order to start from an optimum (minimum energy) structure. Given the force on all atoms (obtained from the force field expression), the coordinates will be updated for every step of minimization. The process can be achieved by the **steepest descent** algorithm which simply moves each atom a short distance in direction of decreasing energy (force is the negative gradient of energy, Eq. 2.4). However, the steepest descent algorithm can be slow and sometimes an alternative algorithm is used such as the **conjugate gradient** algorithm. Briefly, conjugate gradient uses information from previous first derivatives to determine the optimum direction for a line search.

After minimization of the system, molecular dynamics is performed by integrating Newton's equations of motion (Eq. 2.4, Eq. 2.5).

$$\mathbf{F}_i = -\frac{\partial V(\mathbf{r}_1, \dots, \mathbf{r}_N)}{\partial \mathbf{r}_i} \quad (2.4)$$

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \mathbf{F}_i \quad (2.5)$$

The new coordinates are then used to evaluate the potential energy in a new iteration step, as shown in the flowchart of figure 2.2 (page 54). The basic idea is to generate structures by calculating potential function and integrating Newton's equations of motion, structures which are then used to evaluate equilibrium properties of the system. A typical time step is in the order of 1 or 2 femtoseconds.

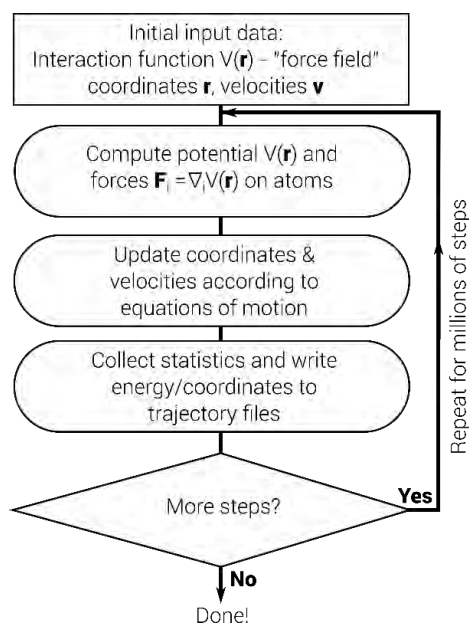


Figure 2.2. Simplified flowchart of a typical molecular dynamics simulations.

When moving from the ‘real’ to the simulation world, even the smallest chemical sample is far too large to represent in a computer. For this reason, molecular simulations often uses **periodic boundary conditions (PBC)** to avoid surface artifacts, so that a water molecule that exits to the right reappears on the left side in the system. Using this technique, small ‘boxes’ of solute and solvent can be used which, at the same time, are surrounded by copies of itself resulting in an infinite periodic system. If the box is sufficiently large the molecules will not interact with their periodic copies (see example in figure 2.3).

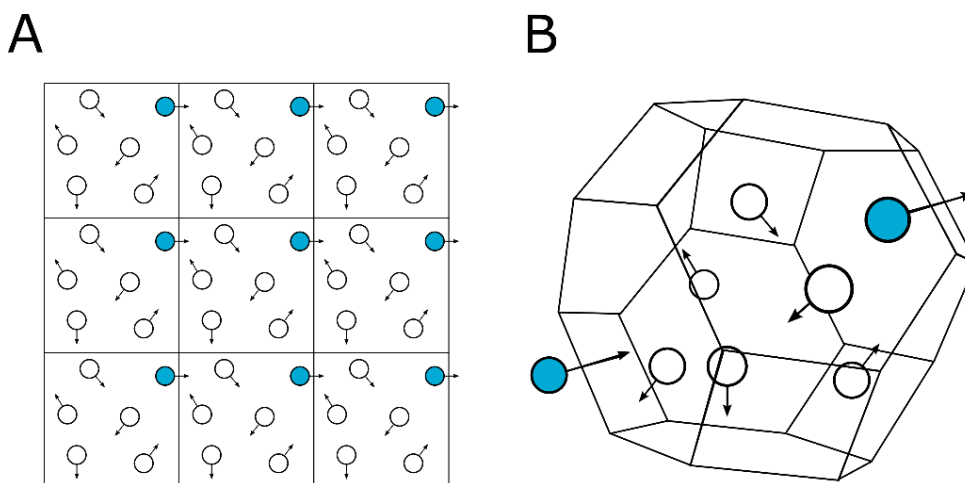


Figure 2.3. Periodic boundary conditions (A) in a squared two-dimensional system and (B) in a three-dimensional truncated octahedron.

Non-bonded interactions ideally should be summed over all neighbors in the periodic system. The current method of choice to simulate this long-range interactions is to use **Particle-**

Mesh-Ewald summation (PME)^{6,7} to calculate the infinite electrostatic interactions by splitting the summation into short- and long-range parts.

Simulations with large number of molecules usually require a **cutoff** in order to reduce the amount of interactions that need to be computed, since non-bonded interactions imply that millions of pairs have to be evaluated. However, cutoffs and rounding errors can lead to drifts in energy, which will cause the system to heat up during simulation. As the potential energy (V) of the structure decreases during the simulation, the kinetic energy (K , i.e., **temperature**) would increase if the total system energy (E) was constant (Eq. 2.6). In order to control the temperature, the system is normally coupled to a thermostat that scales velocities during the simulation to maintain the desired temperature. Similarly, the total **pressure** in the system can be adjusted through scaling the simulation box size.

$$E = K + V \quad (2.6)$$

Nonetheless, the most demanding part of simulations is the computation of non-bonded interactions. Extending the time step between integrations is an important way to improve simulation performance, but several errors are introduced already at 1 fs because of bond vibrations. Luckily, in most simulations these bond vibrations are not of interest per se and can be removed by introducing **bond constraint** algorithms such as SHAKE⁸. These constraints make it possible to extend time step to 2 fs.

2.2.4. ENHANCED SAMPLING TECHNIQUES

Extensive conformational sampling has been a long-standing issue in computational sciences. For complex systems such as proteins and large nucleic acids, the enormity of the number of energy minima couple with the time needed to escape from each of these sometimes not-so-minor conformational traps. Imagine, for example, a complex between a small molecule and a receptor; the potential pathways for complex formation are numerous and the time required for sampling each of them may be enormous. Thus, a reduction of the time scale simulation is required by enhancing the sampling with non-conventional MD techniques. This include metadynamics, parallel tempering MD, accelerated MD (aMD), and steered MD (SMD), among others (figure 2.4, page 57).

REPLICA-EXCHANGE MOLECULAR DYNAMICS

The replica-exchange molecular dynamics (REMD) algorithm of Sugita and Okamoto⁹ has become a widely-used tool for molecular simulation. REMD arises by applying a parallel tempering method to MD. It consist of running multiple isothermal MD simulations in parallel at a sequence of increasing temperatures (T_0, T_1, \dots, T_n) and intermittently attempting to swap simulations between temperatures: each L steps, two structures j and k with positions \mathbf{q} and associated momenta \mathbf{p} are randomly chosen, and the proposed swap is accepted with probability given by the Metropolis ratio (Eq. 2.7).

$$\min \left\{ 1, \frac{\pi_{T_k}(\mathbf{q}_j, \mathbf{p}_j) \pi_{T_j}(\mathbf{q}_k, \mathbf{p}_k)}{\pi_{T_j}(\mathbf{q}_j, \mathbf{p}_j) \pi_{T_k}(\mathbf{q}_k, \mathbf{p}_k)} \right\} \quad (2.7)$$

The distribution $\pi_{T_k}(\mathbf{q}_j, \mathbf{p}_j)$ is the Boltzmann distribution for replica j at temperature T_k given in Eq. 2.8.

$$\pi_k(x) = \frac{e^{-E(x)/(k_B T)}}{\int_{(\mathbf{q}, \mathbf{p})} e^{-E(x)/(k_B T)}} \quad (2.8)$$

where k_B is the Boltzmann constant. The resulting REMD algorithm is a stochastic dynamic system that enables the crossing of large energy barriers. Several studies evidenced that parallel tempering simulations equilibrate dramatically faster than classical MD.^{10–15}

ACCELERATED MOLECULAR DYNAMICS

Accelerated molecular dynamics (aMD) provides a straightforward and effective way to simulate infrequent events required for a macromolecule conformational change without previous knowledge of conformational states, potential energy wells, or barriers. This method modifies the amount of computational time a system spends in a given energy minima by adding a bias potential $[\Delta V(\mathbf{r})]$ to the ‘true’ potential in such a way that potential energy surfaces in vicinity of the minima are raised, while those closer to the barrier or saddle point are not affected. The nonnegative continuous bias potential function $\Delta V(\mathbf{r})$ is defined such that when the potential energy of a system, $V(\mathbf{r})$, falls below a specified boost energy, E , the simulation is carried out using the modified potential, $V^*(\mathbf{r})$ (Eq. 2.9).¹⁶

$$V^*(\mathbf{r}) = \begin{cases} V(\mathbf{r}), & V(\mathbf{r}) \geq E \\ V(\mathbf{r}) + \Delta V(\mathbf{r}), & V(\mathbf{r}) < E \end{cases} \quad (2.9)$$

In the simplest form, the bias potential is given by Eq 2.10.

$$\Delta V(\mathbf{r}) = \frac{(E - V(\mathbf{r}))^2}{\alpha + E - V(\mathbf{r})} \quad (2.10)$$

where α is the acceleration factor. As α decreases, the energy surface is flattened more and biomolecular transitions between low-energy states are increased. Typically two versions of aMD, which provide different acceleration levels, are used: dihedral-boost and dual-boost. In dihedral-boost method all dihedrals in the system experience a boost potential with input parameters E_{dihed} and α_{dihed} . In dual-boost, a total boost potential is applied in addition to the dihedral boost ($E_{dihed}, \alpha_{dihed}, E_{total}, \alpha_{total}$, Eq 2.11).

$$\begin{aligned} E_{dihed} &= V_{dihed_{avg}} + 3.5N_{res}, & \alpha_{dihed} &= (3.5N_{res})/5 \\ E_{total} &= V_{total_{avg}} + 0.175N_{atoms}, & \alpha_{total} &= 0.175N_{atoms} \end{aligned} \quad (2.11)$$

where N_{res} is the number of macromolecule residues, N_{atoms} is the total number of atoms, and $V_{dihed_{avg}}$ and $V_{total_{avg}}$ are the average dihedral and total potential energies extracted from short conventional MD simulations.

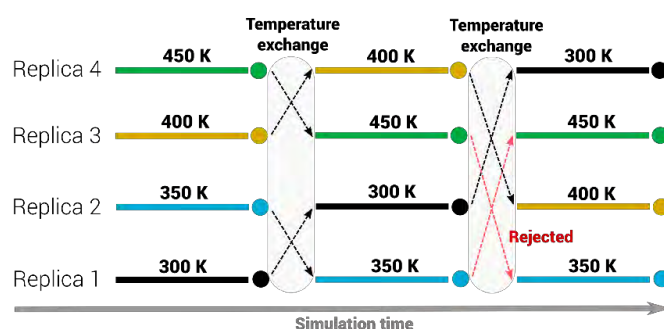
STEERED MOLECULAR DYNAMICS

Steered molecular dynamics (SMD) involves applying external forces to a system to explore its mechanical responsiveness. This approach relies on user defined forces most accurately applied given prior knowledge of the conformational state of interest. This method can be used then to drive a physical process such as ion diffusion, conformational changes and many other applications. By integrating the force over time (or distance), a generalized work can be computed using the so-called Jarzynsky equality¹⁷. This method states that the free energy difference between two states A and B can be calculated as in Eq. 2.12.

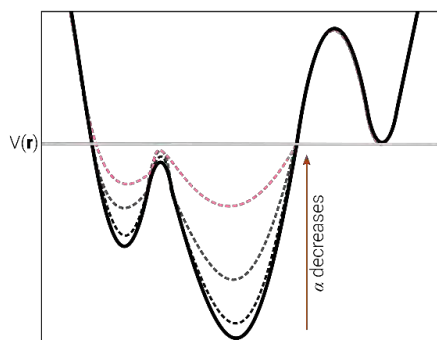
$$e^{-\frac{\Delta G}{k_B T}} = \langle e^{-\frac{W}{k_B T}} \rangle_A \quad (2.12)$$

This means that by computing the work between the two states in question, and averaging over the initial state, equilibrium free energies can be extracted from non-equilibrium dynamics.

REMD



aMD



SMD

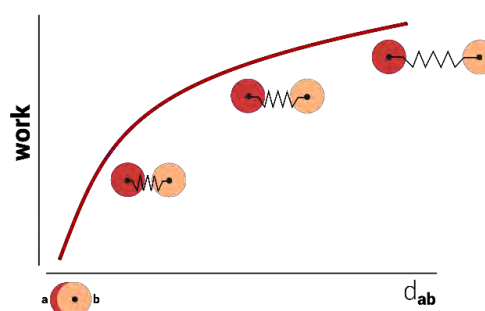


Figure 2.4. Comparison between molecular mechanics-based methods: replica-exchange molecular dynamics (REMD) exchanges temperature between replicas during the simulation with a certain probability to be accepted; accelerated molecular dynamics (aMD) softens the potential energy barriers as α increases; steered molecular dynamics (SMD) computes the work necessary to separate two atoms (**a** and **b**) at a certain distance d_{ab} .

2.2.5. SIMPLIFYING THE MODELS: ELASTIC NETWORKS

Over the last decade, elastic network models (ENMs) have emerged as an alternative to MD techniques for the study of macromolecular dynamics.^{18–24} At its core, ENMs provide a simplified representation of the potential energy function of a system near equilibrium. The nodes of the

network are the building blocks – atoms, or nucleotides, or amino acids – from which the system is composed. Each node is usually represented as a point particle, and the edges of the networks, or the springs joining these nodes, represent harmonic restraints on displacements from the equilibrium structure of the macromolecule. Thus, ENMs provide an intuitive and quantitative description of behavior near equilibrium: the starting or equilibrium conformation resides at the bottom of the harmonic potential well; any deviations from equilibrium will increase the energy and result in a linear net force directed toward restoring the system to its lowest energy state.²⁵

The most attractive feature of ENMs is that they provide a wealth of conformational information at low computational cost. Construction of the elastic network is a matter of defining and linking nodes if information about the structure is available. The standard technique for determining dynamics or statistical distributions from an ENM is to conduct a mode decomposition using spectral graph theory and methods (Gaussian Network Model, GNM)²⁶ or normal mode analysis (NMA)²⁶ with uniform harmonic potentials, both of which provide analytical solutions to the equations of motions without sampling the conformational space.

The most common ENMs are anisotropic network models (ANM)²⁷ that use the $3N$ mass-weighted coordinates of the nodes as generalized coordinates defined as follows: $\mathbf{r} = (\Delta x_1, \Delta y_1, \Delta z_1, \dots, \Delta x_N, \Delta y_N, \Delta z_N)$, where $\Delta x_i = x_i - x_i^0$ is the x -component of the displacement of node i from its equilibrium position \mathbf{r}_i^0 . An interaction matrix \mathbf{K} is the $3N \times 3N$ Hessian matrix, \mathcal{H} , of mixed second derivatives of the potential with respect to the coordinates of the residues. This Hessian matrix might be thought of as an $N \times N$ hypermatrix of 3×3 matrices, each of which describes the energetic contribution from the interaction of two nodes. Each element of \mathcal{H} can be computed from the potential energy function (Eq. 2.13):

$$V = \frac{1}{2} \sum_{ij} \gamma_{ij} (R_{ij} - R_{ij}^0)^2 \quad (2.13)$$

where γ_{ij} is the spring constant between nodes i and j , R_{ij} is their distance, and R_{ij}^0 is their equilibrium distance. The second derivatives of the potential function have the general form (Eq. 2.14):

$$\frac{\partial^2 V}{\partial x_i \partial y_j} = -\frac{\gamma_{ij}(x_j - x_i)(y_j - y_i)}{R_{ij}^2} \quad (2.14)$$

where x_i and y_j are the x - and y -coordinates of nodes i and j . The off-diagonal super-elements of \mathcal{H} are (Eq. 2.15):

$$\mathcal{H}_{ij} = -\frac{\gamma_{ij}}{R_{ij}^2} \begin{bmatrix} x_{ij}^2 & x_{ij}y_{ij} & x_{ij}z_{ij} \\ x_{ij}y_{ij} & y_{ij}^2 & y_{ij}z_{ij} \\ x_{ij}z_{ij} & y_{ij}z_{ij} & z_{ij}^2 \end{bmatrix} \quad (2.15)$$

and the diagonal super-elements satisfy (Eq. 2.16):

$$\mathcal{H}_{ii} = \sum_{j,j \neq i} \mathcal{H}_{ij} \quad (2.16)$$

Diagonalization of \mathcal{H} yields $3N-6$ normal modes, each of which has a 3-vector component for every mode. The remaining 6 modes have zero eigenvalue and correspond to rigid-body rotations and translations of the system. Notice that the spring constants γ_{ij} are the only adjustable parameters in this model, and a variety of methods are used to select their values.

2.3. COMPUTER-AIDED DRUG DESIGN

The start of intense interest in the potential for computer-aided drug design (CADD) is set on October 5th, 1981, when *Fortune* magazine published a cover article entitled the “Next Industrial Revolution: Designing Drugs by computer at Merck”.²⁸ In the past decades, CADD has emerged as a way to significantly decrease the number of compounds necessary to screen while retaining a good level of lead compound discovery. Compounds predicted to be inactive can be skipped, and those predicted to be active can be prioritized. For this reason, CADD is capable of increasing the hit rate of novel drug compounds because it uses a much more targeted search than traditional high-throughput screening (HTS) and combinatorial chemistry.

2.3.1. DRUG DESIGN IN THE DISCOVERY PIPELINE

CADD not only aims to explain the molecular basis of therapeutic activity of compounds but also to predict possible derivatives that would improve activity. In a drug discovery campaign, CADD is usually used for three major purposes: (1) filter large compound libraries and predict active compounds that can be tested experimentally; (2) guide optimization of lead compounds, by increasing its affinity or optimizing drug metabolism and pharmacokinetics properties; (3) design novel compounds, either by *de novo* design with sequential functional groups or by piercing together fragments into novel chemical entities.

CADD can also be classified into two major categories: structure-based drug design (SBDD) and ligand-based drug design (LBDD). SBDD relies on the knowledge of the biological target's structure to calculate interaction energies for the tested compounds, whereas LBDD exploits the knowledge of known active and inactive compounds through chemical similarity searches or quantitative structure-activity relationships (QSAR). On the one hand, SBDD is generally preferred where high-resolution structure data of the crystallized target is available. On the other hand, LBDD is preferred when no or little structural information is available (e.g. membrane proteins).²⁹

2.3.2. STRUCTURE-BASED DRUG DESIGN

The core hypothesis of SBDD is that molecule's ability to interact with a specific target and exert a desired biologic effect depends on its ability to favorably interact with a particular binding site on that target. Thus molecules with similar favorable interactions will exert similar biological effects. For this reason novel compounds can be elucidated through careful analysis of target's binding sites. However, structural information about the target is a prerequisite for any SBDD project. Advances in genomics and proteomics have led to the discovery of a large number of candidate drug targets. The usage of X-ray crystallography and NMR spectroscopy has led to the elucidation of a number of 3D structures of human and pathogenic proteins since the early 80s. For instance, the Protein Data Bank (PDB) has over 108,000 protein and nucleic acid structures (May 2015).³⁹ Nonetheless, drug discovery campaigns have sped up the discovery process and have led to the development of several clinical drugs. SBDD computational methods allow fairly rapid screening of large binders through modeling and/or simulation and visualization techniques.

HOMOLOGY MODELING

An experimentally determined target structure using either X-ray crystallography or NMR techniques and deposited in the PDB is the ideal starting point for conventional SBDD. However, in the absence of experimentally determined structures, several successful screening campaigns have been reported based on comparative models of target proteins. Homology modeling is a specific type of comparative modeling in which the template and target share the same evolutionary origin. Thus, similar sequences may have a similar structure. This process involves the following steps: (1) identification of related targets to serve as template structures, (2) sequence alignment of the target and template, (3) copying coordinated for confidently aligned regions, (4) constructing missing atom coordinates of target structure, and (5) model refinement and evaluation.

BINDING SITE IDENTIFICATION

Target-ligand interaction is a prerequisite for drug activity. Often possible binding sites for small molecules are known from co-crystal structures of the target or a closely related target with a natural or non-natural ligand. However, the ability to identify putative binding sites on proteins and nucleic acids is important if the binding site is unknown or if new binding sites are to be identified. For this reason, computational methods for binding site identification are often used. These methods are generally classified in three major classes: (1) geometric algorithms to find shape concave invaginations in the target, (2) methods based on energetic considerations, and (3) methods considering dynamics of the structure. The latter is based on the premise that the dynamic nature of biomolecules sometimes makes it insufficient to use a single static structure to predict putative binding sites. For instance, many proteins regulate their structure by binding an effector molecule at a site other than the active site. Thus, multiple conformations of target are often used to account for structural dynamics of the target. For instance, MD simulations can be used for obtaining an ensemble of structures which may help identify potential binding sites.

Three basic methods are commonly used to represent target and ligand structures: atomic, surface, and grid representations. Atomic representation of the surface target is usually used when evaluation of potential energy functions - derived from a particular force field - is important. Surface methods represent the topography of molecules using geometric features. This surface is generated by mapping part of van der Waals surface of atoms that is accessible to a probe sphere. For the grid representation, the target is encoded as physicochemical features of its surface.

Molecular docking is a method which predicts the preferred orientation of one ligand to the target when bound to each other to form a stable complex. The 'recognition' method depends on the molecular representation. For instance, docking may be guided by a complementary alignment of ligand and binding site surfaces. Earliest implementation of DOCK³¹ software used a set of non-overlapping spheres represent invaginations of target surface and the surface of the ligand. Geometric matching begins by systematically pairing one ligand sphere a_1 with one receptor sphere b_1 . This is followed by pairing a second set of spheres, a_2 and b_2 . This move is accepted if the change in atomic distances is less than an empirically determined cutoff value, which specifies the maximum allowed deviation between ligand and receptor internal distance. Grid methods digitize molecules using a 3D discrete function that distinguishes the surface from the interior of the target molecule. Molecules are scanned in several orientations in three-dimensions and the extent of overlap between molecules is determined using a correlation function. Other software, such as AUTODOCK³², are based on the atomic representation hence a scoring function based on an atomic force field is required.

All docking methods can be classified as rigid-body docking and flexible docking depending on the degree of flexibility of the target during the docking process. Rigid body docking methods consider only static complementarities between ligand and target and ignore flexibility and induced-fit recognition models. Other algorithms consider several possible conformations of ligand and/or target according to the conformational selection paradigm. Some of the most popular flexible approaches include systematic enumeration of conformations, MD, Monte Carlo search algorithms with Metropolis criterions (MCM), and genetic algorithms. Genetic algorithms introduce flexibility through recombination of 'parent' to 'child' conformations. During this evolutionary process the best scoring conformations are kept and transmitted for another round of recombination, hence the best possible solutions evolve by retaining favorable features from one generation to another (elitism). The conformation or pose of the ligand is described by state variables, i.e., the genotype. This may include a set of values describing translation, orientation, conformation, number of hydrogen bonds, etc. The resulting structural model is called the phenotype. Moreover, the genetic operators may swap regions of parent's genes or randomly mutate the value to give rise to new individuals.

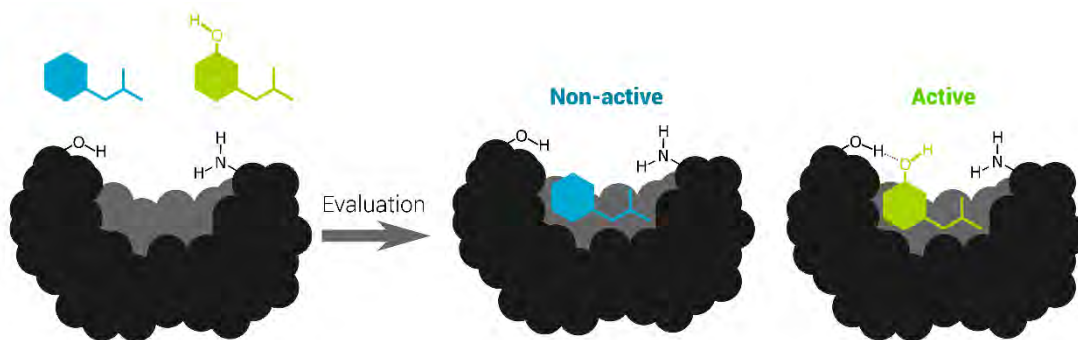


Figure 2.5. Schematic illustration of a docking process. Rigid/flexible ligands (blue and green) are docked into a rigid/flexible target. The best scoring compounds will be selected.

The last component required for a molecular docking evaluation is the scoring function. Due to the typical high number of compounds used in virtual screening campaigns, docking applications need to rapidly assess whether the complex ligand-target is feasible through a scoring function which approximates the energy of interaction. These scoring functions can be classified into four types: (1) force field scoring functions, (2) empirical scoring functions, (3) knowledge-based scoring functions, and (4) consensus scoring functions. For instance, DOCK uses the AMBER force fields in its complex evaluation. Standard force fields are however biased to select highly charged ligands due to the solvent treatment during calculations. For this reason, terms from empirical scoring functions are often added to the force field functions to treat solvation and electronic polarizability bias. That is the case of AUTODOCK which scores the complex by using a semi-empirical force field to evaluate the contribution of water surrounding the complex in the form of empirical enthalpic and entropic terms. Empirical scoring functions fit parameters to experimental data. For example, VINA³³ software expresses binding energy as a weighted sum of explicit hydrogen bond interactions, hydrophobic contact terms, desolvation effects, and entropy. Knowledge-based scoring functions use the information contained in experimentally determined complex structures. This function uses the information contained in experimentally determined complex structures and are formulated under the assumption that interatomic distances occurring more often than average distances represent favorable contacts. An example for this type of scoring function is LigandRNA³⁴. More recently, consensus scoring functions have demonstrated to improve accuracy of prediction through combination of advantages of basic scoring functions. The results obtained from each scoring function can be then weighted, used in a voting strategy, or ranked by average normalized score values.

PHARMACOPHORE MODELING

Although pharmacophore modeling can be classified either as a SBDD or LBDD method, structure-based strategies are often more exhaustive. A pharmacophore model summarizes steric and electronic features needed for optimal interaction of a ligand with a target. Most common features to define pharmacophores are hydrogen bond acceptors, hydrogen bond donors, basic groups, acidic groups, partial charge, aliphatic hydrophobic moieties, and aromatic hydrophobic moieties. On the one hand, structure-based pharmacophore methods are developed based on an analysis of the target binding site or based on a target-ligand complex structure. On the other,

ligand-based pharmacophore assume that the ligand must contain specific features that confer its specificity and potency. Pharmacophore models are a fast and reliable method for large database screening. Those molecules that contain all these pharmacophore specific features – or pharmacophoric keys – in a particular three-dimensional disposition will be assumed to have similar or higher activity to the template compound. Other dimensions may be considered such as the conformational landscape of the molecules, solvent effects or conformational rearrangements inside the target, which lead to multi-dimensional pharmacophoric features.

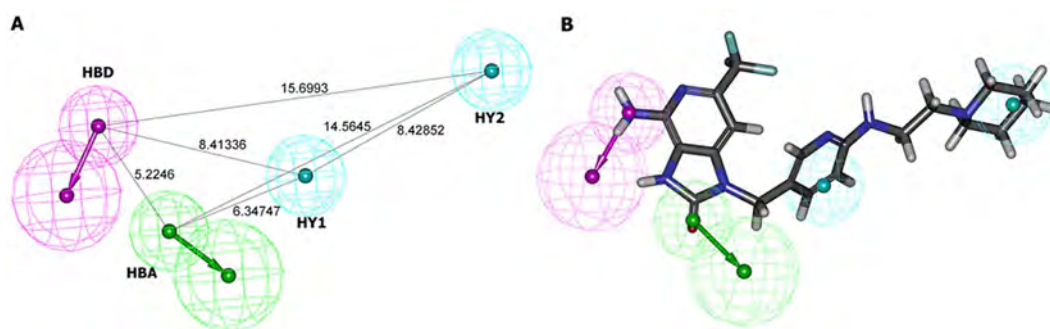


Figure 2.6. (A) Representation of a set of pharmacophoric keys disposed in a three-dimensional space with hydrogen bond donor (HBD), hydrogen bond acceptor (HBA) and hydrophobic groups (HY) pharmacophoric keys. (B) Superposition of a new chemical entity that fits into the pharmacophore features.

2.3.3. LIGAND-BASED DRUG DESIGN

The LBDD approach involves the analysis of ligands known to interact with a target of interest. LBDD methods use a set of reference structures and analyze their 2D or 3D structures. The main goal is to represent these compounds in such a way that the physicochemical properties important for the desired interactions/activity are retained. It is considered an indirect approach in drug discovery in that it does not necessitate prior knowledge of the target structure. There exist two main approaches in LBDD: (1) selection of compounds based on chemical similarity to known actives using a similarity metric, or (2) construction of a QSAR model that predicts biologic activity from chemical structure.

MOLECULAR DESCRIPTORS

The main difficulty of LBDD techniques is the method for describing features of small molecules using computational algorithms. Molecular descriptors can be structural as well as physicochemical and can be classified on multiple levels of complexity. These descriptors can contain information such as molecular weight, geometry, volume, surface area, rotatable bonds, atom types, electronegativities, symmetry, topological charge indices, functional group compositions, and many others. All of them are generated through knowledge-based, graph-theoretical methods, molecular mechanical, or quantum mechanical tools and are generally classified according to the dimensionality of the chemical representation: 1D, scalar, physicochemical properties such as molecular weight; 2D, molecular contribution-derived

descriptors; 2.5D, molecular configuration-derived descriptors; 3D, molecular conformation-derived descriptors.

SIMILARITY SEARCHES

In many situations, 2D similarity searches of databases are performed using structural information from first generations hits, which can lead to modifications that can be evaluated computational or ordered for *in vitro* testing. Similar compounds usually share a central backbone, also called scaffold or chemotype. However, the notion of chemical similarity is one of the most important concepts in chemoinformatics. Chemical similarity can be either described by distance metrics (e.g. Euclidean and Eclidean squared distances), fingerprint-based methods or common subgraph-based measures. In particular, fingerprints consider all parts of the molecule equally and avoid focusing only on parts of a molecular that are thought to be most important for activity. They consist of bit string representations of molecular structure and/or properties and usually encode various molecular descriptors as predefined bit settings as 1 or 0 (where 1 means descriptor is present or 0 if not). This allows chemical identity to be unambiguously assigned the presence or absence of features. Such binary attributes can be then compared between molecules by using metrics such as Tanimoto coefficient (Eq. 2.17). The Tanimoto coefficient uses the ratio of the intersecting set (c) of the attributes of two molecules (a and b) to the union set as the measure of similarity.

$$T(a, b) = \frac{N_c}{N_a + N_b - N_c} \quad (2.17)$$

QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP

Quantitative structure-activity relationship (QSAR) models describe a mathematical relation between structural attributes of the ligands and target response of a set of chemicals (e.g. IC₅₀, EC₅₀, potency, etc.) Classical QSAR involves the correlation of various electronic, hydrophobic, and steric features with biological activity. The general workflow of QSAR in drug discovery is to first collect a group of active and inactive ligands and calculate a set of descriptors for each compound (training set). A mathematical models is then generated to identify the relationship between those descriptors and their experimental activity, maximizing the predictive accuracy. Finally, the constructed model is applied to predict activity for a library of untested compounds that were encoded with the same descriptors (test set). Therefore, success of QSAR depends on the quality of the initial set of the training set compound database and the choice of the descriptors.

One of the most important considerations of QSAR models is that they will be dependent on the sampling space of the initial set of compounds with known activity, the chemical diversity. Thus, divergent scaffolds or functional groups not represented in the training set will not be represented in the final model, and any potential hits within the test set to be screened that contain these groups will likely be missed. Hence, extensive formal validation is required. A common methods to internally validate the proposed QSAR model is cross-validation. This process repeats the regression many times on subsets of data. Usually, each molecule is left out once in turn, and the correlation coefficient is computed using the predicted values of the missing

molecule (leave-one-out method, LOO). Sometimes, more than one molecule (leave-many-out, LMO) are left out at a time. The cross-validated correlation coefficient is used as a diagnostic tool to evaluate the predictive power of an equation.

As a mathematical model, linear and nonlinear QSAR models can be chosen. Linear models include multivariate linear regression analysis (MLR), principal component analysis (PCA), or partial least square analysis (PLS). MLR computes biological activity as a weighted sum of descriptors or features. PCA increases the efficiency of MLR by extracting information from multiple variables into a smaller number of uncorrelated variables (principal components, PCs). PLS combines MLR and PCA and extracts the dependent variable (biological activity) into new components to optimize correlations. Unfortunately, the major drawback is that chemical structure often relates with biological activity in a nonlinear fashion.

The most representative nonlinear model is based on artificial neural networks (ANNs). This method originates in efforts to produce computer models of the information processing that takes place in the brain. The neural network learns the relationship between descriptors and biological activity through iterative prediction and improvement cycles. ANN techniques may be classified into supervised and unsupervised methods. Supervised methods are trained by giving them sets of inputs patterns and associated target patterns. Through the iterative process, the internal representation of the data is modified until the predicted results are as close as desired to the targets. In contrast, unsupervised methods are performed in the absence of any *a priori* targets. Networks of this kind may be used to give information on clusters of compounds based solely on the coordinates of the compounds in the feature space of its variables.

The architecture of a neural network is determined by the way in which the outputs of the neurons are connected to the other neurons. The most popular network used in ANN-QSAR applications is the multi-layer feed-forward network. In this type of architecture, the neurons are arranged into groups called layers – an input layer, an output layer and a number of hidden layers. The input layer should contain as many neurons as variables in the data set; the output layer's neurons number should coincide with the number of responses. However, the number of hidden layers and the number of neurons per layer required is rather inconsistent when it comes to characterizing networks and a trial and error procedure is usually required. Each pair of neurons is connected with a strength, or weight, and biases to each neuron which are both determined during the ANN-QSAR training phase (figure 2.7). The input variables are multiplied by the connection weights w_{ij}^k between the input and hidden layer. The hidden neurons sum the weighted signals from the input neurons and then project this sum on a non-linear activation function (f_h). The resulting activations of the hidden neurons are weighted by the connections between the hidden and output neurons and sent to the output neuron(s). The output neurons also perform a summation and projection on its activation function f_o . The output of these neurons is the estimated response.³⁵ The non-linear model can be improved following an iterative process that optimizes a cost function C (usually, the root-mean-square error, RMSE), such as the backpropagation algorithm.

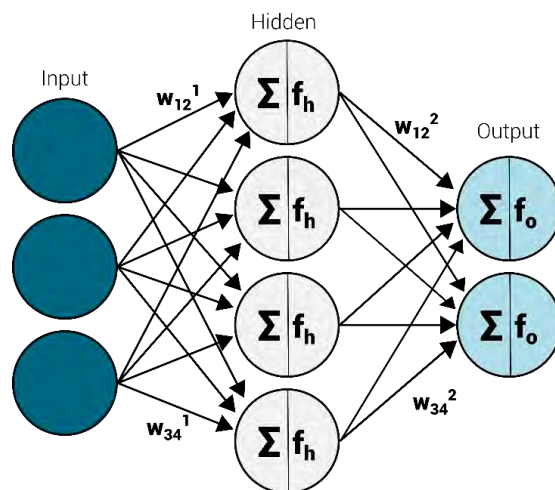


Figure 2.7. Representation of an artificial neural network with 3 inputs, 1 hidden layer and 2 outputs. Neurons are weighted by a w_{ij}^k factor and project the sum over an activation function f .

REFERENCES

1. Schlick, T. *Molecular Modeling and Simulation: An Interdisciplinary Guide*. (2010). doi:10.1007/978-1-4419-6351-2
2. Hehre, W. J. *A Guide to Molecular Mechanics and Quantum Chemical Calculations*. (2003).
3. Van Gunsteren, W. F. *et al.* Biomolecular modeling: Goals, problems, perspectives. *Angewandte Chemie - International Edition* **45**, 4064–4092 (2006).
4. Kukol, A. *Molecular Modeling of Proteins*. (2008). doi:10.1007/978-1-59745-177-2
5. D.A. Case, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B., X. Wu and, V. B. & D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B., X. W. and P. A. K. AMBER 14. (2014).
6. Toukmaji, A., Sagui, C., Board, J. & Darden, T. Efficient particle-mesh Ewald based approach to fixed and induced dipolar interactions. *J. Chem. Phys.* **113**, 10913–10927 (2000).
7. Sagui, C., Pedersen, L. G. & Darden, T. a. Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations. *J. Chem. Phys.* **120**, 73–87 (2004).
8. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. . C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
9. Sugita, Y. & Okamoto, Y. Replica exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
10. Roe, D. R., Bergonzo, C. & Cheatham, T. E. Evaluation of enhanced sampling provided by accelerated molecular dynamics with hamiltonian replica exchange methods. *J. Phys. Chem. B* **118**, 3543–3552 (2014).
11. Butterfoss, G. L. *et al.* De novo structure prediction and experimental characterization of folded peptoid oligomers. *Proceedings of the National Academy of Sciences* **109**, 14320–14325 (2012).
12. Nevin Gerek, Z. & Banu Ozkan, S. A flexible docking scheme to explore the binding selectivity of PDZ domains. *Protein Sci.* **19**, 914–928 (2010).
13. Bergonzo, C. *et al.* Multidimensional replica exchange molecular dynamics yields a converged ensemble of an RNA tetranucleotide. *J. Chem. Theory Comput.* **10**, 492–499 (2014).

14. Isard, B. P. C. Theory and Practice in Replica-exchange Molecular Dynamics Simulation. (2008).
15. Zhang, Y., Zhao, X. & Mu, Y. Conformational transition map of an RNA GCAA tetraloop explored by replica-exchange molecular dynamics simulation. *J. Chem. Theory Comput.* **5**, 1146–1154 (2009).
16. Pierce, L. C. T., Salomon-Ferrer, R., Augusto F. De Oliveira, C., McCammon, J. A. & Walker, R. C. Routine access to millisecond time scale events with accelerated molecular dynamics. *J. Chem. Theory Comput.* **8**, 2997–3002 (2012).
17. Jarzynski, C. A nonequilibrium equality for free energy differences. *II* (1996). doi:10.1103/PhysRevLett.78.2690
18. Fuglebakk, E., Tiwari, S. P. & Reuter, N. Comparing the intrinsic dynamics of multiple protein structures using elastic network models. *Biochim. Biophys. Acta - Gen. Subj.* **1850**, 911–922 (2015).
19. Yang, L., Song, G. & Jernigan, R. L. Protein elastic network models and the ranges of cooperativity. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 12347–12352 (2009).
20. Lyman, E., Pfaendtner, J. & Voth, G. a. Systematic multiscale parameterization of heterogeneous elastic network models of proteins. *Biophys. J.* **95**, 4183–4192 (2008).
21. Lezon, T. R., Shrivastava, I. H., Yang, Z. & Bahar, I. Elastic Network Models For Biomolecular Dynamics: Theory and Application to Membrane Proteins and Viruses. *Handb. Biol. Networks* 129–158 (2009).
22. Setny, P. & Zacharias, M. Elastic network models of nucleic acids flexibility. *J. Chem. Theory Comput.* **9**, 5460–5470 (2013).
23. Das, A. *et al.* Exploring the Conformational Transitions of Biomolecular Systems Using a Simple Two-State Anisotropic Network Model. *PLoS Comput. Biol.* **10**, (2014).
24. Romo, T. D. & Grossfield, A. Validating and improving elastic network models with molecular dynamics simulations. *Proteins Struct. Funct. Bioinforma.* **79**, 23–34 (2011).
25. Prof, L., Helms, V., Gu, W. & Park, Y. *Biomolecular Simulations.* **924**, (1998).
26. Bahar, I., Atilgan, a R. & Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **2**, 173–181 (1997).
27. Doruker, P., Atilgan, A. R. & Bahar, I. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins* **40**, 512–524 (2000).
28. Drie, J. H. Computer-aided drug design: The next 20 years. *J. Comput. Aided. Mol. Des.* **21**, 591–601 (2007).
29. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **66**, 334–395 (2014).

30. Bernstein, F. C. *et al.* Protein Data Bank: A Computer-based Archival file for Macromolecular Structures. *J. Mol. Biol.* **185**, 584–591 (1978).
31. Lang, P. T. *et al.* DOCK 6: combining techniques to model RNA-small molecule complexes. *RNA* **15**, 1219–1230 (2009).
32. Morris, G. & Huey, R. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
33. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.* 455–461 (2010).
34. Philips, A., Milanowska, K., Lach, G. & Bujnicki, J. M. LigandRNA: computational predictor of RNA-ligand interactions. *RNA* **19**, 1605–16 (2013).
35. Salt, D. W., Yildiz, N., Livingstone, D. J. & Tinsley, C. J. The use of artificial neural networks in QSAR. *Pestic. Sci.* **36**, 161–170 (1992).

CHAPTER III. STRUCTURAL COMPLEXITY OF THE RNA

3.1. BACKGROUND OF THE STUDY

RNA flexibility is both a blessing and a curse. It creates unique folds or binding pockets suitable for targeting, but small molecules must be identified so that recognize these folds selectively. Moreover, the number of small molecules required to effectively target RNA is higher than the set required to comprehensively target protein cavities. RNA hairpins containing $N \times N$ internal loops are involved in many diseases and they have caught the attention of drug designers.¹⁻⁵ On the one side, RNA does not have clear pockets and cavities, as the ones observed in proteins, hampering efforts to achieve selectivity. On the other side, many investigators provided evidence that repeated internal loops (e.g. rCUG) can be used as unique anchoring points so molecules containing repeated subunits with a spacer achieve better selectivity values.^{6,7} It is therefore clear that having a good description of (i) the structure of the internal loops and (ii) the intrinsic dynamics of the nucleotides involved, would improve the molecular design of RNA targeted drugs.

Computational methods for predicting the motions of proteins have been thoroughly applied, but dynamics of many RNA relevant structures still remains elusive. However, molecular dynamics (MD) require extensive computational resources and are subjected to force field limitations.

The final aim of this thesis is to apply computer-aided drug design (CADD) methods in order to design novel chemical entities for targeting specific RNA targets from a rational perspective. Thus, the essential elements required for a structure-based drug design (SBDD) is a three-dimensional RNA structure of the target and a deep understanding of the dynamics behind it. In this chapter, the dynamic behavior of CUG repeats (rCUG) and AUUCU repeats (rAUUCU) will be explored using molecular mechanics-based techniques such as molecular dynamics (MD), enhanced-sampling MD methods and anisotropic network models. Sections 2 and 3 describe the dynamics of rCUG fragments and its U-U 1x1 internal loop using two distinct computational methods to represent its intrinsic flexibility; particularly, elastic network models (ENM) and MD. In the last section (section 4) the nature of a particular UCU 3x3 internal loop is investigated using computational analyses of X-ray data.⁸

3.2. UNVEILING THE DYNAMICS OF CNG REPEATS

In chapter 1 myotonic dystrophy type 1 (DM1) has been described as a disease which is related with the formation of large non-coding CUG expansions, produced from non-coding CTG repeats. These structures promote an aberrant sequestration of nuclear RNA-binding proteins, mainly MBNL1 muscleblind-like family protein, triggering a deregulation of several RNA splicing events.

The expanded CUG transcripts present a characteristic hairpin motif which is responsible of providing its cytotoxic activity. Loops are considered to be an indispensable topology upon RNA hairpin folding. Among the most important hairpin loops, tetraloops are highly structured and present characteristic signatures. Even though rCUG^{exp} sequences are experimentally observed as long tract hairpins, they are metastable and present a “slippery” nature that determines their secondary structure, which is not unique. For instance, it was observed that r(CUG)₅ repeats do not present any detectable secondary structure while r(CUG)₁₁, r(CUG)₂₁ and r(CUG)₄₉ presented increasing hairpin stability.

Within this scope, the conventional MD (cMD) trajectories of three different CUG containing sequences have been analyzed. A crystallographic double stranded RNA structure reported in literature consisting of CUG repeats has been compared with two purely CUG hairpins models of different lengths, including the loop (Figure 3.1). Following this strategy, the effects of the number of neighboring CUG repeats and the presence of a loop over the non-canonical U-U pairs have been evaluated by using cMD in DM1 structure-based approaches.

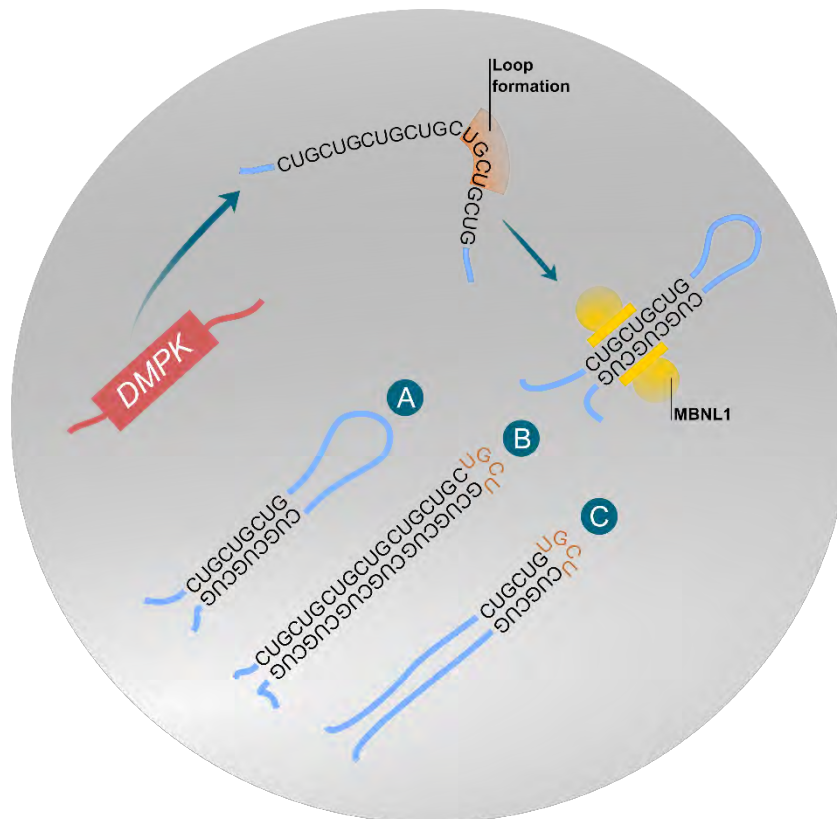


Figure 3.1. *DMPK* mRNA repeated expansions folds into a metastable hairpin and causes DM1 by sequestering MBNL1. The three models proposed in this study are a double stranded ds[(CUG)₆]₂ fragment (A), a purely rCUG^{exp} hairpin rich in CUG repeats (B) and a shortened analogue hairpin (C).

3.2.1. CONFORMATIONAL SAMPLING THROUGH MOLECULAR DYNAMICS

First, a double stranded (ds) ds[(CUG)₆]₂ model system has been studied through molecular dynamics simulations and the dynamic properties of non-canonical U-U pairs have been determined. The simulation was carried out in explicit solvent and minimal net-neutralizing salt using the AMBER ff10 force field⁹ (computational methods are detailed in page 106). This model includes initially a total of six non-canonical base-pairs adopting the stretched U-U wobble conformation. After the cMD simulation a total of seven possible U-U pair conformations were identified by means of cluster analysis which featured different hydrogen bonding patterns. Figure 3.2 (page 76) shows the process of U-U conformational extraction from the cMD trajectory. In fact, the clustering is reduced to five conformations if structural symmetry correction is introduced (conformations **b** - **c**, and **d** - **e** respectively can be considered symmetrical). Among the most representative conformations the presence of 0, 1 or 2 direct hydrogen bonds were observed, some of which formed 1 or 2 water-mediated hydrogen bonds at the same time. These water-mediated hydrogen bonds spend as long as 1.4 ns, considering they are dynamically formed. The population of U-U clustered conformations for each base pair are reported in the table 3.1 (page 76). It can be observed that each pair goes through nearly three or four conformations although some of them are for less than 5% of simulation time. From highest to lowest population cMD results suggest that U-U pairing formation involves 2, 0 or 1 hydrogen bonds, whereas water mediated hydrogen bonds represent only the 2% of total simulation time.

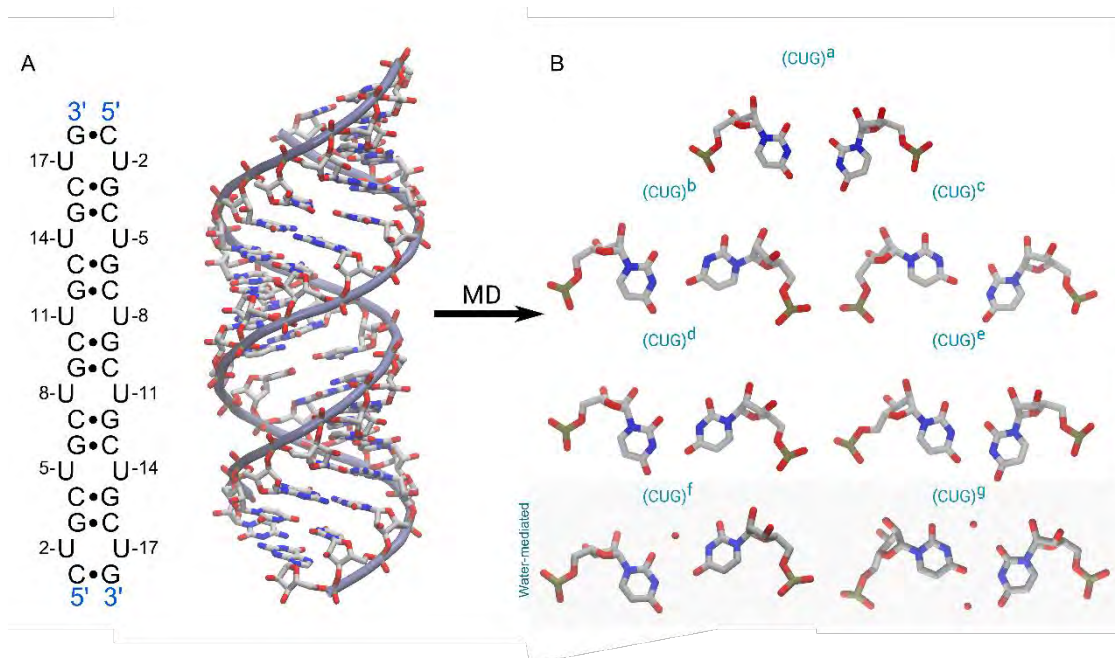


Figure 3.2. (A) Schematic representation and three-dimensional model of $ds[(CUG)_6]_2$. (B) Clustered conformations of non-canonical U-U pairs after the cMD simulation representing 0 ((CUG)^a), two 1((CUG)^b and (CUG)^c), two 2 ((CUG)^d and (CUG)^e) and two water mediated hydrogen bond patterns ((CUG)^f with one water molecule and (CUG)^g with two water molecules).

Table 3.1. Frequency of U-U pairs conformations observed in the $ds[(CUG)_6]_2$ trajectory.							
	Base-paired conformation				Water mediated		
	(CUG) ^a 0HB	(CUG) ^b 1HB	(CUG) ^c 1HB	(CUG) ^d 2HB	(CUG) ^e 2HB	(CUG) ^f 1WHB	(CUG) ^g 2WHB
U2-U17	-	15%	-	71%	10%	4%	-
U5-U14	35%	43%	-	16%	5%	-	1%
U8-U11	52%	13%	12%	-	23%	-	-
U11-U8	77%	-	-	-	20%	3%	-
U14-U5	29%	36%	21%	10%	-	4%	-
U17-U2	3%	1%	-	95%	-	1%	-
Overall	32%	24%		42%		2%	

These hydrogen bond patterns have already been reported in a combined NMR and cMD study of $ds(CCGCUGCGG)_2$ where, after sampling all the possibilities, the authors reported that the lowest-energy motif presented a single hydrogen bond.¹ In contrast to these results, our simulation covers all of the reported U-U conformations, but they are more equally distributed. Interestingly, the terminal U-U pairs of $ds[(CUG)_6]_2$ model form 2 hydrogen bonds during almost all the simulation while the central and adjacent pairs preferentially show the o and r motif respectively. Furthermore, this cMD model system contains 6 CUG successive repeats in each slide with dynamic U-U conformation interconversion, thus a modification of any U-U pair may correlate with other adjacent pairs.

3.2.2. VALIDATION OF THE STRUCTURAL PARAMETERS

Structural parameters of crystallographic structures (refer to figure I.II, page 20) were compared with those obtained from the cMD model. Helical parameters from the different base pairs of the cMD ensemble were obtained using 3DNA software¹⁰ and extracted every 20 ps. Convergence of the helical parameters along the trajectory was evaluated by means of Student's

t-test to assess whether the averages were different from the reference structure 3gm7¹¹ (see table 3.2). The analysis of the overall structure along the cMD trajectory revealed that only opening was statistically significant from the crystal structure at the 95% confidence level.

Table 3.2. Two-tailed t-test analysis of intra- and inter-base pair parameters of the ds[(CUG)₆]₂ trajectory compared to the reference structure (PDB ID 3gm7).

Helical parameter (Å)	Simulation avg/sd	X-ray ds[(CUG) ₆] ₂ value/sd	P-Value
x-displacement	-5.04 ± 3.67	-5.08 ± 2.32	0.96
y-displacement	0.1 ± 3.79	-0.17 ± 1.51	0.76
Shear	-0.17 ± 1.73	0.19 ± 1.42	0.37
Stretch	-0.3 ± 0.87	-0.65 ± 0.48	0.09
Stagger	-0.16 ± 0.67	0.03 ± 0.27	0.23
Shift	-0.06 ± 1.63	0.01 ± 0.87	0.87
Slide	-1.84 ± 0.65	-1.90 ± 0.32	0.70
Rise	3.35 ± 0.48	3.25 ± 0.19	0.35
Helical rise	2.77 ± 1.18	2.49 ± 0.56	0.32
Helical parameter (°)			
Inclination	13.78 ± 16.79	16.98 ± 10.08	0.42
Tip	1.61 ± 14.24	-0.29 ± 4.93	0.57
Buckle	0.40 ± 11.98	-0.04 ± 4.67	0.87
Propeller	-9.26 ± 11.78	-10.47 ± 3.84	0.66
Opening	2.22 ± 17.19	-8.03 ± 9.83	0.01*
Tilt	-0.80 ± 6.47	0.33 ± 3.07	0.46
Roll	7.03 ± 7.3	9.16 ± 4.89	0.21
Twist	29.88 ± 10.99	32.77 ± 10.38	0.27
Helical twist	32.42 ± 10.2	34.54 ± 9.99	0.38

Further non-canonical U-U pair structural analysis was performed in accordance with Coonrod et al. observations.¹² The authors noticed that six U-U conformations were present in all X-ray and NMR available structure, if classified by their number of hydrogen bonds and inclination (not inclined, or inclined towards the major or minor groove). This six types of U-U pairs are named following previously established criteria: type I non-canonical U-U pairs contains 2 hydrogen bonds, with a shortened Cr'-Cr' distance and inclined towards the minor groove. Type II forms 1 hydrogen bonds and is also inclined towards the minor groove. Type III do not state significant inclination and do not form any hydrogen bond. The most observed conformation is type IV which forms 1 hydrogen bond and inclines towards the major groove. Types V and VI also inclines towards major groove but contains 2 or 0 hydrogen bonds respectively. According to this description, a classification was performed using the average structure of the cMD trajectory, which would reproduce the most relevant conformation for each pair (see table 3.3, page 78). A qualitative analysis suggests that each of the six average U-U pairs represent one of those types. The predominance of these substates has been thoroughly discussed in literature. Notice that the cluster analysis previously described captured two additional conformations (1 or 2 water mediated hydrogen bonds) which are not reflected into the average structure due to its low residence time.

3.2.3. SIMULATION OF rCUG HAIRPIN MODELS

Instead of evaluating the stability of the tetraloop, the interest of this study was to describe the structural effects over the hairpin stem. Since the double stranded model showed that each non-canonical U-U pair preferred a different conformation depending on the sequence position, a different sampling conformational distribution was expected.

Table 3.3. Comparison of structural parameters of U-U mismatches. The same tabular format as Coonrod et al. has been used for comparison purposes (Coonrod et al. (2012). Biochemistry, 51, 8330–8337).

Structure	Pair	C1'-C1' (Å)	1 st HB (Å)	2 nd HB (Å)	λ I (°)	λ II (°)	Incline	Shear (Å)	Stretch (Å)	Stagger (Å)	Buckle (°)	Propeller (°)	Opening (°)	Type
ds[(CUG) ₆] ₂ MD	U2-U17	8.8	2.8	3.1	41.0	74.4	minor	-2.56	-1.92	0.10	2.69	-16.31	4.74	I
	U5-U14	10.1	3.1	-	58.3	44.3	major	1.25	-0.85	-0.28	-4.48	-9.02	-9.66	IV
	U8-U11	9.9	-	-	56.3	56.1	none	-0.09	-0.62	-0.27	4.10	-9.16	-0.10	III
	U11-U8	10.8	-	-	73.9	49.2	major	0.91	1.58	-1.00	14.69	-13.68	27.94	VI
	U14-U5	10.6	2.8	-	53.1	63.9	minor	0.06	-1.44	-0.36	-8.97	-10.19	3.60	II
	U17-U2	8.8	2.8	3.0	75.4	41.8	major	2.54	-1.90	0.14	-1.66	-15.64	6.53	V
r(CUG) ₁₆	U5-U44	10.5	2.9	-	38.2	52.2	minor	-1.27	-0.97	-1.01	3.32	-16.82	-21.90	II
	U8-U41	8.8	2.8	3.0	42.6	74.4	minor	-2.42	-1.87	0.11	-1.00	-13.36	6.08	I
	U11-U38	10.2	2.9	-	64.2	30.8	major	2.93	-1.48	-0.12	-3.35	-12.40	-17.36	IV
	U14-U35	8.6	2.8	-	75.7	44.8	major	2.38	-1.91	0.05	7.10	-17.21	9.41	IV
	U17-U32	8.7	2.9	-	44.4	76.0	minor	-2.43	-1.88	0.13	-6.31	-17.38	9.19	II
	U20-U29	8.7	2.8	2.8	74.7	43.9	major	2.31	-1.84	0.33	-6.61	-18.20	8.88	V
r(CUG) ₈	U5-U26	10.9	3.0	-	31.5	12.4	major	3.06	0.01	-0.36	-3.21	10.18	-36.83	IV
	U8U23	10.6	3.1	-	63.3	29.9	major	3.10	-1.24	-0.60	1.56	10.96	-18.98	IV
	U11-U20	9.3	2.8	-	40.8	71.8	minor	-2.32	-1.77	0.56	3.61	-10.21	-0.48	II
3gm7	U2-U17	10.3	2.7	-	34.1	56.7	minor	-2.32	-1.38	-0.18	0.89	-9.98	-21.23	II
	U5-U14	10.6	3.0	-	53.8	33.1	major	2.27	-1.17	-0.36	-6.03	-15.03	-22.06	IV
	U8-U11	10.3	2.7	-	32.4	56.9	minor	-2.50	-1.48	0.45	2.36	-11.27	-21.02	II
	U11-U8	10.0	2.9	-	57.8	40.4	major	1.71	-1.29	-0.44	-1.86	-12.20	-11.93	IV
	U14-U5	10.5	2.7	-	55.9	32.4	major	2.65	-1.29	-0.23	6.92	-11.44	-28.01	IV
	U17-U2	10.3	3.0	-	58.2	39.2	major	2.04	-1.11	-0.15	-4.69	-14.81	-14.80	IV
(CUG) ₂	U2-U34	8.6	2.8	2.8	44.8	75.9	minor	-2.47	-1.89	0.26	-7.12	-16.40	11.60	I
	U5-U31	10.6	2.7	-	28.2	55.8	minor	-2.63	-1.33	-0.10	5.62	-7.12	-25.98	II
Disney (NMR)	U5-U14 (0)	10.7	-	-	71.2	32.0	major	3.57	-0.95	-0.05	-13.48	-9.09	-5.62	VI
	U5-U14 (1)	10.6	2.9	-	61.1	28.9	major	3.03	-1.14	-0.09	-22.07	-6.46	-24.30	IV
	U5-U14 (2)	8.9	2.9	2.9	73.3	48.3	major	2.24	-1.74	0.02	12.56	-17.25	4.58	V
Kiliszek (A-B)	U3-U6	10.7	2.8	-	26.2	57.9	minor	-2.82	-1.19	-0.47	12.51	-8.17	-31.50	II
	U6-U3	10.3	2.9	-	31.6	67.7	minor	-3.03	-1.31	-0.13	14.01	-8.36	-16.37	II
Kiliszek (C-D)	U3-U6	10.9	2.8	-	56.1	22.6	major	2.98	-1.22	-0.44	-4.51	-8.69	-34.42	IV
	U6-U3	10.2	2.6	-	59.2	32.4	major	2.59	-1.51	-0.28	0.57	-15.81	-21.16	IV
Kiliszek (E-E*)	U3-U6	10.6	2.6	-	52.7	29.1	major	2.38	-1.31	-0.48	-4.07	-11.84	-30.56	IV
	U6-U3	10.6	2.6	-	29.1	52.7	minor	-2.38	-1.31	-0.48	4.07	-11.84	-30.56	II
Disney (3SYW)	U8-U14	8.8	3.0	3.0	46.3	72.4	minor	-2.20	-1.70	0.21	0.27	-14.34	10.06	I
	U11-U11	9.9	-	-	50.8	50.8	none	0.01	-1.17	0.07	0.11	-5.73	-8.25	III
	U14-U8	8.8	2.9	3.0	72.5	46.3	major	2.20	-1.71	0.21	-0.76	-14.31	10.00	V
Disney (3SZX)	U8-U14	10.5	2.7	-	53.9	35.3	major	1.85	-1.16	-0.11	-5.24	-8.68	-29.19	IV
	U11-U11	10.0	-	-	46.0	53.0	none	0.14	-0.99	0.18	1.98	-10.74	-0.21	III
	U14-U8	10.5	-	-	58.8	48.1	major	1.30	-0.70	-0.30	-2.99	-11.12	-8.97	VI

Two rCUG hairpin structural models were designed as follows: (1) a r(CUG)₁₆ system representing a predicted purely rCUG hairpin; (2) a r(CUG)₈ structure which contains a shorter CUG repeat system capped by C•G pairs (figure 3.3). Both models were obtained by homology modeling using a UUCG tetraloop as model system template for the loop (see details in page 106).

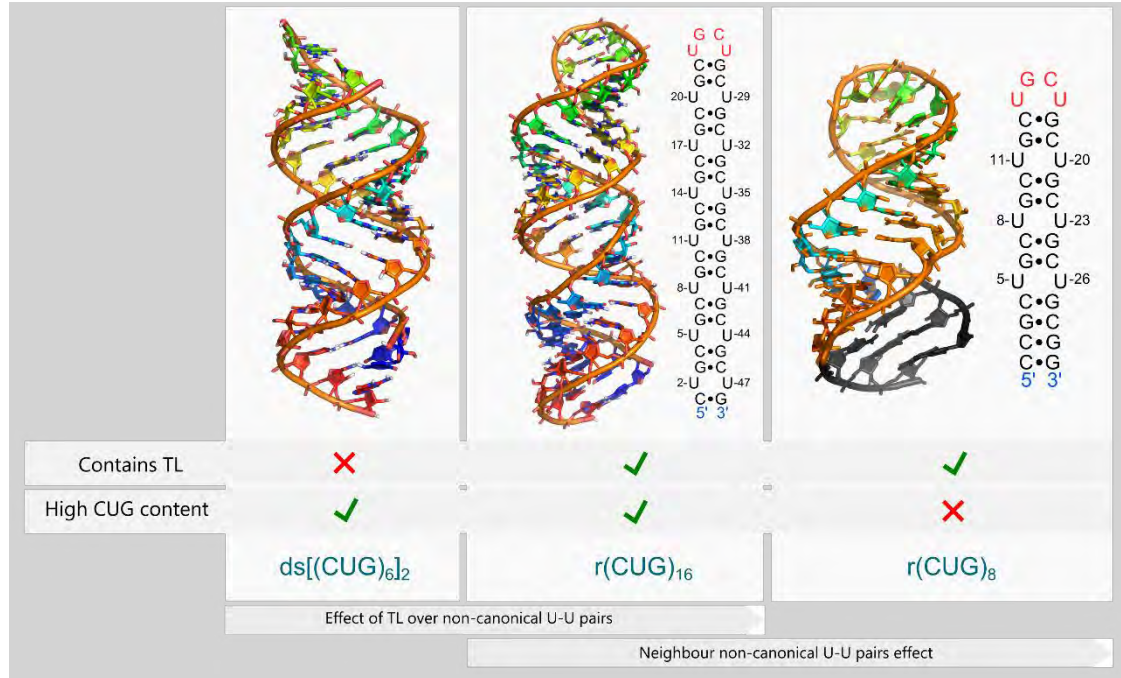


Figure 3.3. Three-dimensional representation of ds[(CUG)₆]₂, r(CUG)₁₆, r(CUG)₈. Each model is classified by presence or not of tetraloop (TL) and high CUG content (considering high as more than 3 repeats). C•G capping region of r(CUG)₈ is black colored. Comparison between ds[(CUG)₆]₂ and r(CUG)₁₆ provides the effect of the TL over the hairpin while the comparison between r(CUG)₁₆ and r(CUG)₈ gives information about the neighbor U-U pairs effect.

Compared to ds[(CUG)₆]₂, r(CUG)₁₆ contains two additional CUG repeats (thus, one additional non-canonical U-U pair) and a cUGCUG tetraloop. According to the previous ds[(CUG)₆]₂ observations, it was hypothesized that the tetraloop would change the U-U hydrogen bonding pattern distribution because one side of the terminal nucleotides is capped. As for r(CUG)₈, the system model contains only three non-canonical U-U pairs besides the tetraloop with a terminal C•G cap. Thus, a combination of a shorter number of repeats, the presence of tetraloop and a capping region would affect the CUG behavior.

Since the interest was focused on the non-canonical U-U pair sampling and its implications on the structural parameters, simulations were stopped once every U-U pair visited each observed conformation at least twice. Comparative analysis of these systems included the previous helical parameter analysis and stiffness comparison using the nonlocal rigid base model. Figure 3.3 roughly illustrates the interpretation of the proposed analysis: on the one side, differences between ds[(CUG)₆]₂ and r(CUG)₁₆ should provide the effect of the tetraloop over the non-canonical U-U pairs. On the other side, r(CUG)₈ is capped by two fragments which involve both the C•G cap and the tetraloop; thus this structure resembles more to the X-ray and NMR structures whose number of CUG adjacent repeats is limited.

3.2.4. U-U CONFORMATIONAL SAMPLING OF r(CUG)_n STRUCTURES

The first notable feature was that both simulations did not provide any tetraloop instability during the simulation in terms of RMSD values ($< 2.5 \text{ \AA}$). However, it is worth of attention the non-negligible effect of the end bases over the r(CUG)₁₆ structure. Attending to the hydrogen bond breaks of the terminal region and α values of U₂ (figure 3.4) the end pair is unable to keep a well-balanced conformation. The α torsion values of U₂ adopts preferentially a gauche g^- ($\sim 300^\circ$) conformational state but it has been observed a non-negligible gauche g^+ ($\sim 60^\circ$) flip that leads to a rotation of the ζ' -end. Thus, this region frays causing an unexpected C ζ' -end inclination towards nucleotide U₃₈ for 6 ns. After that, the end bases recover their original conformation and no further fraying has been observed. However, it is not a priori clear how this effect could be propagated into the adjacent nucleotides. For this reason, the non-canonical U-U pairs were analyzed starting from the second U-U pair (i.e. U₅-U₄₄) in order to avoid undesirable alterations caused by the fraying effect, as previous studies stated.¹³

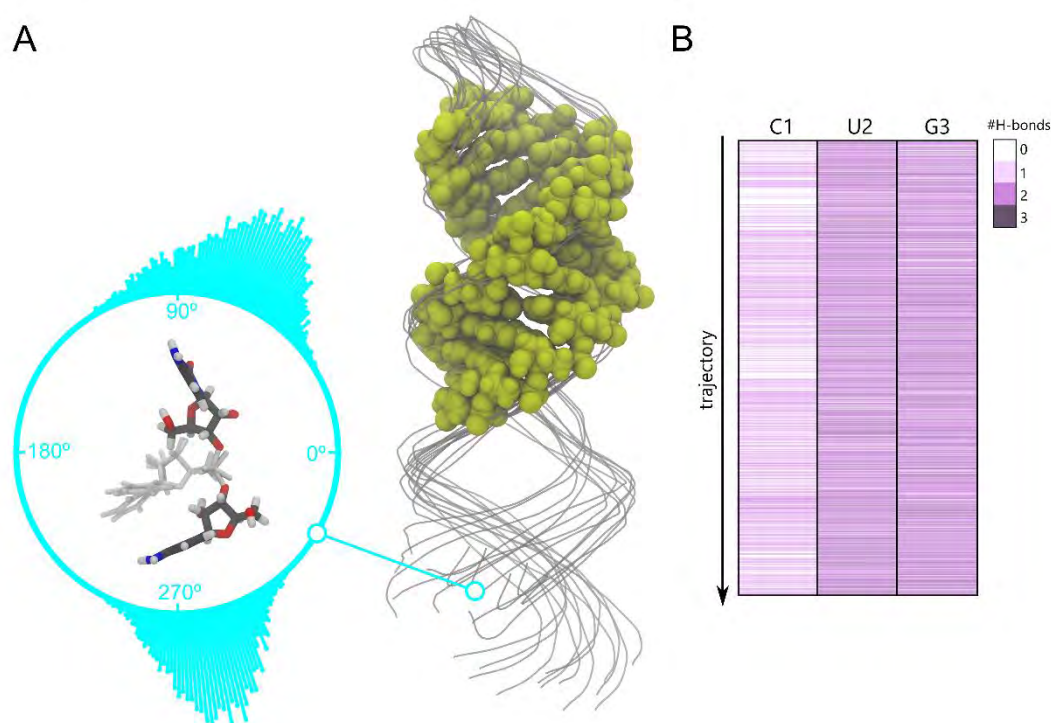


Figure 3.4. (A) Representation of the 6 ns-long fraying and the U₂ conformational states in terms of α values. (B) Number of hydrogen bonds of the first CUG repeats during the trajectory fragment.

U-U pair conformation sampling was performed using the same procedure previously described (see data in table A1 of Annexes, page 199). r(CUG)₁₆ non-canonical pairs prefer a τ hydrogen bond conformation (68% of total simulation time), followed by ζ (22%) and σ (10%) hydrogen bond patterns. In this case, water mediated hydrogen bonds were not found to be significant. All types of U-U conformations were observed except type III and VI because of the low presence of non-canonical pairs in a σ hydrogen bond substrate.

r(CUG)₈ resembles more to experimental structures because of the prevalence of τ hydrogen bond in all the non-canonical U-U hence the average conformations appear as two type II and two type IV which are the most commonly observed types in experimental structures.

In both systems, the CUG stem adopt essentially an A-form helix which is reflected into the helical parameters. For instance, z_p (which is the mean z -coordinate of the two P atoms in the mean reference frame of a dinucleotide step) for $r(\text{CUG})_{16}$ is 1.91 Å, slightly lower than the A-form (2.24 Å), but noticeably different from $r(\text{CUG})_8$ whose value (2.48 Å) is closer to published structures. Helical rise and helical twist for $r(\text{CUG})_{16}$ (2.45 Å / 33.6° respectively) and for $r(\text{CUG})_8$ (2.91 Å / 32.2°) are close to the expected A-form (2.83 Å / 32.5°). Not surprisingly, $r(\text{CUG})_{16}$ values are closer to the reference structure 3gm7 (2.49 Å / 34.5°).

3.2.5. STIFFNESS ANALYSIS OF $r(\text{CUG})_n$ MODELS

In accordance with the importance of conformational flexibility in RNA biological function, the stiffness of these models was studied using Lankaš definition of elastic energy (see Methods section, page 108).¹⁴ Only the central stem region or near the loop was considered which involves C4-G15 to G15-C4 for $ds[(\text{CUG})_6]_2$, C10-G39 to G21-C28 for $r(\text{CUG})_{16}$ and C4-G27 to G12-C19 for $r(\text{CUG})_8$. Diagonal entries of the stiffness matrix correspond to the elastic constants associated to the helical definitions of each residue (see figure 3.5, Annexes table A2, page 200). A straightforward interpretation of these values would be that they are the stiffness constants for deformations when only the corresponding coordinate changes, while the remaining parameters remain in equilibrium.

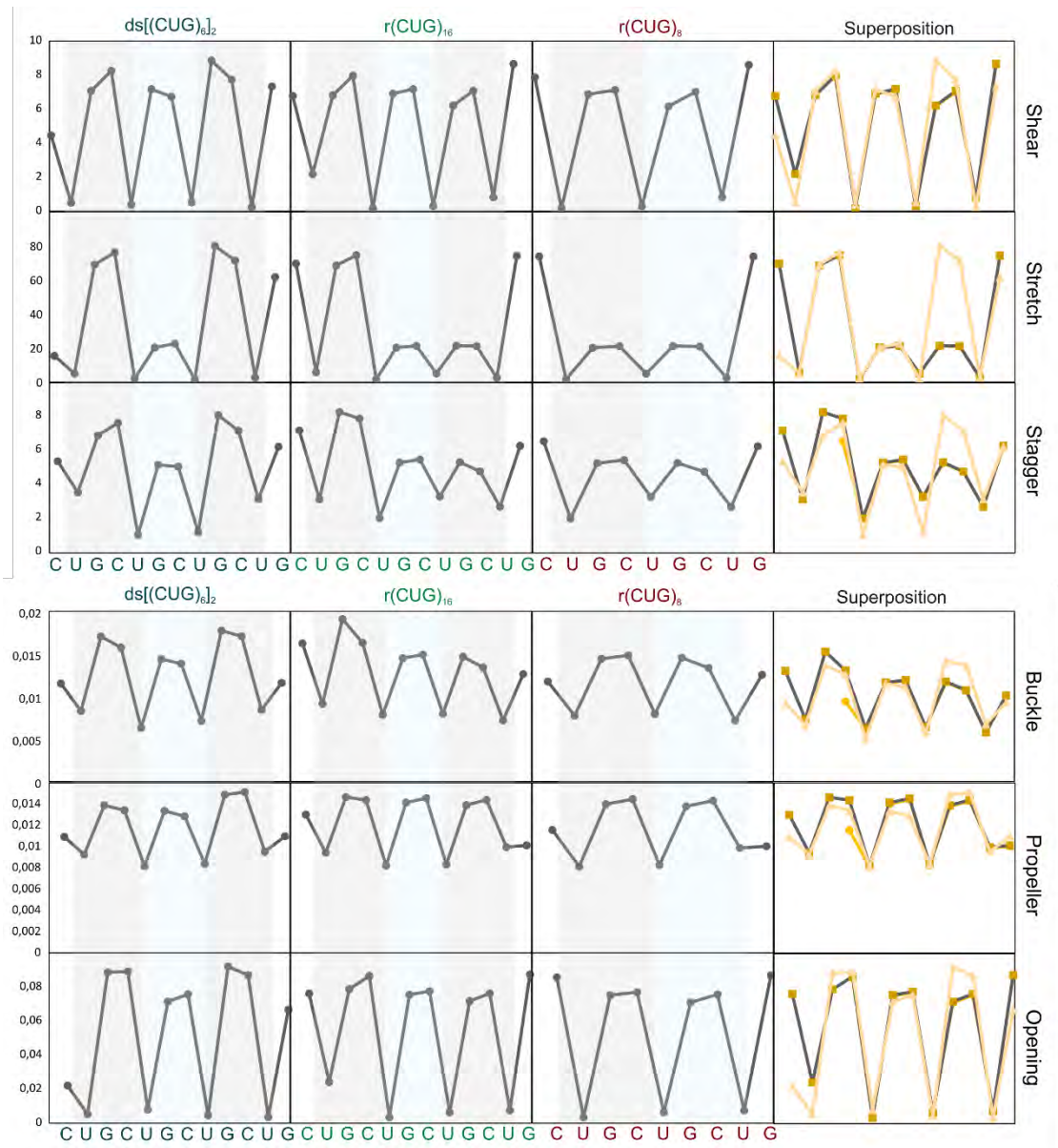
Because the sequence of all models consist mainly of repeated trinucleotides we could expect that the stiffness constant would also be repeated every three positions if no neighbor effects were present. However, $ds[(\text{CUG})_6]_2$ parameters remain constant (specially shear, stretch and stagger) resulted in a typical shape of palindromic sequences, thus the first two repeats are almost mirroring the other two. In this system, the central 5'-UGCU-3' fragment is less stiff except for propeller which remains virtually invariable with respect to the first 5'-UGCU-3' fragment. In good agreement with experimental observations, non-canonical U-U pair constants present very low values if compared to its neighbor base pairs. It has been proved that replacing U-U mismatches with Watson-Crick base pairs almost completely abolishes MBNL1 binding so the relative flexibility of non-canonical U-U pairs and the adjacent C•G pairs may play an essential role in binding transition.¹⁵

Regarding the $r(\text{CUG})_{16}$ structure it is observed a loss of symmetry, especially in shear, which becomes more similar to the expected behavior. If compared with $ds[(\text{CUG})_6]_2$, it is obvious the significant stiffness decrease into the translational constants of the third 5'-UGCU-3' fragment, the one closer to the tetraloop. These results pointed to a flexibility increase in the proximal region of the tetraloop that did not affect the adjacent 5'-GC-3' fragment. On the contrary, non-canonical U-U pairs present higher stiffness constants in the proximal loop region.

The effect of the number of adjacent non-canonical U-U pairs was studied by comparison of $r(\text{CUG})_{16}$ and $r(\text{CUG})_8$ results. The last two 5'-UGCU-3' fragments of $r(\text{CUG})_{16}$ are almost identical to the $r(\text{CUG})_8$ profile. The stiffness of the $r(\text{CUG})_8$ model can be roughly explained by two hypothetic effects: the additional flexibility of the tetraloop and the lower CUG content. The superposition of all three stiffness profiles provides a good convergence of the first two 5'-UGCU-3' fragments of $ds[(\text{CUG})_6]_2$ and $r(\text{CUG})_{16}$ parameters, but not as excellent as the

convergence of the last two fragments of $r(\text{CUG})_{16}$ and $r(\text{CUG})_8$. Indeed, $r(\text{CUG})_{16}$ conserve and merge the characteristic stiffness profile of each system.

Figure 3.5. Stiffness constants along the sequence of the different model fragments and their superposition. The $ds[(\text{CUG})_6]_2$



model exhibits an almost symmetric stiffness profile whereas the inclusion of a tetraloop in model $r(\text{CUG})_{16}$ induces a symmetry loss. The short CUG model ($r(\text{CUG})_8$) present the highest flexible symmetric profile.

3.2.6. CLUSTER ANALYSIS AND HYDRATION EFFECT

A visual inspection of the $r(\text{CUG})_{16}$ simulation suggested the subdivision of the simulation into two possible states according to its backbone conformation, likely due to A-RNA to A'-RNA transition. These possible states have been classified in consonance with two pseudobonds: one from P to C4' and one from C4' to P of the contiguous nucleotide, which are characterized by torsions η ($\text{C4}'_{n-1}-\text{P}_n-\text{C4}'_n-\text{P}_{n+1}$) and θ ($\text{P}_n-\text{C4}'_n-\text{P}_{n+1}-\text{C4}'_{n+1}$). It has been observed that this simple description allows to classify and quantify the different structural motifs and it

also enables the observation of possible conformation changes produced during binding of drugs and/or macromolecules. The classification can be observed by a Ramachandran-analogue plot, as the location of a nucleotide on a η/θ plot corresponds closely with its conformational state.

The density map of the most frequent η/θ states during the cMD dynamics was computed allowing the observation of the two possible conformations including a high populated *t/a*- ($164^\circ/216^\circ$) backbone conformational state with three less frequent conformational substates. A RMSD clustering by the average-linkage algorithm lead to two main clusters with significant differences. In good agreement with the analysis previously described, the first cluster presents a less stretched structure while the second reveals a major groove widening (figure 3.6). The most significant difference observed between both states derived from the tetraloop U₂₃-U₂₆ base pair where both nucleotides reorganize into a more compact and stretched loop.

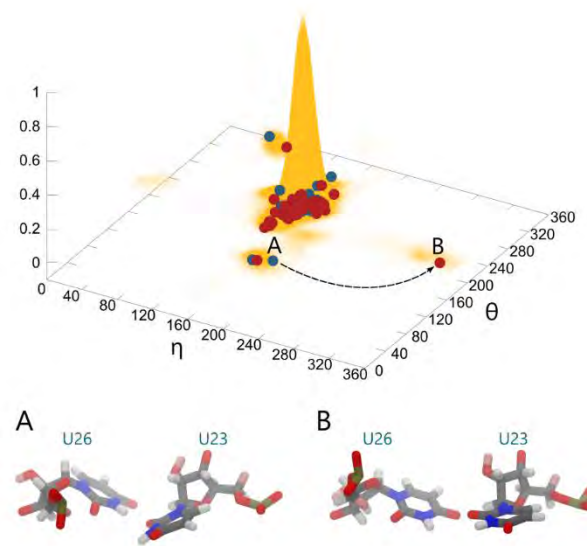


Figure 3.6. Density plot of η/θ states observed along the trajectory. Blue dots represent the distinct nucleotides of cluster 1 and red dots the ones from cluster 2. Observe the transition from **A** to **B** that produces a significant rearrangement of the U₂₃-U₂₆ pair.

Hydration is an important factor that affects in great measure the RNA conformation. U-U mismatches have been fully characterized by the number of intra-residue hydrogen-bonds (typically 0-3) and the presence of water-mediated hydrogen-bonds (henceforth named W-U-U). NMR data suggest that U-U mismatches in RNA hairpins exhibit an H₃(U₂)-O₄(U₁) hydrogen bond and water-mediated hydrogen bonding between H₃(U₁)-O₂(U₂).¹ Previous structural studies have demonstrated the U-U internal loop flexibility, showing a variability in C1'-C1' distances range from 8.8 to 10.5 Å depending on the RNA conformation and, in fact, closer C1'-C1' may result in local pair contractions.^{16,17} Furthermore, the crystal structures described in literature are highly dependent from crystal packing degree and intermolecular contacts.

According to interatomic U-U pair distances and the number of close water molecules, computed trajectory may present 3 states: (1) the formation of direct U-U pairing, (2) water-mediated hydrogen bonding or (3) water-mediated hydrogen bonding accompanied by minor groove hydration. The latter may present significant connotations since a grooved water mediates

some ligand and/or protein interactions. In order to assign the correct states to each traced interatomic distance quantum-mechanics (QM) techniques have been applied.

Brandl et al. computed the optimal geometry of U-U and W-U-U pairs by QM methods replacing the sugar moieties by hydrogen atoms.¹⁸ The different inter-residue distances may reproduce the optimal values for U-U and W-U-U pairs. However, this approximation implies a total neglect of the stacking interactions and backbone fluctuations. Despite stacking interactions are not a major contribution for the base pair geometry, backbone geometry depends on the neighboring base-pair. As exposed above, the effects of U-U mismatches on CUG repeats extend to GpC steps and increases the overall elasticity. For this reason, QM calculations were performed over the most representative U-U pairs of our simulation of the two previous clusters. Three U-U pairs of cluster 1 were considered (U17-U32 pair, W-U8-U41 containing one H₂O molecule and W-U14-U35 containing two H₂O molecules). Energy optimization was conducted and subsequent computed distances were computed for clusters 1 and 2 with HF/6-31G(d,p) and DFT/6-31G(d,p). Because the sugar moieties coordinates were frozen, the presented Cr'-Cr' values matched those obtained by cMD. Thus, Cr'-Cr' distances are quite close to the expected range (8.8 to 10.5 Å) but it is predicted that U-U pairs are more contracted structures compared to W-U-U. These results pointed out the fact that water may act as a spacer that widens the major groove oriented part of the nucleic acid and prevents the formation of direct hydrogen bonds.

This data allowed the monitoring of U-U/W-U-U states by computing the atomic distances from different U-U pairs. Evolution of N₃-O₂ bonds fit quite well for this purpose and reveals periodic transitions between both states. Obviously, the QM data, which have been obtained *in vacuo*, contains no transferability to our cMD trajectory but it is sufficient to suggest a qualitative assignation (figure 3.7). According to this, U₂₀-U₂₉ is 97.6% of the trajectory in W-U-U hydrated stated while U₁₇-U₃₂ is mainly (72.3%) in the U-U state. U₁₁-U₃₈ and U₁₄-U₃₅ present regular transitions between W-U-U and hydrated W-U-U states, being the latter the most frequent (60.9% and 68.2% respectively).

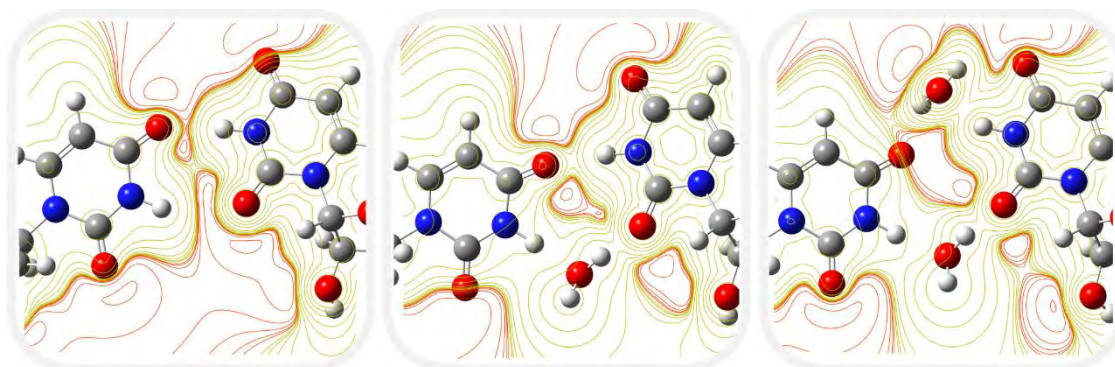


Figure 3.7. Isodensity contour plot of cluster 1 pairs (C1). From left to right: U17U32, U8U41 and U14U35 pairs. The perturbation effect in charge distribution induced by proximal water molecules is clearly observed in the contour lines.

3.2.7. SIGNIFICANCE OF DYNAMICS FOR DRUG DESIGN

Targeting rCUG^{exp} sequences has been proved to be the most appropriate approach to face DM1, since RNA acts as the causative agent, whereas MBNL1 keep unmodified. Structure-based drug design relies on understanding the target structure and its dynamic properties as the structural conformations are not unique in time. The rCUG^{exp} are characterized by their ability to adopt a large set of conformations each of which has been observed in X-ray or NMR resolved structures. However, it is not clear how existing non-canonical U-U pairs behave during time and if they depend on the structure of adjacent CUG repeats. This set of conformations has been classified using the number of hydrogen bonds and their orientation towards the RNA groove.

Herein, the cMD trajectories of three CUG containing structures corresponding to a double stranded RNA with 6 CUG repeats in each strand (ds[(CUG)₆]₂) and two homology modelled hairpins, r(CUG)₁₆ and r(CUG)₈ were analyzed. On one side, the r(CUG)₁₆ system represents a 16-repeats long hairpin (containing 7 non canonical U-U pairs and a UGCU tetraloop). On the other, the r(CUG)₈ structure contains a shorter CUG repeat system (8-repeats long with a total of 3 non canonical U-U pairs and a UGCU tetraloop) capped by C•G pairs.

These findings suggest that the central pairs of a stem region containing multiple CUG repeats are the most flexible pairs. This fact can be observed in the ds[(CUG)₆]₂ cMD trajectory: the two central non-canonical U-U pairs prefer type III and VI conformations in a 0 hydrogen bond state. As the non-canonical pair stands away from the center the number of hydrogen bonds increases. The same effect is also observed in terms of stiffness analysis, where the stiffness constants suffer a significant decrease in the central region. The inclusion of a tetraloop, corresponding to the r(CUG)₁₆ model, breaks the symmetry of hydrogen bond distribution reducing the flexibility of the stem region. If the number of CUG repeats is shortened, as in the r(CUG)₈ model, the whole CUG system gains in flexibility. Globally, modifications in the flexibility of the 5'-UGCU-3' fragments could be related to the different conformations that non-canonical U-U pairs can adopt, which are summarized in figure 3.8 (page 86). However, it is not clear how much the conformation type of the U-U pairs contribute to the rCUG^{exp} overall flexibility because the A-form RNA is constant in every structural model and rCUG hairpins models have similar thermodynamic stabilities.

The usage of several rCUG^{exp} structures has been applied in structure-based drug design using different simulation protocols (mainly molecular docking and MD). However, the different types of non-canonical U-U pairs are dynamically interconverted and it is hypothesized that this effect depends on the number of surrounding CUG repeats and the structural modifiers, such as the tetraloop. Hence, the full sampling of the targeted U-U pairs is not always available, depending on the selected rCUG^{exp} system model and, presumably, the simulation conditions. This conformational heterogeneity leads to a poor stacking of the uridines within the helix, although the A-form of the whole hairpin is well preserved. This fact has interesting implications for the interactions of rCUG^{exp} with ligands and MBNL1 sequestering. It would be expected that the binding affinity for a repetitive structure would grow proportionally to the number of repeats, whereas each CUG repeat exhibits differential stiffness behavior. The druggability of non-canonical pairs is not yet well-established and their repertoire of possible conformations may force to study the system as a large set of possibilities (this concept will be further developed in chapter 4).

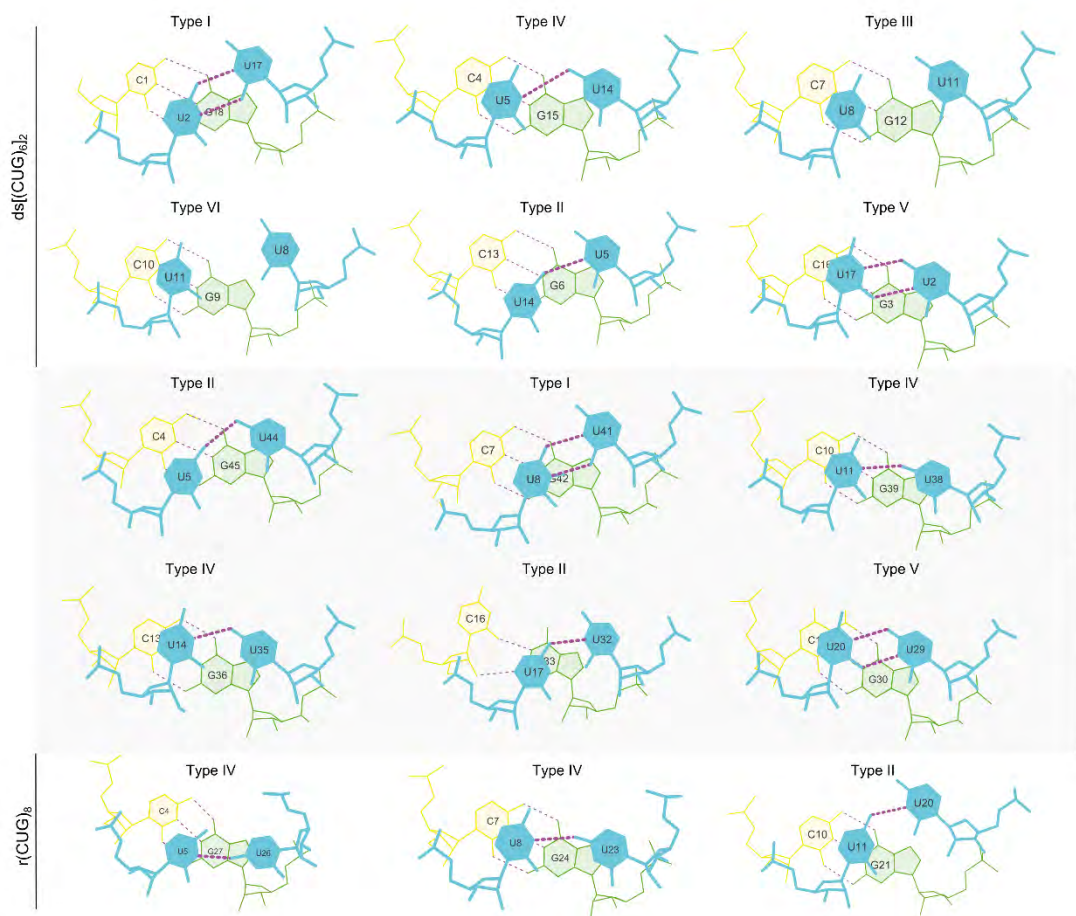


Figure 3.8. Classification of the average CUG repeats obtained from the obtained cMD systems. $ds[(CUG)_6]_2$ model sampled all the experimentally observed U-U pair types. $r(CUG)_{16}$ hairpin model prefers type II and type IV featuring 1 hydrogen bond. $r(CUG)_8$ only presents the 1 hydrogen bond types (type II and type IV).

Nonetheless, MD based methods require large computing resources and a high computational cost in time. Thus, in the next section a simplification of the system will be assessed by using elastic network models (ENM) and, in particular, anisotropic network models (ANM).

3.3. SIMPLIFYING THE MODEL: ELASTIC NETWORK MODELS AS RNA CONFORMATIONAL SAMPLING TOOLS

Anisotropic network models (ANM)¹⁹ are based on the assumption that a system can be described as particles connected by springs. Therefore, the intrinsic flexibility and dynamics of the macromolecule are treated as a set of normal modes of an oscillating system. Recent studies assessed the viability of ANM for conveying a set of apparent motions in RNA and DNA ensembles and succeeded in reproducing the ones observed in experimental data.^{20–22} By following these approaches, we sought to investigate a particularly relevant RNA structure such as CUG trinucleotide repeat overexpansion.

Many RNA structures including one or more CUG triplets in its sequence have been reported by different investigators. Thus, we considered static X-ray dataset as an ensemble of snapshots representative of the intrinsic dynamics of the RNA. The ensemble can be analyzed using principal component analysis (PCA) to extract the principal modes of structural variations (i.e. principal component, PC) which, in turn, can be compared to ANM soft modes. Nevertheless, the number and type of RNA motions are subjected to the number of conformations that have been experimentally resolved.

3.3.1. ANISOTROPIC NETWORK MODEL OF THE rCUG ENSEMBLE

The available structural models of rCUG transcripts in the RCSB Protein Data Bank should be sufficient for the dynamics simulations using PCA although these models are quite sensitive to small changes in the structures. We performed the PCA on a collection of rCUG structures with a number of repeats ranging from 3 to 6. First, a (CUG)₃ ensemble was constructed by

aligning all the possible (CUG)₃ fragments from the PDB X-ray structures (see details in Methods, page 109). Then, the ability of ANM to capture variations within the structural ensemble was evaluated. Previous studies suggested the inclusion of the ribose atoms into the coarse-grained model, which improved the level of experimental agreement if compared to the utilization of only the P atoms.²⁰ Following this criteria, we decided to test an all-atom model and two levels of coarse-graining: P, C2', C4' atoms, and P, C2', C4 and N3 atoms. We considered important to capture the hydrogen-bonding patterns and the U-U mobility, thus we sought to capture this information through the N3 and C4 atom while P and C2' should describe the backbone dynamics. At the same time, γ constant was optimized using a negative exponent weighting approach (refer to figure 3.9A). Different cutoffs ranging from 5 to 15 Å were tested. Then, the generated ANM models were compared to PCs obtained from the PCA of the ensemble.

The best ANM model was achieved using a cutoff of 9 Å and a coarse-grained model represented by atoms P, C2', C4 and N3. No improvement was achieved by using an all-atom model in terms of overlapping modes (see figure 3.10); hence, some loss of description level occurs even though the essential motions of the RNA are correctly described. When comparing ANM modes with PCs we observed that the second NMA mode exhibits an acceptable overlap with PC1 mode (71%), see figure 3.9B. In other words, the second softest ANM theoretical mode is confirmed by the first experimental PC. Visual inspection of the dominant motions extracted from coarse-grained PCA shows that PC1 deformation vector corresponds to the bending of the RNA structure from end to end, opening and closing the major groove (figure 3.9C). Moreover, PC1 and ANM2 show an excellent correlation (0.98), as shown in figure 3.9D, meaning that the crystal structures have low dispersion around this pair of PCA and ANM modes. More interestingly, PC2 and PC3 represent the movement of the base pairs along the *xy* plane. We observed that C•G and G•C pairs shear in opposite directions along the plane, but U-U pairs cooperatively move in the same direction exhibiting a base pair opening. Compared to the backbone movement, this modes are overlapped with the theoretical modes by 48% and 57% respectively.

The principal structural changes can be described through a set of low-frequency ANM modes. The cumulative overlap can be interpreted as the extent to which this set of soft modes can predict a PCA mode. In figure 3.11, we report the cumulative overlap of the first three PC and the percentage of captured variance. Notice that 20 ANM modes can explain ~80% of PC1 and PC3. In fact, the ANM1-12 provide a good description, while higher modes diminish its contribution. Nonetheless, the distribution of motions is captured by the collectivity degree (κ) which provides a measure of the extent of distribution of motions across the structure. The collectivity degrees for the first three PCs are 0.43, 0.64 and 0.74 correspondingly, meaning that the first PC motions are less distributed along the structure. In fact, PC2 and PC3 are mainly represented by the shear and stretching of all the base pairs, especially of the U-U internal loops; thus, these two modes are considered to be highly collective and more relevant for the present study.

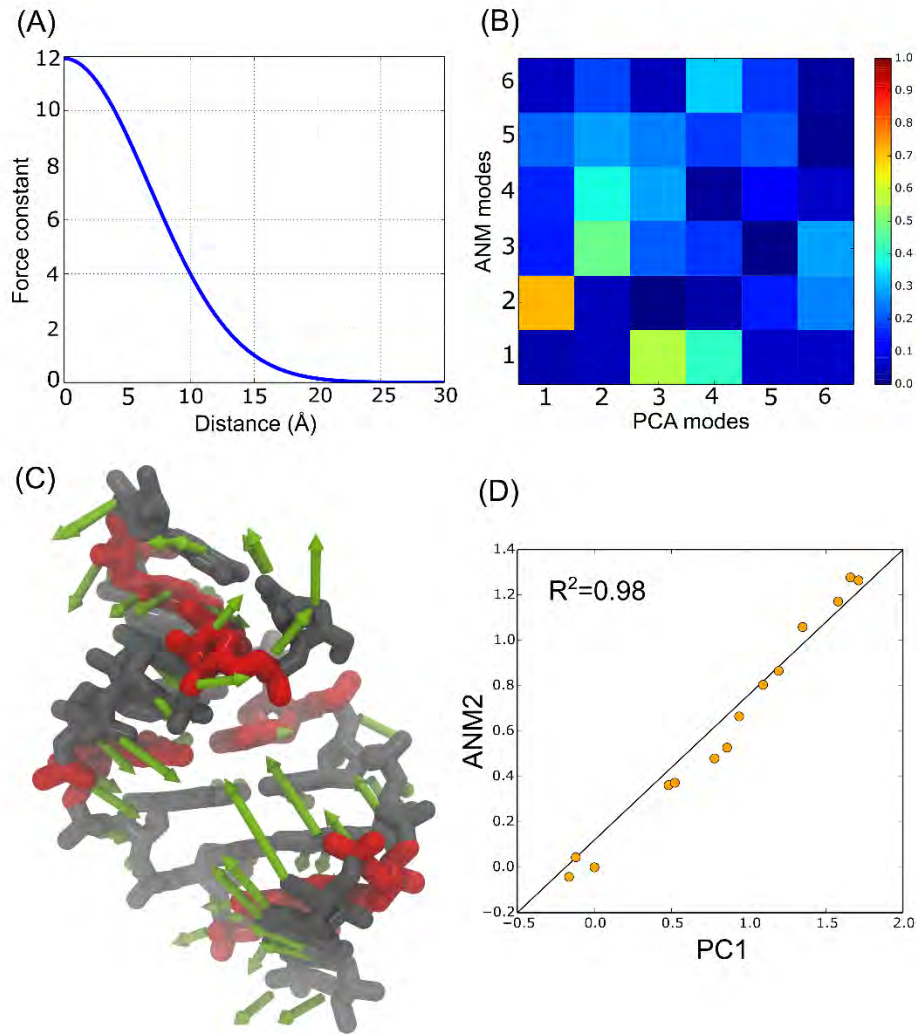


Figure 3.9. (A) Optimized distance dependent force constant (γ). Closest nodes are weighted by 12 (arbitrary units), and the weighting decays exponentially until 25 Å. (B) Overlap between the top six PCA modes and the softest six NMA modes. The second softest ANM mode exhibits the highest overlap with PC1. (C) All-atom representation of (CUG)₃ fragment and PC1 normal mode vectors. U-U pairs are represented in red. Normal mode vectors (in green) show the structural variations along this mode. (D) Representation of the dispersion of the examined PDB structures along the PC1 and ANM2.

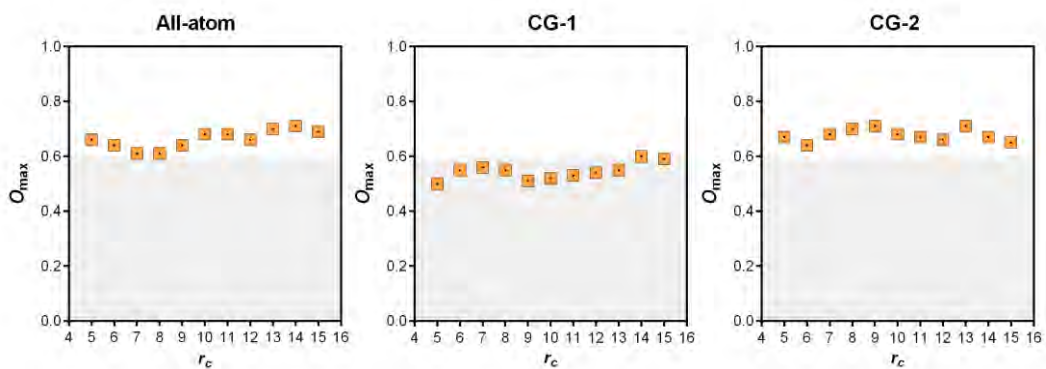


Figure 3.10. Maximum overlapping achieved by using different cutoffs and residues description: all-atom, CG-1 (atoms P, C2', C4') and CG-2 (atoms P, C2', C4', N3). A similar overlapping is achieved with an all-atom model and $r_c=14$ and CG-2 and $r_c=9$.

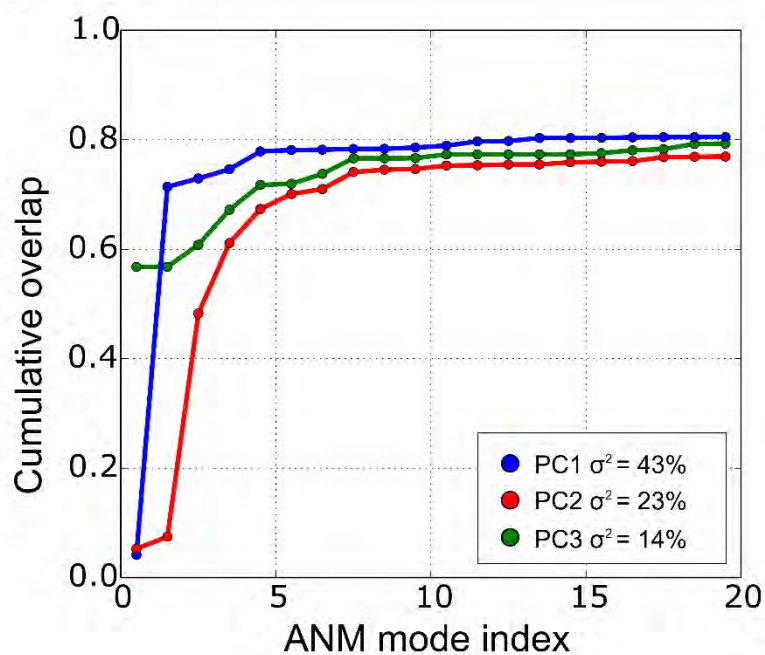


Figure 3.11. Cumulative overlap of ANM soft modes of PC1 to PC3. The legend contains the percentage of variance (σ^2) explained by the corresponding PC. Notice that 20 ANM modes explain ~80% of structural variations along the first and third PCs.

3.3.2. CONVENTIONAL MD OF THE rCUG SYSTEM MODEL

In good agreement with previous reports, our approach shows that ANM can achieve an acceptable level of description of global RNA dynamics. However, ANM are not able to correctly describe local dynamics and the available set of structures do not necessary represents all accessible structural changes. For this reason, further rCUG dynamics were assessed using conventional MD (cMD) simulations. All the MD analysis was performed according to the observed U-U conformations along the trajectory, which are the most relevant local changes, and compared to the experimentally resolved structures.

Kiliszek *et al.* noticed that some uridines involved in U-U pairs tilt towards the minor groove, breaking the palindromic symmetry in a seemingly random manner over the structure.¹¹ The number of CUG repeats determines the number of available three-dimensional U-U structures hence, in longer RNA chains, there must exist a vast repertoire of U-U pairs in terms of available conformations. Kumar *et al.* reported two rCUG X-ray structures providing different 1x1 nucleotide U-U internal loop conformations and considered that the small molecule drug design should take into account all the available U-U conformations.²

As described in the previous section, herein we studied the behavior of a (CUG)₂ model through cMD simulations and determined the dynamic properties of these particular non-canonical U-U pairs. Our system model system consist of two CUG repeats capped by C•G pairs, which should increase the overall stability during the simulation (see figure 3.12A). The non-canonical base-pairs of the system model adopt the stretched U-U wobble conformation; this

conformation forms interactions with only one hydrogen bond between the carbonyl O4 and the N3 imino group of the second residue.

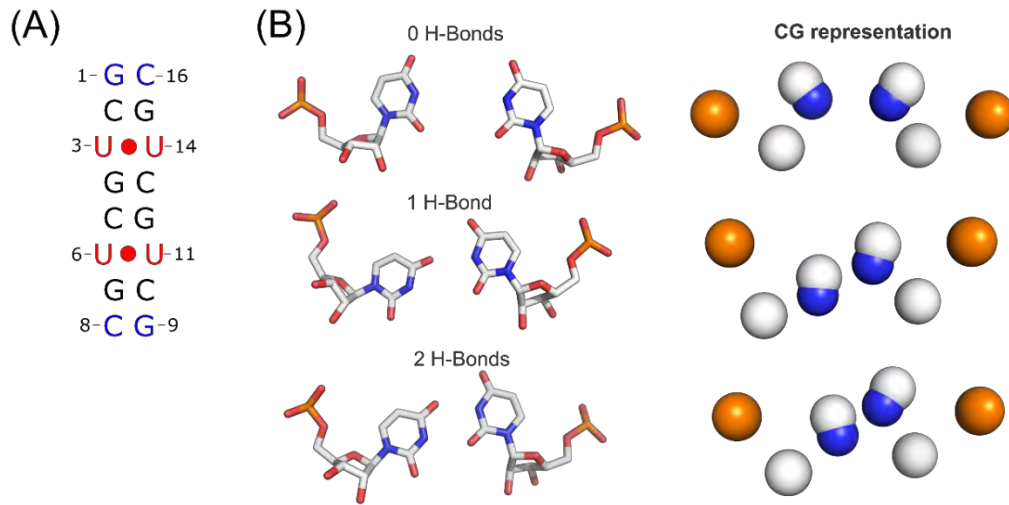


Figure 3.12. (A) Schematic representation of the system model used for the cMD. (B) Representative U-U pair types observed along the trajectory and coarse-grained (CG) schematic representation.

After the cMD simulation we identified a total of 4 possible U-U pair conformations by means of cluster analysis, which featured different hydrogen bonding patterns. Among the most representative conformations we observed the presence of 0, 1 or 2 direct hydrogen bonds, some of which formed 1 or 2 water-mediated hydrogen bonds at the same time. During the simulation each pair goes through nearly two or three distinct conformations although some of them are for less than 5% of simulation time (figure 3B). From highest to lowest population MD results suggest that U-U pairing formation involves 1, 0 or 2 hydrogen bonds, whereas water mediated hydrogen bonds represent a 20% of total simulation time. The analysis of the overall structure along the MD trajectory revealed that only the helical opening was statistically significant from the crystal structure at the 95% confidence level.

A qualitative analysis suggested that each of the two average U-U pairs stayed in type IV and II configurations along the simulation. The predominance of these substates has been thoroughly discussed in literature and types IV and II are the most predominant configurations among the crystal structures. Through examination of all rCUG NMR and crystal structures it becomes clear that U-U pairs can flex between many different conformations. Nevertheless, looking at our MD results (table A3, Annexes, page 201) we observed very close Cr'-Cr' distances, helical averages and standard deviation with experimental data, which suggest that our MD trajectory was able to explore the most relevant experimentally observed U-U conformations, but not all of them.

3.3.3. ACCELERATED MD OF THE rCUG SYSTEM MODEL

cMD allow to access time scales on the order of hundreds of nanoseconds. However, advanced sampling techniques have been developed to explore structural changes in shorter time scales (e.g. replica-exchange molecular dynamics, metadynamics and accelerated molecular dynamics, among others). Accelerated molecular dynamics (aMD) is an enhanced

conformational sampling technique that provides access to events beyond the ones obtained by conventional simulations. For instance, our cMD was run in a sub-microsecond scale and it was not able to explore all the available U-U configuration space, so it became clear that a more exhaustive exploration was required. We decided to assess whether enhanced sampling techniques such as aMD conformational sampling improved the results yielded by cMD.

Surprisingly, the aMD trajectory tends to drift away from that observed in cMD. As shown in figure 3.13 (also see Annexes, figures A1-A2, page 202-203), helical parameters from both simulation are within the same range but large deviations occur in the 3'-end of the aMD without affecting the second U-U pair. This effect was observed for less than 15% of the simulation (see Annexes, figure A2, page 203). We should note that this effect might be a force field artifact caused by improper description of non-bonded interactions or backbone definition. In sharp contrast with the cMD results, the aMD trajectory samples a completely different region of the U-U conformational space. For instance, types I, III and V are preferred along the simulations. In both cases, the main motions are governed by the backbone heavy atoms displacement and the base pair opening. This observation agrees with the previous PCA so further comparison with essential dynamics analysis (EDA) was accomplished.

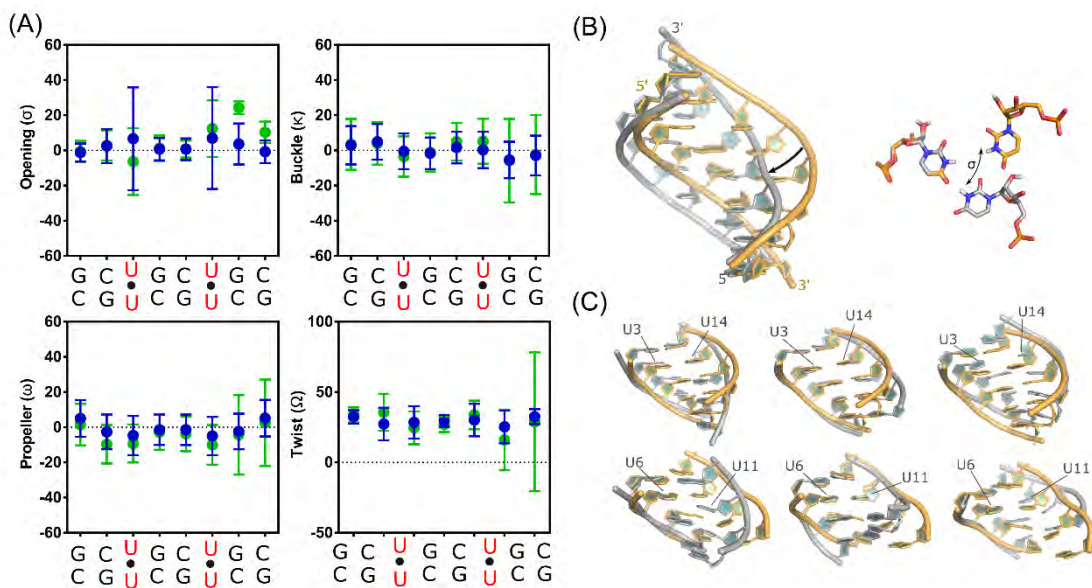


Figure 3.13. Structural analysis of the cMD and aMD simulations. **(A)** Average and standard deviation for each helical parameter (base pair opening, buckle, propeller and helical twist). cMD and aMD results are colored in blue and green respectively. Structures were aligned to all heavy atoms and represented with PyMOL. Notice that the 3'-end from the aMD simulation yields significant deviation from the cMD. **(B)** Cartoon representation of two clustered structures from the cMD simulation and opening effect (α) of the U-U pair. The two main observed motions along the simulation correspond to the backbone expansion-compression (opening and closing of the major groove) and the base pair opening of the uridines. **(C)** Cartoon representation of the first and second CUG fragments from the aMD simulation. The main distortions are observed in the 3'-end, as stated by the helical parameter values.

3.3.4. COMPARISON OF ANM, PCA AND EDA MODES

Once the MD simulations were analyzed, we proceeded to compare the EDA of the generated trajectories with the *reference* modes obtained from the crystal structures. A previous report benchmarked the sampling of MD protein simulations against ANM and PCA using 20 ns trajectories.²³ The authors concluded that generating conformers using the softest ANM modes covered a more comprehensive subspace than the MD ensembles. Moreover, ANM requires very low computational resources compared to conventional methods.

In this study, we asked whether RNA small systems should span similar conformational coverage using any of the aforementioned methods. That is to say, we compared the conformational sampling of experimental structures and simulations, all projected onto the principal subspace spanned by the first three PCs. First, we represented each ensemble (the PDB ensemble and the cMD and aMD trajectory snapshots) onto the PC1-3 subspace (figure 3.14A).

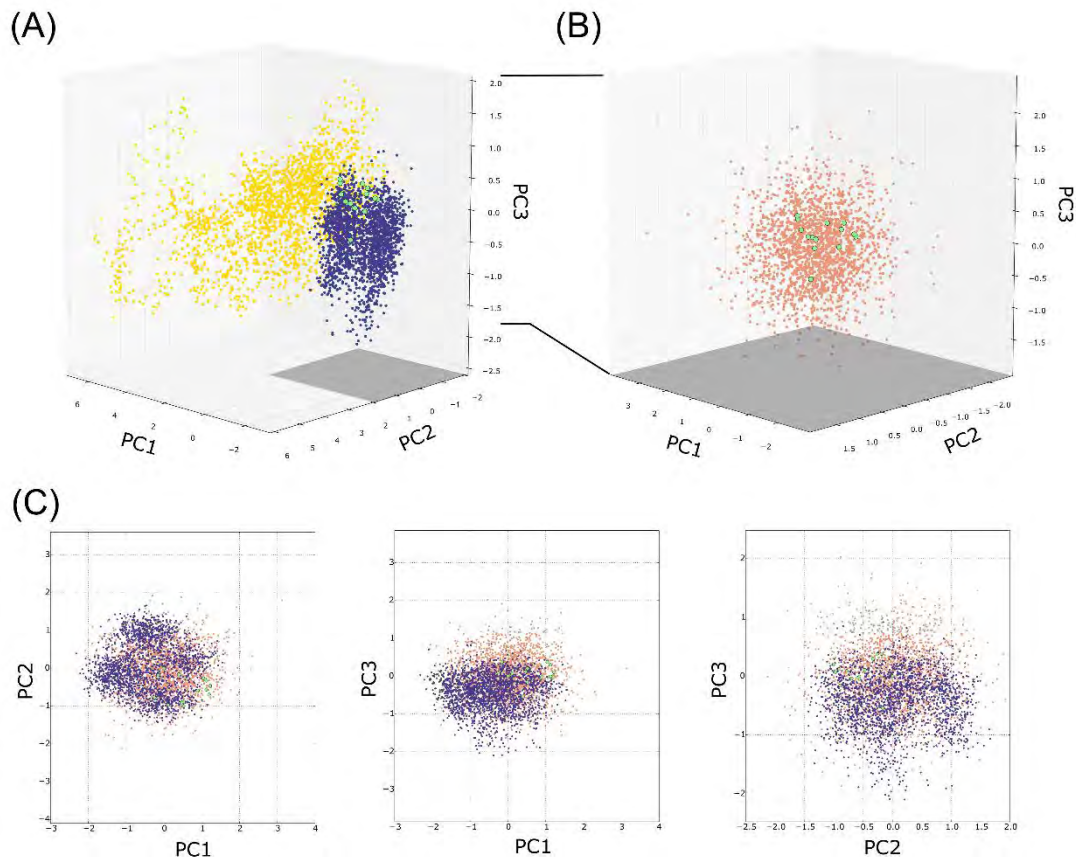


Figure 3.14. Projection over the PC1-3 subspace of (A) the cMD (blue) and aMD (yellow) snapshots, and the PDB ensemble (green). (B) Comparison between the original PDB ensemble and a 2000 conformers (salmon) generated using the softest three modes. The perspective is the same in both panels, but the ranges differ. (C) Two-dimensional projection of the aMD, ANM generated ensemble and PDB ensemble over the first three PCs.

In agreement with our previous analysis of the *reference* or experimental modes, it is clear that the highest collective modes are PC2 and PC3. Compared to PC1, changes are less localized and less pronounced along the other two PCs. In agreement with previous studies,^{23,24} conformers

generated during cMD and aMD simulations encompass only part of the crystal structure and explores the surrounding subspace. For instance, cMD and aMD conformers drift away from the *references* and reproduce only half of the experimental structures, which is reflected in its low essential space overlap (50% and 49% respectively). In sharp contrast, the essential subspace overlap between cMD and aMD simulations account for a total of 77% of subspace overlap. It is unclear whether the explored aMD conformers correlate events which occur on longer time scales. However, the initial and final aMD frames are located in the same subspace region, meaning that the structure ‘visits’ several substates but is capable of returning to the initial structure. From a structural point of view, MD simulations spanned a wide range of U-U base pairs, but also the transition between them and the different combinations of the possible base pairing types. Likewise, the collectivity of the principal modes that describe these motions suggest cooperative dynamics along the structure (see table 3.4).

Table 3.4. Variance (σ^2) and collectivity (κ) for each ensemble system: PDB ensemble and molecular dynamics (cMD and aMD).

Mode	PDB ensemble		cMD		aMD	
	% σ^2	κ	% σ^2	κ	% σ^2	κ
1	43.2	0.43	32.7	0.39	25.9	0.22
2	22.9	0.64	24.4	0.63	17.4	0.50
3	13.89	0.74	11.6	0.73	11.2	0.70

A previous work showed a remarkable coverage of the reference space by ANM predictions. For this reason, we generated 2000 snapshots by deforming the structure along the softest three ANM modes and compared the subspace coverage with that of the cMD. As reported by Bakan and Bahar,²³ deformation along the ANM modes exhibit an excellent coverage of the *references*. Superposition of our generated conformers with the *references* (Figure 5B) show that the ANM modes permit to explore a wider range of conformations along the modes. However, we noticed that the first principal mode was strongly favored without a significant loss in terms of collectivity.

As we can see from figure 3.15, the *references* global fluctuations are in qualitative agreement with the cMD simulation. Not surprisingly, aMD experiences the largest relative displacements. Due to limitations of the ANM models, interactions between beads depend on their distance and not the type of interaction, hence ANM fluctuations of the nucleobases are prone to be highly constrained. A differentiated correlation is observed when only backbone or nucleobase beads are considered. Figure 3.15 shows a clear correlation between P-P fluctuations for each ensemble; thus, backbone global fluctuations are within the same range of native state fluctuations. On the contrary, no correlation exists when only N₃ beads, which mainly represents the U-U configurational state, are considered due to a differentiated conformational exploration of the bases. Table 3.5 summarizes the Pearson correlation coefficients.

Table 3.5. Pearson correlation coefficients between CG set (P, C2', C4 and N3), P beads and N3 beads fluctuations extracted from the *references*, cMD and aMD ensembles.

	CG	P	N3
Ref cMD	0.75	0.95	0.34
Ref aMD	0.28	0.74	0.11
cMD aMD	0.37	0.85	-0.23

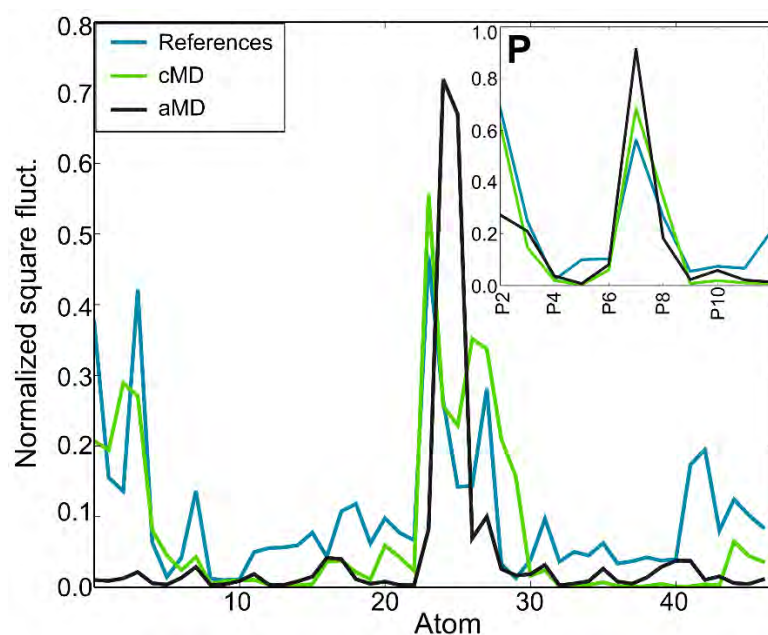


Figure 3.15. Atomic global fluctuations extracted from the reference structures, and the cMD and aMD ensembles. Only P, C2', C4 and N3 atoms are considered. Computed fluctuations for the P atoms only are plotted in the box at the top right corner of the figure.

3.3.5. THE SIMPLER THE BETTER (OR NOT)

RNA plays critical roles in cellular biology hence it is an incredibly important target for small molecule therapeutics. Unfortunately, RNA-small molecule interactions knowledge is still scarce and insights into the intricacies of the dynamics of RNA are essential to provide novel therapeutic scaffolds. In particular, targeting rCUG sequences has been proved to be the most appropriate approach to face DM1, since RNA acts as the causative agent, whereas MBNL1 keeps unmodified. Structure-based drug design relies on understanding the target structure and its dynamic properties as the structural conformations are not unique in time. The rCUG are characterised by their ability to adopt a large set of conformations each of which has been observed in X-ray or NMR resolved structures. The druggability of non-canonical pairs is not yet well-established and their repertoire of possible conformations may force to study the system as a large set of possibilities.

Herein we assessed two of the most relevant techniques to explore the rCUG conformational landscape, which include elastic network models (ENM) and molecular dynamics (MD). In

agreement with previous studies, the use of ENM with a simplified coarse-grained representation is able to reproduce the global motions observed in experimental structures. In particular, the information gathered from the slow modes obtained from ANM is identical to the one contained into the PDB ensemble. The best parameterization for our RNA model was obtained by a coarse-grained model represented by atoms P, C2', C4 and N3. Optimal performance was achieved with a 9 Å cutoff and a distance dependent force constant. In this regard, we noticed that ANM parametrization is able to correctly describe the global fluctuations of a highly dynamic RNA structure such as rCUG, but it required the consideration of both backbone and nucleotide coordinates to attain a good level of description.

In order to assess the viability of ANM methods to capture local motions derived from U-U pairs we proceeded to investigate their motions as described by MD simulations. MD is probably the most accurate computational method for the theoretical study of large-scale dynamics, since it is based on rigorous physical formalisms and quantum-mechanical and experimental parameterizations. However, its high computational cost still limits all-atom simulations to the microsecond scale. Comparison of cMD and aMD simulations demonstrated a clear difference of conformational space exploration. The softening of energy barriers that provides aMD allowed to explore a higher number of internal loop conformations than cMD in the same time scale. For instance, the aMD trajectory analysis concluded that types I, III and V are preferred along the simulation. From the point of view of the rCUG local conformational landscape, the U-U pairs adopt preferentially a type IV conformation (1 hydrogen bond inclined towards the major groove) which is the most experimentally observed conformation. Nonetheless, MD techniques allowed to explore a higher myriad of conformations which induces high local fluctuations onto the structure. In line with these results, studies demonstrated that MBNL1 binding to U-U pairs induces local melting of the RNA structure. That recognition step depends on the punctual loss of the hydrogen bonding patterns of the internal loops. Nevertheless, both MD methods showed that the main motions, global and local, highly depend on the backbone heavy atoms displacement and the base pair opening effect. In additions, the local information is affected by the precision of the force field, which is compounded by inaccuracies in its parametrization. Thus, the conformational analysis is not void of errors that can compromise the reliability of the data extrapolated from MD simulations. However, the short nanosecond time scale used in this study and the agreement in terms of overlapping global space with experimental structures permitted confident comparison between ENM and MD techniques.

It is reasonable to assume that a deformation along the ANM modes explores all the possibilities, without assessing its intrinsic stability, whereas force field parametrization of MD simulations guides the system through a more comprehensive conformational landscape. As mentioned above, ANM fail to properly describe local motions which can be easily discerned from the U-U pairing sampling. All U-U conformations found in the ANM ensemble correspond to type IV, the same as the ones in the *reference* structure. Therefore, the bulk of results presented demonstrate that, if insights into U-U dynamics and transitions are relevant, MD simulations are required. Nonetheless, ANM succeeds in describing the global dynamics of complex RNA structures with a low computational cost.

Traditional rigid docking fails to describe small molecule - RNA interactions due to the lack of RNA adaptation. In fact, a fast and practical approach to improve molecular docking in

proteins is to generate an ensemble of conformations obtained from experimental structures. However, our results suggest that ANM are not suitable for structure-based drug because high local fluctuations are not efficiently captured. Several studies succeeded in applying a molecular dynamics approach for drug-design of rCUG binders,^{25,26} but time and computational limitations for exploring small molecule–RNA interactions exist and large virtual screening campaigns cannot be performed. Other authors suggested the combined use of ENM and MD,²⁷ which can favour particular modes observed into the MD simulations over the global backbone motions, improving the reproduction of both local and global modes.

In conclusion, our work gives a comprehensive comparative analysis of ANM and MD methods for assessing small scale and large scale events along a highly dynamic RNA structure. However, these results are subjected to improvements implemented in other RNA force fields, which are in constantly revision. Further analyses will be conducted to study and compare the accuracy of the force field revisions and their effect on the local and global configurations of the RNA. These studies should provide useful insights that could be exploited for computer-aided drug design strategies.

3.4. THE CONTEXT MATTERS: STRUCTURED rAUUCU

Spinocerebellar ataxia type 10 (SCA10) is caused by a pentanucleotide repeat expansion of rAUUCU within intron 9 of the ATXN10 pre-mRNA. The RNA causes disease by a gain-of-function mechanism in which it inactivates proteins involved in RNA biogenesis. Spectroscopic studies showed that rAUUCU repeats form a hairpin structure, however, there were no high-resolution structural models until recently. The crystal structure reported by Park et al. demonstrated that the rAUUCU tracts adopt an overall A-form geometry in which 3×3 nucleotide 5'UCU3' / 3'UCU5' internal loops are closed by AU pairs. Moreover, helical parameters of the refined structure as well as the corresponding electron density map reflect dynamic features of the internal loop.

In order to capture structural characteristics of rAUUCU repeats we have determined the dynamics of this structure with MD simulations. The results indicate rAUUCU repeats form a metastable A-form RNA and the dynamic characteristic is attributed to the internal 5'UCU3' / 3'UCU5' loop pairs. This internal loop contains one C-C mismatch and two U-U pairs which span a different range of conformations, if compared to the ones observed in DM1 due to the biochemical context. Overall the results presented here provide structural evidence of pathogenic mechanism of SCA10 caused by repeat expansion of rAUUCU. This structure may also provide valuable information to guide the design of therapeutic modalities that target this RNA to ameliorate the disease.

3.4.1. DYNAMICS OF THE 5'UCU3'/3'UCU5' MODEL SYSTEM

Understanding the dynamics of biomolecules is often a computationally challenging process. Moreover, classical molecular dynamics (MD) protocols are subject to sampling limitations, which are hard to overcome. In the former section, aMD simulations were performed in order to assess the dynamics of U-U internal loops contained in the rCUG structures. However, larger systems require more exhaustive methods such as Replica Exchange Molecular Dynamics (REMD). REMD method simulates several replicas at different temperatures with a certain probability to swap periodically from one temperature replica to another to improve thermodynamic sampling in a simple and effective manner. In order to achieve thorough sampling of the 3×3 internal loop conformational space, REMD simulations were employed with the hope that higher temperatures would destabilize the loop and provide possibilities to analyze the dynamic features of this 3×3 RNA internal loop. A total of 40 replicas spanning a temperature range close to 300 K were used in order to study the intrinsic dynamics of the 3×3 internal loop. The initial coordinate of 5'UCU3'/3'UCU5' model system was identical to the central fragment of the L1 crystal structure (PDB: 5btm) except additional C-G and G-C pairs flanking the internal loop in order to enhance the structural stability.

The RMSD analysis was completed for each U-U and C-C pair contained in the 300 K replica individually (figure 3.16). The RMSD values were plotted with respect to the average conformation and each color represents an individual clustered conformation with a residence time higher than 10% of the total simulation (figure 3.16C). In good agreement with the crystal structure, the two U-U pairs in the model stayed in a two hydrogen bond conformation along the 300 K trajectory. However, some differences were observed in terms of stacking. For instance, the stacking surface area of 5'U4U53'/3'U18U193' was 0.56 Å² higher than 5'U7A83'/3'U15U163' so the stacking contribution is greatly diminished in the latter.

The C-C pair is clearly the most dynamic mismatch, whereby a large number of conformations are equally distributed during the simulation. Among which, a total of 3 different relevant conformations were found for the C-C pair while U5-U18 and U7-U16 visited only 1 relevant conformations respectively (figure 3.16B). Notice that this feature is significantly different to rCUG repeats, where the U-U pairs oscillated between three different conformations. Interestingly, the C-C pair showed a zero hydrogen bond conformation during a 34% of the simulation (C1, figure 3.16B) while 2 different one hydrogen bond conformations were present during a 64% of the total time (C2-4, figure 3.16B). This dynamic behavior of the C-C pair during the simulation is in line with the mainly one hydrogen bond state in the crystal structure as well as the poorly defined electron density. The previous sections analyses of rCUG repeats and other studies defined the dynamic nature of U-U pairs, thus it is not surprising the poor stacking of the C-C pairs with the flanking uridines and the concurrent C-C dynamic behavior. Several transitions between *cis* and *trans* Watson-Crick conformations of the C-C pair are observed along the simulations which altogether clearly affects the stiff-ness of the whole internal loop. The C1'-C1' distances of the loop nucleotides remain in the 8.8 – 8.9 Å range on average, as observed in the crystal structure, so no major backbone deviations were produced during the simulation.

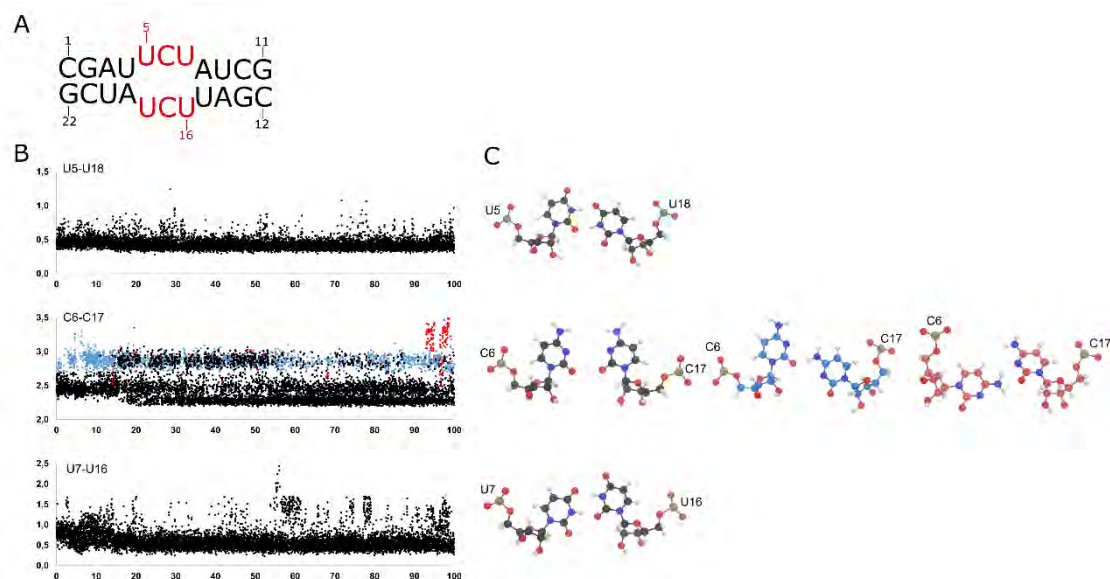


Figure 3.16. (A) Sequence of the model rAUUCU used in the computational studies, (B) symmetry corrected RMSD analysis of the U5-U18, C6-C17 and U7-U16 pairs, and (C) clustered conformations of the internal loop base pairs. Each data point in panel (B) corresponds to a cluster which is represented by a color code displayed in panel (C). Only conformations with more than a 10% residence time are selected for the analysis.

Influence of overall flexibility of the rAUUCU by the UCU internal loop was analyzed using both stiffness constants of a helical space defined by Lankaš et al. and the helical parameters of $3DNA$. Most of the base pair and helical parameters are close to those in the crystal structure except for buckle, propeller and slide which showed significant deviations (see table A4 in Annexes, page 204). However, the fluctuations are considered to be small so the effect of the Jacobian factor was neglected in the stiffness analysis. In general terms, the internal loop presents the lowest stiffness constant values along the structure (figure 3.17), specially rise and tilt, which can be explained by the lack of optimal stacking into this region. Altogether, the labile stacking capability of the internal loop and its large charge density may confer to this RNA a unique binding region.

3.4.2. STABILIZATION OF C-C PAIRS BY WATER-MEDIATED HYDROGEN BONDS

Although the C-C pair remains mainly in a one hydrogen bond conformation, the MD results suggest presence of a stable zero hydrogen bond state. Moreover, two-dimensional potential of mean force analysis (2D-PMF) revealed that the zero hydrogen bond state (C_I in figure 3.16C) is the most stable among the clustered conformations (figure 3.18). The PMF map was constructed using slide and helical twist as variables due to the high variability of these parameters in the crystal structure and during the simulation. C_I coordinates are close to a local minima (-1.9, -23) and deviated from the coordinates of the crystal structure in chain A (-1.1, 36.2) and in chain B (-0.9, 40.1). This indicates an additional force should stabilize C_I in order to keep a zero hydrogen bond as a preferred conformation among all the sampled phase space. A closer inspection of this state suggest that the hydration pattern of the non-canonical pair plays an essential role in its stabilization. Effect of the structure and thermodynamics of the solvent were investigated using GIST²⁸ (Grid Inhomogeneous Solvation Theory), which analyzes the three-

dimensional water density around a certain area referred to as voxels with associated solvent properties. Indeed, a highly occupied volume appears to be close to the C-C pair. Surprisingly, the water molecule occupies a space between C₁₇, U₇ and U₁₆ and coordinates a potential hydrogen bond between atoms O₄ of U₇ and N₃ of C₁₇ (figure 3.19A). At the same time, the U₇-U₁₆ pair inclines towards the C-C pair and breaks the hydrogen bonding pattern, yielding a high negative charged cavity into the RNA minor groove (figure 3.19B). This phenomenon is produced repeatedly and about one third of the time in the MD simulation.

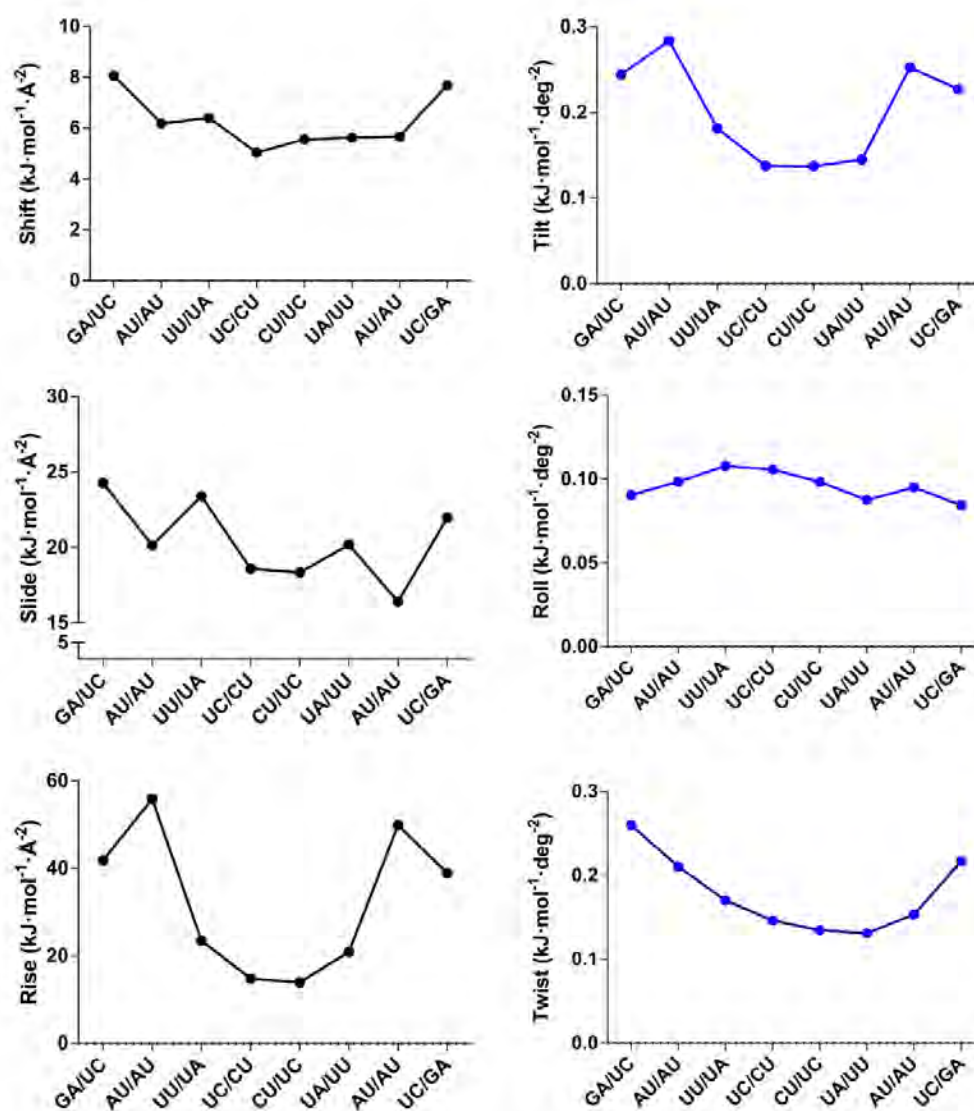


Figure 3.17. Stiffness force constants of the helical space (shift, slide, rise, tilt, roll, twist) for the different nucleotides contained in the [GAUUCUAU] fragment. Highest flexible regions correspond to those with associated low stiffness force constants. Overall, flexibility is especially pronounced into the [UCU] region as observed in the slide, rise and tilt stiffness profiles.

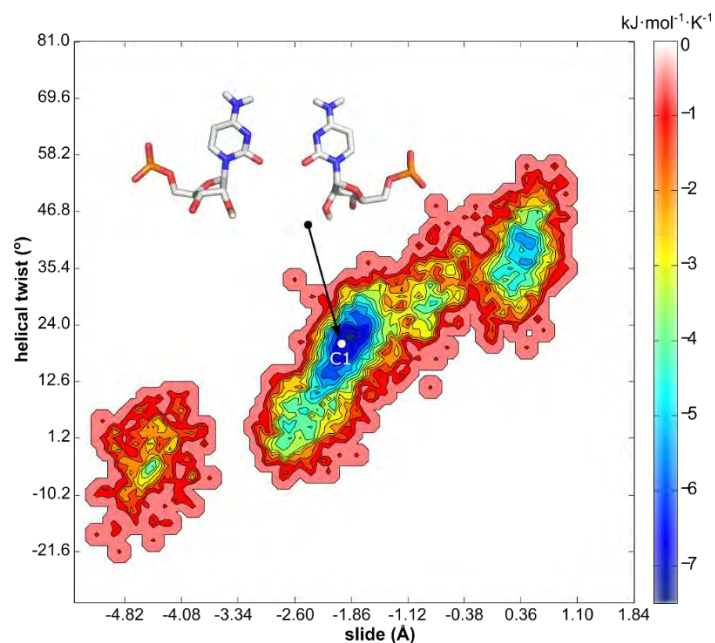


Figure 3.18. 2D-PMF surface showing the exploration of the C-C conformations along the MD simulation. The x- and y-axis represent the slide and helical twist values of the non-canonical CC pair. The zero hydrogen bond C-C conformation (C1 in figure 3.15) is displayed as a white dot in the figure that corresponds to a local minima conformation.

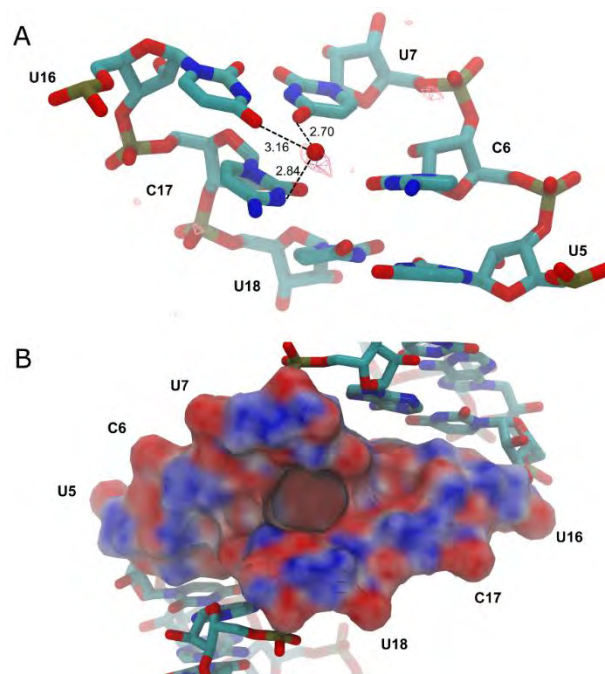


Figure 3.19. (A) Stick model representation of the internal loop and potential hydrogen bond with an explicit water molecule. GIST computed water probability contours are represented as a grid. (B) Charge distribution of the minor groove at the loop is shown as a surface model. Electrostatic potential was calculated using APBS and contoured at ± 10 kT/e.

3.4.3. INHERENT INSTABILITY OF THE 5'UCU3'/3'UCU5' INTERNAL LOOP

The crystallographic and computational results presented here corroborate with findings by Handa et. al. that the 5'UCU3'/3'UCU5' internal loop is metastable. Therefore it has been hypothesized that the RNA unwinding can start through internal loop destabilization. Under this premise, we investigated the stability of each pair individually within the 3×3 internal loop context using Steered Molecular Dynamics (SMD). SMD is a powerful technique used to get insights about several mechanisms, such as unfolding pathways of macromolecules. This procedure exploits non-equilibrium sampling by applying time-dependent biasing forces to guide the system transformation. During SMD experiments, several pulls are simulated in one direction. Previous SMD studies have been proved successful to compute free energy profiles on realistic bimolecular systems.

Herein 25 successive SMD pullings per pair were conducted with a constant velocity of 0.14 Å/ns and a spring constant of 10 kcal·mol⁻¹·Å⁻² and computed force of pulling and cumulative work profile for each system (figure 3.20A). Each plot represents the work required to pull away each pair from bonded state of ~2.5-3.5 Å COM distance to 6.0 Å where hydrogen bond interaction is no longer operative. Pulling force peaked at around 4 Å for U7-U16 and U5-U18 and 5 Å for C6-C17. The mean work performed in each pair clearly reflects that the central C-C pair is the most probable starting point of unwinding ($W = 7.88$ kcal·mol⁻¹). Indeed, the U-U pulling experiments yield very close work values (12.21 and 12.13 kcal·mol⁻¹ for U5-U18 and U7-U16 respectively).

Interaction energies of non-canonical pairs vanished smoothly and had relatively marginal impact over the conformation of the neighboring pairs; thus, the overall RNA structure remained intact during all the pulling experiments. While both U-U pairs did not change the orientation during the steering process, the C-C pair experienced a rotation along the α torsion that flipped C6 out of the RNA duplex. The rotational movement of C6 base was consistently observed in all SMD trials. Therefore, it was concluded that this should be the lowest energy pathway for breaking the C-C pair. The thermodynamic stability of an RNA structure depends not only on base pairing but also on stable stacking interactions. Therefore, the observed unstacking of a cytosine base can further destabilize the 5'UCU3'/3'UCU5' internal loop.

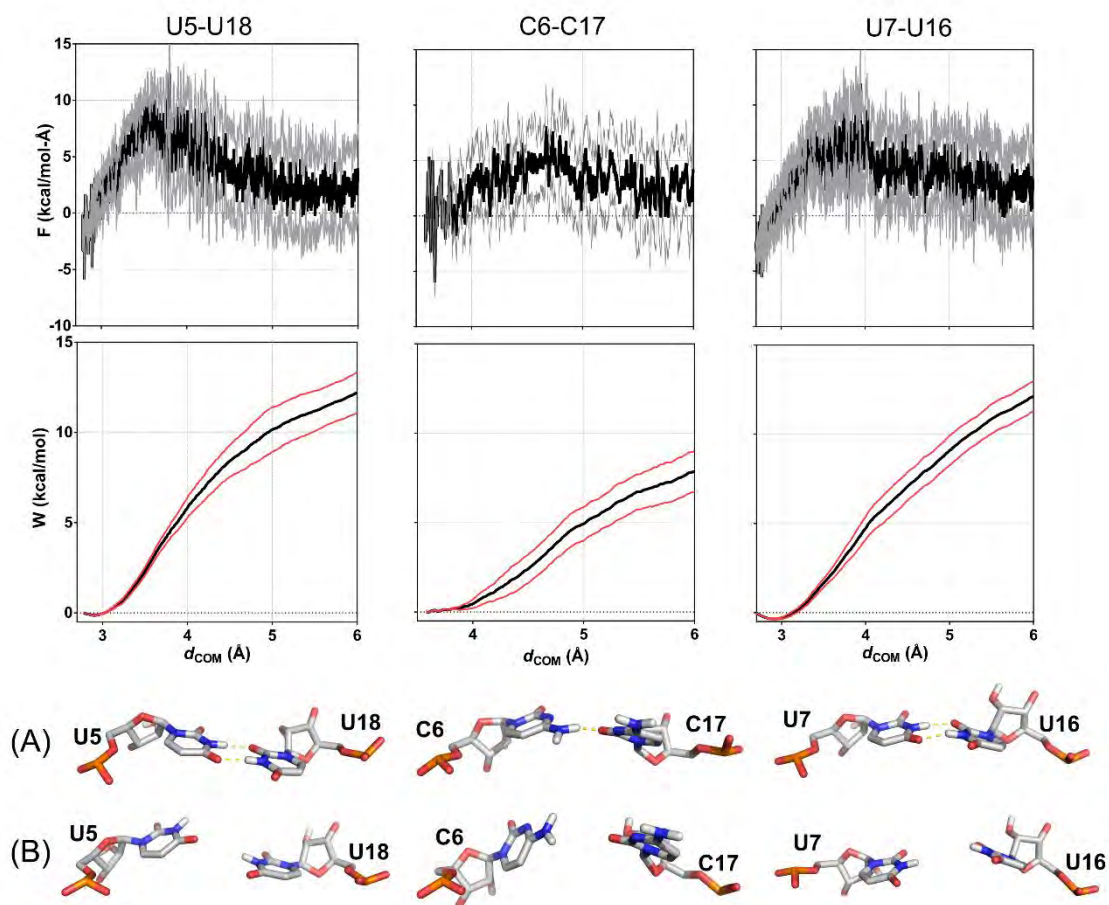


Figure 3.20. Pulling force and cumulative work distribution profiles vs distance from the center of mass (d_{COM}). For each base pair initial (A) and final (B) representative conformations are depicted. The mean force and work are represented as black lines; standard deviation of force and work profiles are depicted as grey and red lines respectively.

3.5. PROTOCOLS

3.5.1. EQUIPMENT

The following computational methods and procedures were accomplished with in-house systems and high performance computing resources (HPC):

In-house equipment

- 4x Intel Xeon 8-core at 3.50GHz, 32GB of RAM memory. NVIDIA Quadro K4000. 2TB filesystem storage.
- Intel Core 2 Quad Q8200 2.33 GHz, 4GB of RAM memory. 2TB filesystem storage.

HPC resources

- **MINOTAURO-BSC:** Red Española de Supercomputación (RES), Barcelona Supercomputing Center (BSC). 126 compute nodes and 2 login nodes. Each node has 2x Intel Xeon E5620 6-core at 2.53GHz. 24GB of RAM memory, 12MB of cache memory. 250GB local disk storage (SSD). 2x NVIDIA M2090 with 512 CUDA cores and 6GB of GDDR5 Memory. GPFS parallel filesystem disk storage (~2PB).
- **TIRANT-UV:** Red Española de Supercomputación (RES). 256 JS20 compute nodes and 5 p515 servers. Each blade has 2 IBM Power4 at 2.0GHz with 4GB of RAM memory. 36GB of local disk storage. GPFS parallel filesystem disk storage (10TB).
- **ATLANTE-ITC:** Red Española de Supercomputación (RES). 336 compute nodes PPC64 (IBM Power PC 970MP at 2.3GHz). 672GB of RAM total memory. GPFS parallel filesystem disk storage (6TB).
- **GARIBALDI:** The Scripps Research Institute (La Jolla Campus). 2848 cores. 250TB of high-performance disk space available from the Data Direct Networks (DDN) SFA10K storage unit.
- **SEPA:** The Scripps Research Institute (Florida Campus). 456 processors. 456GB of distributed memory with a distributed batch queuing system “Sun Grid Engine”. Filesystem disk storage (2.4PB).

3.5.2. PROTOCOLS

MODEL SYSTEMS PREPARATION

r(CUG): The molecular structure of a double stranded RNA with 6 CUG repeats in each slide was taken from high resolution X-ray data (PDB ID 3gm7). This model, ds[(CUG)₆]₂, was prepared using tleap module from the AMBER molecular dynamics package.⁹ The AMBER ff10 force field (which includes the χ torsional potential correction to the ff99b_{sco} force field)²⁹ was used. A total of 34 Na⁺ counterions were added to neutralize the system charge using Joung and Cheatham parameters.³⁰ The system was solvated in a truncated octahedron with a spacing of 12 Å around the RNA using the TIP3P water model,³¹ thus a total of 12,065 water molecules were included.

Purely CUG hairpins (r(CUG)₁₆ and r(CUG)₈) were built by homology modelling. Several X-ray structures containing CUG repeats exist (PDB IDs 3gm7, 3syw, 3szx, 4fnj, 3glp); however no tetraloop is present in any of them. For this reason, we selected a cUUCGg tetraloop (PDB ID 2koc) as a loop template whose high stability has been proved by high-resolution NMR. The same procedure was applied for obtaining a r[CCG(CUG)₈CGG] (henceforth named r(CUG)₈) which includes two capping regions, the tetraloop and a trinucleotide C•G cap. Both systems were prepared following the same protocol as ds[(CUG)₆]₂, using tleap and the AMBER ff10 force field. A total of 47 Na⁺ counterions and 19,770 TIP3P water molecules were added to r(CUG)₁₆. As for r(CUG)₈, 29 Na⁺ counterions and 10,565 TIP3P water molecules were necessary. A truncated octahedron with a spacing of 12 Å between the walls and the solute was used in both cases.

(CUG)₂: the molecular structure of a double stranded RNA with 2 CUG repeats in each slide was taken from high resolution X-ray data (PDB ID: 3gm7) and it was edited by capping the structural model with C•G pairs. The resulting model was prepared using tleap module from the AMBER molecular dynamics package.

rAUUCU: a model rAUUCU structure containing one 3×3 5'UCU³/5'UCU³ internal loop was prepared by ho-mology modeling. A symmetric system was designed with the sequence of r(CG AUUCUAUCG)₂ where C•G and G•C flanking pairs were included to increase the overall structural stability in the MD simulations. The system was prepared with ModeRNA software by extracting the (AUUCUAU)₂ fragment from the crystal structure (PDB ID: 5btm) and adding standard CG and GC base pairs from the rnaDB2005 database fragment library. The system was neutralized with 20 Na⁺ ions and solvated with 5095 TIP3P water molecules in a 12 Å truncated octahedron box.

CONVENTIONAL MOLECULAR DYNAMICS (cMD)

The AMBER force field with revised χ and α/γ torsional parameters for all simulations was used. Minimum net-neutralizing counterions were added to the system using Joung and Cheatham parameters. The system was solvated in a truncated octahedron with a spacing of 12 Å around the RNA using the TIP3P water model.

Prior to the production phase each system was prepared using the following protocol: the RNA was constrained and the solvent and counterions were minimized during a 2,500-step

minimization stage. Next, a second minimization 15,000-step long was run without constraints. After system minimization, the backbone was constrained and the system was heated at 300 K within 100 ps followed by a 200 ps long MD at 300 K with decreasing force constraints. After relaxation, a 2 ns long MD was performed under NpT ensemble ($p = 1$ bar, $T = 300$ K) allowing density balance. The production trajectories were obtained at NVT conditions at 300 K. In all the simulations the Particle Mesh Ewald (PME)³² method for treatment of electrostatic interaction was used under periodic boundary conditions. A 9 Å short-range cut-off was applied and the SHAKE algorithm was used to fix all hydrogen atom positions.³³ The time step was fixed to 2 fs and coordinates were stored every 1 ps.

ACCELERATED MOLECULAR DYNAMICS (aMD)

Accelerated molecular dynamics (aMD) simulations were conducted using the same protocol as equilibration and production runs of conventional molecular dynamics. The first 30 ns of the previous production trajectory were used to calculate the parameters related to the average total potential energy and the average dihedral energy parameters, required for aMD parametrization (i.e. EthreshD = 483 kcal/mol, alphaD = 11 kcal/mol, EthreshP = -36257 kcal/mol, alphaP = 2192 kcal/mol).

REPLICA EXCHANGE MOLECULAR DYNAMICS (REMD)

The simulation was carried out with the AMBER force field with revised χ and α/γ torsional parameters.³⁴ The system was minimized in two steps: first, all residues except solvent were held fixed with a restraint force of 500 kcal/mol-Å² and minimized with 2,500 steepest descent steps followed by 2,500 conjugate gradient steps. A second minimization step was performed without positional restraints using 10,000 steepest descent and 10,000 conjugate gradient steps. REMD simulation was carried out under periodic boundary conditions, using 40 replicas at constant volume. The temperature range was determined using the parallel tempering temperature predictor. A temperature range from 272 K to 363 K was spanned with uniform ratios for exchange of ~30% between neighboring replicas. Each replica was slowly heated to its corresponding replica temperature in 150 ps while constraining the solute with a force gradient of 8.0 to 0 kcal/mol-Å². Langevin dynamics with a collision frequency of 1 ps⁻¹ was used. A 20 ps of pressure equilibration step was then applied with isotropic scaling at 1 atm. Production runs were carried out at constant volume using an exchange frequency of 2 ps. Chemical bonds involving hydrogen atoms were constrained with SHAKE algorithm, which allowed an integration step of 2 fs in the production runs. PME was used in all calculations with a 9 Å long-range cutoff. A total of 5.03 μ s of accumulated simulation time was obtained for the rAUUCU system model.

STEERED MOLECULAR DYNAMICS (SMD)

To describe the instability of the RNA internal loop, we performed three independent SMD experiments by pulling away sequentially the U-U and C-C pairs. SMD simulations were performed with the same protocol described for the equilibration step of REMD, and initiated

with the final structure obtained from the equilibration step. First, a combination of velocities ranging from 0.01 and 1.0 Å/ns and spring constant ranging from 5 to 20 kcal/mol-Å² were tested using a 32 factorial design with 10 repeats. The pulling force was applied to the center of mass (COM) of the O₂, O₄ and N₃ atoms of the U-U pairs, and O₂, N₃ and N₄ atoms of the C-C pair, until a separation of 6 Å was achieved. Once the variables were optimized, a total of 25 successive SMD pulling experiments per pair were conducted with a constant velocity of 0.14 Å/ns and a spring constant of 10 kcal·mol⁻¹·Å⁻².

CLUSTERING AND STRUCTURAL ANALYSES

Conformations and frequencies of the different non-canonical U-U pairs were analyzed on each trajectory using the clustering tool from AMBER cpptraj module.⁹ Clustering was performed over each non-canonical U-U pair heavy atoms using the average-linkage algorithm and specifying a minimum distance cutoff ϵ of 3.0 Å.

Helical parameters were monitored using the 3DNA software and extracted at intervals of 20 ps, including intra-base pair parameters (shear, stretch, stagger, buckle, propeller and opening), inter-base pair parameters (shift, slide, rise, tilt, roll and twist) and backbone torsions.

REMD CLUSTERING

Each replica window was analyzed with the cpptraj AMBER module. Since the force field was parametrized at 300 K only the replica at this temperature was subjected to further analysis. Structures were clustered using the average-linkage hierarchical agglomerative method with a distance cutoff ϵ of 3.0 Å. Symmetry-corrected RMSD clustering was performed on a subset of atoms using the AMBER mask syntax (:1-22@P,C3',C4',C5',O3',O5').

STIFFNESS ANALYSIS

All RNA structural parameters were calculated using the 3DNA suite for the structures extracted from the MD trajectory at 20 ps intervals. Stiffness constants were computed according to Lankaš et al.¹⁴ Once the covariance matrix (\mathbf{C}) of the helical parameters was generated, the stiffness matrix (\mathbf{K}) was computed with the following equation:

$$\mathbf{K} = k_B T \times \mathbf{C}^{-1} \quad (3.1)$$

POTENTIAL OF MEAN FORCE

Potential of Mean Force (PMF) as a function of slide and helical twist was constructed as described previously for other biomolecular systems using observed and reference probability distributions.^{35,36}

GRID INHOMOGENEOUS SOLVATION THEORY (GIST)

Grid Inhomogeneous Solvation Theory (GIST)²⁸ analysis of solvent was performed using the algorithm incorporated in AMBER 14. Prior to the analysis Na⁺ atoms were removed from the system because GIST considers all non-solute atoms as solvent molecules. The grid was centered at (27, 27, 29) and dimensioned to 20 x 20 x 40 Å with a spacing of 0.50 Å. The output was analyzed with VMD.

ANM/PCA/EDA FROM EXPERIMENTS AND SIMULATIONS

The system is represented by a set of nodes (one node per atom if all-atom representation; one or several nodes per residue if coarse-grained representation). The general form for the ANM harmonic potential is:

$$V_{ANM} = -\frac{\gamma}{2} \left[\sum_{i=1}^{N-1} \sum_{j=i+1}^N (s_{ij} - s_{ij}^0)^2 \Gamma_{ij} \right] \quad (3.2)$$

where s_{ij} and s_{ij}^0 are the instantaneous and equilibrium distances of atoms i and j respectively, γ correspond to a homogeneous force constant, and Γ_{ij} is the ij^{th} element of the Kirchhoff matrix. In this study we have tested several distance dependent γ weighted by a negative exponential function before finding the best suitable one. The second order derivative of the potential energy function are collected in the Hessian matrix \mathbf{H} which can be decomposed in $3N - 6$ nonzero λ_i eigenvalues and their corresponding eigenvectors \mathbf{u}^i :

$$\mathbf{H} = \sum_{i=1}^{3N-6} \lambda_i \mathbf{u}^i \mathbf{u}^{iT} \quad (3.3)$$

Principal modes were obtained by decomposing the covariance matrix (\mathbf{C}) for the conformers:

$$\mathbf{C} = \sum_{i=1}^n \sigma_i \mathbf{p}^i \mathbf{p}^{iT} \quad (3.4)$$

where σ_i and \mathbf{p}^i correspond to the i^{th} eigenvalue and eigenvector of \mathbf{C} respectively, and n is the number of non-zero eigenvalues. The ANM covariance matrix is directly related with Hessian matrix (i.e. $\mathbf{C}_{ANM} \propto \mathbf{H}^{-1}$), thus the PCA σ_i is the counterpart of $1/\lambda_i$ and \mathbf{u}^i is the counterpart of \mathbf{p}^i .

For experimental validation of the ANM/PCA methods a (CUG)₃ ensemble was constructed by aligning (CUG)₃ fragments from available experimental structures (PDB IDs: 1zev³⁷, 3gm7¹¹, 3syw², 3sxx², 4e48¹⁶, 4fnj¹²). A total of 15 fragments were aligned with VMD and saved as a single rCUG PDB ensemble file. ANM analysis, which uses a single structure, were performed over one of the central CUG fragment of 3gm7. All ANM and PCA calculations were conducted with the ProDy suite.³⁸

ESSENTIAL DYNAMICS ANALYSIS (EDA)

Essential dynamics were based on the cross-correlation between the fluctuation of the P, C2', C4 and N3 atoms observed during the molecular dynamics trajectory. Essential modes were obtained by decomposing the C matrix for 2000 equally distributed snapshots extracted from the simulations.

COMPARISON METRICS FOR DOMINANT MODES

The overlap between ANM and PCA modes is given by the dot product of the corresponding eigenvectors:

$$O_{ij} = \mathbf{p}^i \cdot \mathbf{u}^j \quad (3.5)$$

The cumulative overlap was used to measure the correlation between predicted and experimental modes. The cumulative overlap is the extent to which a set of ANM soft modes can predict a PCA mode, hence it measures how well a subset of J ANM modes reproduces the i^{th} PCA mode:

$$CO_i^J = \left[\sum_{j=1}^J (O_{ij})^2 \right]^{1/2} \quad (3.6)$$

The essential subspace overlap between two subspaces spanned by top K modes is evaluated as:

$$SO^K = \left[\frac{1}{K} \sum_{i=1}^K \sum_{j=1}^K (O_{ij})^2 \right]^{1/2} \quad (3.7)$$

Finally, the degree of collectivity (κ), which provides a measure of the extent of distribution of motions across the structure, was computed using the definition proposed by Brüschweiler:

$$\kappa_k = \frac{1}{N} \exp \left\{ - \sum_{i=1}^N u_{ik}^2 \log(u_{ik}^2) \right\} \quad (3.8)$$

All these metrics were computed using the ProDy suite.³⁸

3.6 REFERENCES

1. Parkesh, R., Fountain, M. & Disney, M. D. NMR spectroscopy and molecular dynamics simulation of r(CCGCUGCGG)₂ reveal a dynamic UU internal loop found in myotonic dystrophy type 1. *Biochemistry* **50**, 599–601 (2011).
2. Kumar, A. *et al.* Myotonic dystrophy type 1 RNA crystal structures reveal heterogeneous 1 X 1 nucleotide UU internal loop conformations. *Biochemistry* **50**, 9928–9935 (2011).
3. Tran, T. & Disney, M. D. Two-dimensional combinatorial screening of a bacterial rRNA A-Site-like motif library: Defining privileged asymmetric internal loops that bind aminoglycosides. *Biochemistry* **49**, 1833–1842 (2010).
4. Klinck, R., Sprules, T. & Gehring, K. Structural characterization of three RNA hexanucleotide loops from the internal ribosome entry site of polioviruses. *Nucleic Acids Res.* **25**, 2129–2137 (1997).
5. Velagapudi, S. P., Seedhouse, S. J., French, J. & Disney, M. D. Defining the RNA internal loops preferred by benzimidazole derivatives via 2D combinatorial screening and computational analysis. *J. Am. Chem. Soc.* **133**, 10111–10118 (2011).
6. Jahromi, A. H. *et al.* Developing bivalent ligands to target CUG triplet repeats, the causative agent of myotonic dystrophy type 1. *J. Med. Chem.* **56**, 9471–9481 (2013).
7. Coonrod, L. a. *et al.* Reducing levels of toxic RNA with small molecules. *ACS Chem. Biol.* **8**, 2528–2537 (2013).
8. Park, H. *et al.* Crystallographic and Computational Analyses of AUUCU Repeating RNA That Causes Spinocerebellar Ataxia Type 10 (SCA10). *Biochemistry* **54**, 3851–3859 (2015).
9. D.A. Case, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B., X. Wu and, V. B. & D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B., X. W. and P. A. K. AMBER 14. (2014).
10. Lu, X. J. & Olson, W. K. 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31**, 5108–5121 (2003).
11. Kiliszek, A., Kierzek, R., Krzyzosiak, W. J. & Rypniewski, W. Structural insights into CUG repeats containing the ‘stretched U-U wobble’: Implications for myotonic dystrophy. *Nucleic Acids Res.* **37**, 4149–4156 (2009).
12. Coonrod, L. A., Lohman, J. R. & Berglund, J. A. Utilizing the GAAA tetraloop/receptor

- to facilitate crystal packing and determination of the structure of a CUG RNA helix. *Biochemistry* **51**, 8330–8337 (2012).
13. Galindo-Murillo, R., Roe, D. R. & Cheatham, T. E. Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochim. Biophys. Acta - Gen. Subj.* **1850**, 1041–1058 (2015).
 14. Lankaš, F., Šponer, J., Langowski, J. & Cheatham, T. E. DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations. *Biophys. J.* **85**, 2872–2883 (2003).
 15. Warf, M. B. & Berglund, J. A. MBNL binds similar RNA structures in the CUG repeats of myotonic dystrophy and its pre-mRNA substrate cardiac troponin T. *RNA* **13**, 2238–2251 (2007).
 16. Tamjar, J., Katorcha, E., Popov, A. & Malinina, L. Structural dynamics of double-helical RNAs composed of CUG/CUG- and CUG/CGG-repeats. *J. Biomol. Struct. Dyn.* **30**, 505–523 (2012).
 17. Yildirim, I., Chakraborty, D., Disney, M. D., Wales, D. J. & Schatz, G. C. Computational Investigation of RNA C U G Repeats Responsible for Myotonic Dystrophy 1. *J. Chem. Theory Comput.* **11**, 4943–4958 (2015).
 18. Brandl, M., Meyer, M. & Sühnel, J. Water-Mediated Base Pairs in RNA: A Quantum-Chemical Study. *J. Phys. Chem. A* **104**, 11177–11187 (2000).
 19. Doruker, P., Atilgan, A. R. & Bahar, I. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. *Proteins* **40**, 512–524 (2000).
 20. Setny, P. & Zacharias, M. Elastic network models of nucleic acids flexibility. *J. Chem. Theory Comput.* **9**, 5460–5470 (2013).
 21. Zimmermann, M. T. & Jernigan, R. L. Elastic network models capture the motions apparent within ensembles of RNA structures. *RNA* **20**, 792–804 (2014).
 22. Pinamonti, G., Bottaro, S., Micheletti, C. & Bussi, G. Elastic network models for RNA: a comparative assessment with molecular dynamics and SHAPE experiments. *Nucleic Acids Res.* **43**, gkv708 (2015).
 23. Bakan, A. & Bahar, I. Computational generation inhibitor-bound conformers of p38 map kinase and comparison with experiments. *Pac. Symp. Biocomput.* 181–192 (2011). doi:10.1016/j.humpath.2011.03.002
 24. Meireles, L., Gur, M., Bakan, A. & Bahar, I. Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins. *Protein Sci.* **20**, 1645–1658 (2011).
 25. Childs-Disney, J. L. *et al.* Induction and reversal of myotonic dystrophy type 1 pre-mRNA splicing defects by small molecules. *Nat. Commun.* **4**, 2044 (2013).
 26. Wong, C. H. *et al.* Targeting toxic RNAs that cause myotonic dystrophy type 1 (DM1)

- with a bisamidinium inhibitor. *J. Am. Chem. Soc.* **136**, 6355–6361 (2014).
27. Orellana, L. *et al.* Approaching elastic network models to molecular dynamics flexibility. *J. Chem. Theory Comput.* **6**, 2910–2923 (2010).
 28. Nguyen, C. N., Young, T. K. & Gilson, M. K. Grid inhomogeneous solvation theory: hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *J. Chem. Phys.* **137**, 044101 (2012).
 29. Zgarbová, M. *et al.* Refinement of the Cornell *et al.* Nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.* **7**, 2886–2902 (2011).
 30. Joung, I. S. & Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **112**, 9020–9041 (2008).
 31. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926 (1983).
 32. Toukmaji, A., Sagui, C., Board, J. & Darden, T. Efficient particle-mesh Ewald based approach to fixed and induced dipolar interactions. *J. Chem. Phys.* **113**, 10913–10927 (2000).
 33. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. . C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
 34. Yildirim, I., Stern, H. A., Tubbs, J. D., Kennedy, S. D. & Turner, D. H. Benchmarking AMBER force fields for RNA: Comparisons to NMR spectra for single-stranded r(GACC) are improved by revised χ torsions. *J. Phys. Chem. B* **115**, 9261–9270 (2011).
 35. Deng, N.-J. & Cieplak, P. Free Energy Profile of RNA Hairpins: A Molecular Dynamics Simulation Study. *Biophys. J.* **98**, 627–636 (2010).
 36. Yildirim, I., Park, H., Disney, M. D. & Schatz, G. C. A dynamic structural model of expanded RNA CAG repeats: A refined X-ray structure and computational investigations using molecular dynamics and umbrella sampling simulations. *J. Am. Chem. Soc.* **135**, 3528–3538 (2013).
 37. Mooers, B. H. M., Logue, J. S. & Berglund, J. A. The structural basis of myotonic dystrophy from the crystal structure of CUG repeats. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 16626–16631 (2005).
 38. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).

CHAPTER IV. RNA DRUG DESIGN STRATEGIES FOR DM1

4.1. BACKGROUND OF THE STUDY

Myotonic dystrophy type 1 (DM1) is mainly induced by pre-mRNA splicing defects caused by sequestration of proteins that regulate alternative splicing (e.g. MBNL1).¹ Disruption of the rCUG-MBNL1 complex has been established as the best therapeutic approach for DM1, hence two main strategies may be proposed: protein-targeting of MBNL1² or RNA-targeting of rCUG^{exp}.³⁻⁵ Unfortunately, recent studies showed that targeting MBNL1 causes dysregulation of alternative splicing, suggesting that MBNL1 is not suitable for therapeutic targeting.²

The rCUG^{exp} structure has been successfully targeted using several approaches such as *D*-amino acid hexapeptides,⁶ antisense oligonucleotides or small molecules.⁷ The potential for targeting RNA using small molecules is overwhelming and several scaffolds have been proved to reverse MBNL1 sequestering (e.g. bis-benzimidazoles).⁸ In an effort to explore the available chemical space, Chen et al. performed a quantitative high-throughput screening (qHTS) of more than 300,000 compounds and identified those molecules which inhibited r(CUG)₁₂ – MBNL1 complex formation.⁹ *In vitro* testing of some of these compounds showed distinct molecular mechanisms and divergent results in missplicing rescue. Among these, lomofungin and its dimer, dilomofungin, were identified as potential inhibitors due to its rCUG^{exp} binding potency but unexpected effects on RNA decay were reported.¹⁰

Zimmerman and coworkers followed a multivalent ligand strategy based on a previously reported triaminotriazine-acridine conjugate for increasing affinity through cooperative binding to rCUG^{exp}.¹¹ Bivalent ligands exhibited greater inhibition potency than its monomer in DM1 model cells. However, inherent cytotoxicity of intercalators led to the design of groove-binding ligands carrying two triaminotriazine units. In this study, Wong et al. reported partial relief of MBNL1 sequestration and a reversing of the eye phenotype in DM1 *Drosophila* model.¹² The identification of pentamidine and Hoechst 33258 as privileged scaffolds for disrupting rCUG^{exp} -

MBNL1 interaction led to the discovery of a new set of compounds similar to these in shape and chemical features. Using chemical similarity searching, Parkesh et al. identified a Hoechst derivative that improved DM1-associated splicing defects in cellular and animal models.¹³

In summary, targeting toxic RNA is a feasible and promising strategy for reducing nuclear MBNL sequestration, but selective targeting of nucleic acids is still challenging. Classical molecular modelling approaches include, among others, library searching and selection, quantitative structure-activity relationships (QSAR), molecular docking and molecular dynamics. These approaches have proved successful, especially structure-based techniques which try to simulate the RNA-ligand complex formation and its dynamic behavior.¹⁴ However, the lack of NMR and X-ray structural models of RNA-small molecule complexes limits the ability *in silico* approaches to perform and validate intensive virtual screening protocols.

Without minimizing the importance of the classical lock-and-key or complementary-shaped bodies model, it is increasingly becoming clear that RNA molecular recognition requires to consider the flexibility of the macromolecule and potential structural changes. Monod, Wyman and Changeux postulated the conformational selection model, which considers that the macromolecule conformations pre-exist in dynamic equilibrium prior to ligand binding, and the ligand shifts the equilibrium to a bounded state by recognizing and stabilizing one such conformation.¹⁵

Herein both ligand-based and structure-based drug design approaches are applied in order to identify novel classes of inhibitor scaffolds with promising activities. In particular, we combined essential dynamics analysis (EDA), which predicts RNA intrinsic dynamics using experimentally resolved rCUG^{exp} structures,¹⁶ and molecular docking, whose ability to predict bound conformations have been extensively probed.^{17,18} In addition, ligand-based methods such as pharmacophore screening and quantitative activity-structure relationships (QSAR) will also be assessed. Importantly, this chapter provides a fast and rational approach for designing chemical entities which may potentially and selectively recognize rCUG^{exp} structures.

4.2. NOVEL SCORING FUNCTION FOR NUCLEIC ACIDS – SMALL MOLECULE COMPLEXES

RNA is becoming increasingly important target as some diseases are mediated by its deregulation.^{19–21} Many targets including trinucleotide repeat expansions, small non-coding microRNAs (miRNAs) or HIV-1 TAR RNAs, among others, present well-defined secondary and tertiary structure. For instance, myotonic dystrophy type 1 is a paradigm of trinucleotide repeat expansions because of the entire deregulation of the muscular cell differentiation by an RNA gain-of-function mechanism.²² The pathogenic RNA is known to reproduce a metastable hairpin structure.²³ Other targets such as miRNAs have been observed to have several structural and functional characteristics similar to proteins.²⁴ With the increasing evidence of the importance of RNA, the knowledge of its structural diversity is increasing as well. RNA can form a myriad of well-defined structures which confers its druggability potential. DNA is a fountain of genetic information with a well-known and studied structure, which makes it a frequent target in drug design.^{25,26}

Probably, the most applied technique in conventional structure-based drug design (SBDD) is molecular docking, which have been historically designed for protein targeting. For this reason, many scoring functions are not parametrized for nucleic acids and cannot recognize nucleic acids-ligand interactions, hence docking methods become unavailable for this type of studies. Many approaches have been proposed during recent years due to the increase of X-ray and NMR complexed structures that allowed to create new knowledge-based scoring functions.^{27–30} Due to wide popularity of these methods, comparative assessments of docking protocols must be conducted. In order to develop most suitable docking methods, it is necessary to understand the current limitations of these methods.

Within this scope, herein a new scoring function is proposed using an artificial neural network (ANN) rescoring function which should be transferable to both RNA and DNA complexes. In addition, a benchmarking of several scoring functions for RNA targeting is presented, in particular DOCK6, rDock, Glide and Vina. An additional rescoring using LigandRNA has also been applied in combination with DOCK6 as a previous study suggested.²⁹

4.2.1. ARTIFICIAL NEURAL NETWORKS FUNCTION

Artificial neural networks (ANN) origin in efforts to produce computer models of the information processing that takes place in the brain. Supervised ANN learns the relationship between descriptors and biological activity through iterative prediction and improvement cycles. ANNs are trained by giving them sets of inputs patterns and associated target patterns. Through the iterative process, the internal representation of the data is modified until the predicted results reach an acceptable value (low error, low number of iterations, etc.). Its architecture is determined by the way in which the outputs of the neurons are connected to the other neurons.³¹

The most popular network used in ANN applications is the multi-layer feed-forward network. In this type of architecture, the neurons are arranged into groups called layers – an input layer, an output layer and a number of hidden layers. The input layer contains as many neurons as variables in the data set; the output layer's neurons number should coincide with the number of responses. However, the number of hidden layers and the number of neurons per layer required is rather inconsistent when it comes to characterizing networks and a trial and error procedure is usually required. Each pair of neurons is connected with a strength, or weight, and biases to each neuron which are both determined during the neural network training phase. The input variables are multiplied by the connection weights between the input and hidden layer. The hidden neurons sum the weighted signals from the input neurons and then project this sum on an activation function. The resulting activations of the hidden neurons are weighted by the connections between the hidden and output neurons and sent to the output neuron(s). The output neurons also perform a summation and projection on its activation function. The output of these neurons is the estimated response.³¹ The model optimizes the cost function by minimizing the root-mean-squared error (RMSE) using an iterative process such as the backpropagation learning algorithm.

In designing a neural network to analyze a complex, an adequate description of the system is needed so that the neural network can infer conclusions on its own. Herein we propose the construction of a new ANN scoring function (ANNScore) based on nucleic acid complexes from X-ray and NMR data. In this study, we used the current description of steric and electrostatics terms, hydrophobicity and hydrogen bonding included in Vina. The initial structure-based set was compiled from coordinates deposited in the Protein Data Bank (PDB). The training and test sets are described in Annexes (table A5, page 205). A total of 49 nucleic acids – ligand complexes with experimental binding data were used for constructing a combined DNA/RNA training and test sets (comprising 35 DNA complexes and 14 RNA complexes). Figure 4.1 shows the distribution of main physicochemical properties of the training and test data sets. Aminoglycosides were included in the training and test sets because they are an important class of RNA binders. However, as a result of these large molecular scaffolds, the number of rotatable bonds in the data set covered a very wide range (0 to 41 rotatable bonds). Hydrogen bond acceptors and donors range from 0 to 9, and highly charged compounds are found (between -2 and 6). Molecular weights range from 175.1 to 676 amu, and $xlogP$ ($-13.2 \leq xlogP \leq 4.3$) and topological polar surface area ($25.1 \text{ \AA}^2 \leq TPSA \leq 292 \text{ \AA}^2$) do not conform to conventional drug-like indices.

Procedure of the scoring function is schematized in figure 4.2. A total of 34 complexes (**training set**) were docked with Vina without weighting ($w_i=1$). Non-weighted docking terms (f_i)

were considered the inputs of the ANN and new weights were computed using the ANN iterative process. A **test set** consisting of 15 complexes was used to validate the model and retrieve its recognition and prediction capabilities. The scoring function was parameterized so that ANNScore scores the final docking poses by comparing the predicted and experimental ΔG values. The best ANN model was integrated into Vina, which is open source software since 2010. Finally, a benchmark comparison between Vina and the new ANNScore scoring function was performed with an external dataset containing 12 nucleic acid complexes to assess the improvement of ANNScore against the original Vina software.

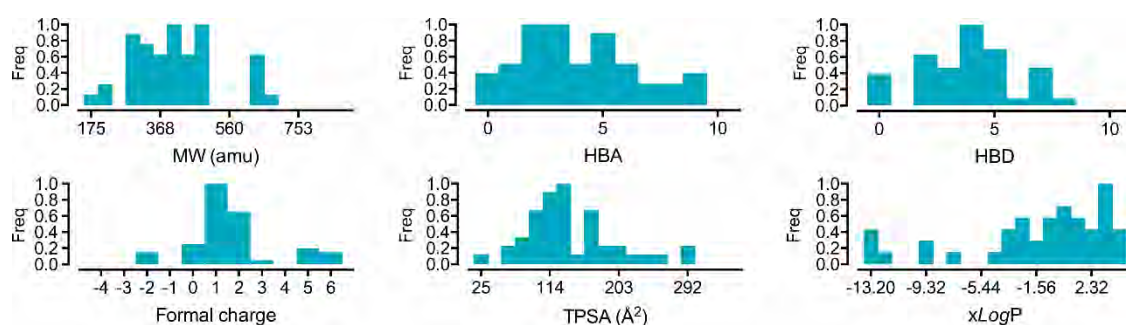


Figure 4.1. Distribution of main physicochemical properties of the training and test sets: molecular weight (MW), hydrogen-bond acceptors (HBA) and donors (HBD), formal charge, topological polar surface area (TPSA) and xLogP.

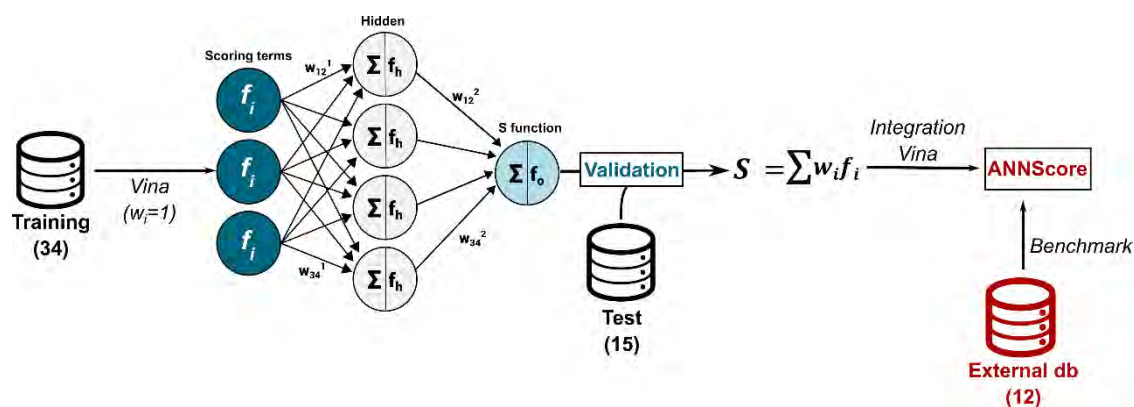


Figure 4.2. Schematic representation of the parameterization of the ANN scoring function. First, 34 molecular complexes were used as training set for the ANN. Several ANN models were completed and tested with a test set of 15 molecules. After validation, the best ANN model was integrated into the Vina's code. A final benchmark of Vina and ANNScore was performed with 12 external RNA complexes.

A total of 53 ANN models were performed and validated according to the test set. As shown in figure 4.3A, the best model was selected according to its root-mean-square error (RMSE) and recognition capability. The best model (red dot) yielded a recognition of 24.2% and RMSE of 0.651. Strong correlation between experimental and predicted data in this model indicates a high

predictive ability ($R^2_{\text{train}} = 0.89$, $R^2_{\text{test}} = 0.71$, figure 4.3B). Modification of the Vina scoring function was performed as described in the Methods section (page 150).

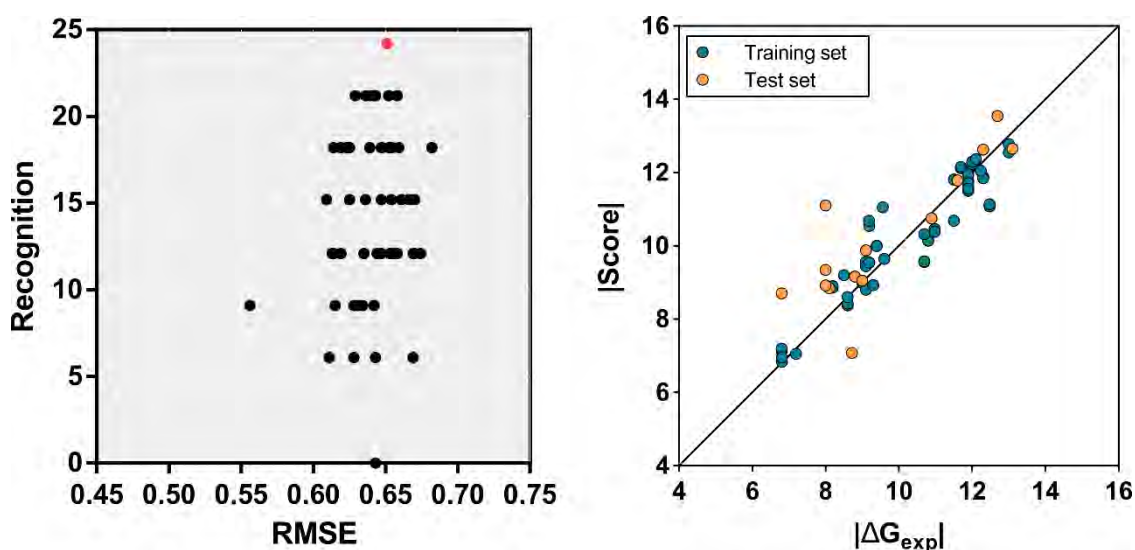


Figure 4.3. (A) Recognition vs root-mean-squared error (RMSE) of the 53 ANN models. Best model is represented as a red dot. (B) Experimental (ΔG_{exp}) vs predicted (Score) interaction energies. Data is separated into training and test sets. Correlation between experimental and predicted data is $R^2_{\text{train}} = 0.89$, $R^2_{\text{test}} = 0.71$.

Finally, a benchmark between the new parametrized scoring function and the original Vina function was conducted with 12 external nucleic acids - ligand complexes. This simple benchmark was intended to characterize the ability of each scoring function to select the lowest RMSD conformation as the highest scoring pose. To explore the variety of conformations generated by each sampling methods, > 100 poses were generated with Vina. Figure 4.4 shows that ANNScore function performs better than Vina's scoring function in most of the cases. In fact, ANNScore scores similar 'native' poses to Vina when $\text{RMSD} \leq 2 \text{ \AA}$. When $\text{RMSD} > 2 \text{ \AA}$, ANNScore yields the lowest RMSD pose except for 2gxm and 2gxj.

Notice that this ANN approach depends on the poses generated by Vina's algorithm, hence the purpose of the herein developed ANNScore is entirely intended for rescoring purposes. For this reason, the ANNScore scoring function should be able to rescore poses generated by any docking software. Nonetheless, docking programs generally use a scoring function that can be seen as an attempt to approximate standard chemical potentials. Thus, superficially physics-based terms like Lenard-Jones potentials and Coulomb energies are used within the scoring terms, which require to be significantly empirically weighted to account for the difference between energies and free energies of binding. On the contrary, Vina scoring terms rely more on a machine learning approach than directly physics-based functions, which greatly simplified its re-parameterization.

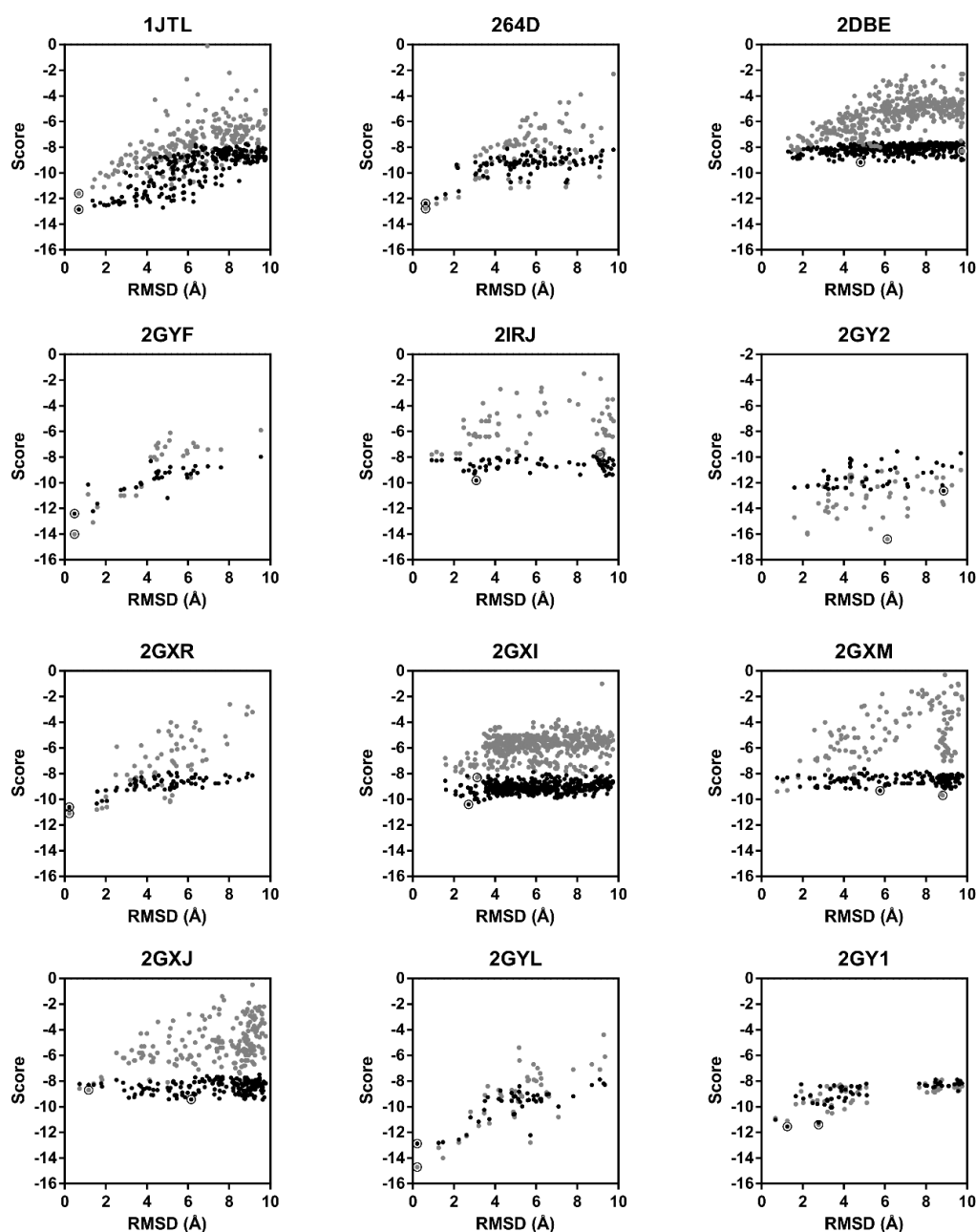


Figure 4.4. Score vs RMSD representation of poses scored with Vina (gray dots) and ANNScore (black dots). The lowest energy poses (best scoring) are circled.

4.2.2. BENCHMARKING SEVERAL DOCKING PROTOCOLS FOR RNA

Users of docking software are typically interested in obtaining a native-like model of a receptor-ligand complex, and for that purpose, a few top-scored poses are usually analyzed. In contrast to the previous presented benchmark, herein we examined the one, three and five top-scoring poses reported by several well-known docking software such as Glide, rDock, Vina and DOCK6. Top-scoring poses are defined as those poses with $\text{RMSD} \leq 2.5 \text{ \AA}$. The test set comprised 56 RNA-ligand complexes, which was used in a previously reported benchmark.²⁸ Table 4.1

contains the lowest RMSD found among a total of 50 generated conformations. Notice that Glide is not able to identify a ‘good’ pose in most of the complexes (only a 21%) while Vina and DOCK6 identified one correct pose at least in a ~50% of the cases. In contrast, rDock outperforms other software and identified a correct pose in 89% of correct conformations.

If only the top-scoring poses were considered, rDock found ‘native’ conformations among the top 5 in 75% of the cases, followed by DOCK6 (40%), Vina (36%) and Glide (20%) (see figure 4.5). If three top-scoring poses are considered, the trend is almost the same except for DOCK6 and Vina which yielded 32% and 33% respectively. However, a generalized huge decrease is observed when only the first top-scoring molecule is considered. Only rDock retrieves > 30% of ‘native’ conformations.

Further, we checked the combined potential of a conventional docking software in combination with a rescoring functions, in particular DOCK6+LigandRNA and Vina+ANNScore. Thus, it was checked whether the rescored poses improved the number of top-scoring conformations. The benchmark (figure 4.5) shows that individual LigandRNA and ANNScore are moderately effective in identifying poses that are close to the experimentally determined structures. LigandRNA increased ~10% of the five top-scoring poses retrieval while ANNScore increased Vina’s retrieval by ~14%. A negative effect is observed by looking the top 1 of Vina vs Vina+ANNScore, which decreases the chances of reporting a ‘native’ conformation. An example of the best pose among the top 5 obtained during the benchmark process is shown in figure 4.6 (structure ifyp).

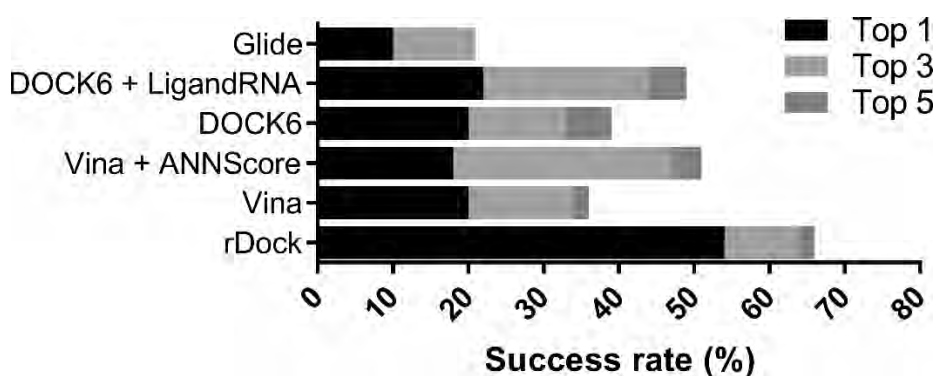


Figure 4.5. Identification of native conformations from one, three and five top-scoring poses using Glide, DOCK6, Vina or rDock. A combination of DOCK6+LigandRNA and Vina+ANNScore were also assessed in order to quantify the potential of rescoring functions.

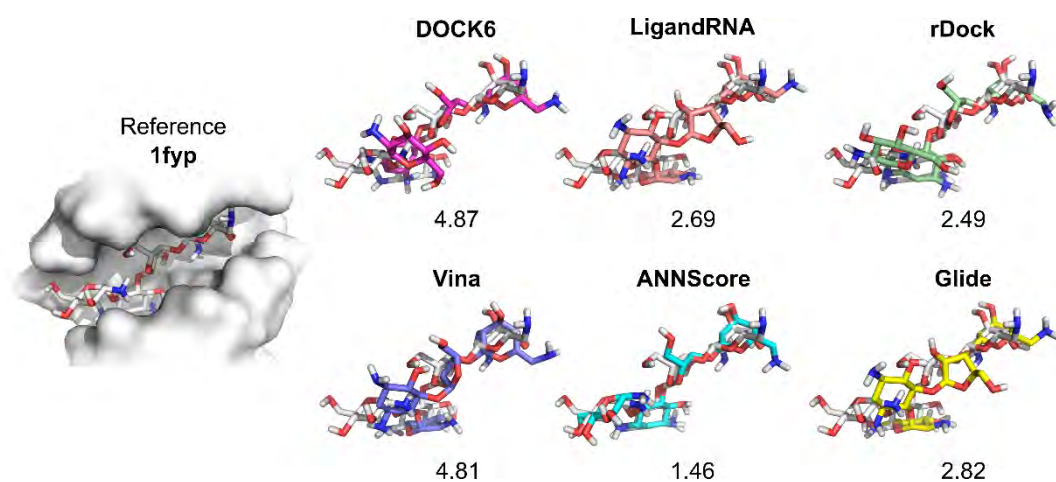


Figure 4.6. Superposition of native conformation (white sticks representation) and docking generated poses (colored sticks representation). Each conformation represents the best pose among the top 5 and includes its RMSD.

In summary, rDock outperforms the rest of docking software when testing RNA-ligand complexes. However, rDock requires higher computational times which compromises its virtual screening potential. Vina, in contrast, improves the speed of docking due to code optimization and multithreading which can efficiently screen hundreds of compounds in less than a day. However, Vina alone do not suffice to retrieve a high percentage of native conformations but ANNScore rescoring modestly improves its binding prediction capability. For this reason, prospective docking virtual screening performed along this thesis is performed with Vina+ANNScore unless stated otherwise.

Table 4.1. Best RMSD (< 2.5 Å) retrieved among all conformations.

PDB	Glide	rDock	Vina	DOCK6
1aju	3.18	1.31	2.97	1.62
1akx	6.44	1.93	2.41	2.16
1am0	4.54	0.86	3.51	4.52
1byj	9.20	1.28	1.74	1.92
1eht	0.45	0.70	3.49	0.70
1ei2	2.21	3.69	6.02	3.95
1f1t	1.93	0.65	0.40	4.00
1f27	10.10	1.43	1.25	1.35
1fmn	4.21	1.29	2.89	1.41
1fyp	2.82	1.46	1.16	2.52
1j7t	4.68	0.73	4.21	5.51
1koc	6.52	1.42	2.78	4.96
1kod	11.95	1.53	2.69	2.42
1lc4	8.74	0.71	1.27	11.21
1lvj	1.58	3.36	6.24	5.57
1mwl	2.29	0.39	4.56	0.36
1nem	3.29	1.02	1.26	0.47
1nta	8.41	1.57	1.67	1.71
1ntb	2.48	1.07	0.88	1.08
1nyi	2.47	1.91	7.40	5.93
1o15	8.98	1.05	0.94	3.54
1o9m	11.75	1.26	2.04	3.02
1pbr	9.71	1.51	0.74	0.63
1q8n	5.35	0.46	4.53	2.22
1qd3	0.49	2.87	5.38	13.24
1tob	8.35	1.10	2.09	1.62
1u8d	9.84	0.30	0.51	0.42
1uts	7.41	1.19	7.65	8.08
1uud	6.03	1.57	5.79	5.71
1uui	0.42	1.64	5.94	7.54
1xpf	9.07	1.84	4.43	4.22
1y26	12.44	0.30	2.39	0.37
1yrj	10.02	1.31	2.02	7.61
2au4	9.98	2.76	3.65	7.55
2be0	9.06	2.00	3.47	7.90
2bee	11.93	2.97	2.66	2.12
2esj	9.69	1.97	6.93	6.81
2et3	1.78	1.03	6.17	7.33
2et4	2.25	1.04	5.38	2.21
2et8	2.05	0.83	0.60	0.93
2f4s	2.61	0.78	0.66	1.38
2f4t	7.86	1.55	7.28	1.78
2f4u	2.77	1.49	1.28	1.66
2fcx	8.57	0.63	5.06	5.40
2fcy	6.69	0.98	5.87	2.87
2fcz	7.84	1.08	5.25	4.57
2fd0	9.76	1.47	5.33	6.32
2g5q	6.88	1.32	1.32	7.31
2juk	6.15	3.35	6.40	4.89
2o3v	5.18	2.02	6.89	6.03
2o3w	9.14	1.71	5.18	3.91
2o3x	8.61	0.85	1.15	2.45
2oe5	7.85	0.96	1.47	1.03
2oe8	8.87	0.74	1.55	1.48
2tob	8.82	0.35	1.95	1.07
3c44	7.15	1.19	1.46	0.70

4.3. LIGAND-BASED SELECTION OF COMPOUNDS FOR DM1

The University of Valencia (UV), in collaboration with Institut Universitari de Ciència i Tecnologia (IUCT) and Institut Químic de Sarrià (IQS) initiated in 2011 a small molecule screening campaign for the treatment of DM1. The screening consisted of a > 300 in-house compound library that were randomly selected and classified by their scaffold. In addition, two ‘reference’ compounds in DM1 were introduced: pentamidine and Höechst 33258. Within this context, it was sought the correction of the phenotype in a DM1 *Drosophila* model previously described. The screening was based on the correction of the *Drosophila* eye and muscle defects associated with the disease. Figure 4.7 shows the average recovery of the *Drosophila* population after treatment with the aforementioned compounds reported by the University of Valencia (personal communication, July, 2012). These values were normalized using a Z-score defined as follows:

$$Z = \frac{\%rec\ compound\ x - \langle \%rec\ all\ compounds \rangle}{\sigma} \quad [Eq. 4.1]$$

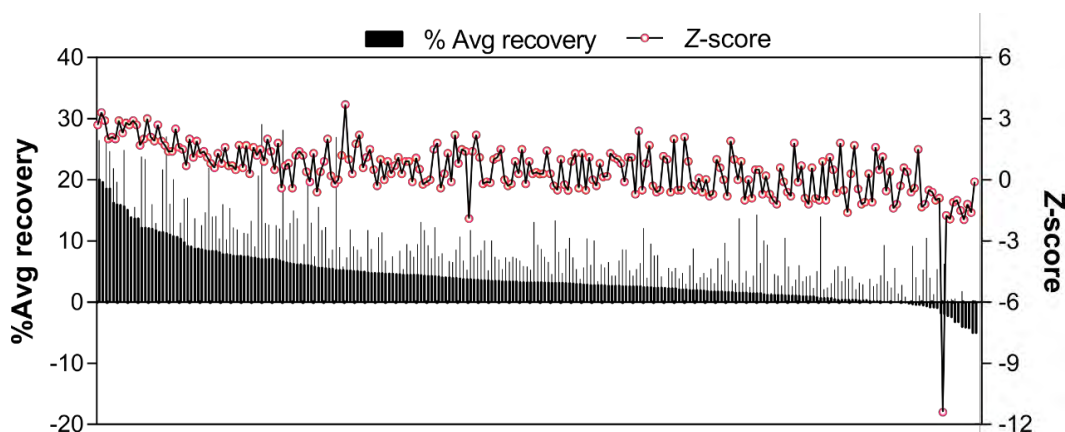


Figure 4.7. Percentage of average recovery of the *Drosophila* population after treatment with each compound in a bar plot representation. Standard deviation is indicated as vertical lines. Z-score is plotted using red circles by a black line.

Only 11% of compounds yielded an acceptable a *Z*-score above pentamidine's score that were mainly represented by small molecular units such as caffeine. As surprising as it may sound, caffeine has been previously described in literature as an intercalative agent for DNA but subsequent somatic instability has been observed.³²

4.3.1. DESCRIPTION OF THE CHEMICAL LIBRARY

Physicochemical properties of the entire chemical library are summarized in figure 4.8A. Molecular weights range between 41.1 and 710.2 amu and hydrogen-bond acceptors and donors between 0 and 10. Highly charged molecules exist among the database in [-4, 4] range. Nevertheless, neutral charged ligands predominate which should increase its selectivity for the RNA at the expense of receptor affinity. Finally, $xlogP$ ($-7.2 \leq xlogP \leq 11.2$) and topological polar surface area ($9.2 \text{ \AA}^2 \leq TPSA \leq 323 \text{ \AA}^2$) values exhibit a normal distribution around 0.5 and 75.3 \AA^2 respectively. Altogether, a 67% of our chemical database conform to the RO5 guidelines.

Next, the activity adjusted frequency of the most common fragments among the screened database was analyzed by using the Retrosynthetic Combinatorial Analysis Procedure³³ (RECAP, see figure 4.8B). Briefly, RECAP identifies fragments from molecule based on chemical knowledge. The highest frequency values are obtained for simple fragments such as piperidinyl, piperazinyl and imidazolyl substituents (7.94%). In fact, the latter fragment is found in many of the reported DM1 bioactive structures, like substituted benzimidazoles. In particular, bis-benzimidazoles have been extensively studied by Disney and co-workers and its bioactive potential has been clearly demonstrated.⁸ Other more complex fragments such as 4,5,6,7-tetrahydro-1*H*-purin-6-amine (1.71%), 3-methyl-3,4,5,7-tetrahydro-1*H*-purine-2,6-dione (1.38%) and *N*-substitued 3-methylbenzamide (0.62%) were also found, but yielded lower frequency values. This in-house chemical library comprises a low diversity of scaffolds mainly represented by substituted pyrido[2,3-*d*]pyrimidines and substituted imidazoles. Interestingly, substituted pyrido[2,3-*d*]pyrimidines are structurally similar 3-phenylpyrimido[5,4-*e*][1,2,4]triazine-5,7(1*H*,6*H*)-dione which is an active scaffold previously reported in literature.

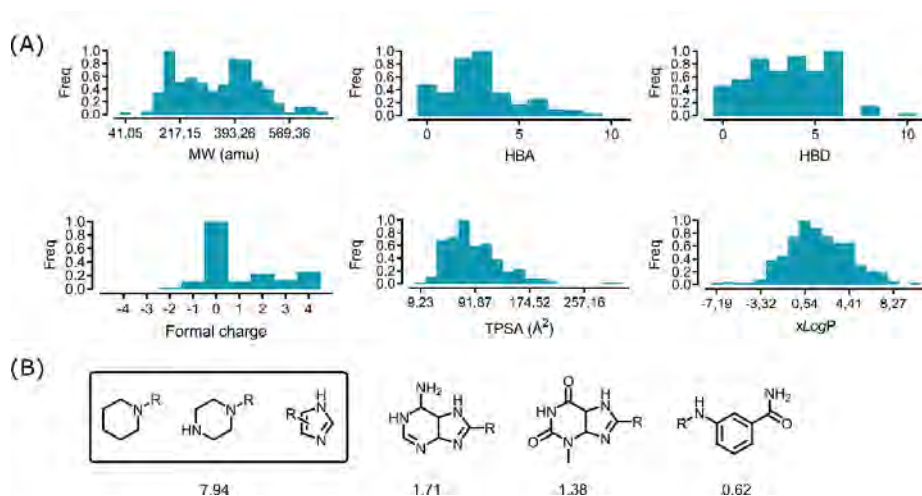


Figure 4.8. (A) Distribution of principal physicochemical properties: molecular weight (MW), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), formal charge, topological surface area (TPSA) and $xlogP$. (B) Most observed fragments ordered by activity adjusted frequency values (see Methods for a detailed description).

4.3.2. ACITIVITY-BINDING CORRELATION

A first approach to rationalize the recovery of the *Drosophila* DM1 model was to assess if any correlation between the observed recovery and *in silico* binding prediction of the tested small molecules over an rCUG structure existed. Figure 4.8 shows the receiver operating characteristic (ROC) plot obtained from the virtual screening of the in-house database. The ROC method differentiates two populations so that it can be applied to separate the active ligands against the decoys. The ROC curve represent the rate of true positives (TPR) versus the rate of false positives (FPR). True positives were described as those compounds whose bioactivity and docking score were equal or higher to Höchst 33258. Similarly, false positives presented high docking scores but their bioactivity was lower than the threshold. The red line indicates the fraction of the selected active ligands versus that of the selected 'decoys' or inactive molecules. If the docking program works as it should, the fraction of selected ligands will always be larger than its corresponding fraction of selected decoys, and the red line will be always above the black diagonal line. Thus, the plot shown in figure 4.9 demonstrates that either the docking method is not able to correctly classify between active and inactive molecules or the bioactivity data does not completely correlate with rCUG^{exp} binding. The flexibility of rCUG^{exp} structures may also affect molecular recognition. In addition, the training set is substantially different from our chemical library, hence the method may fail in recognizing novel types of interaction. In either case, animal models are subject to many ligand side effects and DM1 phenotype's complexity makes difficult the exploration of new chemical entities through structure-based methods. For this reason, we proceeded with ligand-based drug design methods.

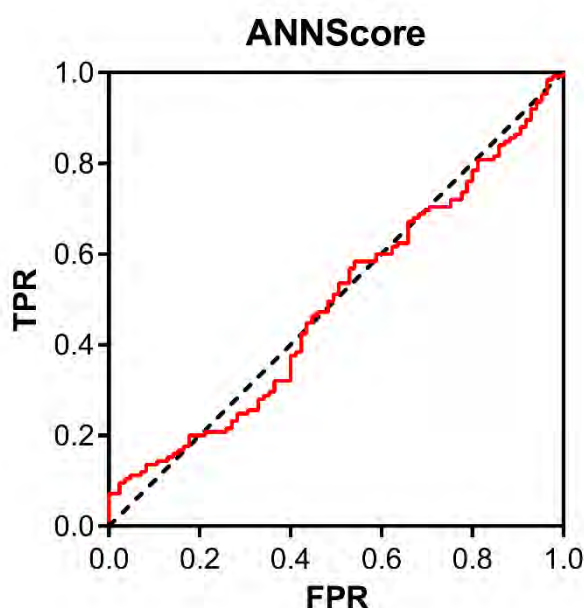


Figure 4.9. ROC plot of docking results vs *Drosophila* model data (Z-score). The plot represents the rate of false positives (FPR) and the rate of true positives (TPR). True positives are defined as molecules that scored higher than Höchst 33258 in the docking and present higher Z-score. False positives correspond to molecules with docking scores higher than Höchst 33258 but Z-scores lower than the threshold.

4.3.3. DE NOVO DRUG DESIGN: THEOPHYLLINE DIMERS

Based on previous results, we decided to search new chemical entities using two different approaches: de novo drug design and pharmacophore screening. The former approach was supported on the fact that caffeine is a bioactive molecule that tightly binds RNA but with low selectivity which confers its high toxicity. Thus, new molecules were designed within the scope of this project following a linker strategy: due to the repeated structure of the rCUG^{exp}, small molecular units connected by a linker should bind two sequential repeats (see structures in figure 4.10A). Our main hypothesis was that two theophylline monomers (methylated caffeine) should stack with two sequential U-U pairs in the minor groove side. Thus, a dimer should perform a double stacking while the linker accommodate hydrophobic interaction in the minor groove. Although theophylline's stacking cannot be simulated through conventional docking, the linkers' length was optimized according to docking results of the monomer which bind through the minor groove (figure 4.10C). Synthesis and selection of these compounds was based on synthetic availability (figure 4.10B).

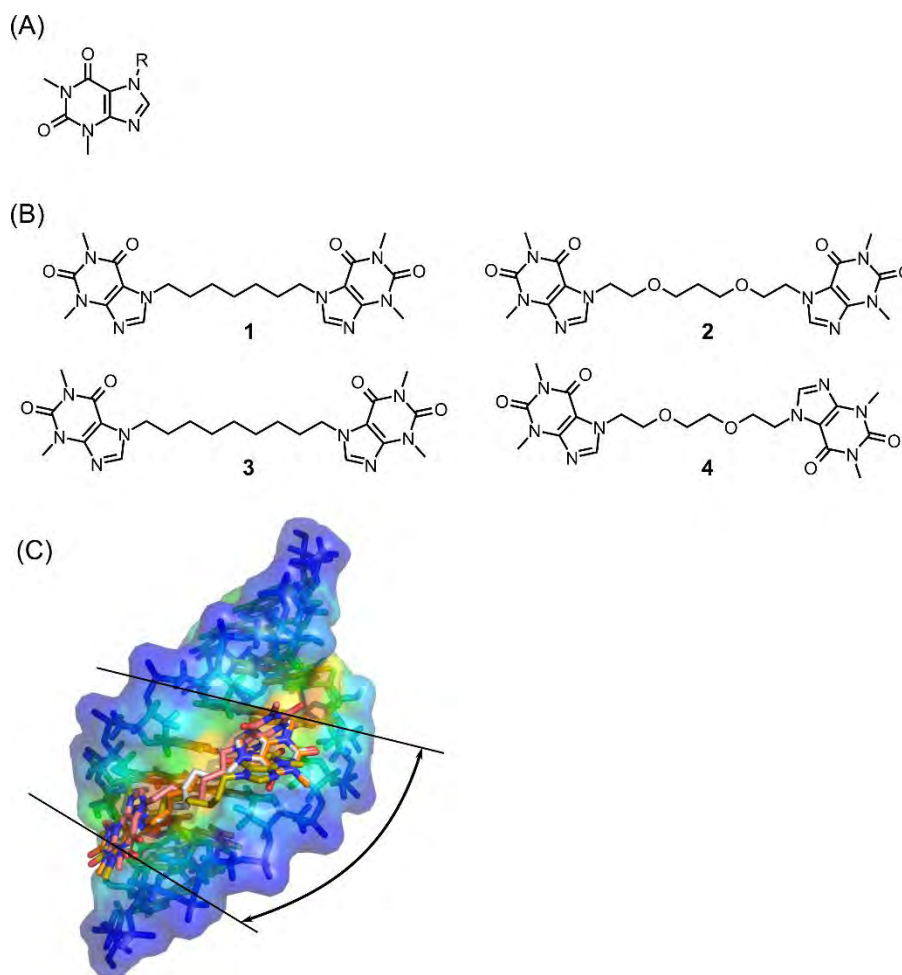


Figure 4.10. (A) Molecular structure of theophylline. (B) Molecular structure of theophylline derivatives. (C) Three-dimensional representation of compound **3** binding through the major groove. Multiple conformers are shown. Black lines and arrow indicate U-U pair planes and curvature of the minor groove.

Bioactivity of these compounds was assessed by Dr. Artero's group at University of Valencia. Relative fluorescence polarization (FP) informs on molecular orientation and mobility processes that change upon ligand binding to an rCUG^{exp}. Figure 4.11 shows the FP values for caffeine, theophylline, **1**, **3** and pentamidine at different concentrations. Compounds **2** and **4** showed no binding to rCUG^{exp} hence they were excluded from the analyses. No dose-response effect was observed except for compound **3**, which outperformed pentamidine's binding at sub-micromolar concentrations. Higher concentrations of **3** reversed this effect, probably because a quenching effect. In addition, compounds **1** and **3** were observed to significantly reduce nuclear foci in *in cellulo* studies with patient-derived myoblasts (data not shown).

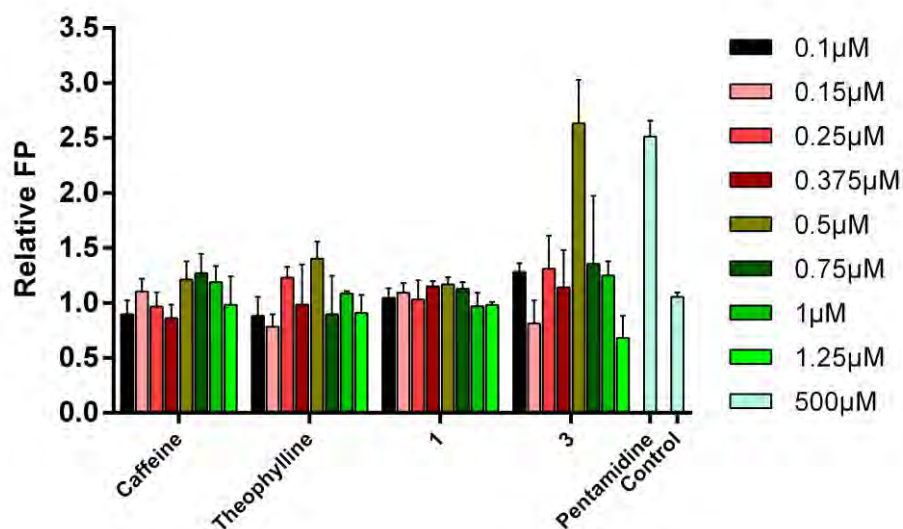


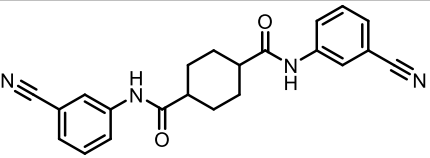
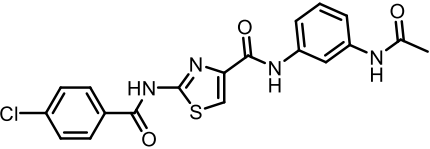
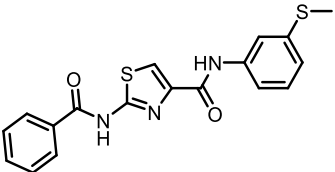
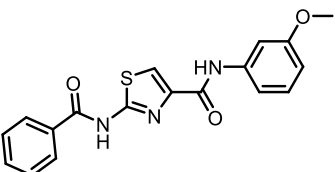
Figure 4.11. Relative fluorescence polarization (FP) for caffeine, theophylline and compounds **1** and **3** at different concentrations. Pentamidine at 500 µM and control (no compound) are included. Data was provided by Dr. Artero's group at University of Valencia.

4.3.4. PHARMACOPHORIC SCREENING: PENTAMIDINE-LIKE COMPOUNDS

Based on the extensive bioactivity data of pentamidine in the literature for DM1 treatment, we decided to enrich our chemical library with pentamidine-like structures. Using a previously described strategy, we selected four compounds by means of chemical and structural similarity.¹³ This method has been successfully applied in previous DM1 studies suggesting that it is an excellent approach to improve a potential hit or even elucidate novel scaffolds. In this study we screened a > 4 million drug-like commercial database. The 50 top scoring compounds were selected and analyzed accordingly. TanimotoCombo scores range between 1.16 and 1.35 (see structures in Annexes, pages 206-208). Next, we completed a diversity selection using our in-house software PRALINS.³⁴ Briefly, PRALINS performs either sparse or full array selection algorithms for rational selection of compounds. In this study, we used a genetic algorithm by which the most diverse molecules were selected using distance-based methods. Thus, the four most dissimilar molecules in terms of physicochemical properties were added to the in-house chemical library as our pentamidine-like subset (compounds MGL, see table 4.2). Table 4.2 also presents the FP values observed for these compounds. FP values were normalized to FP of

pentamidine at maximum concentration (500 μM). Hence the table indicates the increase or decrease of FP induced by each compound tested at 75, 150, 200, 250, 375 and 500 μM .

Table 4.2. Structure and bioactivity of selected MGL compounds. FP indicates normalized fluorescence polarization at distinct concentrations. Bioactivity data was provided by Dr. Artero's group at University of Valencia.

ID	Structure	FP75 μM	FP150 μM	FP200 μM	FP250 μM	FP375 μM	FP500 μM
MGL02		-0,61	-0,66	-0,52	-0,50	-0,53	-0,61
MGL03		-0,51	-0,61	-0,44	-0,64	-0,57	-0,64
MGL05		-0,59	-0,54	-0,54	-0,55	-0,61	-0,61
MGL011		-0,48	-0,55	0,05	-0,46	-0,50	-0,73

Despite the low FP values observed for this family if compared to pentamidine, MGL compounds have been observed to reduce foci and inhibit (data not shown). It has been previously hypothesized that DM1 treatment is not limited to rCUG^{exp} targeting. Wojciechowska et al. previously reported that small molecules that target specific kinases are able to reduce foci in DM1 patient-derived cells hence alternative pathways must be considered.³⁵ For this reason, further studies are currently being conducted due to promising bioactivity of these compounds.

4.3.5. CHEMOINFORMATIC ANALYSES FOR SCAFFOLD IDENTIFICATION

In order to gain more information about the bioactive DM1 compounds reported till date, a general analysis of the scaffolds and properties of these compounds was conducted. A quantitative high throughput screening (qHTS) for inhibitors of MBNL1-(CUG)₁₂ has been previously described for >300,000 compounds, which is the highest chemical library screened for DM1 till date.¹⁰ Among these, the activity of five of these compounds have been characterized in previous studies and showed differentiated effects on *in vitro* and *in cellulo* studies. Equally important, their associated molecular frameworks are also much differentiated but a clear correlation between their structural - physicochemical particularities and activity is still missing. For this database, the most frequent cyclic systems were identified (see figure 4.12). The set can be decomposed into 41 scaffolds which demonstrates a high diverse chemical space for inhibiting

the MBNL₁-rCUG^{exp} interaction. Each inner box in this figure contains the highest complexity scaffold, which contains at the same time lower complexity systems (outer boxes). Lowest complexity scaffolds are represented with its SMILES code.

Compound collection was analyzed in terms of radial fingerprints as implemented in Canvas (Schrödinger 2014). The projection into the PCA space ($\sigma^2 = 41\%$) of its radial fingerprints (see figure 4.13A) clearly demonstrates that active molecules may contain some privileged scaffolds and functional groups that inhibit MBNL₁ interaction with the rCUG target. Overall, the widespread distribution of data points are in agreement with the larger structural diversity of the molecular set. Inactive molecules (red circles) and inconclusive data from the primary screening (orange) are scattered around the PCA space, while active molecules (blue) are centered on the origin. Not surprisingly, the superposition of previously reported active molecules which were not in the qHTS set, such as the triaminotriazine-acridine conjugates and Höchst 33258 among others, fall into the active region of the PCA subspace. Two molecules are slightly off, which correspond to pentamidine (-1.1 and 0.8, for PC₁ and PC₂ respectively) and (*E*)-4-phenyl-2-(3-(thiophen-2-yl)acrylamido)thiophene-3-carboxylic acid (-3.7, 0.5); the latter was hypothesized to inhibit rCUG^{exp}-MBNL₁ interaction through protein inhibition, hence it may be discarded as an RNA binder.²

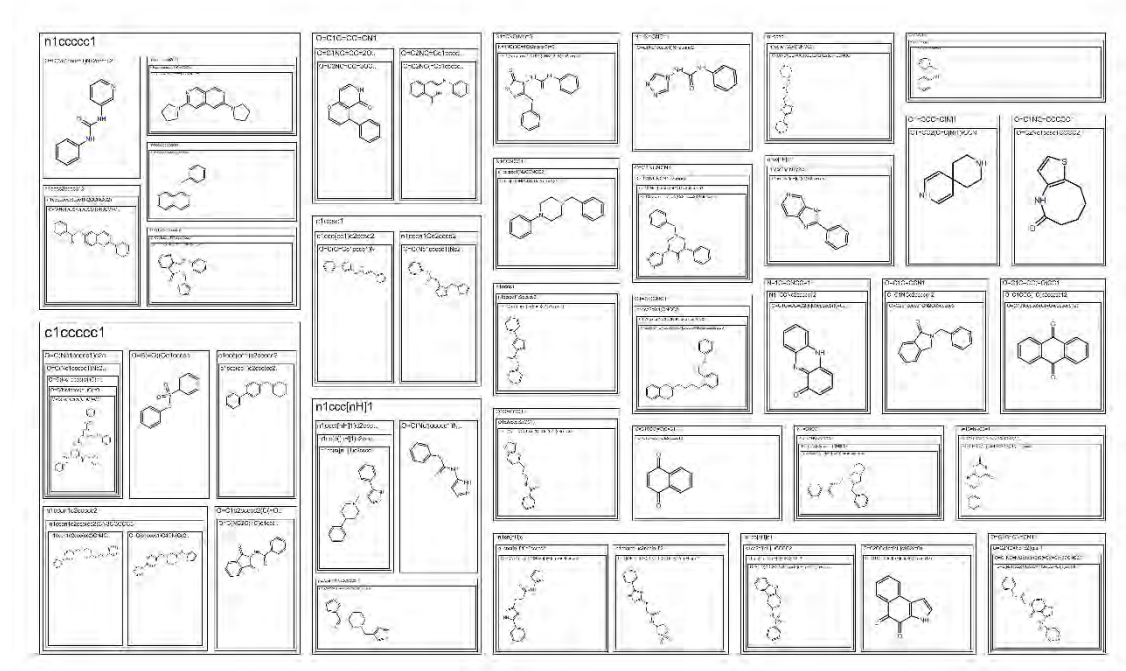


Figure 4.12. Scaffold tree map (molecular frameworks) derived from Scaffold Hunter. Only the highest complexity scaffold is shown for each cluster along with the SMILES code for each substructure. The space of each scaffold is filled according to their activity score.

Figure 4.13B shows the distribution of the main physicochemical properties of the aforementioned active molecules. Overall, molecular weights range between 30 and 1120.8 and charged molecules are not present, being neutral charged molecules the most common among the active data set. Hydrophobicity and polarity are evaluated with $\text{xlog}P$ ($-5.4 \leq \text{xlog}P \leq 8.2$) and topological polar surface area ($0 \text{ \AA}^2 \leq \text{TPSA} \leq 467 \text{ \AA}^2$) respectively, which values exhibit a normal

distribution around 2.9 and 103.7 Å² respectively. Altogether, a 94% of this chemical database conforms to the ‘Rule of five’ (RO₅) guidelines.

4.3.6. PYRIDO[2,3-*d*]PYRIMIDINES AS AN ACTIVE SCAFFOLD

Roughly, our in-house pyrido[2,3-*d*]pyrimidine chemical library can be divided into monomers and dimers, were two pyrido[2,3-*d*]pyrimidines subunits are linked with a spacer. While most of the monomers fall into the “active” region described by the previous qHTS analysis (figure 4.14), the dimers are displaced ‘south’ the active region, mainly because their higher structural complexity. Altogether, both monomers and dimers close to the “active” region that passed the RO₅ filters were selected for subsequent biological testing.

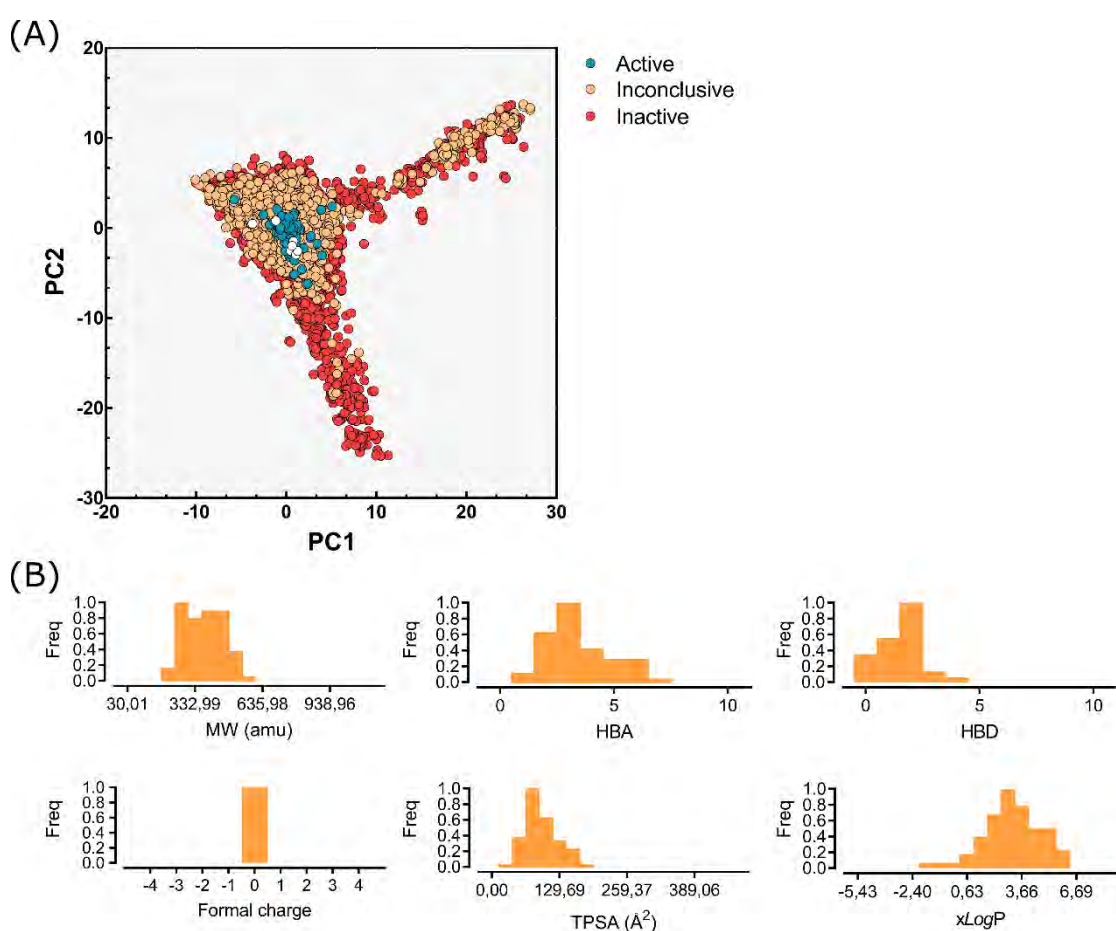


Figure 4.13. (A) Projection over the PC1 and PC2 subspace from the radial fingerprints. Inactive (red), inconclusive (orange) and active molecules (blue) are scattered over the subspace. White dots represent previously reported bioactive molecules. (B) Observed frequencies of main physicochemical properties distribution of the chemical dataset: molecular weight (MW), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), formal charge, topological surface area (TPSA) and xlogP.

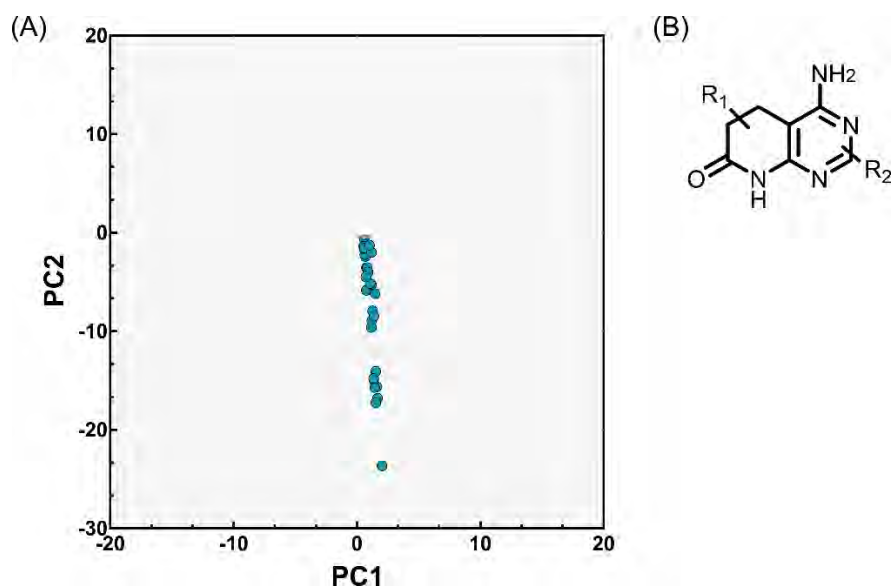


Figure 4.14. (A) Projection over the PC1 and PC2 subspace from the radial fingerprints of pyrido[2,3-*d*]pyrimidine structures. (B) General structure of substituted pyrido[2,3-*d*]pyrimidines.

In addition to these pyrido[2,3-*d*]pyrimidines (henceforth named **LRo6** to **LRu**), a subset of non-pyrimidinic structures that were in the ‘active’ region of the PCA space were also considered (**LRo1** to **LRo5**). Dr. Artero’s group at University of Valencia tested the ability of these compounds to bind an rCUG^{exp} structure using relative FP. Table 4.4 contains the results of the aforementioned compounds. FP values were normalized to FP of pentamidine at maximum concentration (500 μM) as described above. Hence the table indicates the increase or decrease of FP induced by each compound tested at 75, 125, 250 and 500 μM. Notice that only **LRo8** and **LRu** at 500 μM were able to induce an FP increase upon binding to cognate RNA. Nonetheless, this effect was observed also at lower concentrations for these two compounds. No improvement was observed with non-pyrimidinic compounds hence these scaffolds were discarded for prospective screenings (structures not shown). **LRo8** also showed promising results in DM1 patient-derived cells. These results made us confident to select pyrido[2,3-*d*]pyrimidines as a promising bioactive scaffold and, currently, following a hit-to-lead optimization process new derivatives are being tested.

Table 4.4. Relative fluorescence polarization (FP) normalized to pentamidine binding at 500 μM . This binding assay measures the ability of the small molecule to bind to an rCUG^{exp} structure *in vitro*. Values below 0 indicate less binding than pentamidine and values above 0 provide better binding than pentamidine. Compounds were tested at 75, 125, 200 and 500 μM . EC₅₀ (half maximal effective concentration) at 75 μM is also presented. Testings were performed by Artero's group at University of Valencia.

Pyrido[2,3-d]pyridimines							
Compound	R ₁	R ₂	FP (75 μM)	FP (125 μM)	FP (250 μM)	FP (500 μM)	EC ₅₀ (75 μM)
LR06	-Me		-0,23	-0,06	-0,16	-0,20	1.49
LR07	-Me		-0,37	-0,33	-0,37	-0,58	n.a.
LR08	-Me		-0,33	-0,04	0,69	0,66	1.18
LR09	-Me		-0,47	-0,39	-0,33	-0,14	n.a.
LR10	-Ph		-0,67	-0,63	-0,66	-0,60	n.a.
LR11	-Ph		0,22	0,52	1,03	0,23	2.36
Non-pyrimidinic compounds							
LR01			-0,60	-0,43	-0,59	-0,31	n.a.
LR02			-0,56	-0,50	-0,49	-0,30	n.a.
LR03			-0,40	-0,17	-0,42	-0,28	n.a.
LR04			-0,53	-0,49	-0,54	-0,49	n.a.
LR05			-0,52	-0,45	-0,33	-0,15	n.a.

4.4. DRUGGABILITY STUDY OF rCUG STRUCTURES

In an effort to have a better understanding of the chemotype requirements for RNA binding we proceeded to investigate the druggability of the rCUG^{exp} receptor from a structural perspective. RNA-ligand recognition is a complex phenomenon involving many factors, but computational analyses of druggable sites of the receptor have probed to correctly identify and rationalize structure-based drug design.¹² In this study, rationalization of rCUG^{exp} druggability was assessed using a molecular dynamics (MD) approach. Bakan et al. reported an MD based method that enables the identification of binding sites using water and organic molecules as probes.³⁶ Each probe describes a different type of interaction due to the differentiated physicochemical properties of the molecules such as hydrophobic/hydrophilic binding sites and polar or charged areas. Although the protocol was originally designed for proteins, we pursued the study of the druggability of a truncated version of an rCUG^{exp} target under the premise that RNAs may also contain binding pockets. For this study, we investigated two independent systems: **(A)** a system containing an r(CUG)₃ structure in a box of water and Na⁺ counterions, and including a mixture of 70% isopropanol, 10% acetamide (polar group representative) and 20% of sodium acetate – isopropylamine (charged group representatives); **(B)** an equivalent system with a composition of 30% isopropanol, 50% imidazole, 10% acetamide, 5% sodium acetate and 5% isopropylamine. While system **A** corresponds to a ‘standard’ composition for target druggability studies, system **B** composition was optimized for investigating the binding of imidazoles. Imidazole fragments, and in particular benzimidazole, is a recognized chemotype that selectively binds to RNAs containing 1x1 internal loops produced by U-U non-canonical pairs.

4.4.1. DRUGGABILITY ANALYSIS

The analysis of the most probable druggable regions, or hotspots, was provided by three “druggable” solutions per system (see figure 4.15). Red spheres correspond to the lowest energy hotspots (highest density probes) and blue spheres are the highest energy regions (lowest density probes). It was consistently seen that small molecule binding occurs mainly through the major

groove in system **A**. Nonetheless, the hotspots are mainly located around the U-U non-canonical pairs (in particular U₅ and U₁₄ of our model system), which are the most dynamic and accessible regions along the RNA structure. This observation is in line with a recent MD study that describes a pocket along the groove produced by the intrinsic U-U pair dynamics. Assuming that probe molecules reach Boltzmann distribution within the simulation, the inverse Boltzmann relation may provide the binding free energy for each spot. Not surprisingly, the highest affinities are achieved by hydrophobic and charged hotspots which yielded a binding free energy of $-1.66 \text{ kcal}\cdot\text{mol}^{-1}$ and $-2.38 \text{ kcal}\cdot\text{mol}^{-1}$ respectively. The latter was defined by a charged based interaction as shown in solutions 1 and 2 of model A, defined by the interaction with the phosphate group of U8. The maximum achievable affinity for this site was 15 nM, mostly contributed by isopropanol probe interactions. Solutions 2 and 3 of system **A** identified lower affinity binding sites of 23 nM and 80 nM respectively.

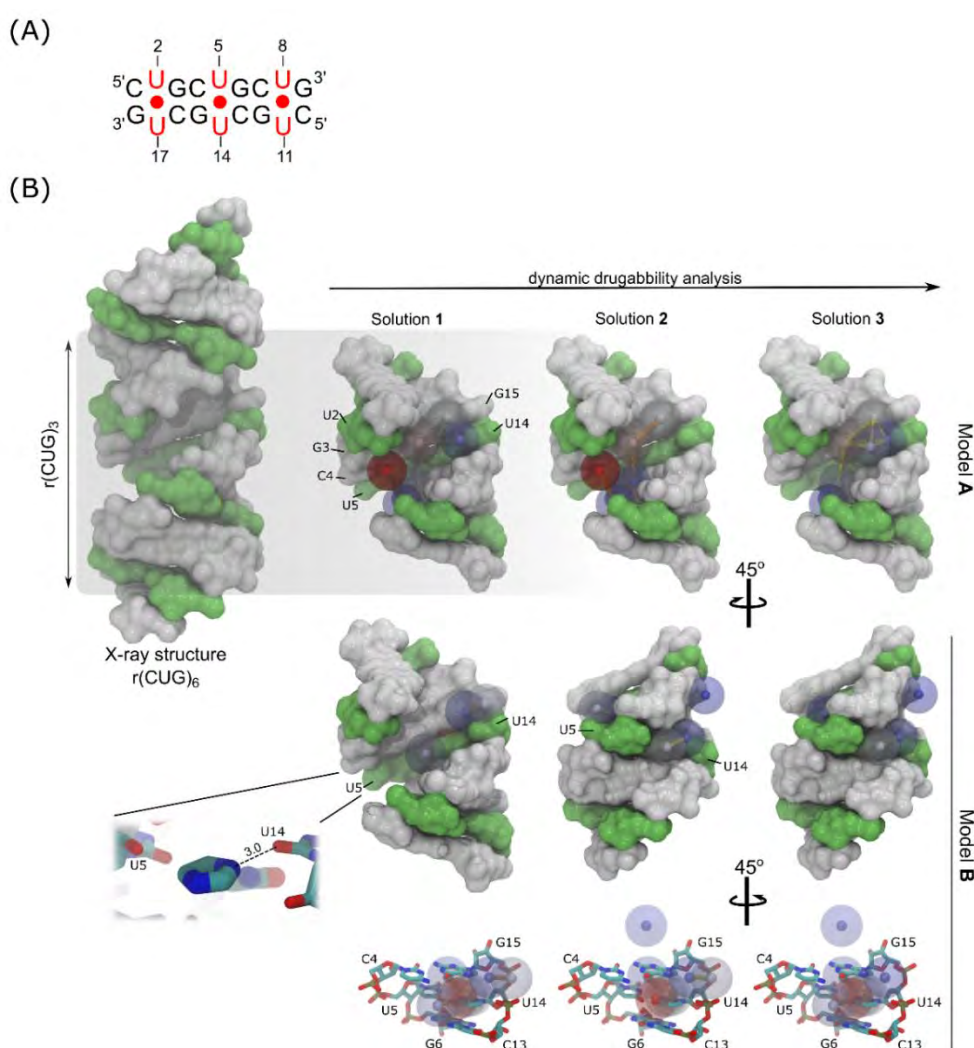


Figure 4.15. (A) Schematic representation of the r(CUG)₃ model system. (B) Druggability analysis description of model systems A and B. RNA is shown as surface representation with C•G and G•C pairs in white and U-U pairs in green. Three druggability solutions were obtained per system. Each druggable region, or hotspot, is represented by a colored sphere (red to blue, from lowest to highest energy of binding respectively). Notice that model **A** has a distribution of druggable sites along the major groove, but mainly located in the U-U pairs. Model **B** offers a stacking interaction pattern provided by the imidazole fragment, which stacks through one U-U pair and form an H-bond with the O4 atom of U14 (3.0 Å). Some features can also be observed along the minor groove.

In sharp contrast with the first system, model **B** was characterized by hotspots stacked into the middle of the U-U pair. The imidazole fragment is more than 40% of the MD simulation stacked between U₅-U₁₄, forming a unique hydrogen bond with the O₄ atom of U₁₄. Moreover, low energy regions are also found along the minor groove which are evenly represented by hydrophobic and charged interactions and, at a lower level, polar interactions. Highest drug-like affinity was found to be lesser than model **A** (37 nM for solution 1) but solutions 2 and 3 yielded regions with predicted affinities of 66 nM and 74 nM respectively, which are in the same range as the former model.

Indeed, both models of druggability provide feasible and high affinity solutions for rCUG^{exp} targeting (< 300 nM) and pointed out that this particular RNA can recognize different types of interaction depending on the small molecule chemotype. Nonetheless, this procedure may help to rationalize the structure-based drug design approach for prospective virtual screening and selection of compounds for its biological evaluation. On one hand, highly charged and hydrophobic compounds should potentially bind through the major groove of the RNA through specific U-U interactions and, on the other hand, H \ddot{o} chst-like compounds or imidazole-containing compounds are prone to stack around the U-U pairs or bind to the RNA minor groove as previously elucidated in a DNA-H \ddot{o} chst X-ray crystal structure.³⁷ Nevertheless, the applicability of this MD based method for assessing druggability could potentially explain the selectivity of specific chemotypes, such as imidazole derivatives, presumably by a combination of stacking interactions and specific hydrogen bonding patterns with the U-U pairs.

4.4.2. MOLECULAR RECOGNITION DEPENDS ON ESSENTIAL DYNAMICS

In silico studies of rCUG^{exp} have been proved successful in previous de novo design strategies.^{8,12,38} Among the most common drug design techniques, docking and molecular dynamics, or a combination of both, are useful for the structural rationalization of the design. However, molecular docking tends to yield poor complex predictions due to the lack of flexibility of the receptor during the process and MD requires from high computational power for assessing relatively short virtual screening campaigns. Although MD is the most reliable method for conformational sampling of macromolecules, it is still challenging to sample large amplitude fluctuation events (such as conformational changes upon ligand binding). On the other hand, Elastic Network Models (ENM) and Essential Dynamics Analysis (EDA) assume that the major collective modes of fluctuation dominate the functional dynamics. This approach has the advantage that dynamics along different modes can be inspected and visualized individually, thereby allowing to analyze local and global fluctuations separately. In this chapter, we apply these assumptions to rationalize from a structural perspective the bioactivity of the previously presented pyrido[2,3-*d*]pyrimidines for the treatment of DM1 by coupling docking and deformation of an r(CUG)₃ target structure along its normal modes.

First, a dynamic ensemble was constructed by deforming the structure along the 20 lowest frequency modes obtained from an all-atom model. Two deformations per mode were performed up to a mass-weighted RMSD of 2 Å, hence a total of 40 RNA conformations were obtained (see Methods for details). By doing so, we explored the conformational space accessible

to this RNA. We assumed that a subset of 20 soft modes would be sufficient to map the most significant changes within the RNA structure. In fact, this method enables ligands to ‘capture’ different RNA-small molecule complexes through conformational selection, thus a better description of the interaction can be achieved.

Computationally inexpensive rigid docking performs poorly when benchmarking the predicted affinities against experimental data. However, simulation-based methods and comprehensive sampling of ligand and RNA conformational space may take into account entropic effects and improve binding affinity predictions. For this reason, RNA flexibility was considered using the dynamic ensemble. A total of 50 RNA conformations of compound **5** were used and each molecule was docked over 40 different RNA conformations. Thus, a total of 2000 potential bindings were obtained. After rescore, ligands were ranked according to the provided ANNScore score. Compounds that scored values above the median of our references (pentamidine and H_öchst 33258) were considered as potentially active. ROC curves shown in figure 4.16 indicate that the EDA flexible docking improves the identification of new ligands using this structure-based strategy. Both enrichment factors (EF) and area under the curve (AUC) are higher if RNA flexibility is considered (see data in table 4.5).

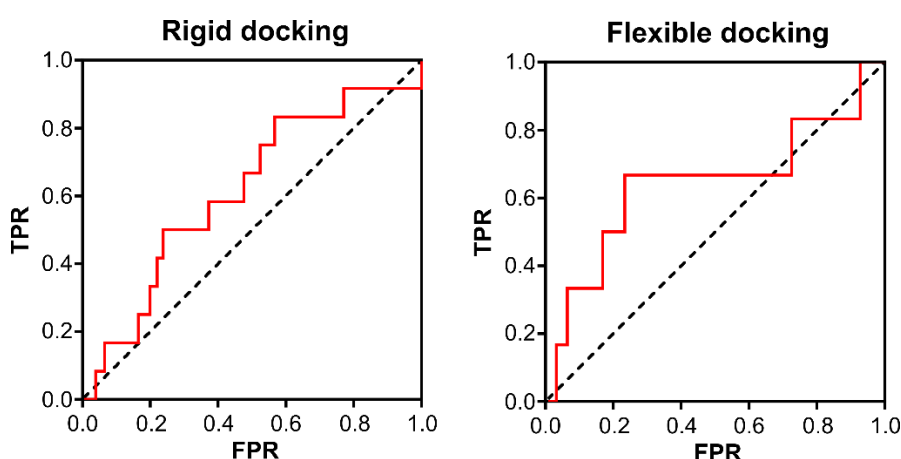


Figure 4.16. ROC curve obtained from the rigid and flexible docking. Area under the curve (AUC) and enrichment factors (EF) are presented in table 4.5.

Table 4.5. Area under the curve (AUC) and enrichment factors at 10%, 50% and 100% of screened database.

	AUC ^{10%}	AUC ^{100%}	EF ^{10%}	EF ^{50%}
Rigid	0.06	0.61	2.8	4.4
Flexible (EDA)	0.04	0.64	5.3	8.2

In good agreement with previous modelling studies, most of the conformations preferentially bind to the major groove of the rCUG^{exp}. More than three conformations are able to represent the lowest energy conformation in different RNA models which enhances the change of success of the conformational selection approach. In particular, the most energetically favorable conformation of compound **LRo8** binds through the RNA minor groove making hydrogen bonds with U₅-U₁₄ and G6 (figure not shown). The potential binding interactions

occur into the non-canonical pairs through polar contacts and hydrogen bonding. Moreover, the placement of the pyrido[2,3-*d*]pyrimidine subunit agrees with a predicted binding site by the druggability analysis for a minor groove binding (model **B**, solution 2, see figure 4.17). Hydrogen bond acceptors and donors of one pyrido[2,3-*d*]pyrimidine fall into the moderate and low affinity regions while the other interacts with the backbone through ribose interactions.

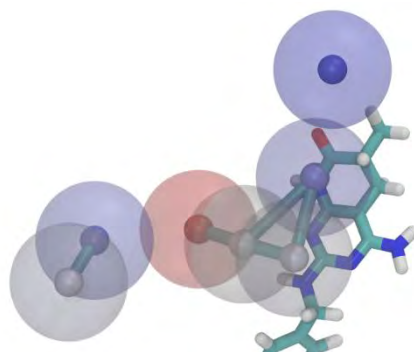


Figure 4.17. Superposition of monomer fragment of **LR08** and the binding hotspots (model **B**, solution 2) predicted with the druggability analysis.

Overall, these results are in line with previous modelling studies (either molecular docking or molecular dynamics) and should shed some light into the key interactions for small molecule RNA recognition. Moreover, we discovered pyrido[2,3-*d*]pyrimidines as a novel potential scaffold for developing new compounds for DM1, and we expect to develop further modifications according to structure-based methodologies.

4.5. QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR)

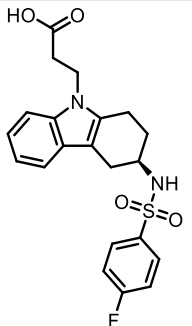
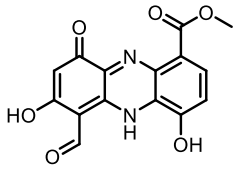
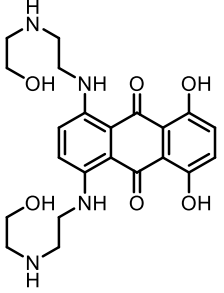
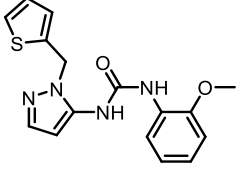
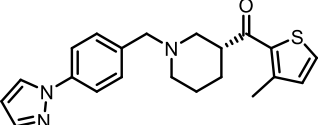
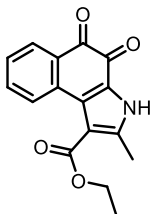
For both financial and social reasons, the drug industry needs to develop new drugs in much shorter times than is possible using a purely experimental approach. The number of compounds that can be obtained by organic synthesis is so high that the probability of choosing a random molecule with sufficient biological activity is almost zero. A method able to predict the biological activities of untested compounds is necessary to evaluate these molecular features in a rapid and inexpensive approach. Among these methods, the most important practice is quantitative structure-activity relationships (QSAR). QSAR is a mathematical equation that correlates the activity of compounds with one or more structural or physicochemical property. The resulting mathematical equation can then be used to predict the activity of other (usually related) compounds. These relationships establish correlations between the descriptors of individual compounds and their activity, which have proved especially productive when coupled with artificial neural networks (ANNs).

ANN-QSAR is a nonlinear modeling technique that has attracted increased attention in recent years. Its dynamic adaptability in different situations relies upon techniques inspired in learning. This behavior determines its ability to test hypotheses, detect statistical patterns and regularities and adjust an implicit model which is implemented in the architecture of the network model. The importance of ANN-QSAR methods is their flexibility and their advantage of providing results with higher speed and lower cost than experiments. In addition, principal component analysis (PCA) can also be coupled with ANN for nonlinear modeling between PCs and biological activity. In that regard, using the aforementioned qHTS database, an ANN-QSAR model has been performed in order to make a binary classification between active (1) and inactive (0) molecules that bind rCUG structures and inhibit its complex with MBNL1.

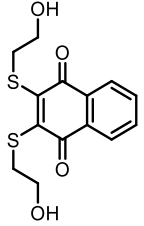
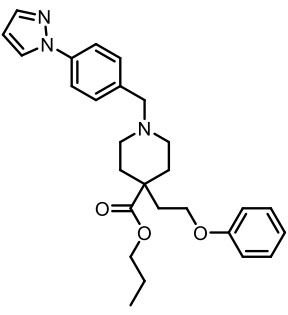
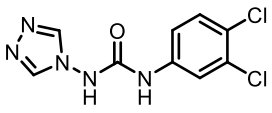
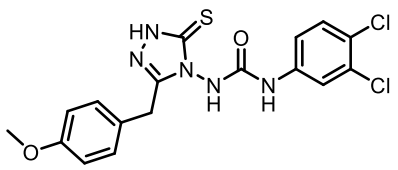
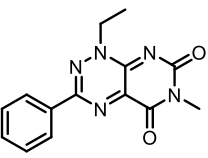
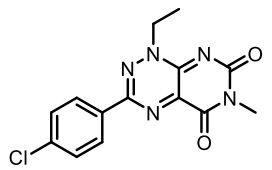
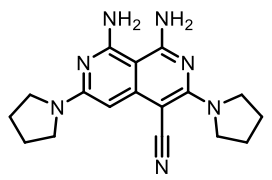
4.5.1. DESCRIPTION OF THE DATASET

The chemical library correspond to the previously described qHTS data set (see page 134). Table 4.6 contains the 13 most active molecules according to their scoring value. The score is a relative value that is scaled to each binding curve class' score range. Active compounds have score values between 40 and 100. Among these, lomofungin (CID 5351222) has been reported as a strong rCUG-binder and reduces rCUG-induced splicing defects.¹⁰

Table 4.6. Active compounds identified in a previously described qHTS screening.¹⁰

CID	Structure	IC ₅₀ (μM)	Efficacy (%)	Score
123879		0.01	106.39	97
5351222		0.29	96.48	90
4212		0.89	117.54	88
2544536		0.65	97.98	87
16189866		18.34	96	85
1736294		2.06	107.43	84

Continuation of table 4.5.

262093		10	86.12	84
16188674		6.51	116.90	83
816658		6.51	95.22	83
2121658		41.05	99.10	82
647501		0.30	96.94	69
460749		0.26	99.18	48
23724040		0.65	98.41	44

However, the disproportion between active (99) and inactive molecules (> 300,000) would compromise its predictive power, hence a previous selection of inactive molecules was performed. A total of 200 inactive molecules were selected according to their physicochemical properties using a diversity-based selection. Main physicochemical properties of the final selected subset (active and inactive, 299 molecules) are shown in figure 4.18.

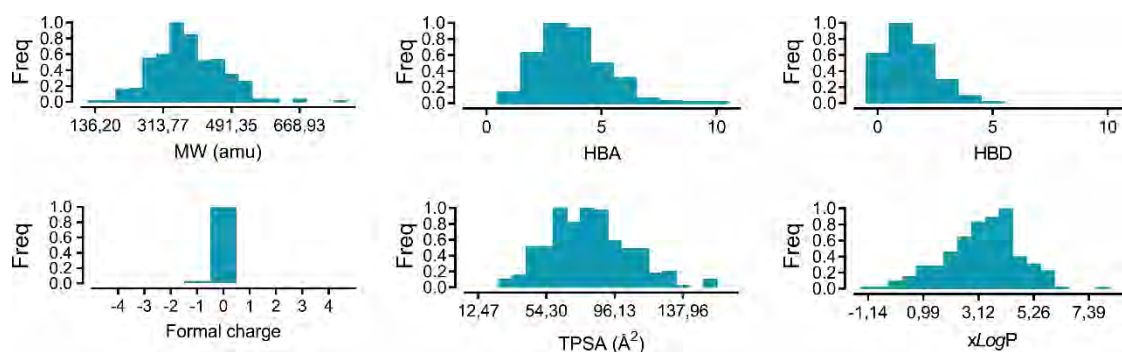


Figure 4.18. Distribution of principal physicochemical properties of the selected molecules: molecular weight (MW), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), formal charge, topological surface area (TPSA) and $x\log P$.

Next, molecules were ‘described’ using a total of 192 2D molecular descriptors. This type of descriptors are defined to be numerical properties that can be calculated from the connectivity table representation of the molecule. Therefore, they are not dependent on the conformation of the molecule and are most suitable for large database studies. Roughly, they can be classified as:

- **Physical properties.** E.g. polarity, density, formal charge, weight, hydrophobicity, etc.
- **Subdivided surface areas.** These descriptors are based on an approximate accessible van der Waals surface area calculation for each atom along with some other atomic property.
- **Atom/bond counts.** Functions of the count of atoms and bonds.
- **Kier & Hall connectivity and kappa shape indices.** They compare the molecular graph with minimal and maximal molecular graphs, and intend to capture different aspects of molecular shape.
- **Adjacency and distance matrix.** The adjacency matrix is defined by 1 when two atoms are bonded and zero otherwise. The distance matrix is defined by the length of the shortest path from two atoms.
- **Pharmacophore features.** Atom types are assigned to each heavy atom (non-hydrogen atom). Pharmacophore features depend on the atom types, and are mainly classified as donor, acceptor, polar, positive (base), negative (acid), hydrophobe and other.
- **Partial charge.** Descriptors that depend on the partial charge of each atom of a chemical structure.

However, an inevitable difficulty when dealing with molecular descriptors is that of collinearity, which may exist between independent variables. Thus, a better predictive model can be obtained by orthogonalization of the variables by means of principal component analysis (PCA). In addition, in order to decrease the dimensionality of the independent variable space, a restricted number of PCs are used. Herein we used the simplest and most frequent methods, which consist on the ranking of the PCs in order of decreasing eigenvalue. As reported in figure 4.19A, 9 PCs accounts for 72% of total variance and no significant improvement is achieved afterwards. Hence these 9 PCs were selected for the ANN-QSAR input. Figure 4.19B shows the normalized weight of each descriptor in each PC. The first PC is mainly governed by the

'petitjean' descriptor which defines the eccentricity of a vertex to be the longest path from that vertex to another vertex in the graph. In other words, it represents a balance between its cyclic and acyclic parts. In fact, most of DNA and RNA minor groove binders depend on shape and curvature in order to optimize their hydrophobic interactions. Among the molecular descriptors, molar refractivity (SMR), hydrophobicity (SlogP) and atomic partial charges (PEOE) are also significant.

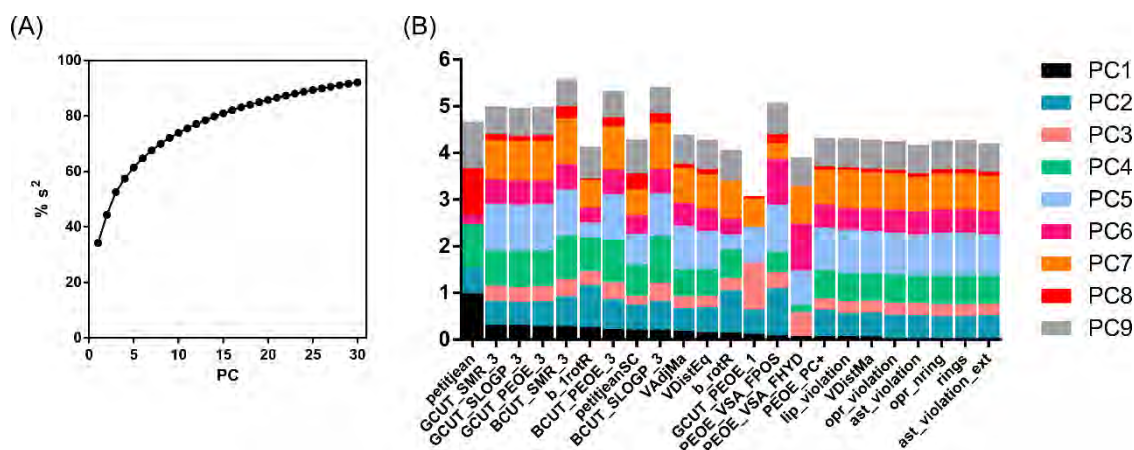


Figure 4.19. (A) Percentage of explained variance (s²) versus number of selected principal components (PCs). (B) Normalized molecular descriptors weights for each PC.

4.5.2. ANN-QSAR MODEL PERFORMANCE

Several conditions were tested for optimizing ANN-QSAR performance. This included fine-tuning of the topology and the learning rate. The learning rate affects the speed at which the ANN arrives at the minimum solution. If the learning rate is too high the system will either oscillate about the true solution or it will diverge. If it is too low, the system will take a long time to converge on the final solution. Defined topologies included one or two hidden layers with 3 to 8 neurons per layer. Learning rates were tested at 0.001, 0.005 and 0.01. The number of iterations was set to 4000 for all the models. The data set was split into a training set (269) and test set (30) by random selection. Each model was run 10 times and their classification capability was assessed according to recognition (percentage of well classified molecules in the training set) and prediction values (percentage of well classified molecules in the test set). Accuracy (fraction of true positives and true negatives in the test set) was assessed and compared to the Cohen's kappa factor (κ) which measures the agreement between experimental and predicted activities as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (4.1)$$

where p_o is the relative observed agreement, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. Complete agreement implies $\kappa = 1$, and no agreement other than would be expected by chance, implies $\kappa \leq 0$.

As shown in figure 4.20 the best classificatory model was selected by Pareto criterion, and yielded a 94.8% of recognition and 83.3% of prediction capabilities with a total RMSE of 0.053. Moreover, the κ factor (0.66) indicated a good agreement between experimental and observed activity. This model was achieved with a single hidden layer of 6 neurons and a learning rate of 0.001. Confusion matrices for the training and test sets are presented in table 4.7.

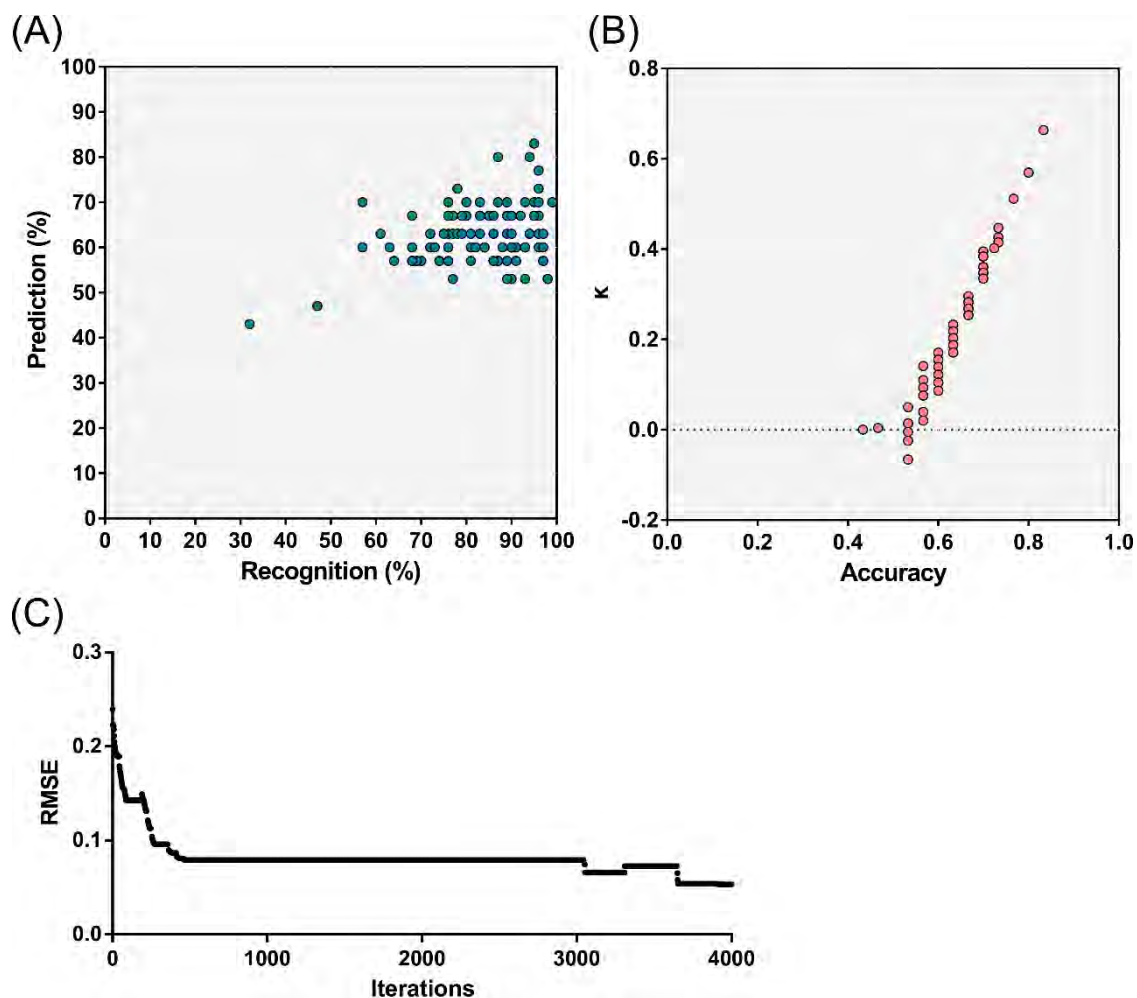


Figure 4.20. (A) Prediction versus recognition capability for each ANN model. (B) Representation of kappa (κ) factor's dependence of accuracy. (C) Evolution of the root-mean-squared error (RMSE) for each iteration for the best ANN model.

Table 4.7. Confusion matrices of the training and test sets.			
Training set			
		Predicted	
		Active	Inactive
Experimental	Active	76	7
	Inactive	10	176
Test set			
		Predicted	
		Active	Inactive
Experimental	Active	14	2
	Inactive	3	11

A total of 10 and 3 false positives were identified in the training and test sets respectively (see structures in figure 4.21). Most of the molecules are characterized by sulfone and sulfonamides, and triple bonds. In contrast, pyridinylureas and imidazolidines are frequently found among the data set.

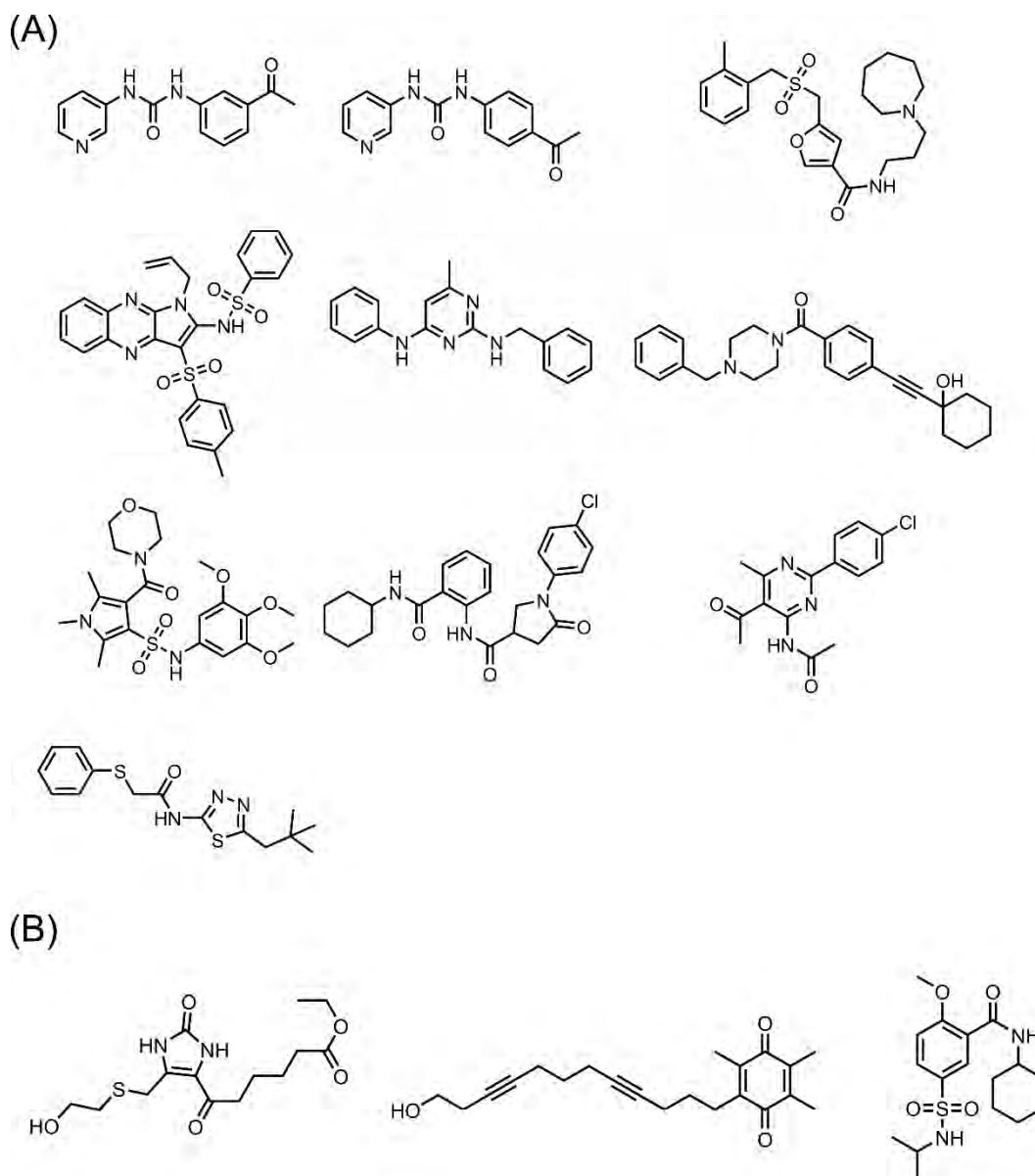


Figure 4.21. False positive molecules obtained in the training set (A) and the test set (B).

In conclusion, our ANN-QSAR model provides an excellent classification between active and inactive molecules that can avert the rCUG^{exp}-MBNL₁ interaction. This model can be applied for prospective virtual screening to identify new chemical entities for DM1. However, as a classificatory model, it cannot quantify bioactivities hence further efforts will be pursued to establish new models that can accurately predict novel molecules' bioactivity. The main strength

of ANNs is that they are nonlinear and so can deal with structure-activity relationships in which there are nonlinear correlations between activity and one or more molecular descriptors. However, the main weaknesses of ANNs are that they do not yield a QSAR equation directly, hence they are difficult to interpret and they do not work well with small data sets.

4.6. PROTOCOLS

4.6.1. EQUIPMENT

The following computational methods and procedures were accomplished with in-house systems:

In-house equipment

- 4x Intel Xeon 8-core at 3.50GHz, 32GB of RAM memory. NVIDIA Quadro K4000. 2TB filesystem storage.
- Intel Core 2 Quad Q8200 2.33 GHz, 4GB of RAM memory. 2TB filesystem storage.

4.6.2. PROTOCOLS

ANN SCORING FUNCTION

ANN scoring function. The ANN scoring function (ANNScore) is based on nucleic acid complexes from X-ray and NMR data. All ANN models were performed with an in-house software, ArIS (Artificial Intelligence Suite), described in Dr. Estrada's thesis.³⁹ Description of steric and electrostatics terms, hydrophobicity and hydrogen bonding from Vina were used. The training and test sets are attached in Annexes (table A5, page 205). A total of 49 nucleic acids – ligand complexes (35 for DNA and 14 for RNA) with experimental binding data were used for constructing a combined DNA/RNA training and test sets. A total of 34 complexes (training set) were docked with Vina without weighting ($w_i=1$). Non-weighted docking terms (f_i) were considered the inputs of the ANN and new weights were computed using the ANN iterative process. A test set consisting of 15 complexes was used to validate the model and retrieve its recognition and prediction capabilities. The scoring function was parameterized so that ANNScore scores the final docking poses by comparing the predicted and experimental ΔG values.

Benchmark of scoring functions for RNA. Benchmarking of six scoring functions was conducted by comparing the number of best poses ($\text{RMSD} \leq 2.5 \text{ \AA}$) among the top 1, top 3 and top 5 scored conformations. The scoring functions used for the benchmark were rDock, DOCK6, MOE (London dG), Vina, LigandRNA (for rescoring of DOCK6 poses) and ANNScore (for rescoring of Vina's poses).

SELECTION OF NEW CHEMICAL ENTITIES

Chemoinformatics analysis of the chemical space. Compound collections were analyzed and compared based on physicochemical properties, scaffolds and radial fingerprints.

Physicochemical properties were computed using FILTER from the OpenEye suite (version 3.0.0).⁴⁰ To obtain a visual representation of the molecular space, a principal component analysis (PCA) was carried out using Canvas from the Schrödinger 2014 suite⁴¹ considering the radial fingerprints properties. Scaffold analysis was conducted using Scaffold Hunter.⁴²

Similarity screening. Similarity screening was performed with ROCS from the OpenEye suite (version 3.0.0).^{43,44} The chemical library was selected from the ZINC subset database (clean lead-like, $250 \leq MW \leq 350$, $xlogP \leq 3.5$ and rotatable bonds ≤ 7). First, 100 conformers per molecule were generated using Omega (version 2.02),⁴³ including the reference molecule pentamidine. Then, the 50 most similar molecules to pentamidine were selected as potential hits as determined by the TanimotoCombo scores (≥ 1.0). A complete list of the selected compounds is available in the Annexes (pages 206-208).

Diversity selection was performed using our in-house software PRALINS.³⁴ First, physicochemical descriptors of the 50 potential hits were computed with MOE (Molecular Operating Environment) software. Then, a diversity selection algorithm was applied with PRALINS and the 4 most dissimilar compounds were selected, defining the Euclidean distance in feature space as metric.

DRUGGABILITY AND BINDING ANALYSES

MD simulations for druggability assessment. Simulations were performed using NAMD⁴⁵ software and the CHARMM⁴⁶ force field. Productive simulation times were 40 ns in all runs. The druggability assessment by MD approach was performed using DruGui³⁶ software. Two different sets of probes were used: (A) a system with a mixture of 70% isopropanol, 10% acetamide and 20% of acetate – isopropilamine; (B) an equivalent system with a composition of 30% isopropanol, 50% imidazole, 10% acetamide, 5% acetate and 5% isopropilamine.

Structural dynamics of the rCUG^{exp} fragment. Available (CUG)_n structures were retrieved from the Protein Data Bank (PDB ids: 1zey, 3gm7, 3syw, 3szx, 4e48, 4fnj).^{23,47-50} Analysis of the small molecules size and previous computational studies suggested that 3 nucleotide repeat expansion should suffice for *in silico* studies. Structures were prepared using PyMOL⁵¹ by retaining only those fragments with three repeats (n=3). For longer repeated fragments, all the possible n=3 combinations were extracted as individual structures. Next, all the structures were superimposed and saved as a pdb ensemble. Essential dynamics analysis (EDA) was completed with ProDy⁵² using a total of 20 deformation modes. The previous r(CUG)₃ ensemble was used for the EDA approach. A total of 40 RNA conformations were obtained by applying two deformations per mode using a mass-weighted RMSD of 2 Å.

Molecular docking. Molecular docking was performed according to a cross-docking approach. First, compound structure was prepared using MOE.⁵³ Next, the structure was minimized using the MMFF94⁵⁴ force field and AM1 charges⁵⁵ were computed. Molecular docking was completed using Vina (version 1.1.2)⁵⁶ and 40 RNA conformations generated by EDA deformation were used as receptor. A total of 50 conformations were generated per run, thus a total of 4000 potential bindings per ligand were obtained. Then, conformations were rescored using ANNScore.

4.7. REFERENCES

1. Gareiss, P. C. *et al.* Dynamic combinatorial selection of molecules capable of inhibiting the (CUG) repeat RNA-MBNL₁ interaction in vitro: Discovery of lead compounds targeting myotonic dystrophy (DM₁). *J. Am. Chem. Soc.* **130**, 16254–16261 (2008).
2. Childs-Disney, J. L. *et al.* Induction and reversal of myotonic dystrophy type 1 pre-mRNA splicing defects by small molecules. *Nat. Commun.* **4**, 2044 (2013).
3. Ketley, A. *et al.* High-content screening identifies small molecules that remove nuclear foci, affect MBNL distribution and CELF₁ protein levels via a PKC-independent pathway in myotonic dystrophy cell lines. *Hum. Mol. Genet.* **23**, 1551–1562 (2014).
4. Coonrod, L. A. *et al.* Reducing levels of toxic RNA with small molecules. *ACS Chem. Biol.* **8**, 2528–2537 (2013).
5. Arambula, J. F., Ramisetty, S. R., Baranger, A. M. & Zimmerman, S. C. A simple ligand that selectively targets CUG trinucleotide repeats and inhibits MBNL protein binding. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 16068–16073 (2009).
6. Garcia-Lopez, A., Llamusi, B., Orzaez, M., Perez-Paya, E. & Artero, R. D. In vivo discovery of a peptide that prevents CUG-RNA hairpin formation and reverses RNA toxicity in myotonic dystrophy models. *Proc. Natl. Acad. Sci.* **108**, 11866–11871 (2011).
7. Mulders, S. A. M. *et al.* Triplet-repeat oligonucleotide-mediated reversal of RNA toxicity in myotonic dystrophy. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 13915–13920 (2009).
8. Childs-Disney, J. L., Hoskins, J., Rzuczek, S. G., Thornton, C. a. & Disney, M. D. Rationally designed small molecules targeting the RNA that causes myotonic dystrophy type 1 are potently bioactive. *ACS Chem. Biol.* **7**, 856–862 (2012).
9. Chen, C. Z. *et al.* Two high-throughput screening assays for aberrant RNA-protein interactions in myotonic dystrophy type 1. *Anal. Bioanal. Chem.* **402**, 1889–1898 (2012).
10. Hoskins, J. W. *et al.* Lomofungin and dilomofungin: Inhibitors of MBNL₁-CUG RNA binding with distinct cellular effects. *Nucleic Acids Res.* **42**, 6591–6602 (2014).
11. Wong, C. H. *et al.* Targeting toxic RNAs that cause myotonic dystrophy type 1 (DM₁) with a bisamidinium inhibitor. *J. Am. Chem. Soc.* **136**, 6355–6361 (2014).
12. Wong, C. H. *et al.* Investigating the Binding Mode of an Inhibitor of the MBNL₁-RNA Complex in Myotonic Dystrophy Type 1 (DM₁) Leads to the Unexpected Discovery of a DNA-Selective Binder. *ChemBioChem* **13**, 2505–2509 (2012).
13. Parkesh, R. *et al.* Design of a bioactive small molecule that targets the myotonic dystrophy type 1 RNA via an RNA motif-ligand database and chemical similarity searching. *J. Am. Chem. Soc.* **134**, 4731–4742 (2012).
14. Disney, M. D., Yildirim, I. & Childs-Disney, J. L. Methods to enable the design of bioactive small molecules targeting RNA. *Org. Biomol. Chem.* **12**, 1029–39 (2014).

15. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12**, 88–118 (1965).
16. Pinamonti, G., Bottaro, S., Micheletti, C. & Bussi, G. Elastic network models for RNA: a comparative assessment with molecular dynamics and SHAPE experiments. *Nucleic Acids Res.* **43**, gkv708 (2015).
17. Rasti, B., Shahangian, S. S., Taghdir, M., Hasannia, S. & Sajedi, R. H. Identification of RNA-binding sites in artemin based on docking energy landscapes and molecular dynamics simulation. *Iran. J. Biotechnol.* **10**, 8–15 (2012).
18. Lang, P. T. *et al.* DOCK 6: combining techniques to model RNA-small molecule complexes. *RNA* **15**, 1219–1230 (2009).
19. Sobczak, K. *et al.* Structural diversity of triplet repeat RNAs. *J. Biol. Chem.* **285**, 12755–12764 (2010).
20. Brook, J. D. *et al.* Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell* **68**, 799–808 (1992).
21. Krzyzosiak, W. J. *et al.* Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nucleic Acids Res.* **40**, 11–26 (2012).
22. Lee, J. E. & Cooper, T. a. Pathogenic mechanisms of myotonic dystrophy. *Biochem. Soc. Trans.* **37**, 1281–1286 (2009).
23. Kiliszek, A., Kierzek, R., Krzyzosiak, W. J. & Rypniewski, W. Structural insights into CUG repeats containing the 'stretched U-U wobble': Implications for myotonic dystrophy. *Nucleic Acids Res.* **37**, 4149–4156 (2009).
24. Wang, Y., Li, Y., Ma, Z., Yang, W. & Ai, C. Mechanism of microRNA-target interaction: Molecular dynamics simulations and thermodynamics analysis. *PLoS Comput. Biol.* **6**, 5 (2010).
25. Vargiu, A. V., Ruggerone, P., Magistrato, A. & Carloni, P. Dissociation of minor groove binders from DNA: Insights from metadynamics simulations. *Nucleic Acids Res.* **36**, 5910–5921 (2008).
26. Tor, Y. Targeting RNA with small molecules. *ChemBioChem* **4**, 998–1007 (2003).
27. Brozell, S. R. *et al.* Evaluation of DOCK 6 as a pose generation and database enrichment tool. *Journal of Computer-Aided Molecular Design* **26**, 749–773 (2012).
28. Ruiz-Carmona, S. *et al.* rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* **10**, e1003571 (2014).
29. Philips, A., Milanowska, K., Lach, G. & Bujnicki, J. M. LigandRNA: computational predictor of RNA-ligand interactions. *RNA* **19**, 1605–16 (2013).
30. Pfeffer, P. & Gohlke, H. DrugScore RNA Knowledge-Based Scoring Function To Predict RNA–Ligand Interactions. *J. Chem. Inf. Model.* **47**, 1868–1876 (2007).

31. Walker, J. M. *Artificial Neural Networks*. **1260**, (Springer New York, 2015).
32. Gomes-Pereira, M. & Monckton, D. G. Chemically induced increases and decreases in the rate of expansion of a CAG·CTG triplet repeat. *Nucleic Acids Res.* **32**, 2865–2872 (2004).
33. Lewell, X. Q., Judd, D. B., Watson, S. P. & Hann, M. M. RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **38**, 511–522 (1998).
34. Pascual, R., Borrell, J. I. & Teixidó, J. Analysis of selection methodologies for combinatorial library design. *Mol. Divers.* **6**, 121–133 (2000).
35. Wojciechowska, M., Taylor, K., Sobczak, K., Napierala, M. & Krzyzosiak, W. J. Small molecule kinase inhibitors alleviate different molecular features of myotonic dystrophy type 1. *RNA Biol.* **11**, 1–13 (2014).
36. Bakan, A., Nevins, N., Lakdawala, A. S. & Bahar, I. Druggability assessment of allosteric proteins by dynamics simulations in the presence of probe molecules. *J. Chem. Theory Comput.* **8**, 2435–2447 (2012).
37. Sriram, M., van der Marel, G. A., Roelen, H. L., van Boom, J. H. & Wang, a H. Conformation of B-DNA containing O6-ethyl-G-C base pairs stabilized by minor groove binding drugs: molecular structure of d(CGC[e6G]AATTCGCG complexed with Hoechst 33258 or Hoechst 33342. *EMBO J.* **11**, 225–232 (1992).
38. Pushechnikov, A. *et al.* Rational Design of Ligands Targeting Triplet Repeating Transcripts That Cause RNA Dominant Disease: Application to Myotonic Muscular Dystrophy Type 1 and Spinocerebellar Ataxia Type 3. *J. Am. Chem. Soc.* **131**, 9767–9779 (2009).
39. Estrada Tejedor, R. Desenvolupament del programari ArIS (Artificial Intelligence Suite): implementació d'eines de cribratge virtual per a la química mèdica. at <<http://www.tdx.cat/handle/10803/51367>>
40. Software, O. S. OEChem. (2010).
41. Duan, J., Dixon, S. L., Lowrie, J. F. & Sherman, W. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* **29**, 157–170 (2010).
42. Wetzel, S. *et al.* Interactive exploration of chemical space with Scaffold Hunter. *Nat. Chem. Biol.* **5**, 581–583 (2009).
43. Boström, J., Greenwood, J. R. & Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graph. Model.* **21**, 449–462 (2003).
44. McGaughey, G. B. *et al.* Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **47**, 1504–1519 (2007).

45. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **26**, 1781–1802 (2005).
46. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614 (2009).
47. Mooers, B. H. M., Logue, J. S. & Berglund, J. A. The structural basis of myotonic dystrophy from the crystal structure of CUG repeats. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 16626–16631 (2005).
48. Kumar, A. *et al.* Myotonic dystrophy type 1 RNA crystal structures reveal heterogeneous 1 X 1 nucleotide UU internal loop conformations. *Biochemistry* **50**, 9928–9935 (2011).
49. Tamjar, J., Katorcha, E., Popov, A. & Malinina, L. Structural dynamics of double-helical RNAs composed of CUG/CUG- and CUG/CGG-repeats. *J. Biomol. Struct. Dyn.* **30**, 505–523 (2012).
50. Coonrod, L. A., Lohman, J. R. & Berglund, J. A. Utilizing the GAAA tetraloop/receptor to facilitate crystal packing and determination of the structure of a CUG RNA helix. *Biochemistry* **51**, 8330–8337 (2012).
51. LLC, S. The PyMOL Molecular Graphics System.
52. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).
53. Chemical Computing Group Inc. MOE: Molecular Operating Environment. (2013)
54. Halgren, T. A. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* **17**, 520–552 (1996).
55. Jakalian, A., Bush, B. L., Jack, D. B. & Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **21**, 132–146 (2000).
56. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.* 455–461 (2010).

CHAPTER V. PROTEIN AND PEPTIDE STRATEGIES

5.1. BACKGROUND OF THE STUDY

RNA-protein complexes have proven difficult targets for conventional small molecule inhibitor design due to the large surface interfaces buried in the complexes. In addition, inherent flexibility of both RNA and protein in the free states seem to favor mutually induced fit mechanisms of binding. Pharmacological disruption of their interaction would generate novel RNA drugs but efforts to discover molecules with sufficient specificity for a particular target have been unsuccessful. Structural mimicry of proteins provide ideal starting points for rational structure-based inhibitor design. Design and optimization of peptides is emerging as an innovative approach to target either proteins or RNA. Molecular recognition is usually mediated by surface exposed secondary structure motifs such as hairpins and helices. Thus, certain binding protein regions can be mimicked in smaller conformationally 'locked' peptides.

In this chapter, structural analyses will be performed on the MBNL₁ protein in order to gain insights on structural recognition of cognate YGCY RNA motifs. Loss of MBNL function is a central pathological event in DM. Yet the MBNL protein family also plays a prominent role in the regulation of alternative splicing (AS) during development. For instance, MBNL₁ and MBNL₂ knockdown increases the expression of key pluripotency genes requires for induced pluripotent stem cell differentiation. In the last decade, several MBNL₁ cognate RNAs have been identified and characterized, and studies revealed that the four zinc-finger domains contained in MBNL₁ recognize GpC steps in YGCY sequence elements (where Y represents any pyrimidine). Despite the sequence and structural similarities between the four binding domains, a differentiated binding and splicing activities have been found between them. Herein we present structural insights into the binding selectivity of the four zinc finger domains using essential dynamics analysis and steered molecular dynamics simulations.

Using a MBNL₁ mimicry strategy, García-López et al.¹ reported a set of 16 D-hexapeptides that specifically targeted rCUG in animal model studies. The most active peptide (ABPr) improved *Drosophila*'s model phenotype and induced a conformational shift of hairpin rCUG without compromising MBNL₁ activity. However, further improvements of these family of peptides require sequence-structure-activity rationalization. For this reason, conformational sampling studies have been performed over the subset of 16 D-hexapeptides and classified according to their secondary structural content.

The same peptide strategy was followed for studying peptides for targeting the transactivator response element (TAR). Tat protein –TAR interactions of HIV virus is an attractive target for the discovery of new antiviral drugs. Pascale et al.² designed polyamide amino acids (PAAs) which bind TAR with similar sub-micromolar affinities. However, their ability to compete against Tat differed which induced to think about two distinct interaction modes. Thus, this thesis offers a structural perspective of these differentiated recognition modes using MD simulations in order to rationalize structure-activity relationships.

5.2. INVESTIGATING THE MBNL1 PROTEIN

Alternative splicing (AS) of pre-mRNA is a determinant mechanism for the regulation of gene expression by which distinct functional protein isoforms are generated from a relatively small number of genes in eukaryotes.³⁻⁵ Sequence changes due to AS may result in isoforms that map different regions of the protein functional space.⁵⁻⁷ AS regulators control the expression of protein isoforms through binding either to splice sites or to other sequences in the pre-mRNA, thus enhancing or repressing the inclusion of alternative exons. By means of AS, different protein isoforms can alter their function or cellular localization, or even regulate protein levels by leading to non-productive splicing RNA turnover.⁴

Intriguingly, sequence and structural studies revealed that AS is primarily found in intrinsically disordered regions, which can sometimes involve the whole protein.^{5,7,8} They are characterized by having highly heterogeneous structural ensembles, thus they do not have a stable structure in this particular region. Given the observed correlation between AS sites and structurally disordered regions, some studies pointed out that the modulation of AS may be driven by the modification of the structural sampling. Romero et al. suggested that small changes in AS sites can modify the specificity of the protein and revealed that a relationship between AS and intrinsic disorder exists.⁵ Barbany et al. investigated the isoforms of two AS proteins and reported that AS is not necessarily related to neither local nor global changes on the size of protein fluctuations;⁷ however, they noticed that AS induce subtle changes in protein dynamics that may explain its specificity for RNA targets. More recently, the same group reported that AS might be modulated through changes in cavity couplings.⁹

Muscleblind-like proteins (MBNL) are AS factors that, in mammals, are encoded by three genes: MBNL₁, MBNL₂ and MBNL₃. MBNL proteins can act as either repressors or activators of splicing in several transcripts.¹⁰⁻¹² They belong to a family of tissue-specific RNA metabolism regulators that have a key role in terminal muscle differentiation. All three family members share

four highly conserved zinc-finger domains (ZnF) for recognizing specific pre-mRNA and mRNA targets.¹¹ These RNA binding domains of the CCCH type are arranged in tandem pairs that are positioned towards the N-terminal region (ZnF₁ and ZnF₂) and in the middle part of the sequence (ZnF₃ and ZnF₄). Each domain contains a different spacing between the zinc-coordinated residues, thereby ZnF₁ and ZnF₃ contain a CX₇CX₆CX₃H motif, whereas ZnF₂ and ZnF₄ displays a CX₇CX₄CX₃H sequence. The RNA binding faces in each domain are arranged back-to-back, creating an anti-parallel alignment of RNA binding to ZnF domains.¹¹

In particular, MBNL₁ has been the main focus of intense studies over the last years due to its implication in Myotonic Dystrophy (DM) pathogenic pathway.¹³⁻¹⁹ Normal splicing pattern in muscle differentiation transcripts and repression of the adult protein isoforms results in this neuromuscular disease. For instance, Myotonic Dystrophy type 1 (DM1) is caused by expansion of RNA CUG repeats which bind and sequester MBNL₁.^{20,21} SELEX experiments determined that the optimal MBNL binding RNA sequence consist of multiple YGCY consensus motifs (where Y is a pyrimidine), explaining the binding to CUG expansions and inactivation of its normal functions.¹⁵ A single crystal structure of MBNL₁ ZnF domains is reported in literature (PDB id: 3d2s)¹¹, which shows the interaction of the ZnF_{3/4} tandem with an RNA fragment. Crystallographic analysis confirmed that aromatic residues Phe202 and Tyr236 intercalate between the GC step and several hydrogen bonds are formed between the GC and the side chains in the protein, which may explain its high affinity for the YGCY motif.

Human MBNL₁ includes twelve exons, ten of which correspond to the coding sequence (exons 1-10) and six (exons 3, 5, 6-9) undergo alternative splicing. Thus, there exist at least seven MBNL₁ mRNA variants which lead to extensive alternative splicing regulation.²² Analysis of MBNL₁ deletion constructs have been conducted and proved that exon 5 and a five amino acid region in exon 6 are essential for nuclear localization.²³ On the contrary, exons 1, 2 and 4 encode for the ZnF domains and exon 7 participates in dimerization and induces the formation of ring-like structures upon binding on RNA.^{22,23} However, relevant questions about MBNL₁ binding sites architecture remain to be addressed. For instance, the amino acid sequences of ZnF_{1/2} and ZnF_{3/4} are very well conserved; however, despite their high structural resemblance, truncated versions of MBNL₁ showed a differentiated binding affinity for target RNAs.^{12,24} Moreover, deletion of either ZnF₁ or ZnF₄ alone greatly diminish their interaction with CUG repeats in vivo.^{12,22} Interestingly, the linker sequence encoded by exon 3, which separates the two tandems, also determines the MBNL₁-RNA interaction.^{11,12}

In that context, the questions we ask here are: how heterogeneous are the MBNL domains between them in terms of conservation, coevolution and mobility? Local and global fluctuations are characteristic for each binding domain? Do these amino acids have any significant effect on RNA binding events? Herein we address these questions using a combination of molecular dynamics (MD) simulations and bioinformatics analyses. Our results provide insights into the structural particularities of each tandem and indicate that local and global motions are not conserved for each RNA-binding domain.

5.2.1. SEQUENCE COEVOLUTION OF THE CCCH DOMAIN

MBNL₁ interacts with YGCY consensus motif RNAs through its four CCCH ZnFs, although some other regions of the protein are hypothesized to interact with the RNA as well or play an additive role.^{12,22,24} Each ZnF adopts a similar fold, as reported for other CCCH-type zinc fingers such as TIS11d.¹¹ Each tandem is constituted by two ZnF subunits that interact with each other through hydrophobic interactions. Thus, the tandem adopts a compact global fold where both ZnFs are approximately symmetrical. Moreover, the linker between them forms an antiparallel β -sheet that enhances this conformation.

The architecture of the protein usually encodes its global motions. The role of this concerted motions correlates with protein function, as reported in recent works.^{25–29} Native folds are controlled by mechanical sites that control the global movements while preserving the stability of the protein core. In fact, recent studies have probed that sequence variability and structural dynamics are tightly related.³⁰ In particular, the CCCH-motif is present in 4739 sequences deposited in PFAM, and its core is highly conserved, especially among the MBNL proteins. Mutations in protein sequences are usually analyzed in terms of mutual information (MI). The mutual information between two variables (or mutations) is a measure of their mutual dependence; hence it measures how similar the joint distribution $p(X,Y)$ is to the products of factored marginal distribution $p(X)p(Y)$, where X and Y represent two amino acids at specific positions. In other words, MI permits to evaluate if a mutation in position X induces a mutation in position Y , thus a structural dependence between both amino acids can exist. Mutual information and conservation analysis (see figure 5.1A) shows in a circos representation that the most conserved amino acids are the constituents of the CCCH motif (red) followed by Phe88, Gly191, Gly196 and Phe202 which are located at a 2 and 3 amino acid distance to the Zn coordinated site. Tridimensional distance analysis shows that these residues are in close contact with the CCCH motif, which indicates a tightly conserved core (see figure 5.1B in a logo representation). The degree of mutual information of a given residue may be evaluated with the cumulative mutual information (cMI), which measures the degree of shared mutual information between pairs of amino acids, and the proximity mutual information (pMI) that informs about the mutual information in the proximity of a residue within 5 Å. Figure 5.1C shows that those residues located near the active sites (black arrows) exhibit coevolutionary trends, as stated by the cMI values. Notice that the pMI values increase along the sequence until the conserved residues Phe88, Gly191, Gly196 and Phe202 (blue arrows) are reached. In fact, the aromatic ring of Phe202 present in the ZnF₃ of MBNL₁ was observed to form stacking interactions with cytosines in the RNA structure and facilitates the macromolecular interaction.¹¹ The high coevolution propensity observed in this (F/Y)GG(F/Y) motif into the RNA binding site is apparent even by examining the MI values, and further mobility analyses could reveal a correlation between coevolution and motions of this region. In particular, inspection of the sequences of MBNL₁ and MBNL₂ shows that the (F/Y)GG(F/Y) is only present in MBNL₁ ZnF₃. On the contrary, the motif in ZnF₁, ZnF₂ and ZnF₄ of MBNL₁ is reduced to (F/Y)G(F/Y).

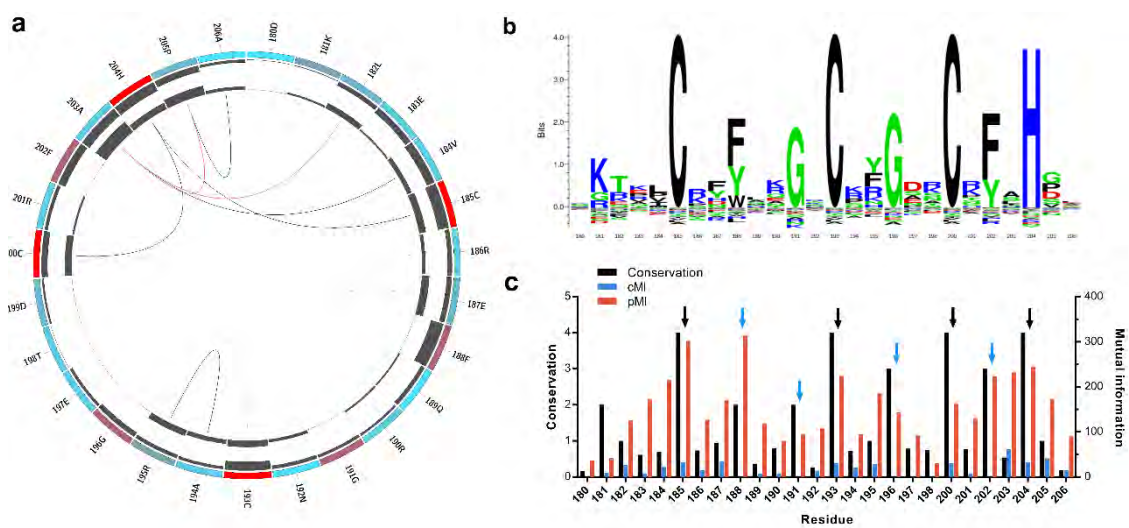


Figure 5.1. (a) Circos representation of the CCCH domain family. Square boxes indicate the KL conservation score (from red-highest to cyan-lowest values). The cMI and pMI scores are represented as histograms in the inner circle. Lines in the center connect pairs with an MI score higher than 6.5. Red lines represent the top 5%, black lines are between 95% and 70%, and gray lines indicate < 70%. (b) KL sequence logo of selected nodes. (c) Histogram representing the conservation per residue, cumulative mutual information (cMI) and proximity mutual information (pMI) within a 5Å threshold. Black arrows represent the CCCH motif and the blue arrows correspond to residues Phe188, Gly191, Gly196 and Phe202. The sequence id numbers correspond to those in the circos representation.

5.2.2. MBNL DOMAINS FROM A STRUCTURAL PERSPECTIVE

Structural studies are less frequent than sequence studies and very few MBNL experimental structures are available. Previous studies of AS proteins suggested that the overlap between AS sites and protein dynamics is clear, hence AS can modulate function by modifying the structural sampling during protein dynamics.^{5,7} Looking more specifically at the tertiary structure of the MBNL family, the peptide backbone topology of the MBNL domains is essentially invariable as observed by a comparison of MBNL₁ ZnF_{1/2}, ZnF_{3/4} and MBNL₂ ZnF_{1/2} (see structures in figure 5.2A, page 166), but information about the inter-domain linker is only partially available for the NMR ensemble of MBNL₂. On the contrary, the intra-domain linker between the zinc finger pairs is not flexible, and maintains a global compact fold. By doing so, both tandems may independently bind different RNA regions and no cooperativity may be expected. However, their binding affinities for RNA targets are different as pointed out by alanine substitution studies.³¹ In fact, RNA binding and splicing activity is higher in a truncated MBNL₁ protein that only contains a ZnF_{1/2} tandem than a truncated version with only ZnF_{3/4}.¹² Previous studies suggested that substrate recognition is assisted by coevolving residue pairs with enhanced global mobility that may improve its partner interactions with cognate binding motifs.³⁰ Thus, despite of their high structural similarities, small changes in charge distribution and hydrogen-bonding potential on each domain should yield a differentiated binding and splicing capabilities.

The electrostatic potential analysis of the two tandems of MBNL₁ (figure 5.2B for ZnF_{1/2} and figure 5.2C for ZnF_{3/4}) highlights the differences between the binding interfaces of each ZnF

domain. ZnF₄ clearly exhibits the most positively charged interface due to an additional Lys residue at position 235. Five positive charges in total are present in this domain in contrast to the four positive charges present in ZnFs 1 to 3. Not surprisingly, ZnF₃ centers its positive charge distribution into the binding pocket upon RNA recognition, which increases the compactness of the protein. Liu et al. observed that polar and charged residues present a high coevolvability and high mobility, specially Ser, Asn, and Lys.³⁰ In line with that observation, the binding regions of the ZnFs are Lys and Arg rich, which may be essential for the recognition of the RNA backbone, and have high cMI and pMI values as shown in figure 5.1c (e.g. positions 186, 190, 195 and 201 corresponding to ZnF₃). Interestingly, high MI values between the (F/Y)GG(F/Y) motif and the charged residues, which are close in sequence (1 or 2 amino acids) and space, suggest a coevolutionary trend in these positions. It is widely accepted that conserved residues within protein families provides the stabilization of the bound ligand. For this reason, RNA binding should entail the conformational adaptability prior to stabilization by conserved interactions. On the one side, Arg and Lys residues are not highly conserved but provide the necessary platform for the recognition of the RNA backbone and nuclear bases through charged and π -stacking interactions. On the other, Phe202 aromatic ring in the conserved (F/Y)GG(F/Y) motif contribute to RNA binding through stacking interactions as reported in previous studies;¹¹ Phe88 stacks with His204 from the CCCH motif and likely contributes to stabilize the core domain.

5.2.3. MBNL2 EXPERIMENTAL FLUCTUATIONS CORRELATE WITH PREDICTED NORMAL MODES

It is clear that protein function is linked with its conformational adaptation but the extraction of the relevant motions is not always straightforward. For instance, an ensemble of static structures obtained by either X-ray or NMR may provide the experimental global fluctuations of the protein through a principal component analysis (PCA). When only single structures are available, anisotropic network models (ANM) have extensively proved to yield and efficient solution for the collective motions of a structure without the need of any simulations (see Methods for details).^{27,29,32,33} To provide a basis for comparison with results from a network model approach, and to gain insights into the global conformational space of MBNL proteins and to test the viability of ANM for the analysis of this protein family, a combined PCA and ANM analysis was performed with MBNL₂, which is the only MBNL NMR ensemble reported till date. A γ spring constant (see Eq. 2.13, page 58) based on the secondary structure and connectivity of the residues of MBNL₂ ensemble was used during all the analyses. Figure 5.3A shows that a correlation between the PC₁ and the fourth ANM mode exist ($r = 0.67$); in other words, the fourth theoretical mode correspond to the first experimental mode obtained from the NMR ensemble of MBNL₂, which proves the predictive capability of the method.

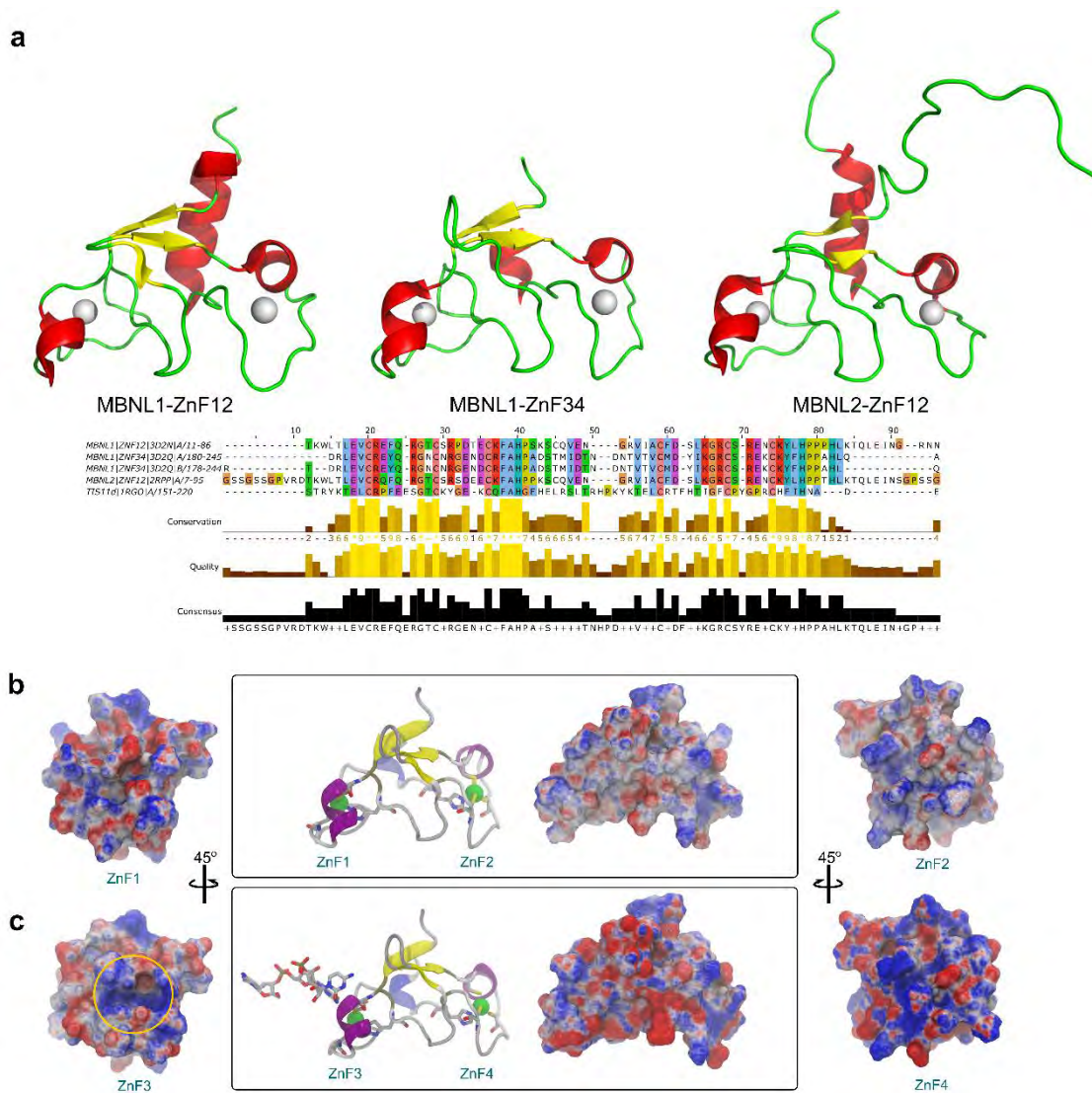


Figure 5.2. (a) Three-dimensional structure of MBNL1 ZnF1/2, MBNL1 ZnF3/4 and MBNL2 ZnF1/2; sequence alignment, conservation, quality and consensus sequence are included. (b) Electrostatic potential surface representation obtained with APBS contoured at ± 10 kT/e for ZnF1/2 (PDB id 3d2n) and (c) ZnF3/4 of MBNL1 binding an RNA fragment (PDB id 3d2s). The ZnF3 binding site region is contoured with a yellow circle.

The PCA of the MBNL2 ZnF1/2 NMR ensemble revealed that global fluctuations (figure 5.3B) are localized in charged and polar residues, mainly characterized by Arg27, Arg31, Ser37 and Glu39 in the ZnF1 domain, and Glu71 in the ZnF2 domain. These results are in line with the coevolvability analysis, but it is particularly interesting the fact that the compactness of the protein is not dynamically conserved in both domains, despite their high structural and sequence similarity. The CCCH motifs present invariably low fluctuations but it surprising the high mobility of the conserved aromatic Phe43 residue due to its potential role in RNA binding. Importantly, Edge et al. pointed out that that alanine substitution in this position of ZnF1 of MBNL1 did affect its ability to activate splicing.¹²

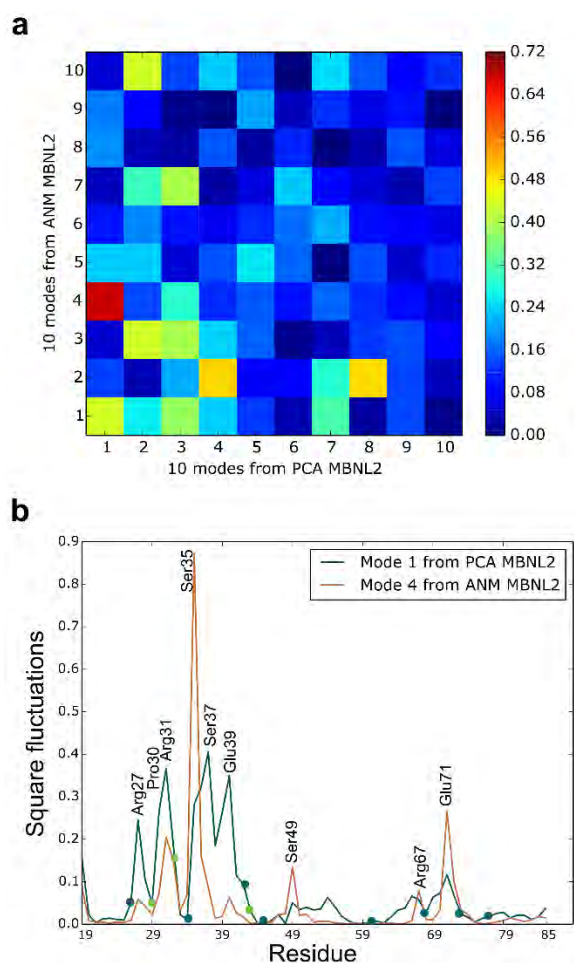


Figure 5.3. (a) Correlation matrix between the first 10 PC modes from the MBNL2 NMR ensemble and the first 10 modes of the ANM analysis of the first MBNL2 structure. The highest correlation is obtained for PC1 and ANM4 ($r = 0.67$). (b) Normalized square fluctuations of the C α of MBNL2 obtained by both PCA and ANM methods. Blue dots indicate the amino acids from the CCCH motifs of the ZnF1 and 2 domains, and green dots indicate Phe29, Gly32 and Phe43 from the (F/Y)GG(F/Y) consensus motif. Indices correspond to those in the PDB id 2rpp (NMR ensemble of MBNL2 ZnF1/2).

5.2.4. ZNF1/2 FROM MBNL1 AND MBNL2 EXHIBIT EQUIVALENT LARGE-SCALE MOTIONS

Due to the lack of a dynamic ensemble of structures of MBNL₁ tandems, herein the intrinsic dynamics of MBNL₁ were studied using a molecular dynamics (MD) approach. Full atomic MD simulations are restricted to the time scale of nanoseconds due to its computational cost in time, but they allow to obtain simulations of many biological processes of interest. First, convergence of the simulation of the MBNL₁ ZnF_{1/2} tandem was assessed by observing the histogram of the projection of PC₁ (accounting for 43% of the total motion during the simulation). Superposition of two halves of the simulation indicates an excellent convergence of the simulation and that no relevant conformational transitions are to be expected along the nanosecond scale (see figure 5.4A).

Next, fluctuations of the tandem were computed using essential dynamics analysis (EDA) and compared to the fluctuations of MBNL₂ ZnF_{1/2}. Interestingly enough, the PC₁ of MBNL₁

and MBNL2 yielded a clear correlation ($r = 0.61$, see figure 5.4A) which suggests that both domains should perform equivalent fluctuations. Indeed, superposition of the fluctuations extracted from both tandems (figure 5.4B) demonstrate that global fluctuations are localized into the same regions and follow a similar trend. Not surprisingly, the inter-domain linker region of MBNL1 also exhibits high fluctuations due to high mobility of these residues along the dynamics simulation, particularly residues Pro73, His74 and Thr77. The highest fluctuations are observed in charged and polar residues next to the CCCH motif, as noticed before. No overall structure deformation was observed along the MD trajectory.

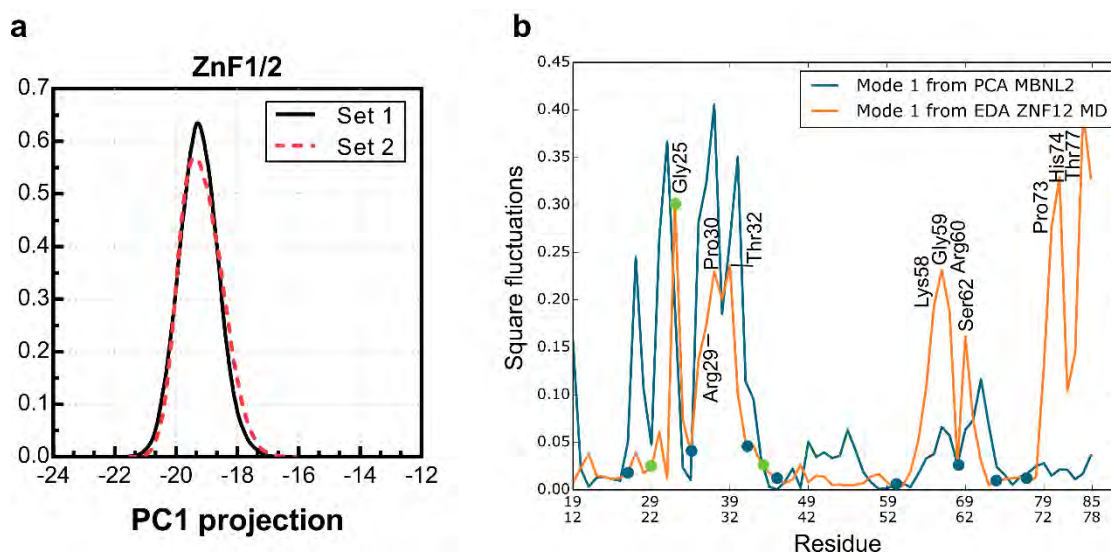


Figure 5.4. (a) PC1 histogram for ZnF1/2 simulation (accounting for 43% of the total motion during the simulation). Set 1 and 2 correspond to two segments of the simulation (set 1: 0-100 ns; set 2: 100-200 ns from the equilibrated trajectory). (b) Superposition of normalized square fluctuations of the Ca of ZnF1/2 of MBNL2 and MBNL1, obtained by PCA and EDA respectively. Only high fluctuating residues observed in MBNL1 ZnF1/2 are labeled. Blue dots indicate the CCCH motifs of the ZnF1 and 2 domains and green dots indicate Phe22, Gly25 and Phe36 from the (F/Y)GG(F/Y) consensus motif. Indices correspond to those in the PDB id 2rpp (indices 19 to 85) and PDB id 3d2q (indices 12 to 78).

5.2.5. GLOBAL DYNAMICS OF ZNF3/4 DIFFER FROM THOSE OF ZNF1/2

Following the dynamics analysis, a similar approach was conducted for ZnF3/4 using MD and ANM analyses. Unfortunately, no reference experimental structure was available for comparison with this tandem. First, the convergence of the simulation was assessed as previously described (see figure 5.5A). Interestingly, albeit having a highly similar sequence and structure, ZnF3/4 dynamics strongly differed from those observed for ZnF1/2. $C\alpha$ rmsd between ZnFs after the simulation was 3.11 Å but no structural rearrangements were observed into the protein core along 200 ns of simulation. Main deviations were located at the 3_{10} helix element. Figure 5.5B shows the superposition of the computed global fluctuations of ZnF3/4 extracted from the MD simulation and the ones obtained from the correlated theoretical modes ($r = 0.49$). ZnF3/4 presents its highest fluctuations onto the ZnF4 domain, in contrast to the first tandem that exhibits higher fluctuations around the ZnF1 domain. Either theoretical modes or MD fluctuations along the trajectory agree in huge increase between the second and third Cys from

the CCCH-motif. High mobility is also localized in the intra-domain linker with fluctuations characterized by Ser208, Thr213 as observed into the MD. These results show that, despite the large noise inherent to the technique, there is a good overlap between the AS control regions.

The same ANM analysis performed over the 3d2s structure (ZnF3/4 tandem binding an RNA unit through the ZnF3) suggests that, upon cognate binding, high fluctuating motions such as Asn194 and Asp215 of ZnF3 are dissipated but, in return, localized motions around ZnF4 are promoted. Indeed, binding and signaling effectivity increases with a tight packing and low residue fluctuations in the global modes of residues in close spatial proximity to the RNA binding site; the restricted mobility of these residues in the global modes upon RNA binding and their high pMI values suggests a coevolutionary trend, which may correlate with binding affinity. Interestingly, global fluctuations from ZnF3/4 remarkably differ from those observed in ZnF1/2 ($r = -0.01$). Further support to our results was provided by the analysis of the essential subspace overlap between ZnF1/2 and ZnF3/4. The conformational coverage between both domains (subspace overlap, see Eq. 3-7, page 110) 0.27 and indicates that global fluctuations are not conserved.

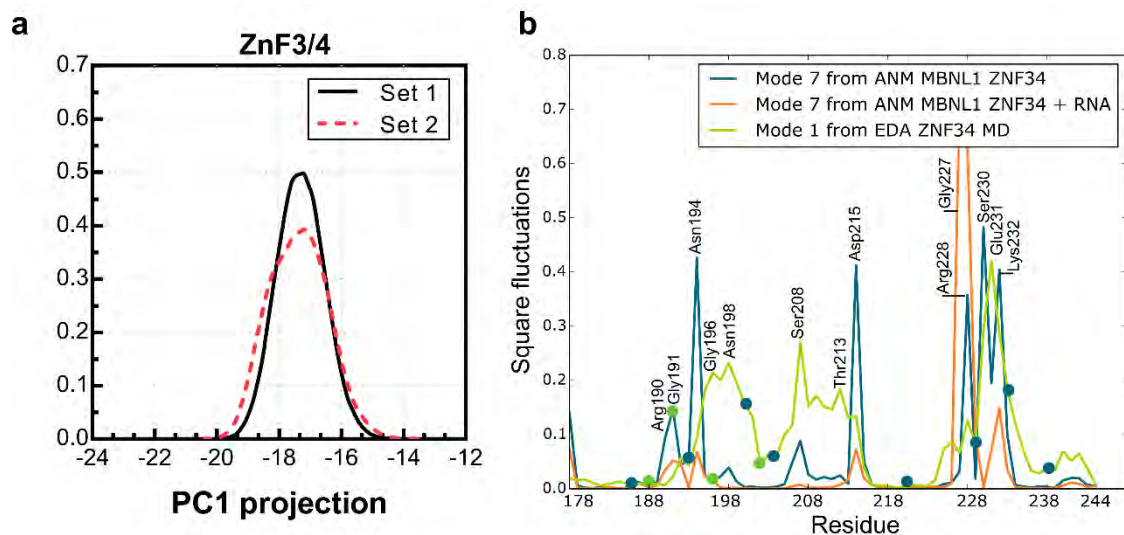


Figure 5.5. (a) PC1 histogram for ZnF3/4 simulation (accounting for 39% of the total motion during the simulation). Set 1 and 2 correspond to two segments of the simulation (set 1: 0-100 ns; set 2: 100-200 ns from the equilibrated trajectory). (b) Superposition of normalized square fluctuations of the Ca of ZnF3/4, obtained by ANM and EDA. ANM fluctuations correspond to ZnF34 RNA-bound and unbound structures (PDB id 3d2q and 3d2s respectively). Only high fluctuating residues observed in MBNL1 ZnF3/4 are labeled. Blue dots indicate the CCCH motifs of the ZnF3 and 4 domains and green dots indicate Tyr188, Gly191, Gly196 and Phe202 from the (F/Y)GG(F/Y) consensus motif. Indices correspond to those in the PDB id 3d2q (indices 19 to 85).

5.2.6. EFFECT OF RNA BINDING OVER LOCAL FLUCTUATIONS

To prove for structural changes into the RNA binding sites of MBNL1, a comparative MD analysis between RNA-bound and unbound complexes was completed. Local fluctuations were measured for each ZnF binding pocket in presence or absence of a UGCU fragment. Figure 5.6 (see page 171) illustrates the main differences between each pair of trajectories. The local analysis and simulations indicate that, in general terms, most of the polar and charged residues

fluctuations are enhanced upon RNA binding; for instance, Ser56, Arg60 and Arg63 in ZnF2, Arg186 in ZnF3 and Lys226 in ZnF4. On the contrary, Arg24 of ZnF1 fluctuations are suppressed due to its anchoring to the RNA backbone through charge based interactions (refer to figure 5.6). In agreement with Teplova et al. structural studies, the GC fragment is strongly bonded to adjoining pockets using concerted charged and π -stacking interactions, and hydrogen-bonding networks. Interestingly, highly conserved Gly (specifically Gly25, Gly59 and Gly191) shows enhanced local mobility, especially in ZnF1 and ZnF3. These amino acids correspond to non-structured regions of the binding pocket adjoining the α helix. Visual inspection of MD trajectories shows that these binding regions fluctuate between 'open' and 'closed' configurations that accommodate the binding partner and enhance local interactions. A significant correlation has been found between bound and unbound ZnF4 local fluctuations (table 5.1), meaning that the binding process does not produce significant alternations into its pocket. ZnF3 yielded modest correlations with ZnF1 - RNA and ZnF3 - RNA trajectories.

More interestingly, local fluctuations are not only modified upon RNA binding but also upon binding to the adjoining tandem domain. Table 1 shows that only ZnF4 is able to maintain its structural configuration during all possible binding events. As a general trend, cognate RNA binding modifies global and local fluctuations of ZnF1-3; hence, the analysis and simulation indicate that RNA binding may modify structural protein couplings in partnering binding pockets.

Table 5.1. Pearson correlation coefficient between RNA-bound and unbound local fluctuations for each ZnF domain. Tandem partner refers to each ZnF-partner into the same tandem (ZnF1 partners ZnF2, and ZnF3 partners ZnF4).

	ZnF1	ZnF2	ZnF3	ZnF4
ZnF1-RNA	0.01	0.05	0.31	0.12
ZnF2-RNA	0.06	0.03	0.21	-0.06
ZnF3-RNA	0.12	0.04	0.36	-0.09
ZnF4-RNA	0.13	-0.04	0.08	0.76
Tandem partner-RNA	0.05	0.11	0.06	0.87

5.2.7. BOTH ZNFs OF MBNL1 HAVE DIFFERENTIATED AFFINITY FOR RNA

To complete the picture of the impact of AS cognate RNA-binding over the structural rearrangements, steered molecular dynamics (SMD) were conducted on each MBNL1 domain. Forces were applied to pull the UGCU fragment in order to describe the putative substrate binding/unbinding process for each domain (see schematic representation in figure 5.7A, page 173). The starting points for this study were 20 randomized frames from each ZnF - RNA complex equilibrated trajectory, which makes a total of 80 pulling simulations. Each simulation computes the work of pulling away the center of mass (COM) of the ZnF binding site and the GC pair by 20 Å. The COM for each system was defined as described in Protocols (CD-ROM). A spring constant of 55.6 nN·Å⁻¹ and a constant velocity of 1.35 Å·ns⁻¹ were used to compute the cumulative work profile for each system. Pulling force is peaked at ~8.5 Å in all system models, with a subsequent rearrangement in the adjoining GC binding pockets. No contribution of U nucleobase to the binding/unbinding process was observed except for the binding of Arg residues to the backbone of the first and fourth nucleotide.

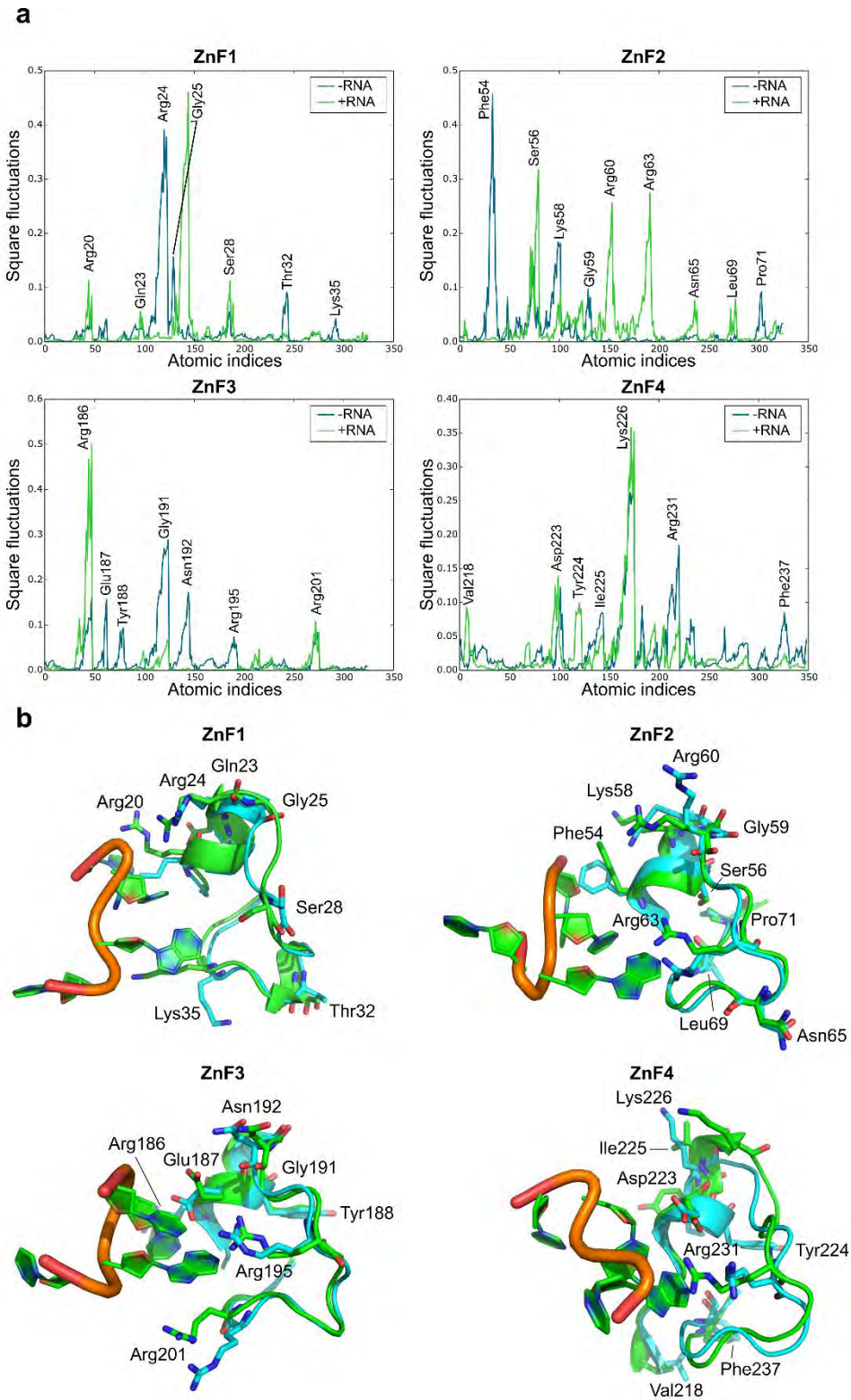


Figure 5.6. (a) Local fluctuations extracted from the MD simulations for each domain alone (-RNA) or bonded to RNA (+RNA). Only the atomic indices of the RNA binding site are represented (ZnF1: 18 to 38; ZnF2: 52 to 72; ZnF3: 184 to 204; ZnF4: 218 to 238). (b) Relevant fluctuating amino acids in an unbound (blue) and bound (green) state.

As shown in figure 5.7B, a statistical significant difference is observed between ZnF₃ and the first and second domains ($p < 0.01$). ZnF₁ and ZnF₂ pulling simulations yielded very close work values (31.14 ± 5.03 and 31.43 ± 2.35 kcal·mol⁻¹ respectively). Furthermore, ZnF₃ ($W = 26.29 \pm 4.52$ kcal·mol⁻¹) clearly exhibits the lowest cumulative work followed by ZnF₄ ($W = 29.10 \pm 5.01$ kcal·mol⁻¹). This observation is in agreement with truncation and point mutation studies that revealed that ZnF_{1/2} domains are required for an effective splicing of the majority of targets.

To elucidate detailed differences between the binding domains, local fluctuations were computed at four equally distributed segments extracted from the SMD process (each segment represents 2 ns). Figure 5.8A shows that local fluctuations are more pronounced in ZnF₂ and ZnF₄ while the other domains yield more localized fluctuations. Arg, Lys, Glu and Asp mainly represent the pocket mobility and, not surprisingly, most of them are reduced upon cognate binding. Notice that half of the fluctuating residues are not at a binding distance to the RNA target (greyed regions in figure 5.8A). Visual inspection of the SMD trajectories from ZnF₁ and ZnF₃ do not provide any evidence about the binding process differences and the C α rmsd between them is maintained around ~ 1.4 Å. However, the contradiction can be reconciled when we consider both local fluctuations and pocket rearrangements. We noticed that residues Gly₂₅ and Gly₁₉₁ exert dynamic control of the binding region, which induce subtle changes in protein dynamics. General differences between the binding pockets are observed in figure 8b. Binding pockets of ZnF₂ and ZnF₄ exhibit a broad volume distribution that agrees with the major number of local fluctuations into these regions. This effect indicates local and large rearrangement of the adjoining pockets upon RNA binding. On the contrary, ZnF₁ and ZnF₃ display similar distributions whose statistical mode is located at 24 Å³ and 48 Å³ respectively. Mode of ZnF₂ and ZnF₄ distributions are close (36 Å³ and 40 Å³ respectively) and their distribution range from 12 Å³ to 172 Å³.

5.2.8. SIGNIFICANCE OF THE STRUCTURAL STUDY

This work was intended to characterize the motions of the four zinc fingers (ZnFs) contained into the MBNL₁ protein and provide a complimentary picture to the experimental observations from the structural point-of-view using conventional and steered molecular dynamics. Coevolutionary couplings also encode for dynamical and functional information, hence secondary and tertiary structure is usually more conserved than the sequence among the same family. Sequence coevolution analysis of the CCCH domain confirmed this view and highly coevolving residues in MBNL₁ present high fluctuations that determine each domains' affinity. Their combined coevolution propensity and conformational mobility suggest that proteins at substrate recognition sites suitably recruit charged and polar amino acids (especially Lys and Arg). Thus, these residues are at the same time specific and flexible enough to mediate substrate selectivity.

MBNL₁ global fold is conserved upon RNA binding but local and global fluctuations are remarkably modified. The ZnF_{3/4} tandem is more resilient to changes in local motions. ZnF₃ fluctuations present a modest correlation with those observed upon RNA binding. However, RNA-binding to ZnF₄ produces a remarkable change into ZnF₃ local motions. An intriguingly finding is that ZnF₄ motions remain invariable during all possible events (free state, RNA-

binding and RNA-binding to ZnF3). Some studies hypothesized that ZnF4 is prone to establish protein-protein interactions (PPI) with another MBNL1 ZnF4 unit by establishing Tyr224, Gln244 and Tyr236 contacts.¹¹ Our simulations suggest that only Tyr224 mobility is enhanced upon RNA binding but no changes are apparent in Tyr236 and Gln244. Recent studies showed that mutation of these aromatic residues did not provide any effect in functional assays.¹² Nonetheless, cavities can be the locus of PPI and their composition and shape complementarity may be related to binding affinity and specificity. Interestingly, ZnF4 is highly charged if compared to the other ZnF domains and the enhanced conformational adaptation of its pocket suggests favorable inter-domain binding features.

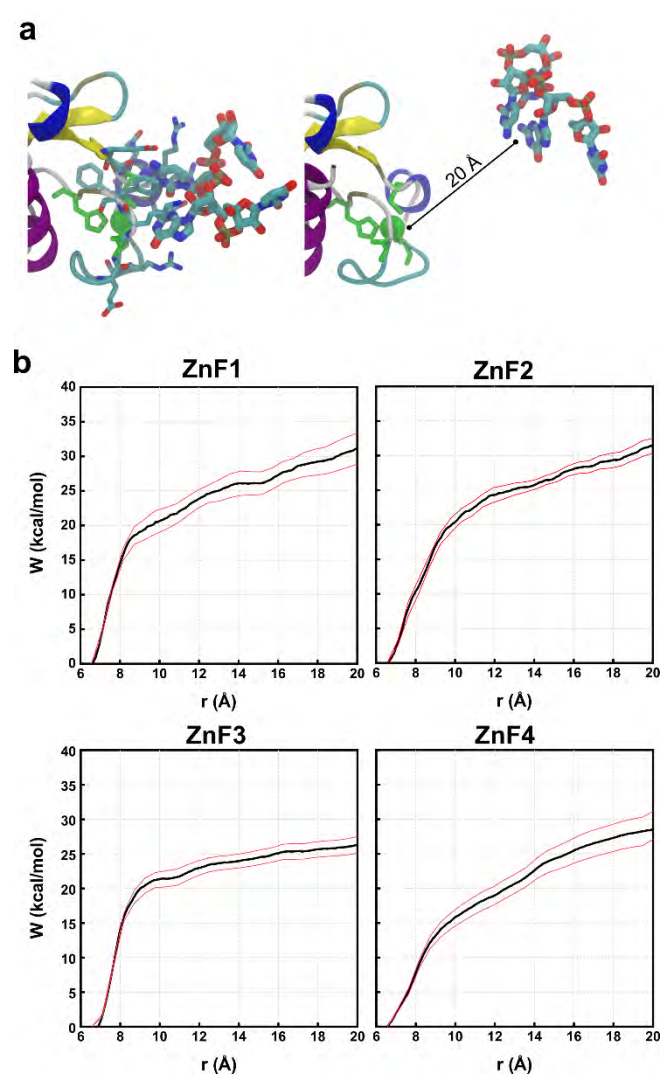


Figure 5.7. (a) Representative snapshots before (left) and after (right) the SMD pulling process. The CCCH motif is colored green. (b) Cumulative work profiles for each ZnF vs distance from the center of mass (r). Black and red lines represent the mean and the standard error of the mean, respectively.

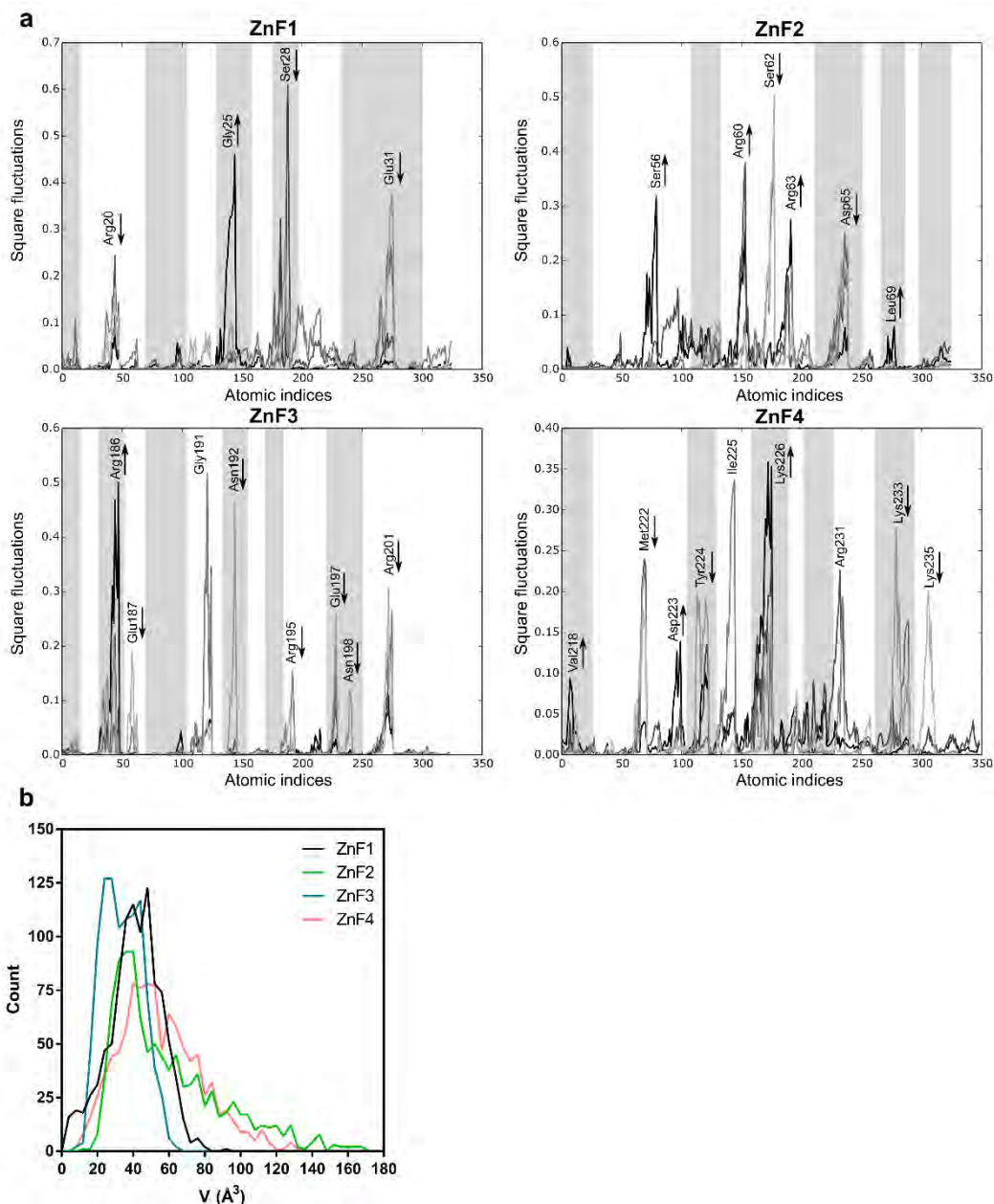


Figure 5.8. (a) Local fluctuations extracted from the SMD simulations at four equally distributed segments of the trajectory. The binding process fluctuations (reversed SMD trajectory) are represented in light gray (unbound) to black lines (bound). Only the atomic indices of the RNA binding site are represented (ZnF1: 18 to 38; ZnF2: 52 to 72; ZnF3: 184 to 204; ZnF4: 218 to 238). The most relevant fluctuations that increases or decreases during the process are indicated with an arrow. Grayed regions are not at RNA-binding distance. (b) Volume of the pocket computed during the SMD process for each ZnF.

As commented before, the four ZnFs in MBNL1 have different RNA-binding and splicing activities, hence global and local fluctuations have been investigated. The simulations, among other things, confirmed that conserved Gly adjoining to the α -helix enhance local mobility of the RNA-binding region. We also found that, within the length of the simulations, global fluctuations are remarkably different in each RNA-binding domain. Albeit the core structure of each domain is preserved in all simulations several effects, alone or in combination, explain the differences in observed local and global motions. Summing up the results, it can be concluded that the four ZnF domains provide distinct RNA-binding platforms in terms of structural

sampling and mobility that may have implications in the differentiated splicing events observed in literature. Although these results are in good agreement with experimental data, they are not conclusive and further research will be required to validate these observations. An important future line of work will be to identify the contributions of not structured regions such as the inter-domain linker.

5.3. D-HEXAPEPTIDES AS DM1 DRUGS

RNA-protein interactions play crucial roles in controlling gene expression (e.g. MBNL1 binding to YGCY). Hence they are becoming important targets for pharmaceutical applications.^{1,34-35} Peptidic ligands are providing attractive drug leads for many targets such as proteins or RNA. However, as it has been discussed in the previous chapter, targeting RNA is challenging due to the highly charged nature and flexibility of the RNA. Moreover, designing RNA-binding molecules is limited by the poor understanding of RNA-ligand recognition events. MD simulations with implicit-solvent representations have shown to provide a relatively good agreement with native-like binding poses.³⁶⁻³⁸ Nonetheless, conformations that a peptide adopts in solution are a function of both their sequence and the surrounding environment.

In 2011, García-López et al.¹ discovered a *D*-amino acids hexapeptide – henceforth named ABP1 or **ppyawe** – that efficiently suppressed CUG-induced phenotypes in *Drosophila* tissues. The authors noticed that ABP1 reduced the number of CUG-RNA foci and caused Mbl (muscleblind-like (*drosophila*) protein) subcellular redistribution in the fly muscle. ABP1 was proved to directly bind rCUG structures and likely relies on the induction of a conformational modification in rCUG secondary structure. More recently, the same group reported that ABP1 preferentially adopts a polyproline helix II (data not published).

Albeit *L*-amino acids give rise to right-handed helices, *D*-amino acids favor left-handed structures. A particularly important contribution to left-handed structures is provided by polyproline, which forms helical structures with two well-characterized conformations. Polyproline helix II (PPII or left-handed polyproline helix) is formed when sequential residues adopt a particular dihedral backbone (ϕ, ψ) = $(-75^\circ, 146^\circ)$ with the prolyl bonds in the *trans*-isomer conformation. However, the designation of PPII helix is somewhat misleading, since proline is not mandatory, though it has the highest occurrence in these structures. In fact, as up to 46% of PPII helices in folded proteins contain no proline. Interestingly, its extended conformation and the absence of regular intra-molecular hydrogen bonds make PPII an ideal structure for a wide range of molecular interactions. Thus, the next step in this thesis was to perform a

conformational study of a subset of 16 *D*-hexapeptides reported by Artero's group using MD simulations in order to find sequence-structure-activity relationships. Activity reported from animal model studies is shown in table 5.2.

Table 5.2. Activity in *Drosophila* model of the 16 *D*-hexapeptides reported by García-López et al.¹

Sequence	Concentration assayed	Emerged treated/emerged control
cpyaqe	80 μ M	0.3
cpyawe	80 μ M	-
cpytqe	80 μ M	0.8
cpytwe	62 μ M	-
cqyaqe	25 μ M	2.0
cqyawe	25 μ M	-
cqytqe	80 μ M	1.4
cqytwe	57 μ M	0.9
ppyaqe	80 μ M	2.0
ppyawe (ABP1)	80 μ M	4.0
ppytqe	80 μ M	0.8
ppytwe	80 μ M	3.0
pqyaqe	80 μ M	0.8
pqyawe	40 μ M	1.8
pqytqe	40 μ M	0.5
pqytwe	38.5 μ M	0.4

5.3.1. PEPTIDE STRUCTURAL SAMPLING

Roughly, these hexapeptides can be separated into two families: those who start with any residue (^DCys or ^DPro) followed by a ^DPro – a [Pre-^DPro•^DPro•General] or PPX scheme – and those not containing a ^DPro into the second position which follows a ‘general’ scheme. The main difference between both schemes is that PPII conformations are more frequently observed when the sequence starts with a ^DPro or Pre-^DPro residue. Although this condition is not necessary true, MD simulations yielded < 8% PPII clustered structures for sequences following a ‘general’ scheme, probably induced by force field parameterizations. Figures 5.9 and 5.10 show the projection of several snapshots of the PPX and ‘general’ peptide groups respectively over a Ramachandran plot. Notice that torsional values have been modified in order to match an *L*-amino acids Ramachandran representation. Hence the represented Ramachandran corresponds to the mirror image for each peptide's original representation. In addition, due to chirality effects, left and right-handed helices are also inverted in the Ramachandran. Non-PPX peptides exhibit a higher propensity for β -sheets than PPX, but helical content is high in all 16 peptides. In fact, PPII helix is close to the β strand in the conformational space, and the transition between both conformations only requires a shift of ϕ . Therefore, minor changes in sequence may trigger the formation of β -sheets from PPII conformations, or vice versa. PPII propensity in PPX peptides ranges from ~10 to ~30% being **ppytqe** the hexapeptide with both highest structured and PPII content. PPX with two ^DPro (or **p**) residues exhibit a higher amount of PPII clusters, as stated by the Proline and Pre-Proline plots in figure 5.9. Notice that peptides containing ^DThr (**t**) in position 4, and ^DTrp (**w**) or ^DGln (**q**) in position 5 explore non-allowed regions which induce a destructure of the peptide.

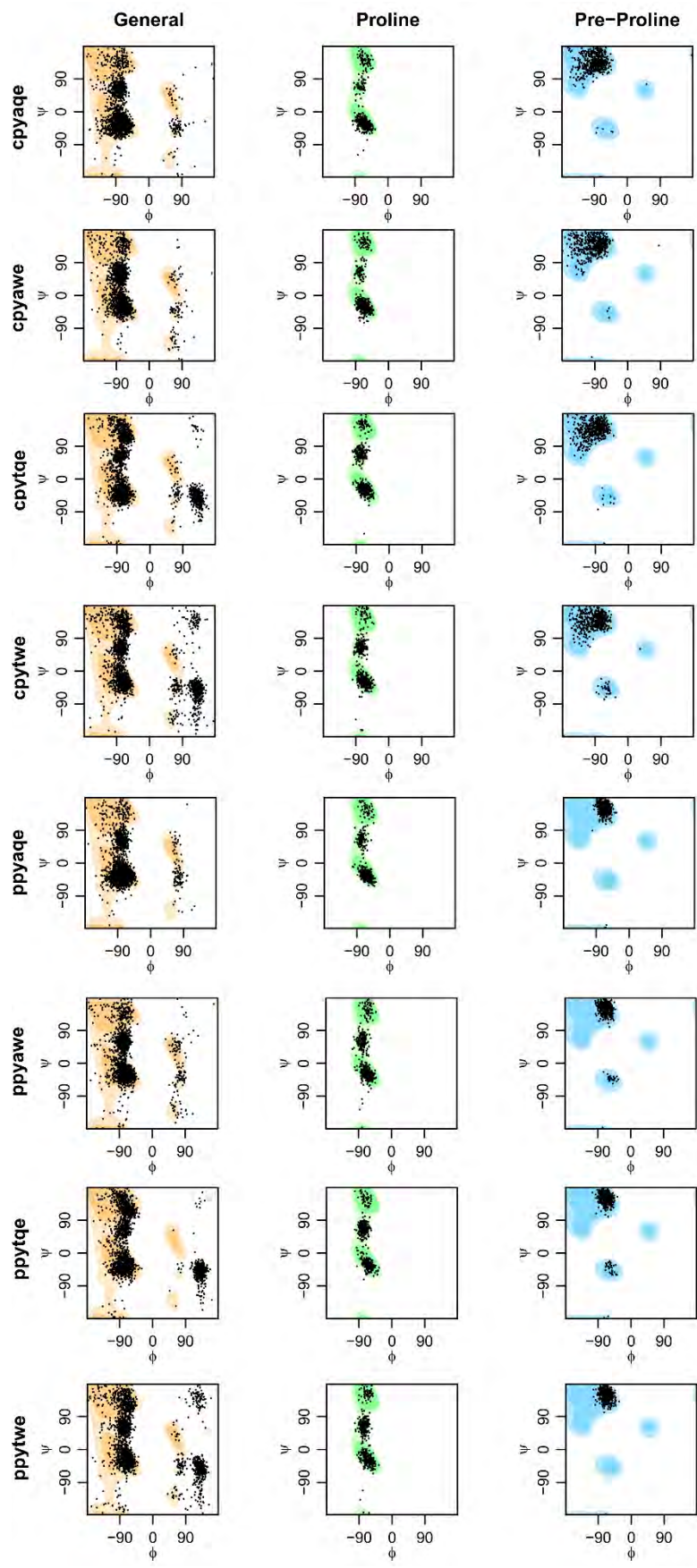


Figure 5.9. PPX type Ramachandran plots extracted from the MD. Three projections are shown per residue: general amino acids (non-Pro and non-Pre-Pro residues), Proline and Pre-Proline.

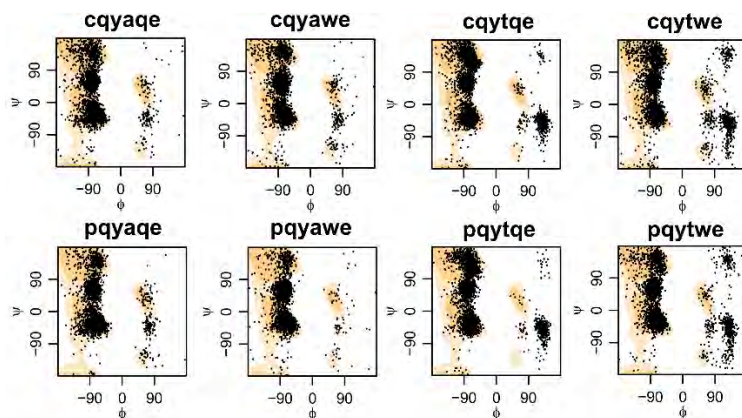


Figure 5.10. General Ramachandran plot representations of the non-PPX peptides.

Nonetheless, the most active peptides (**ppyawe**, **pytwe** and **pyaqe**) are characterized by high PPII content which is mostly provided by the initial **p**'s, whereas peptides containing a **q** in position 2 hardly present a structured form, which agrees with circular dichroism (CD) spectra. As shown in figure 5.11, PPII configuration is mainly conditioned by residues in positions 1 and 2. High helical content (including α , 3_{10} and PPII helices) is found in **pyaqe**, **cpyaqe**, **pyaqe** and **pyawe** hence helices require ^DTyr (**y**) and ^DAla (**a**) into the middle positions in order to maintain a helical global fold.

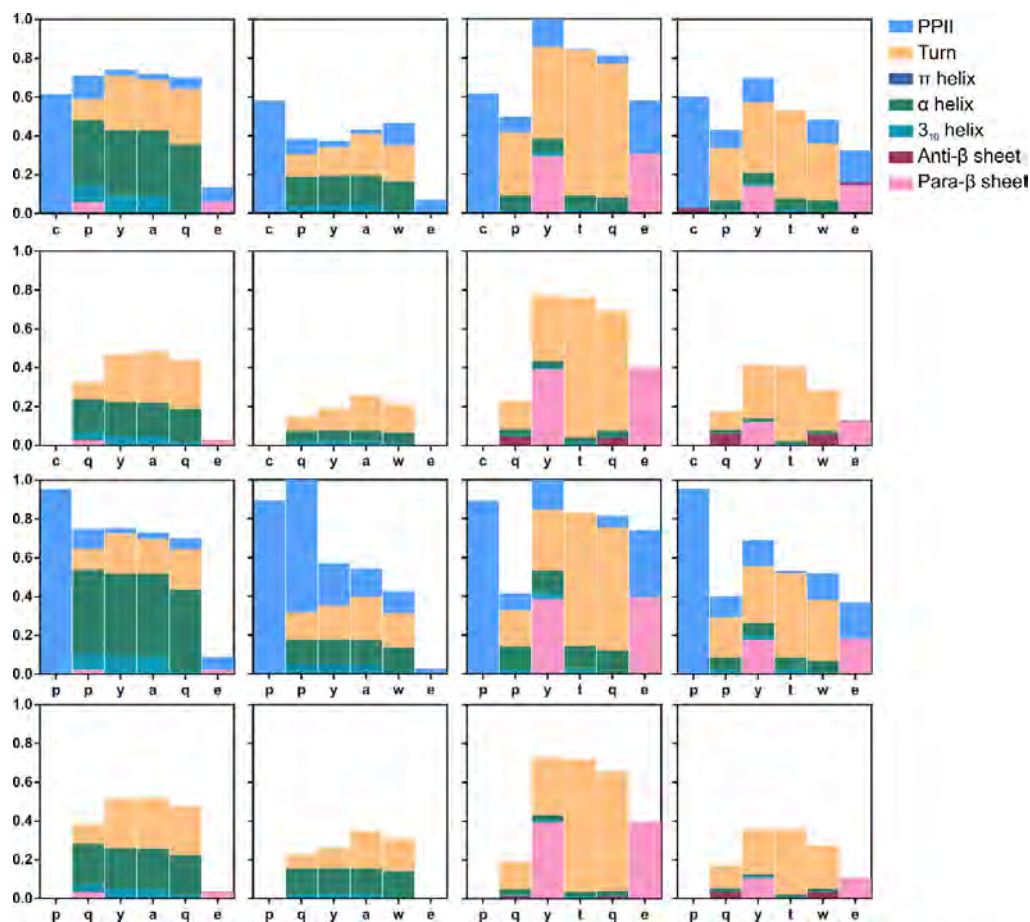


Figure 5.11. Normalized contribution of each amino acid to the overall secondary structure: polyproline II helix (PPII), turn, π helix, α helix, 3_{10} helix, antiparallel β sheet (Anti- β sheet) and parallel β sheet (Para- β sheet).

Interestingly, $^{\text{D}}\text{Cys}$ (c) in position 1 does not alter the overall structural propensity of the peptide but yield lower PPII content values (figures 5.12A-C). In addition, a modest coefficient of determination ($R^2 = 0.58$) is observed between animal model data and PPII content (figure 5.12B). Although several limitations are inherent to the MD method, such as the force field parameterization and the use of implicit solvent for the simulations, this correlation can be confidently attributed to a sequence-structure-activity relationship. Nonetheless, **ppytqe** and **ppyawe** stay in a PPII conformation during a significant time interval of the simulation but transitions between PPII and α -helix are frequently observed. In particular, figure 5.12 also shows the main differences between both conformations of **ppyawe**: PPII conformation maintains **y** and **w** residues at the same side, which may favor stacking interactions with the target RNA (figure 5.12D). On the contrary, α -helix **ppyawe** adopts a polarized configuration which may favor polar based interactions with the RNA through the backbone and the side chains (figure 5.12E).

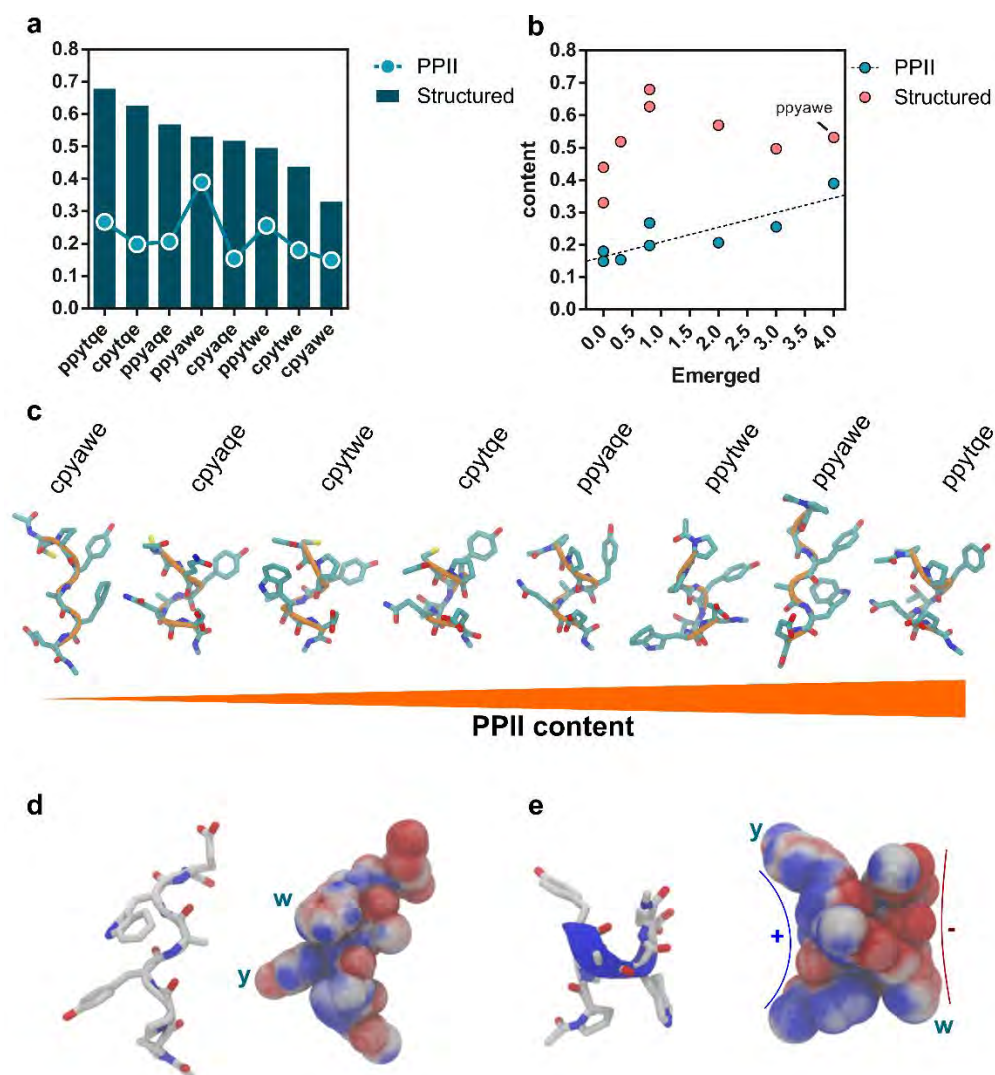


Figure 5.12. (A) Total structural propensity and PPII content found along the MD trajectory for each PPX. (B) Correlation between PPII content and biological activity. (C) Clustered conformations for each PPX. (D) PPII and (E) α helix configurations and charge distribution of ABP1 (ppyawe). The electrostatic potential has been computed using ABPS and contoured at ± 10 kT/e.

5.3.2. PEPTIDE MODIFICATIONS

Second generation ‘peptide-like’ compounds have been proposed using two distinct methods. On the one hand, the first modification proposed by Artero’s group was the change of the **ppyawe**’s sequence to its peptoid homolog (thus bonding the side chain of each amino acid into the N atom instead of the C α , figure 5.13A). These oligo *N*-substituted glycines have chemical properties similar to peptides and have been the subject of interest for their biomimetic properties and drug-delivery. The peptoid backbone is achiral, but specific patterns of chiral *N*-substituted or bulky sidechains can be engineered in order to populate specific conformations. On the other, it is expected that a linker of length *n* between the first and last residue would ‘lock’ a specific conformation of the peptide. Specifically, **ppyawe** was linked with a C $_8$ chain as shown in figure 5.13B. In that regard, MD simulations of these modification were conducted as previously described in order to study their structure sampling. The main goal of this peptoid and peptide modification research is to improve the selectivity of the aforementioned peptides. The main hypothesis is that well-folded oligomers may adopt specific conformations or conformations beyond helix and loops.

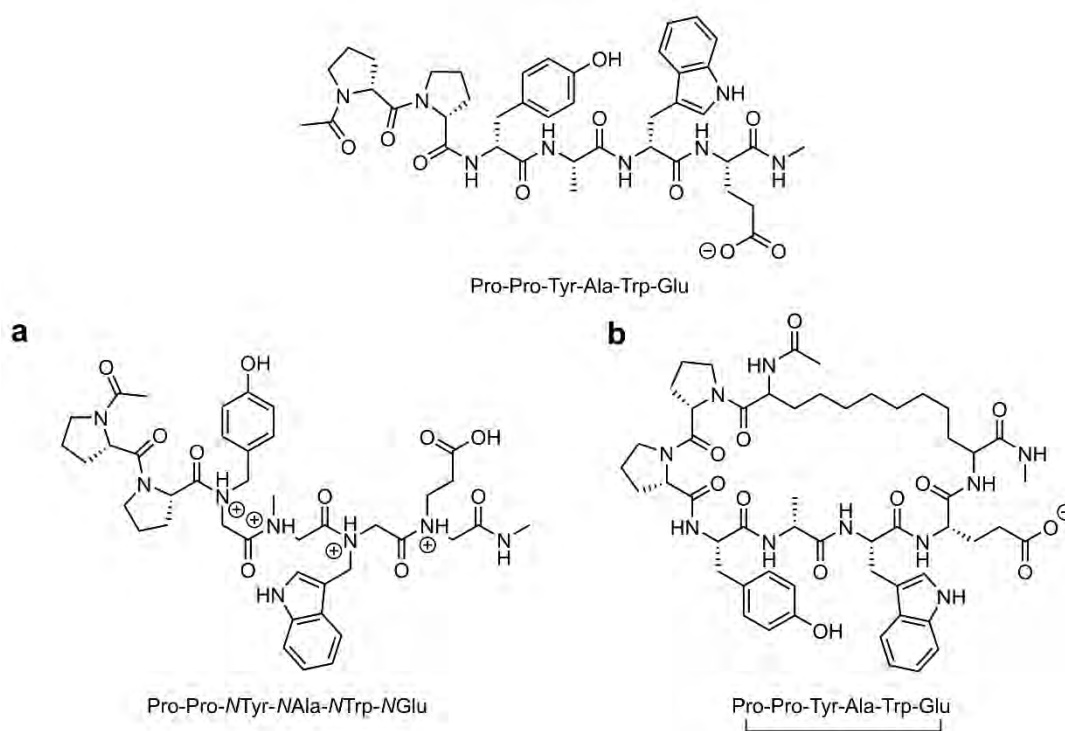


Figure 5.13. ABP1 or ppyawe proposed modifications: (A) peptoid analog and (B) D-hexapeptide linked with a C $_8$ chain.

In agreement with previous MD results, peptoid’s simulation reward compact structures and solvent exposed charges, especially in a PPII configuration. Notice that the peptoid contains 4 additional positive charges as well as 4 less chiral centers which should induce a totally different conformational sampling. Surprisingly, high content of both left and right-handed PPII structures were found during the sampling (38% of total simulation time). As reported previously for the PPX peptides, PPII conformation of the peptoid is controlled by the initial **p**’s. Not

surprisingly, the Ramachandran plot offers symmetric representation due to the chirality loss (figure 5.14A).

In contrast, the addition of the linker to the peptide restricts the conformations to a 3_{10} helix and turn configuration (figure 5.14B). This outcome was attributed to two independent factors: (i) the linker design and (ii) the peptides starting structure. However, we hypothesize that the linker is not long and flexible enough to allow for a conformational reorganization of the peptide and longer linkers should yield higher PPII and α helix propensities. Nonetheless, both peptoid and modified peptide contain non-conventional building blocks which may compromise the system and introduce additional force field inaccuracies.

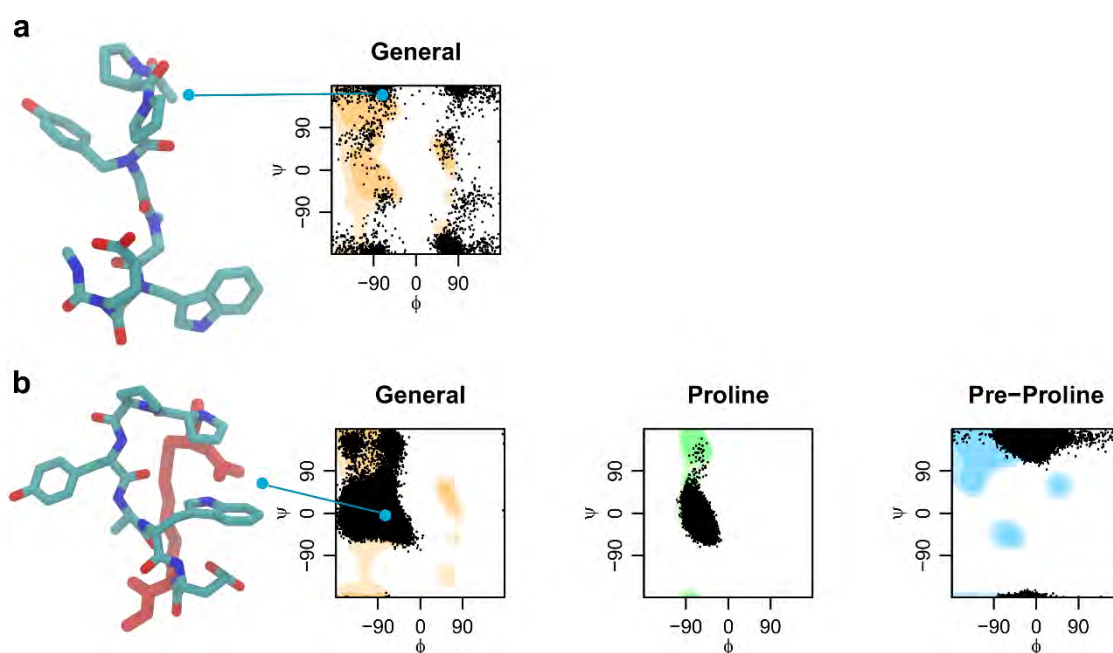


Figure 5.14. Structural sampling and representative conformation of (A) the peptoid analog of ABP1 and (B) ABP1 linked with a C_8 (the linker is represented as red sticks).

This strategy is being applied into current project and new peptides are being investigated. We foresee to apply these MD protocols for studying new peptoids or modified peptides in order to have a better understanding of the structural propensities and lead a structurally rationalized peptide/peptoid drug design. Future studies should include a fine tuned parametrization of the peptoid building blocks and a high understanding of the potential force field inaccuracies. Moreover, peptide's conformations depend on the environment and the PPII/ α -helix content must be fine-tuned for implicit solvent simulations.

5.4. PEPTIDE STRATEGIES FOR HIV-1 TARGETING

Earlier attempts to inhibit the Tat-TAR interactions of HIV using linear peptides or peptoids provided limited structure-activity relationships, mainly because two factors: the flexibility of these molecules in solution and the lack of knowledge of their bound structures. Contrary to rCUG-MBNL₁, recent structural information of the HIV Tat-TAR complex is available and the structure of a peptide bound to TAR has been reported.^{39,40} Recent studies suggest that peptide affinity and selectivity can be grouped into three categories: stacking with residue A₃₅ of the TAR apical loop; hydrophobic contacts that drive formation of a base triple between bulge residues U₂₃ and base pair A₂₇-U₃₈; and stabilization of the base triple by charges residues.⁴⁰

Pascale et al.² recently reported a series of polyamide amino acids (PAAs) that exerted excellent TAR binding affinities. However, thermodynamic studies suggest that their ability to compete against the Tar protein is different, probably because they induce different structural rearrangements upon binding. Studied PAA structures and experimental data are shown in figure 5.15 and table 5.3 respectively.

As shown in table 5.3, **IIIb**, which forms loose complexes with TAR, was shown to be a stronger Tat competitor than those forming tight ones (e.g. **Ic**). In fact, thermodynamic studies highlight an entropy-enthalpy compensation phenomenon which results in similar binding free energies. For a better comprehension on the comparative binding modes of these PAAs subsequent truncation studies were performed (KRFR for **IIIb** and RFFR for **Ic**). Thermodynamic data suggested that Phe-rich interactions are fewer or less optimized than Arg-rich ones. In addition, the F- β -alaninamide moiety slightly destabilize the complex and produces an enthalpy loss and entropy gain. In the previous section, no complex data was available hence it complicated the rationalization of the activity due to the lack of structural models. Notwithstanding the predictive power of peptide structural sampling, MD is also a powerful tool to explore molecular complexes formation at an atomistic level and analyze the conformational energy landscape accessible to these molecules. Such protocols have been applied

in this study in order to predict the interaction modes between C- α -PAA and TAR that may compliment experimental data.

Table 5.3. TAR affinity (K_D), inhibitory constant (IC_{50}) and thermodynamic parameters for selected PAA-TAR complexes.					
	K_D	IC_{50}	ΔG°	ΔH°	$T\Delta S^\circ$
IIIb (KRFRF)	57.7	34.7	-41.3	-13.3	28.0
Ic (RFFRF)	89.5	> 2000	-40.2	-14.4	25.8
KRFR	-	-	-36.98	-56.73	-19.77
RFFR	-	-	-39.73	-20.60	19.13

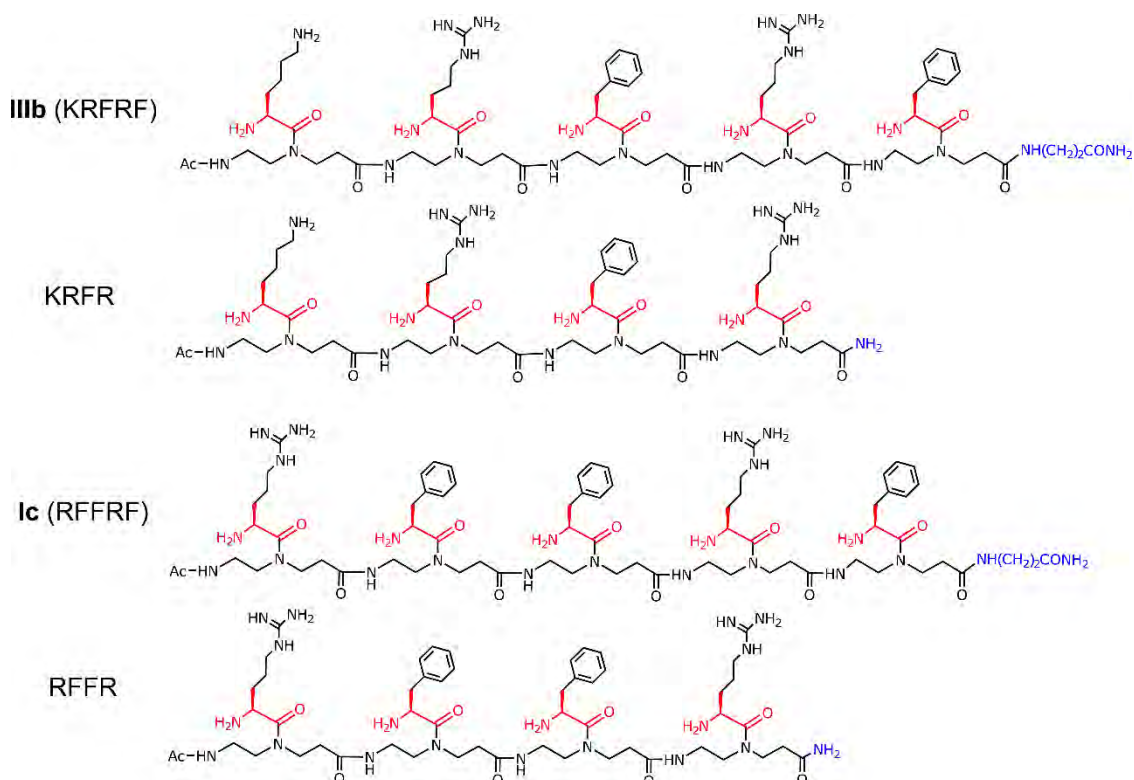


Figure 5.15. Molecular structure of the four selected PAAs. Pentameric structures contain a β -alaninamide fragment (blue) while tetrameric PAAs contain a NH_2 group.

5.4.1. TAR MOLECULAR RECOGNITION OF PENTA-C- α -PAA Ic AND IIIb

First each complex was completed and simulated as described in the Methods section. Visual inspection of MD results is in good agreement with NMR data, which suggested a preferred binding site in the bulge region of TAR for both **Ic** and **IIIb** compounds. However, the orientation differs from both complexes: **IIIb** points towards the interhelical junction of the bulged TAR, whereas **Ic** provides a more packed conformation and buries into the TAR major groove similarly to previously reported structures. The dynamic behavior of **Ic** and **IIIb** complexes along the MD is noteworthy. Analysis of the full trajectory suggest that **Ic** widens the

bulge region up to 15.2 Å while **IIIb** produces a contraction of this region up to 9.0 Å with subsequent RNA helicity loss (figure 5.16A). According to these results, both compounds induce non-negligible TAR conformational changes, as pointed out by CD spectra. In addition to major structural changes of TAR upon ligand binding, MD confirms the positive influence of the N-terminal sequence since F-β-alaninamide terminal moiety remains exposed to solvent during almost all the trajectory. Moreover, no specific interactions have been characterized for this unbound fragment that might contribute to ΔH° .

A closer inspection of the complex trajectories suggests that specificity is clearly achieved by side-chain interactions based on π stacking, cation- π stacking and charge-based interactions as well as the orientation of the compound along the major groove (Figure 5.16B). Compound **Ic** buries into the major groove of TAR by means of intimate charge-based contacts. Interestingly, U25 is flipped out from the RNA complex and drowns down the whole UCU loop. This is counterbalanced by C24 stabilization into the bulge region through a Phe2 π -stacking interaction and Arg4 charge-based interaction with the phosphate group of U23. R4 also yields intramolecular contacts with Phe3 via a cation- π stacking interaction. This interaction is reproduced by R1 with the stack of the guanidium group on top of C30, which significantly contributes to complex stabilization. Despite the sequence dissimilarity between compounds **Ic** and **IIIb**, both provide similar interactions with TAR. For instance, C24 stacks with Arg2 via a cation- π interaction, while the acetyl moiety buries into the major groove of the apical loop and interacts with the 2'OH group of U23 via hydrogen bonding. The same interaction is hypothesized for R4 through the sugar moiety of G33. Moreover, the penta-PAA backbone significantly contributes to binding through electrostatic interactions. For instance, Lys1 provides a charge-based interaction with the backbone and Phe3 anchors to the C30/U31/G34/G36 cluster.

Following the aforementioned MD approach, the study of the computational thermodynamic properties of the tetra- and penta-PAA complexes was conducted using MM/GBSA analysis. This procedure has been extensively used in the prediction of binding free energies by rigorously decomposing the total binding free energy into different interaction terms.^{4†} The explanation for the difference of enthalpy can be derived from the number of intermolecular interactions that must be broken during the dissociation process. Thus, the correlation between experimental and computational thermodynamic data may lead to a feasible description of these interactions. Under that premise, the binding free energy of compounds **Ic**, **IIIb** and their tetra- derivatives was predicted and correlated with experimental data.

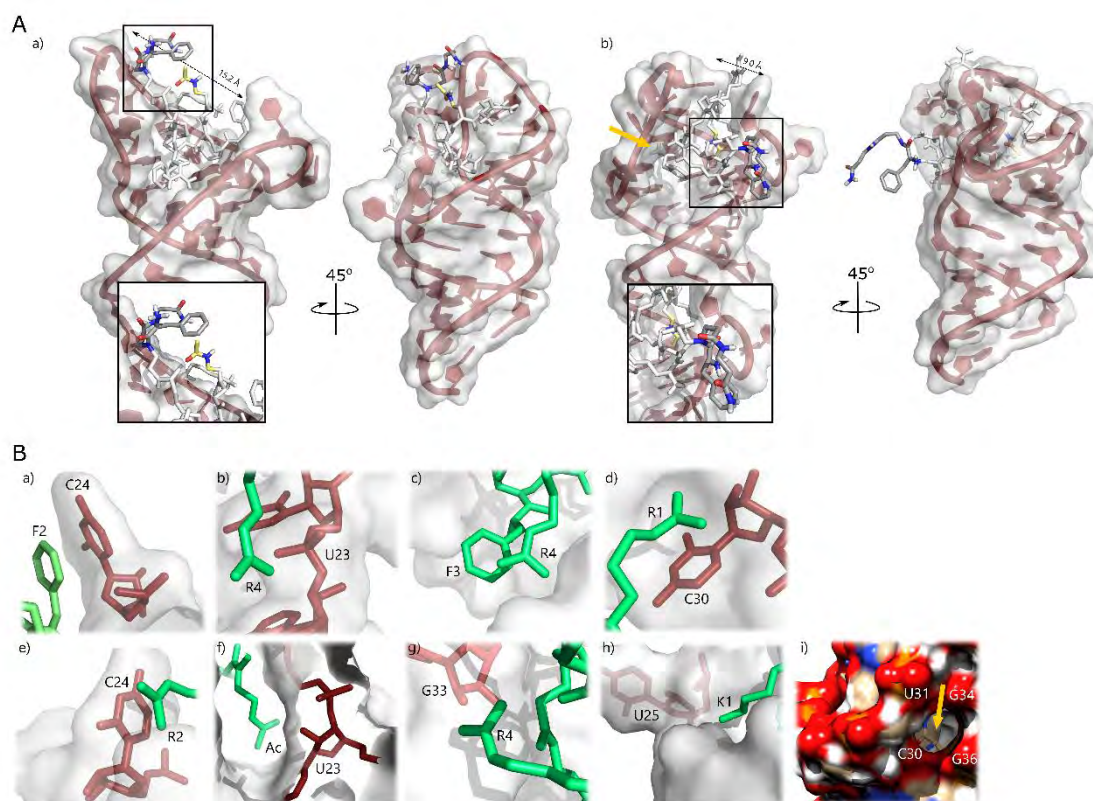


Figure 5.16. (A) Final MD frame representation of **Ic** complex (a,b) and **IIIb** (c,d). F- β -alaninamide (in red) remains exposed to the solvent while the rest of the penta-PAA buries into the TAR major groove. The bulge closure distance has been calculated as N2(G34)-N4(C24) distance for **Ic** and N2(G32)-N4(C24) distance for **IIIb**. Acetyl terminus is depicted as a blue fragment. (B) Main interactions provided by MD simulations for **Ic** (a to d) and **IIIb** (e to i). Henceforth, PAA are numbered as Ac-R1-F2-F3-R4-F5 and Ac-K1-R2-F3-R4-F5. HIV-1 TAR is represented as a white surface with red nucleotides and PAA are represented in green. On the one side, **Ic** interacts (a) with C24 into the bulge region through a Phe2 (F2) π -stacking interaction. Arg4 (R4) interacts at the same time with (b) phosphate U23 group and (c) Phe3 (F3) via intra- residue cation- π stacking interaction. (d) Arg1 (R1) stacks with C30 providing another cation- π interaction which stabilizes the complex. On the other side, **IIIb** (e) stacks with C24 via a Arg2 (R2) cation- π interaction. (f) Ac group buries into the major groove and interacts with the 2'OH group of U23. (g) G33 also provides a hydrogen bond interaction with Arg4 (R4) through the 2'OH group. (h) Lys1 (K1) provides charge-based interactions with the backbone. (i) F3-NH $_3^+$ group (yellow arrow) anchors to the TAR opening, which contains a high negative charge density.

5.4.2. INTERFACIAL WATERS PLAY AN ESSENTIAL ROLE IN RECOGNITION

As shown in figure 5.17, the MM/GBSA binding free energy correlated poorly with experimental data ($R^2 = 0.464$). However, the correlation is dramatically improved when interfacial water molecules are included in the analysis ($R^2 = 0.826$). Waters that remain in the interface between TAR and PAA for more than 10% of the total simulation time are selected in this analysis (see figure 5.18). Inspection of the structures shows a total of 1 and 3 water molecules involved in the **Ic** and RFFR complexes respectively, and 3 and 1 water molecules involved in the **IIIb** and KRFR respectively. These patterns of water-mediated interactions between TAR and PAA show an improvement of predictive power of classical MM/GBSA resulting in better estimates of binding free energy. Nevertheless, MM/GBSA is only an approximation to the 'real' binding free energy because the lack of the configurational entropy in its calculation. In order to address this issue, the entropic term was computed using the quasi-harmonic approximation. Unfortunately, the protocol fails to accurately estimate entropy, as shown in table 5.4. Configurational entropy of KRFR and RFFR is overestimated due to the absence of the F- β -

alaninamide moiety. In contrast, the presence or absence of the β -alaninamide chain did not result in significant changes in the orientation of the backbone. Although computational results agree with experimental data, more accurate and computationally intensive protocols should be applied in order to obtain a better entropy estimation.

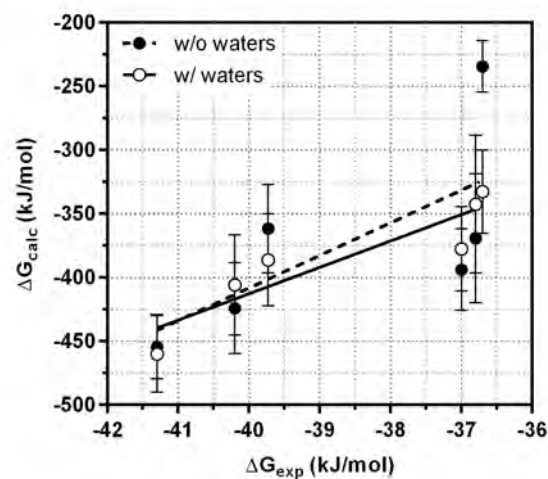


Figure 5.17. Correlation between experimental and calculated ΔG of binding not including ($r^2=0.464$) and including ($r^2=0.826$) interfacial waters in the calculation.

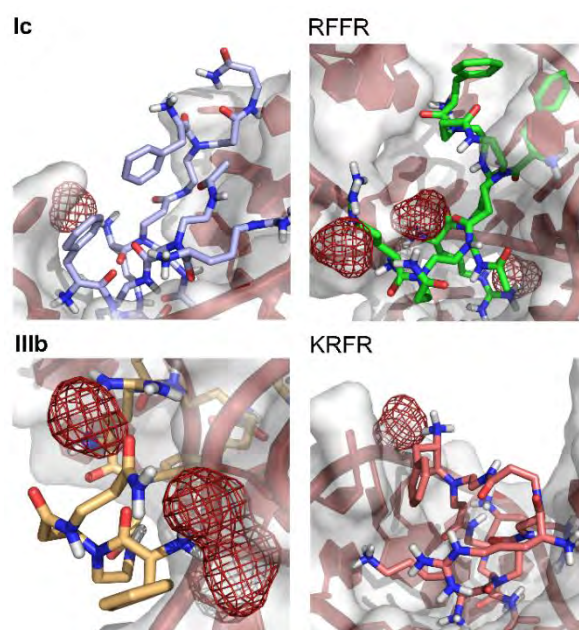


Figure 5.18. Interfacial density maps for Ic, RFFR, IIIb and KRFR.

Table 5.4. Relative free energy of binding considering interfacial waters, conformational entropy and corrected free energy of binding using the quasi-harmonic approximation.

	ΔG°	$T\Delta S^\circ_{\text{conf}}$
KRFRF	-459.93 ± 29.90	-346.66
RFFRF	-405.83 ± 39.21	-338.42
KRFR	-377.53 ± 33.04	-421.07
RFFR	-386.23 ± 36.04	-407.48

5.5. PROTOCOLS

5.5.1. EQUIPMENT

The following computational methods and procedures were accomplished with in-house systems and high performance computing resources (HPC):

In-house equipment

- 4x Intel Xeon 8-core at 3.50GHz, 32GB of RAM memory. NVIDIA Quadro K4000. 2TB filesystem storage.
- Intel Core 2 Quad Q8200 2.33 GHz, 4GB of RAM memory. 2TB filesystem storage.

HPC resources

- **MINOTAURO-BSC:** Red Española de Supercomputación (RES), Barcelona Supercomputing Center (BSC). 126 compute nodes and 2 login nodes. Each node has 2x Intel Xeon E5620 6-core at 2.53GHz. 24GB of RAM memory, 12MB of cache memory. 250GB local disk storage (SSD). 2x NVIDIA M2090 with 512 CUDA cores and 6GB of GDDR5 Memory. GPFS parallel filesystem disk storage (~2PB).
- **TIRANT-UV:** Red Española de Supercomputación (RES). 256 JS20 compute nodes and 5 p515 servers. Each blade has 2 IBM Power4 at 2.0GHz with 4GB of RAM memory. 36GB of local disk storage. GPFS parallel filesystem disk storage (10TB).

5.5.2. PROTOCOLS

CONSERVATION AND COEVOLUTION ANALYSES OF THE CCCH DOMAIN

The multiple sequence alignment (MSA) was retrieved from the Pfam database (PF00642, containing Zinc finger CX₈CX₅CX₃H type and similar).⁴² Mutual information (MI) between amino acids at the *i*th and *j*th positions were computed with MISTIC.⁴³ This server uses an APC corrected MI to reduce background mutual information and translate them into *Z*-scores. The cumulative MI (cMI) and proximity MI (pMI) were computed for each residue with *Z*-score > 6.5 as previously described.

MBNL1 MODEL SYSTEMS PREPARATION

Structural models of ZnF1/2 and ZnF3/4 in a free state were retrieved from the Protein Data Bank (PDB ids 3d2n and 3d2q respectively). Short molecular dynamics (MD) simulations indicated that the short α -helix in the C-terminal region of ZnF3/4 was unusually unstable, thus

residues 245 to 253 were homology modeled using ZnF1/2 as a model template in order to extend the C-terminal region. A ZnF3/4-r(CG CUGUG) system was retrieved from the PDB (id 3d2s)¹¹ and prepared as follows: the RNA model was reduced to a 4-nt long sequence and mutated to UGCU (corresponding to a YGCY motif present in CUG repeats). This system model corresponds to the binding of RNA to ZnF3. Hence ZnF1, ZnF2 and ZnF3 bound to the RNA fragment were prepared by sequentially aligning the binding regions with the ZnF3 template model. Short MD were run in order to guarantee the stability of the modeled systems. Each system was neutralized with Cl⁻ ions and solvated with TIP3P water molecules⁴⁴ in a 12 Å truncated octahedral box. The Amber ff13 force field was used in all the simulations.⁴⁵

MOLECULAR DYNAMICS OF MBNL1

Each system was minimized using a conventional two-step process. First, all residues except solvent were held fixed with a restraint force of 100 kcal·mol⁻¹·Å⁻² and minimized with 2,500 steepest descent steps followed by 2,500 conjugate gradient steps. A second minimization step was performed without positional restraints using 10,000 steepest descent and 10,000 conjugate gradient steps. Each system was slowly heated to 300 K in 150 ps while constraining the solute with a force gradient of 8.0 to 0 kcal·mol⁻¹·Å⁻². Langevin dynamics with a collision frequency of 1 ps⁻¹ was used. A 20 ps of pressure equilibration step was then applied with isotropic scaling at 1 atm. Production runs were carried out at constant volume and chemical bonds involving hydrogen atoms were constrained with SHAKE algorithm,⁴⁶ which allowed an integration step of 2 fs in the production runs. Particle Mesh Ewald (PME)⁴⁷ was used in all calculations with a 9 Å long-range cutoff.

STEERED MOLECULAR DYNAMICS OF ZNFs-RNA

To describe the binding/unbinding MBNL1-RNA process, we performed four independent SMD experiments by pulling away the center of mass (COM) of the GC binding pair from the COM of each binding pocket (see input example and COM definition in CD-ROM). SMD simulations were performed with the same protocol described for the equilibration step, and initiated with the final structure obtained from the equilibration step. The pulling force was applied to the COMs defined in the Supporting Information from ~6.5 Å until a separation of 20 Å was achieved. A total of 20 successive SMD pulling experiments per complex were conducted with a constant velocity of 1.35 Å·ns⁻¹ and a spring constant of 55.6 nN·Å⁻¹. The total cumulative time for the SMD experiments was 800 ns.

STRUCTURAL ANALYSES OF MBNL1 DOMAINS

Anisotropic network models (ANM) and principal component analysis (PCA) were used to study the dynamic properties of experimental structures. ANM represent the system model by a set of nodes connected by springs defined by a harmonic potential. PCA was completed by decomposing the covariance matrix as previously described. Likewise, PCs from MD trajectories were extracted using essential dynamics analysis (EDA). EDA modes were obtained by decomposing the covariance matrix for 10,000 equally distributed snapshots extracted from each simulation. Local and global fluctuations were computed from the PC1 extracted from either

PCA or EDA. Global fluctuations analyses were based only on the C α atoms of the model system. Overlap between ANM and PCA modes was calculated by the dot product of the corresponding eigenvectors. All ANM, PCA and EDA analyses were completed with ProDy software.⁴⁸ The cpptraj module of Amber 14⁴⁵ was used for root-mean-square deviation (RMSD) and process the MD trajectories. The output was analyzed with VMD.

SIMULATIONS OF D-HEXAPEPTIDES AND MODIFICATIONS

Peptides were capped with ACE and NME blocking groups and chirality was modified using the tleap module from AMBER.⁴⁵ Peptide modifications were prepared with *Antechamber* by parametrizing the peptoid and linker building blocks. AM1-bcc partial charges⁴⁹ were assigned to each modification. MD simulations were conducted in implicit solvent with an infinite cutoff and the salt concentration was set to 0.300 M. Each system was slowly heated to 300 K in 1 ns while constraining the backbone with a force gradient of 8.0 to 0 kcal·mol⁻¹·Å⁻². Production runs were completed using Langevin dynamics with a collision frequency of 1 ps⁻¹ during 1 μ s. Structural analyses were performed with cpptraj using the 'secstruct' and 'multidihedral' functions.⁴⁵

MOLECULAR DYNAMICS OF PAAs

AutoDock Vina was used for generating 40 initial complex conformations for each PAA using PDB ID 2kx5⁴⁰ as the receptor structure. Starting from the best scored docking poses each molecular dynamics system was prepared using explicit solvent for its equilibration and free energy evaluation. AM1-bcc partial charges were assigned to each PAA using *Antechamber* from the AMBER suite.⁴⁵ Then, the complexes were prepared with tleap using the standard ff10 Amber force field. Each structure was minimized in two steps: first, all residues except solvent were held fixed with a restraint force of 100 kcal·mol⁻¹·Å⁻². Steepest descent minimization followed by conjugate gradient was performed using 2,500 steps in both cases. The same minimization protocol was applied in 10,000 steps without positional restraints. After minimization, the temperature was raised from 0 to 300 K in 100 ps using constant volume dynamics. A restraint force of 10 kcal·mol⁻¹·Å⁻² was applied to the complex and SHAKE was turned on for bonds involving hydrogen atoms. Then, 200 ps of constant pressure dynamics were applied for density equilibration. Finally, each production run was performed using Langevin dynamics with a collision frequency of 1 ps⁻¹. An atom-based long range cutoff of 9 Å was applied during all the simulations.

MM/GBSA THERMODYNAMIC ANALYSIS

In order to quantify the binding of each PAA with TAR, relative binding free energies were computed for each binding mode with MM/GBSA approach.⁵⁰ Explicit waters were included as part of the RNA structure in the calculations using the closest command in cpptraj from the AMBER suite. Specifically, water molecules whose oxygen atom were within 3.5 Å of an RNA heavy atom and a PAA heavy atom were selected. The interaction cutoff distance was 999 Å and the salt concentration was set to 0.300 M. MM/GBSA energies were calculated for each of the last 2 ns of the complexes, and then averaged to obtain the overall calculated binding energy.

5.6. REFERENCES

1. Garcia-Lopez, A., Llamusi, B., Orzaez, M., Perez-Paya, E. & Artero, R. D. In vivo discovery of a peptide that prevents CUG-RNA hairpin formation and reverses RNA toxicity in myotonic dystrophy models. *Proc. Natl. Acad. Sci.* **108**, 11866–11871 (2011).
2. Pascale, L. *et al.* Deciphering structure-activity relationships in a series of Tat/TAR inhibitors. *J. Biomol. Struct. Dyn.* **1102**, 1–54 (2015).
3. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* **6**, 386–398 (2005).
4. Black, D. L. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annu. Rev. Biochem.* **72**, 291–336 (2003).
5. Romero, P. R. *et al.* Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 8390–8395 (2006).
6. Zheng, S., Chen, Y., Donahue, C. P., Wolfe, M. S. & Varani, G. Structural Basis for Stabilization of the Tau Pre-mRNA Splicing Regulatory Element by Novantrone (Mitoxantrone). *Chem. Biol.* **16**, 557–566 (2009).
7. Barbany, M. *et al.* Characterization of the impact of alternative splicing on protein dynamics: The cases of glutathione S-transferase and ectodysplasin-A isoforms. *Proteins Struct. Funct. Bioinforma.* **80**, 2235–2249 (2012).
8. Malhotra, S. & Sowdhamini, R. Sequence search and analysis of gene products containing RNA recognition motifs in the human genome. *BMC Bioinformatics* **15**, 1–11 (2014).
9. Barbany, M. *et al.* Molecular Dynamics Study of Naturally Existing Cavity Couplings in Proteins. *PLoS One* **10**, e0119978 (2015).
10. Warf, M. B. & Berglund, J. A. MBNL binds similar RNA structures in the CUG repeats of myotonic dystrophy and its pre-mRNA substrate cardiac troponin T. *RNA* **13**, 2238–2251 (2007).
11. Teplova, M. & Patel, D. J. Structural insights into RNA recognition by the alternative-splicing regulator muscleblind-like MBNL1. *Nat. Struct. & Mol. Biol.* **15**, 1343–1351 (2008).
12. Edge, C., Gooding, C. & Smith, C. W. Dissecting domains necessary for activation and repression of splicing by muscleblind-like protein 1. *BMC Mol. Biol.* **14**, 29 (2013).
13. Klinck, R. *et al.* RBFOX1 Cooperates with MBNL1 to Control Splicing in Muscle, Including Events Altered in Myotonic Dystrophy Type 1. *PLoS One* **9**, e107324 (2014).
14. Wong, C. H. *et al.* Investigating the Binding Mode of an Inhibitor of the MBNL1-RNA Complex in Myotonic Dystrophy Type 1 (DM1) Leads to the Unexpected Discovery of a DNA-Selective Binder. *ChemBioChem* **13**, 2505–2509 (2012).

15. Goers, E. S., Purcell, J., Voelker, R. B., Gates, D. P. & Berglund, J. A. MBNL₁ binds GC motifs embedded in pyrimidines to regulate alternative splicing. *Nucleic Acids Res.* **38**, 2467–2484 (2010).
16. Laurent, F. X. *et al.* New function for the RNA helicase p68/DDX5 as a modifier of MBNL₁ activity on expanded CUG repeats. *Nucleic Acids Res.* **40**, 3159–3171 (2012).
17. Warf, M. B., Diegel, J. V., von Hippel, P. H. & Berglund, J. A. The protein factors MBNL₁ and U2AF65 bind alternative RNA structures to regulate splicing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9203–9208 (2009).
18. Yuan, Y. *et al.* Muscleblind-like 1 interacts with RNA hairpins in splicing target and pathogenic RNAs. *Nucleic Acids Res.* **35**, 5474–5486 (2007).
19. Childs-Disney, J. L., Hoskins, J., Rzuczek, S. G., Thornton, C. a. & Disney, M. D. Rationally designed small molecules targeting the RNA that causes myotonic dystrophy type 1 are potently bioactive. *ACS Chem. Biol.* **7**, 856–862 (2012).
20. Philips, A. V., Timchenko, L. T. & Cooper, T. a. Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. *Science* **280**, 737–741 (1998).
21. Michalowski, S. *et al.* Visualization of double-stranded RNAs from the myotonic dystrophy protein kinase gene and interactions with CUG-binding protein. *Nucleic Acids Res.* **27**, 3534–3542 (1999).
22. Grammatikakis, I., Goo, Y. H., Echeverria, G. V. & Cooper, T. a. Identification of MBNL₁ and MBNL₃ domains required for splicing activation and repression. *Nucleic Acids Res.* **39**, 2769–2780 (2011).
23. Tran, H. *et al.* Analysis of exonic regions involved in nuclear localization, splicing activity, and dimerization of muscleblind-like-1 isoforms. *J. Biol. Chem.* **286**, 16435–16446 (2011).
24. Purcell, J., Oddo, J. C., Wang, E. T. & Berglund, J. a. Combinatorial Mutagenesis of MBNL₁ Zinc Fingers Elucidates Distinct Classes of Regulatory Events. *Mol. Cell. Biol.* **32**, 4155–4167 (2012).
25. Lezon, T. R., Shrivastava, I. H., Yang, Z. & Bahar, I. Elastic Network Models For Biomolecular Dynamics: Theory and Application to Membrane Proteins and Viruses. *Handb. Biol. Networks* 129–158 (2009). at <papers://publication/uuid/9B230FBB-A40F-4630-9421-7FF883030646>
26. Frappier, V. & Najmanovich, R. J. A Coarse-Grained Elastic Network Atom Contact Model and Its Use in the Simulation of Protein Dynamics and the Prediction of the Effect of Mutations. *PLoS Comput. Biol.* **10**, (2014).
27. Das, A. *et al.* Exploring the Conformational Transitions of Biomolecular Systems Using a Simple Two-State Anisotropic Network Model. *PLoS Comput. Biol.* **10**, (2014).
28. Fenwick, R. B., Orellana, L., Esteban-Martín, S., Orozco, M. & Salvatella, X. Correlated motions are a fundamental property of β -sheets. *Nat. Commun.* **5**, 4070 (2014).
29. Bakan, A. & Bahar, I. The intrinsic dynamics of enzymes plays a dominant role in

- determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 14349–14354 (2009).
30. Liu, Y. & Bahar, I. Sequence evolution correlates with structural dynamics. *Mol. Biol. Evol.* **29**, 2253–2263 (2012).
 31. Fu, Y., Ramisetty, S. R., Hussain, N. & Baranger, A. M. MBNL₁-RNA Recognition: Contributions of MBNL₁ Sequence and RNA Conformation. *ChemBioChem* **13**, 112–119 (2012).
 32. Yang, L., Song, G. & Jernigan, R. L. Protein elastic network models and the ranges of cooperativity. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 12347–12352 (2009).
 33. Nevin Gerek, Z. & Banu Ozkan, S. A flexible docking scheme to explore the binding selectivity of PDZ domains. *Protein Sci.* **19**, 914–928 (2010).
 34. Alaybeyoglu, B., Akbulut, S. & Ozkirimli, E. A novel chimeric peptide with antimicrobial activity. *J. Pept. Sci.* (2015). doi:10.1002/psc.2739
 35. Liang, G., Yang, L., Kang, L., Mei, H. & Li, Z. Using multidimensional patterns of amino acid attributes for QSAR analysis of peptides. *Amino Acids* **37**, 583–591 (2009).
 36. Ho, B. K. & Dill, K. A. Folding Very Short Peptides Using Molecular Dynamics. *PLoS Comput. Biol.* **2**, e27 (2006).
 37. Mu, Y. & Stock, G. Conformational Dynamics of RNA-Peptide Binding: A Molecular Dynamics Simulation Study. *Biophys. J.* **90**, 391–399 (2006).
 38. Guardiani, C., Signorini, G. F., Livi, R., Papini, A. M. & Procacci, P. Conformational landscape of N-glycosylated peptides detecting autoantibodies in multiple sclerosis, revealed by hamiltonian replica exchange. *J. Phys. Chem. B* **116**, 5458–5467 (2012).
 39. Athanassiou, Z. *et al.* Structural Mimicry of Retroviral Tat Proteins by Constrained β -Hairpin Peptidomimetics: Ligands with High Affinity and Selectivity for Viral TAR RNA Regulatory Elements. *J. Am. Chem. Soc.* **126**, 6906–6913 (2004).
 40. Davidson, A., Patora-Komisarska, K., Robinson, J. a. & Varani, G. Essential structural requirements for specific recognition of HIV TAR RNA by peptide mimetics of Tat protein. *Nucleic Acids Res.* **39**, 248–256 (2011).
 41. Hou, T., Wang, J., Li, Y. & Wang, W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* **51**, 69–82 (2011).
 42. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
 43. Simonetti, F. L., Teppa, E., Chernomoretz, A., Nielsen, M. & Marino Buslje, C. MISTIC: Mutual information server to infer coevolution. *Nucleic Acids Res.* **41**, 8–14 (2013).
 44. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926 (1983).

45. D.A. Case, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B., X. Wu and, V. B. & D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B., X. W. and P. A. K. AMBER 14. (2014).
46. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. . C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
47. Toukmaji, A., Sagui, C., Board, J. & Darden, T. Efficient particle-mesh Ewald based approach to fixed and induced dipolar interactions. *J. Chem. Phys.* **113**, 10913–10927 (2000).
48. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).
49. Jakalian, A., Bush, B. L., Jack, D. B. & Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **21**, 132–146 (2000).
50. Rastelli, G., Del Rio, A., Degliesposti, G. & Sgobba, M. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *J. Comput. Chem.* **31**, 797–810 (2010).

CONCLUSIONS

1. The dynamic behavior of trinucleotide (rCUG) and pentanucleotide-repeat RNA overexpansions (rAUUCU), involved in myotonic dystrophy type 1 and spinocerebellar ataxia type 10 respectively, has been investigated using molecular dynamics. It has been observed that the conformational sampling greatly depends on the structural context such as the number of repeats or the nucleotides surrounding the mismatches. Such features confer to this type of RNA structures unique druggable sites that can be exploited for structure-based drug design.
2. Elastic Network Models, and in particular Anisotropic Network Models, are able to describe the global dynamics of the RNA by using both backbone and nucleotide atomic information. These methods can provide a fast and reliable platform for generating new RNA conformations surrounding the local minima of experimentally resolved structures. However, the oversimplification of the model limits its reliability to inform about other structural minima of the conformational space, and the lack of structural models limits its verification.
3. A novel small molecule – nucleic acids scoring function has been developed using the open source docking software Vina. This scoring function, ANNScore, relies on an artificial neural networks methodology which improves the prediction of hit scoring poses if compared to the original Vina's scoring function. This method outperformed some small molecule – RNA scoring functions such as DOCK6, Glide, LigandRNA and Vina.
4. *De novo* and ligand-based drug design strategies allowed the development and identification of three novel molecular frameworks (named as TFL, MGL and LR families) for the treatment of myotonic dystrophy type 1. Some of these compounds retrieved an acceptable level of rCUG binding and improved MBNL1 liberation and foci reduction.
5. A quantitative structure-activity relationship model was developed based on previously reported data of compounds for the treatment of myotonic dystrophy type 1. This model relies on artificial neural algorithms for the classification of molecules between potentially active or inactive by averting the rCUG-MBNL1 interaction.
6. The MBNL1 protein's dynamics and cognate RNA binding have been investigated using conventional and steered molecular dynamics. In agreement with experimental data, MBNL1 dynamics of the four homologous binding domains are significantly different, which explains its differentiated RNA binding and domain fluctuations. Moreover, sequential steered molecular dynamics revealed that local fluctuations affect their RNA binding properties.
7. Molecular dynamics studies have been conducted for the study of RNA-binding peptides. The conformational sampling of 16 *D*-hexapeptides previously reported for the treatment of myotonic dystrophy type 1 have shown to modestly correlate with their ability to produce a polyprolyne type II conformation. This observation is in agreement with circular dichroism data and prospective peptide's design will be performed based on this observation. Following a similar strategy, studies of the interaction of 4 polyamide polyamide amino acids with the RNA TAR binding domain of HIV-1 rationalized the experimentally observed thermodynamic properties.

ANNEXES

Table A1. Frequency of U-U pairs conformations observed in the r(CUG)₁₆ and r(CUG)₈ trajectory.

r(CUG) ₁₆							
	(CUG) ^a 0HB	Base-pair conformation				Water mediated	
		(CUG) ^b 1HB	(CUG) ^c 1HB	(CUG) ^d 2HB	(CUG) ^e 2HB	(CUG) ^f 1WHB	(CUG) ^g 2WHB
U5-U44	9%	90%	-	-	-	1%	-
U8-U41	-	100%	-	-	-	-	-
U11-U38	42%	25%	-	26%	5%	2%	-
U14-U35	5%	-	8%	-	85%	1%	-
U17-U32	2%	93%	-	5%	-	-	-
U20-U29	1%	-	89%	-	11%	1%	-
Overall	10%	68%		22%		Not significant	
r(CUG) ₈							
	(CUG) ^a 0HB	Base-pair conformation				Water mediated	
		(CUG) ^b 1HB	(CUG) ^c 1HB	(CUG) ^d 2HB	(CUG) ^e 2HB	(CUG) ^f 1WHB	(CUG) ^g 2WHB
U5-U26	35%	41%	18%	6%	-	-	-
U8-U23	34 %	44%	-	18%	2%	2%	-
U11-U20	-	-	83%	8%	9%	-	-
Overall	23%	62%		15%		Not significant	

Table A2. Stiffness constants using Lankas' definition for the three $r(\text{CUG})^{\text{exp}}$ models.

		Shear	Stretch	Stagger	Buckle	Propeller	Opening
$ds[(\text{CUG})_{62}]$	C4-G15	4.46	15.38	5.39	0.012	0.011	0.022
	U5-U14	0.50	5.03	3.57	0.008	0.009	0.005
	G6-C13	7.07	69.02	6.91	0.017	0.014	0.089
	C7-G12	8.25	76.21	7.65	0.016	0.013	0.089
	U8-U11	0.38	2.07	1.06	0.006	0.008	0.008
	G9-C10	7.17	20.30	5.18	0.014	0.013	0.072
	C10-G9	6.72	22.58	5.07	0.014	0.013	0.076
	U11-U8	0.52	1.69	1.22	0.007	0.008	0.005
	G12-C7	8.85	80.05	8.10	0.018	0.015	0.092
	C10-G6	7.74	71.47	7.18	0.017	0.015	0.087
	U11-U5	0.24	2.72	3.18	0.008	0.010	0.004
	G12-C4	7.33	61.69	6.25	0.012	0.011	0.067
$r(\text{CUG})_{16}$	C7-G42	6.77	69.64	7.20	0.016	0.013	0.076
	U8-U41	2.18	5.80	3.16	0.009	0.009	0.024
	G9-C40	6.82	68.50	8.29	0.019	0.015	0.079
	C10-G39	7.96	74.56	7.90	0.016	0.014	0.087
	U11-U38	0.17	1.71	2.05	0.008	0.008	0.004
	G12-C37	6.91	20.31	5.31	0.015	0.014	0.076
	C13-G36	7.17	21.27	5.49	0.015	0.015	0.078
	U14-U35	0.30	5.06	3.33	0.008	0.008	0.006
	G15-C34	6.20	21.41	5.34	0.015	0.014	0.072
	C16-G33	7.08	21.06	4.79	0.013	0.014	0.076
	U17-U32	0.82	2.46	2.72	0.007	0.010	0.008
	G18-C31	8.65	74.15	6.30	0.013	0.010	0.087
$r(\text{CUG})_8$	C4-G27	7.87	73.93	6.56	0.012	0.012	0.086
	U5-U26	0.16	1.69	2.03	0.008	0.008	0.003
	G6-C25	6.88	20.24	5.28	0.014	0.014	0.075
	C7-G24	7.12	21.09	5.46	0.015	0.014	0.077
	U8-U23	0.30	5.02	3.30	0.008	0.008	0.006
	G9-C22	6.16	21.22	5.29	0.015	0.014	0.071
	C10-G21	7.01	20.85	4.77	0.013	0.014	0.076
	U11-U20	0.81	2.44	2.70	0.007	0.010	0.008
	G12-C19	8.60	73.80	6.27	0.013	0.010	0.087

Table A3. Base pair parameters calculated with 3DNA, C1'-C1' and hydrogen bonding distance and inclination of U-U pairs.

Structure	Pair	C1'-C1' (Å)	1st HB (Å)	2nd HB (Å)	λ I (°)	λ II (°)	Incline	Shear (Å)	Stretch (Å)	Stagger (Å)	Buckle (°)	Propeller (°)	Opening (°)
CUG centroids	U3-U14	10,2	2.9	-	67,5	31,4	Major	3.10	-1.53	-0.05	-2.92	-10.39	-11.58
	U3-U14	10,9	2.7	-	25,2	57,4	Minor	-2.98	-1.14	0.18	3.87	-5.26	-27.76
	U6-U11	10,9	2.8	-	20,4	62,6	Minor	3.37	-1.72	-0.05	-1.68	-1.60	-15
	U6-U11	10,1	2.8	-	70,6	30	Major	-3.33	-1.36	-0.67	1.35	-13.86	-13.93
Disney NMR	U5-U14 (0)	10.7	-	-	71.2	32	Major	3.57	-0.95	-0.05	13.48	-9.09	-5.62
	U5-U14 (1)	10.6	2.9	-	61.1	28.9	Major	3.03	-1.14	-0.09	-22.07	-6.46	-24.3
	U5-U14 (2)	8.9	2.9	2.9	73.3	48.3	Major	2.24	-1.74	0.02	12.56	-17.25	4.58
Kiliszek et al.	U3-U6 (A-B)	10.7	2.8	-	26.2	57.9	Minor	-2.82	-1.19	-0.47	12.51	-8.17	-31.5
	U6-U3 (A-B)	10.3	2.9	-	31.6	67.7	Minor	-3.03	-1.31	-0.13	14.01	-8.36	-16.37
	U3-U6 (C-D)	10.9	2.8	-	56.1	22.6	Major	2.98	-1.22	-0.44	-4.51	-8.69	-34.42
	U6-U3 (C-D)	10.2	2.6	-	59.2	32.4	Major	2.59	-1.51	-0.28	0.57	-15.81	-21.16
	U3-U6 (E-E*)	10.6	2.6	-	52.7	29.1	Major	2.38	-1.31	-0.48	-4.07	-11.84	-30.56
	U6-U3 (E-E*)	10.6	2.6	-	29.1	52.7	Minor	-2.38	-1.31	-0.48	4.07	-11.84	-30.56

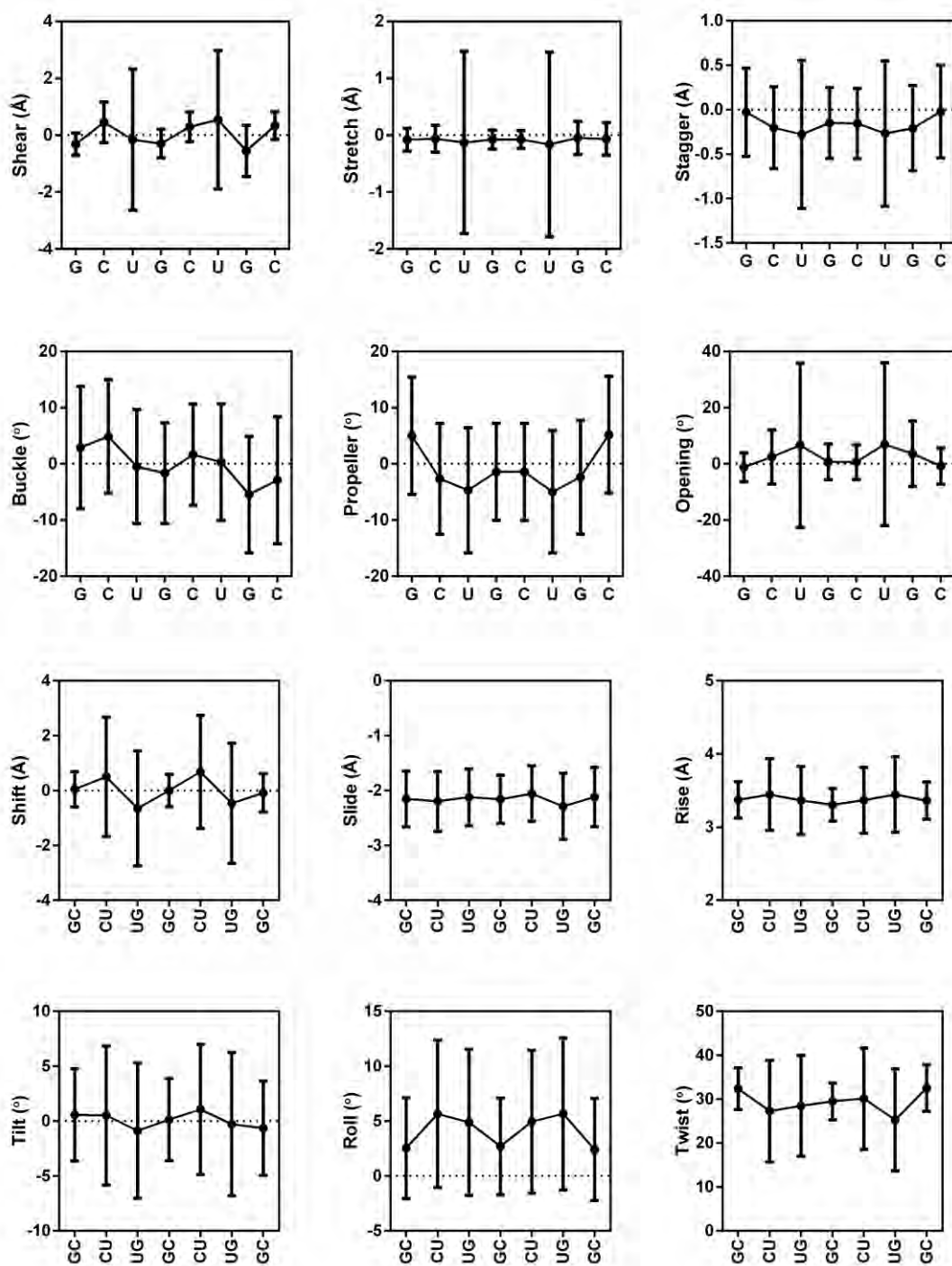


Figure A1. Base pair (shear, stretch, stagger, buckle, propeller and opening) and base step parameters (shift, slide, rise, tilt, roll and twist) extracted from the cMD simulation. Average and standard deviation are indicated.

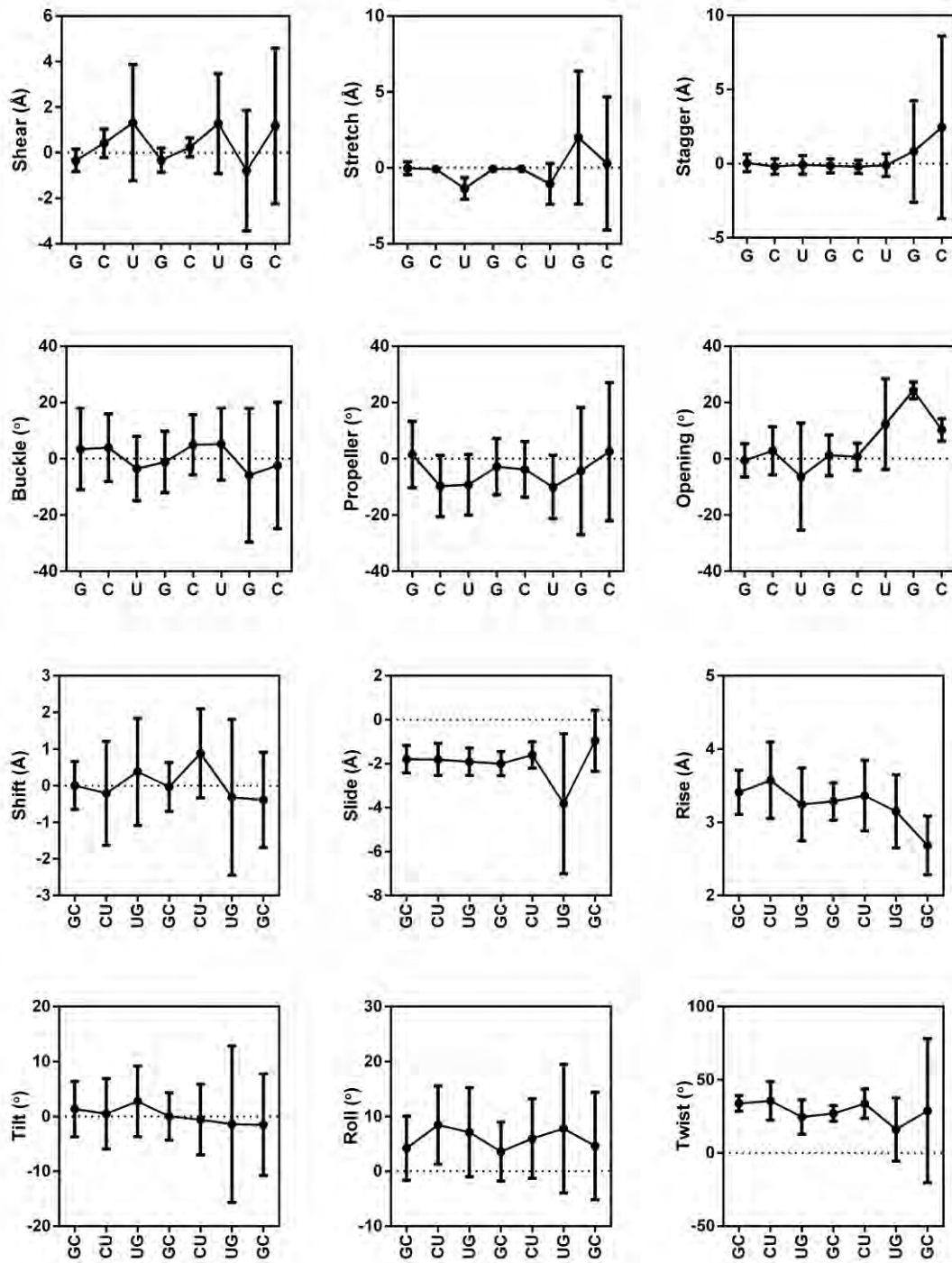


Figure A2. Base pair (shear, stretch, stagger, buckle, propeller and opening) and base step parameters (shift, slide, rise, tilt, roll and twist) extracted from the aMD simulation. Average and standard deviation are indicated.

Table A4. Helical parameters for different base pair steps averaged over the rAUUCU MD trajectory.

Local base pair step parameters							Local base pair helical parameters					
Average												
Step	Shift (Å)	Slide (Å)	Rise (Å)	Tilt (°)	Roll (°)	Twist (°)	X-disp (Å)	Y-disp (Å)	h-Rise (Å)	Incl. (°)	Tip (°)	h-Twist (°)
UA/UU	0.05	-1.83	3.22	-0.6	5.54	29.02	-4.65	-0.29	2.69	10.78	1.51	30.59
AU/AU	0.40	-2.08	3.26	0.93	3.41	26.64	-5.17	-0.62	2.88	6.95	-2.01	27.88
UU/UA	0.46	-1.59	3.65	-1.15	10.52	38.55	-3.7	-0.88	3.03	15.38	1.84	40.81
UC/CU	1.00	-2.12	3.14	0.42	7.4	25.38	-4.67	-2.82	2.1	15.12	-5.35	27.18
CU/UC	-1.09	-2.15	2.97	2.34	7.21	22.51	-3.3	3.54	2.05	12.71	6.74	23.3
UA/UU	-0.33	-1.65	3.49	2.24	10.4	33.92	-4.17	0.86	2.71	16.9	-4.99	36.75
AU/AU	-0.45	-2.00	3.25	-0.57	4.39	26.8	-5.18	0.82	2.78	8.93	1.41	28.22
Standard deviation												
UA/UU	0.72	0.85	0.35	4.7	6.34	5	2.47	2.02	0.63	12.75	9.94	4.74
AU/AU	0.94	0.96	0.25	4.63	5.91	5.13	2.05	2.53	0.58	12.47	10.13	5.12
UU/UA	0.89	1.01	0.5	5.62	6.36	6.8	2.86	2.26	0.67	9.7	9.55	6.62
UC/CU	1.77	2.06	0.71	8.11	6.49	22.06	9	6.94	1.68	22.97	24.14	23.45
CU/UC	1.68	2.12	0.69	8.06	6.73	22.8	9.62	5.5	1.55	23.97	25.85	25.44
UA/UU	1.09	1.26	0.58	6.38	6.7	10.52	3.64	3.5	0.99	13.77	13.49	10.16
AU/AU	1.03	1.14	0.27	5	5.99	6.62	2.47	2.84	0.65	12.84	11.36	6.69

Table A5. Small molecule – nucleic acids used as training and test sets for ANNScore. Original parameter values retrieved from Vina are included.

	PDB	GAUSS1	GAUSS2	REPULSION	HYDROPH	HBOND	$\Delta G(\text{exp})$
TRAINING SET	2GXV	289,911	157,674	1856,358	7,425	4,503	11,9
	2IRL	313,135	127,241	1870,004	3,62	1,984	10,8
	2GYH	419,486	157,417	1841,974	3,567	6,169	12,3
	2GXP	390,687	128,887	1316,202	5,32	6,831	9,1
	1RAW	150,859	1585,564	40,233	0	9,365	6,8
	2GYE	298,773	154,179	1907,798	4,335	8,501	11,9
	2GXY	224,628	110,602	1682,003	3,841	2,466	10,7
	109D	554,973	166,911	1895,528	4,559	8,687	12
	1TOB	83,41	1636,333	6,256	1,182	6,724	12,48
	264D	449,945	154,048	1997,172	4,413	5,793	11,7
	2GY6	453,991	158,718	1768,199	4,645	8,342	11,9
	2IRJ	211,117	72,642	1153,693	1,308	3,189	8,6
	2GXK	284,934	101,146	1261,615	2,479	3,894	9
	2GYM	119,817	92,749	2401,125	2,563	2,755	9,4
	1BYJ	82,816	1498,274	1,447	0,496	2,835	10,98
	1PRP	144,308	89,964	1526,934	2,954	1,502	8,2
	2GYJ	596,606	220,976	2325,534	7,21	10,889	13
	1AM0	146,851	1335,056	11,945	0	8,305	6,81
	2GXJ	344,866	91,883	1189,339	2,923	3,504	9,1
	1JTL	198,028	175,944	2250,34	5,234	3,285	11,5
	261D	195,815	176,366	2011,629	6,585	2,809	11,5
	2GYF	494,758	169,018	1933,092	4,185	7,732	12,1
	2GXI	133,04	98,899	1580,892	3,202	2,618	8,5
	2GY1	308,681	126,677	1832,234	4,539	3,403	9,2
	2GXN	421,113	89,103	1247,768	1,757	3,32	9,2
	2GXR	391,901	128,933	1536,263	2,634	3,643	10,7
	1PBR	106,496	2069,545	1,943	0,585	5,508	9,19
	2GY8	331,551	146,443	1787,99	3,619	6,941	11,9
	1NEM	143,828	2198,765	10,411	0,193	13,125	9,6
	2DBE	190,075	86,759	1329,587	1,978	1,961	8,6
	2TOB	138,163	1861,687	15,863	1,343	14,506	12,24
	227D	228,396	94,425	1489,96	2,466	1,125	9,3
	1LVJ	175,109	1433,718	44,474	8,067	0	9,56
1FMN	187,031	1616,073	5,22	0	2,917	7,2	
TEST SET	2IRK	224,587	99,933	1445,225	4,121	1,77	8,1
	2GY2	482,032	187,29	2868,313	6,384	12,125	12,7
	2GXO	351,388	109,357	1381,896	2,485	4,471	9,1
	2GXM	469,583	75,251	1166,293	0,92	1,53	8,8
	1EHT	128,004	1001,159	9,466	0	2,14	8,72
	2GYL	469,23	212,016	2282,302	6,752	6,867	13,1
	2GYG	522,495	176,078	2045,72	4,745	8,927	12,3
	1QD3	139,268	2252,168	19,215	0	13,481	8
	1KOC	64,203	841,962	6,49	0	1,392	6,8
	2GXT	429,314	137,751	1688,446	4,201	3,54	10,9
	2GY0	179,217	109,835	1598,597	3,991	0,588	9
	2GY4	273,404	128,261	2463,972	2,701	5,501	11,6
	1D63	195,001	107,809	1345,057	2,55	3,83	8
	1KOD	90,886	875,694	335,756	4	12,52	6,8
	1E12	99,546	1693,592	7,938	1,216	6,886	8

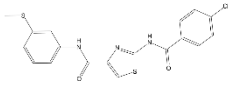
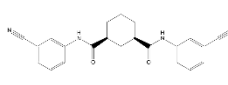
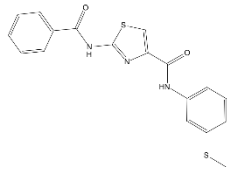
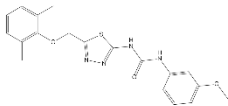
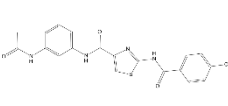
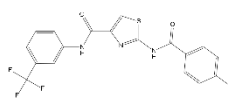
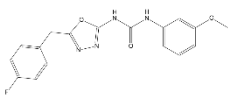
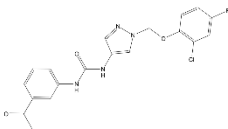
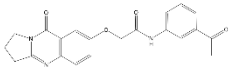
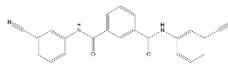
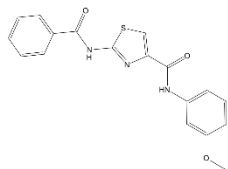
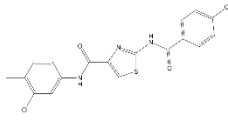
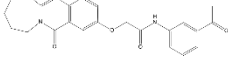
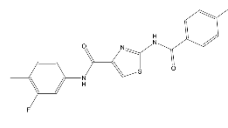
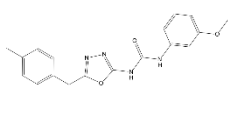
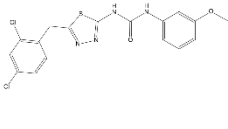
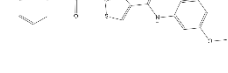
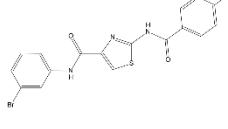
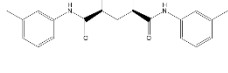
	<p>ID: IQS-DM1-MGL01</p>  <p>SlogP: 5.0230 sol (mM): 1.2250</p>	<p>ID: IQS-DM1-MGL02</p>  <p>SlogP: 3.8136 sol (mM): 6.2976</p>	<p>ID: IQS-DM1-MGL03</p>  <p>SlogP: 4.3696 sol (mM): 2.5528</p>
<p>ID: IQS-DM1-MGL04</p>  <p>SlogP: 4.6529 sol (mM): 5.4525</p>	<p>ID: IQS-DM1-MGL05</p>  <p>SlogP: 4.2595 sol (mM): 2.7590</p>	<p>ID: IQS-DM1-MGL06</p>  <p>SlogP: 5.6314 sol (mM): 1.1827</p>	<p>ID: IQS-DM1-MGL07</p>  <p>SlogP: 3.4521 sol (mM): 6.0044</p>
<p>ID: IQS-DM1-MGL08</p>  <p>SlogP: 4.8251 sol (mM): 11.4032</p>	<p>ID: IQS-DM1-MGL09</p>  <p>SlogP: 3.1863 sol (mM): 9.3062</p>	<p>ID: IQS-DM1-MGL10</p>  <p>SlogP: 3.9346 sol (mM): 2.4134</p>	<p>ID: IQS-DM1-MGL11</p>  <p>SlogP: 3.6563 sol (mM): 6.7411</p>
<p>ID: IQS-DM1-MGL12</p>  <p>SlogP: 5.2629 sol (mM): 1.3903</p>	<p>ID: IQS-DM1-MGL13</p>  <p>SlogP: 3.9665 sol (mM): 6.2161</p>	<p>ID: IQS-DM1-MGL14</p>  <p>SlogP: 4.7486 sol (mM): 2.1573</p>	<p>ID: IQS-DM1-MGL15</p>  <p>SlogP: 3.9298 sol (mM): 3.1256</p>
<p>ID: IQS-DM1-MGL16</p>  <p>SlogP: 5.0883 sol (mM): 1.7573</p>	<p>ID: IQS-DM1-MGL17</p>  <p>SlogP: 4.3564 sol (mM): 4.6512</p>	<p>ID: IQS-DM1-MGL18</p>  <p>SlogP: 5.0636 sol (mM): 1.1433</p>	<p>ID: IQS-DM1-MGL19</p>  <p>SlogP: 4.6870 sol (mM): 4.9243</p>

Figure A3. List of MGL compounds ordered from highest to lowest TanimotoCombo score. SlogP and solubility in water are included.

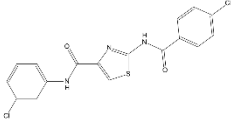
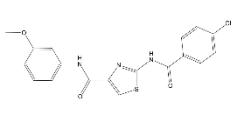
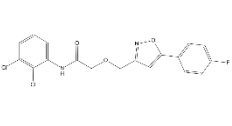
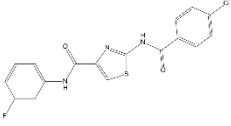
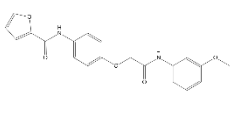
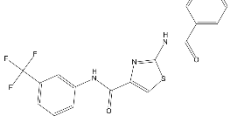
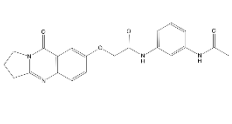
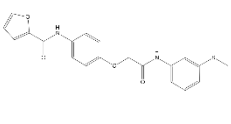
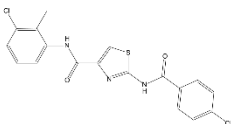
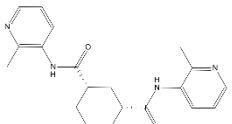
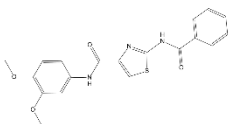
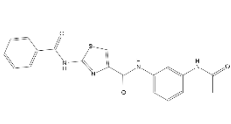
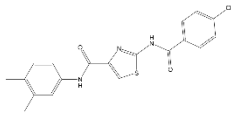
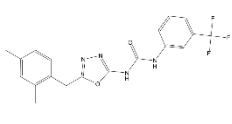
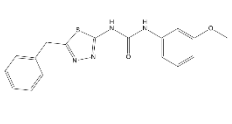
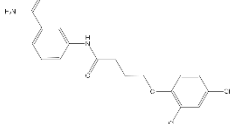
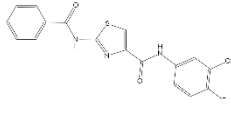
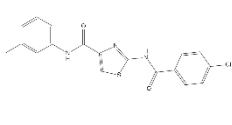
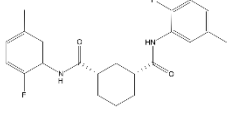
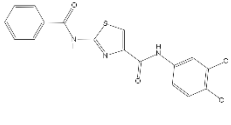
<p>ID: IQS-DM1-MGL20</p>  <p>SlogP: 4.9545 sol (mM): 1.6324</p>	<p>ID: IQS-DM1-MGL21</p>  <p>SlogP: 4.3097 sol (mM): 3.2347</p>	<p>ID: IQS-DM1-MGL22</p>  <p>SlogP: 5.2092 sol (mM): 1.5718</p>	<p>ID: IQS-DM1-MGL23</p>  <p>SlogP: 4.4402 sol (mM): 2.5328</p>
<p>ID: IQS-DM1-MGL24</p>  <p>SlogP: 3.5580 sol (mM): 5.4966</p>	<p>ID: IQS-DM1-MGL25</p>  <p>SlogP: 4.9780 sol (mM): 2.4647</p>	<p>ID: IQS-DM1-MGL26</p>  <p>SlogP: 2.9421 sol (mM): 10.3140</p>	<p>ID: IQS-DM1-MGL27</p>  <p>SlogP: 4.2713 sol (mM): 2.0815</p>
<p>ID: IQS-DM1-MGL28</p>  <p>SlogP: 5.2629 sol (mM): 1.3903</p>	<p>ID: IQS-DM1-MGL29</p>  <p>SlogP: 3.4770 sol (mM): 84.0595</p>	<p>ID: IQS-DM1-MGL30</p>  <p>SlogP: 3.6649 sol (mM): 6.4099</p>	<p>ID: IQS-DM1-MGL31</p>  <p>SlogP: 3.6061 sol (mM): 5.7497</p>
<p>ID: IQS-DM1-MGL32</p>  <p>SlogP: 4.9179 sol (mM): 1.3185</p>	<p>ID: IQS-DM1-MGL33</p>  <p>SlogP: 5.2515 sol (mM): 1.1428</p>	<p>ID: IQS-DM1-MGL34</p>  <p>SlogP: 3.7815 sol (mM): 7.6321</p>	<p>ID: IQS-DM1-MGL35</p>  <p>SlogP: 3.8900 sol (mM): 5.3211</p>
<p>ID: IQS-DM1-MGL36</p>  <p>SlogP: 4.4402 sol (mM): 2.5328</p>	<p>ID: IQS-DM1-MGL37</p>  <p>SlogP: 4.6095 sol (mM): 2.1178</p>	<p>ID: IQS-DM1-MGL38</p>  <p>SlogP: 4.9652 sol (mM): 2.7298</p>	<p>ID: IQS-DM1-MGL39</p>  <p>SlogP: 4.9545 sol (mM): 1.6324</p>

Figure A3. (continuation)

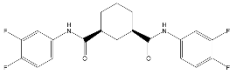
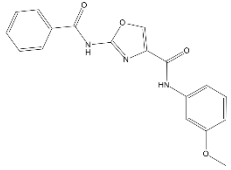
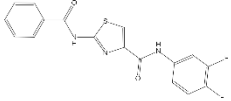
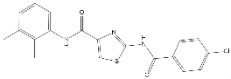
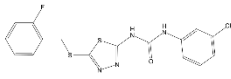
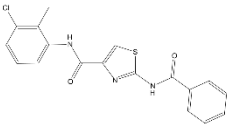
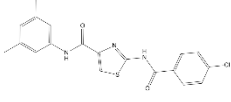
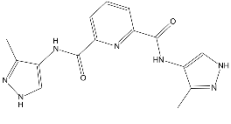
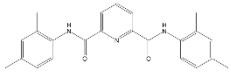
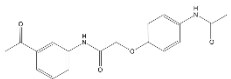
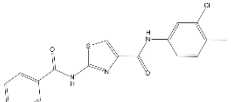
<p>ID: IQS-DM1-MGL40</p>  <p>SlogP: 4.6266 sol (mM): 3.9044</p>	<p>ID: IQS-DM1-MGL41</p>  <p>SlogP: 3.1878 sol (mM): 7.1230</p>	<p>ID: IQS-DM1-MGL42</p>  <p>SlogP: 3.9259 sol (mM): 3.9300</p>	<p>ID: IQS-DM1-MGL43</p>  <p>SlogP: 4.9179 sol (mM): 1.8039</p>
<p>ID: IQS-DM1-MGL44</p>  <p>SlogP: 5.5333 sol (mM): 0.4952</p>	<p>ID: IQS-DM1-MGL45</p>  <p>SlogP: 4.6095 sol (mM): 2.8975</p>	<p>ID: IQS-DM1-MGL46</p>  <p>SlogP: 4.9179 sol (mM): 1.3185</p>	<p>ID: IQS-DM1-MGL47</p>  <p>SlogP: 1.6492 sol (mM): 153.9050</p>
<p>ID: IQS-DM1-MGL48</p>  <p>SlogP: 4.8199 sol (mM): 3.5481</p>	<p>ID: IQS-DM1-MGL49</p>  <p>SlogP: 2.8651 sol (mM): 19.1892</p>	<p>ID: IQS-DM1-MGL50</p>  <p>SlogP: 4.6095 sol (mM): 2.8975</p>	

Figure A3. (continuation)

PUBLICATIONS

Crystallographic and Computational Analyses of AUUCU Repeating RNA That Causes Spinocerebellar Ataxia Type 10 (SCA10)

HaJeung Park,^{*,†} Àlex L. González,^{||} Ilyas Yildirim,[⊥] Tuan Tran,[§] Jeremy R. Lohman,[§] Pengfei Fang,[‡] Min Guo,[‡] and Matthew D. Disney^{*,§}

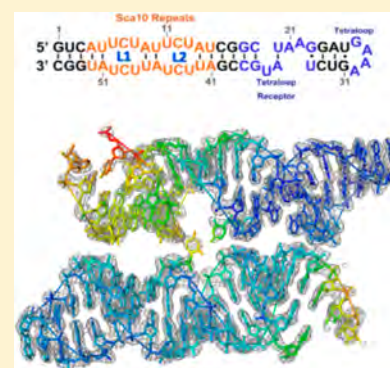
[†]Translational Research Institute, [‡]The Department of Cancer Biology, and [§]The Department of Chemistry, The Scripps Research Institute, Scripps Florida, 130 Scripps Way, Jupiter, Florida 33458, United States

^{||}Grup d'Enginyeria Molecular (GEM), Institut Químic de Sarrià (IQS)-Universitat Ramon Llull (URL), Barcelona 08017, Spain

[⊥]Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Supporting Information

ABSTRACT: Spinocerebellar ataxia type 10 (SCA10) is caused by a pentanucleotide repeat expansion of r(AUUCU) within intron 9 of the *ATXN10* pre-mRNA. The RNA causes disease by a gain-of-function mechanism in which it inactivates proteins involved in RNA biogenesis. Spectroscopic studies showed that r(AUUCU) repeats form a hairpin structure; however, there were no high-resolution structural models prior to this work. Herein, we report the first crystal structure of model r(AUUCU) repeats refined to 2.8 Å and analysis of the structure via molecular dynamics simulations. The r(AUUCU) tracts adopt an overall A-form geometry in which 3 × 3 nucleotide 5'UCU^{3'}/3'UCU^{5'} internal loops are closed by AU pairs. Helical parameters of the refined structure as well as the corresponding electron density map on the crystallographic model reflect dynamic features of the internal loop. The computational analyses captured dynamic motion of the loop closing pairs, which can form single-stranded conformations with relatively low energies. Overall, the results presented here suggest the possibility for r(AUUCU) repeats to form metastable A-form structures, which can rearrange into single-stranded conformations and attract proteins such as heterogeneous nuclear ribonucleoprotein K (hnRNP K). The information presented here may aid in the rational design of therapeutics targeting this RNA.



RNA repeat expansions cause various neuromuscular diseases, including spinocerebellar ataxia type 10 (SCA10), myotonic dystrophy (DM), Huntington's disease (HD), and frontotemporal dementia/amyotrophic lateral sclerosis (FTD/ALS). Repeat modules are generally three to six nucleotides in length.¹ For example, DM1 and HD are caused by triplet repeats (CUG and CAG, respectively) while longer repeats cause SCA10 (AUUCU). Repeat length scales with disease severity. For example, in SCA10, healthy individuals typically have <50 repeats while those afflicted with disease have up to ~5000 repeats.²

Studies have shown that the pathology of repeat expansion disorders is predominantly caused by two modes of RNA toxicity. Repeats bind to and sequester proteins involved in RNA biogenesis, leading to downstream defects in RNA processing, termed RNA gain of function. Also, expanded repeats initiate translation without the use of a start codon. Termed repeat-associated non-ATG (RAN) translation, this mode produces toxic homopolymeric proteins that accumulate as inclusion bodies and induce apoptosis.^{3,4} In SCA10, r(AUUCU)^{exp} sequesters heterogeneous nuclear ribonucleoprotein K (hnRNP K), inducing translocation of protein kinase Cδ to mitochondria and caspase-3-mediated apoptosis of neuronal cells via RNA gain of function.²

Structural studies have been reported for various repeating transcripts^{5–7} and revealed common structural features. For example, they adopt an overall A-form geometry, with variations in base pair and helical parameters. It is possible that repeating RNAs with longer repeat modules (>3) share similar features; however, high-resolution information for these RNAs is scarce. A biophysical study by Handa et al. suggests r(AUUCU)₉ forms a structured A-form helix via circular dichroism (CD) and nuclear magnetic resonance (NMR) analysis.⁸ Their NMR studies revealed evidence of A-U and U-U base pairing, suggesting that r(AUUCU) repeats harbor 3 × 3 nucleotide 5'UCU^{3'}/3'UCU^{5'} internal loops with two U-U noncanonical pairs and one C-C noncanonical pair.⁸

To capture structural characteristics of r(AUUCU) repeats, we have determined a crystal structure of a model RNA containing two copies of 5'AUUCU^{3'}/3'UCUUA^{5'} and thoroughly analyzed the dynamics of this structure with molecular dynamics (MD) simulations. The results indicate r(AUUCU) repeats form a metastable A-form RNA, and the dynamic characteristic is attributed to the internal

Received: May 20, 2015

Revised: June 2, 2015

Published: June 3, 2015

$5'UCU^{3'}/^{3'}UCU^{5'}$ loop pairs. Overall, the results presented here provide structural evidence of the pathogenic mechanism of SCA10 caused by repeat expansion of r(AUUCU). This structure may also provide valuable information to guide the design of therapeutic modalities that target this RNA to ameliorate the disease.

MATERIALS AND METHODS

RNA Synthesis and Purification. A single-stranded DNA template for the AUUCU construct was purchased from Integrated DNA Technologies, Inc. (IDT). A double-stranded template suitable for *in vitro* transcription was generated by polymerase chain reaction as previously described⁸ by using the following primers: forward primer 5'-d(CTAATACGACTCACTATAGCCCCTGCCTGCCTGCAGCTAAGGATG) (where bold nucleotides indicate a T7 RNA polymerase promoter) and reverse primer 5'-d(GCCCAGGCAGGCAGGCAGCATAGACTTTCATCCTTAGCTGCAGGCAGGCAGGAG).⁹ Transcription of the corresponding RNA was completed by runoff transcription with T7 RNA polymerase as previously described followed by purification by denaturing polyacrylamide gel electrophoresis.¹⁰

Crystallization. The RNA sample was dissolved in distilled water to afford a 1 mM solution and was folded by heating at 95 °C for 2 min and then cooled to room temperature. Screening of crystallization conditions was completed with a Nucleix Suite (Qiagen) using a Gryphon nanodrop crystallization robot (Art Robinson) in a 96-well sitting drop format. Large, reproducible crystals were grown in hanging drops with a 2 μ L drop size at room temperature using a precipitant containing 100 mM ammonium acetate, 5 mM magnesium sulfate, 50 mM MES (pH 6.0), and 600 mM NaCl. The crystals were then mounted on a free mounting system (FMS) and dehydrated to a relative humidity of 75% to improve diffraction quality. Dehydrated crystals were coated with perfluoropolyether oil (Hampton Research) and flash-frozen in liquid nitrogen for synchrotron data collection.

Data Collection, Phasing, Refinement, and Analysis. X-ray diffraction data of the SCA10 repeat containing RNA were collected at beamline ID-G of LS-CAT in APS using a Mar300 CCD detector. The highest-quality crystal diffracted to Bragg spacings of 2.75 Å. The data set was then integrated and scaled using iMOSFLM.¹¹ Preliminary analysis of the diffraction data showed the crystal belongs to tetragonal space group *P*422. Detailed analysis, however, revealed that the crystal was merohedrally twinned. Thus, molecular replacement (MR) was conducted in the *P*4 space group using a tetraloop/tetraloop receptor core domain of Protein Data Bank (PDB) entry 4FNJ and Phaser in the Phenix suite.¹² The best MR solution with a log-likelihood-gain (LLG) gain and a translation function *Z*-score (TFZ) of 116.3 and 9.2, respectively, was obtained in space group *P*4₁. The electron density map of the MR solution showed base pair steps outside of the search model. Rigid body refinement against the MR solution also showed R_{work} and R_{free} values of 35 and 38%, respectively. Missing RNA bases were manually modeled using Coot.¹³ Crystallographic refinement of the modeled RNA was performed by using Phenix (phenix.refine) and CCP4 (refmac) suites applying noncrystallographic symmetry (NCS) restraints along with a twin refinement protocol. The final R_{work} and R_{free} values of the structure were 17.8 and 22.4%, respectively. Figures were prepared with PyMol.¹⁴ Helical parameter calculations were completed by 3DNA.¹⁵ Data collection and

refinement statistics are summarized in Table S-5 of the Supporting Information.

Model System Preparation. A model r(AUUCU) structure containing one 3×3 $5'UCU^{3'}/^{3'}UCU^{5'}$ internal loop was prepared by homology modeling. A symmetric system was designed with the sequence of r(CG AUUCUAUCG)₂, where CG and GC flanking pairs were included to increase the overall structural stability in the MD simulations. The system was prepared with ModeRNA¹⁶ by extracting the (AUUCUAU)₂ fragment from the crystal structure and adding standard CG and GC base pairs from the rnaDB2005 database fragment library.¹⁷ The system was neutralized with 20 Na⁺ ions¹⁸ and solvated with 5095 TIP3P water molecules¹⁹ in a 12 Å truncated octahedral box. The AMBER force field²⁰ with revised χ^{21} and α/γ^{22} torsional parameters was used in all the MD simulations.

Replica Exchange Molecular Dynamics (REMD) Simulation. The system was minimized in two steps. First, all residues except solvent were held fixed with a restraint force of 500 kcal mol⁻¹ Å⁻² and minimized with 2500 steepest descent steps followed by 2500 conjugate gradient steps. A second minimization step was performed without positional restraints using 10000 steepest descent and 10000 conjugate gradient steps. REMD simulation was conducted under periodic boundary conditions, using 40 replicas at a constant volume. The temperature range was determined using the parallel tempering temperature predictor.²³ A temperature range from 272 to 363 K was spanned with uniform ratios for exchange of ~30% between neighboring replicas. Each replica was slowly heated to its corresponding replica temperature in 150 ps while the solute was constrained with a force gradient from 8.0 to 0 kcal mol⁻¹ Å⁻². Langevin dynamics with a collision frequency of 1 ps⁻¹ was used. A 20 ps pressure equilibration step was then applied with isotropic scaling at 1 atm. Production runs were conducted at a constant volume using an exchange frequency of 2 ps. Chemical bonds involving hydrogen atoms were constrained with the SHAKE algorithm,²⁴ which allowed an integration step of 2 fs in the production runs. Particle mesh Ewald (PME)^{25,26} was used in all calculations with a 9 Å long-range cutoff. A total of 5.03 μ s of accumulated simulation time was obtained.

Steered Molecular Dynamics (SMD). To describe the instability of the RNA internal loop, we performed three independent SMD experiments by pulling away sequentially the U-U and C-C pairs. SMD simulations were performed with the same protocol described for the equilibration step and initiated with the final structure obtained from the equilibration step. First, a combination of velocities ranging from 0.01 and 1.0 Å/ns and a spring constant ranging from 5 to 20 kcal mol⁻¹ Å⁻² were tested using a 3² factorial design with 10 repeats. The pulling force was applied to the center of mass (COM) of the O2, O4, and N3 atoms of the U-U pairs, and O2, N3, and N4 atoms of the C-C pair, until a separation of 6 Å was achieved. Once the variables were optimized, a total of 25 successive SMD pulling experiments per pair were conducted with a constant velocity of 0.14 Å/ns and a spring constant of 10 kcal mol⁻¹ Å⁻².

Simulation Analysis. Each replica window was analyzed with the *cptraj* module (Amber 14).²⁰ Because the force field was parametrized at 300 K, only the replica at this temperature was subjected to further analysis. Structures were clustered using the average-linkage hierarchical agglomerative method with a distance cutoff ϵ of 3 Å. Symmetry-corrected root-mean-

square deviation (rmsd) clustering was performed on a subset of atoms using the Amber mask syntax (:1–22@P,C3',C4',C5',O3',O5'). All RNA structural parameters were calculated using the 3DNA suite¹⁵ for the structures extracted from the MD trajectory at 20 ps intervals. Stiffness constants were computed according to the method of Lankas et al.²⁷ Once the covariance matrix (**C**) of the helical parameters was generated, the stiffness matrix (**K**) was computed with the following equation (eq 1):

$$\mathbf{K} = k_B T \times \mathbf{C}^{-1} \quad (1)$$

The potential of mean force (PMF) as a function of slide and helical twist was constructed as described previously for other biomolecular systems²⁸ by using observed and reference probability distributions. Grid inhomogeneous solvation theory (GIST)²⁹ analysis of solvent was performed using the algorithm incorporated in Amber 14.²⁰ Prior to the analysis, Na⁺ atoms were removed from the system because GIST considers all nonsolute atoms as solvent molecules. The grid was centered at (27, 27, 29) and dimensioned to 20 Å × 20 Å × 40 Å with a spacing of 0.50 Å. The output was analyzed with VMD.³⁰

RESULTS AND DISCUSSION

Overall Crystal Structure. RNA molecules tend to produce poor-quality crystals because of limited intermolecular

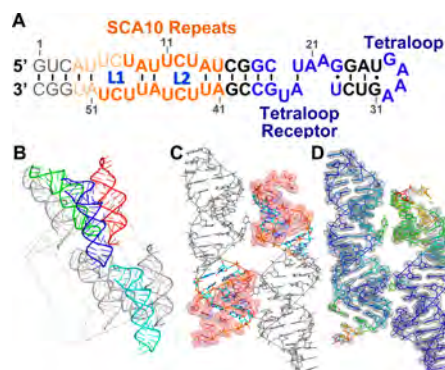


Figure 1. Structure of r(AUUCU) repeat-containing RNA. (A) Secondary structure of the crystallized RNA. The two loops are labeled L1 and L2. Bold fonts represent the modeled region. (B) Crystal packing environment. Two molecules, red and blue, are found in an asymmetric unit. Red and cyan molecules are coaxially aligned. Blue and green molecules interact through the tetraloop/tetraloop receptor interaction. (C) Overall structure of the asymmetric unit. The UCU loops are shown as orange sticks with a transparent sphere. AU closing pairs are colored cyan; remaining regions of the RNAs are shown as gray sticks. (D) Electron density colored gray at a contour level of 1.2σ. RNAs are colored according to temperature factor by rainbow colors from blue (low) to red (high).

contacts. The tetraloop/tetraloop receptor motif, which provides additional contacts, has been utilized to promote crystallization of various RNAs.^{31,32} We employed this motif in our crystallization construct that contains two model SCA10 repeats (Figure 1A). The base pair alignment of the repeats was designed using the method of Handa et al.,⁸ and additional G-C or G-U base pairs were included on both sides of the repeats to provide further stabilization.⁸

A molecular replacement search revealed two independent RNA molecules in the asymmetric unit (chains A and B) with a solvent content of 59% (Figure 1B,C). Although the packing

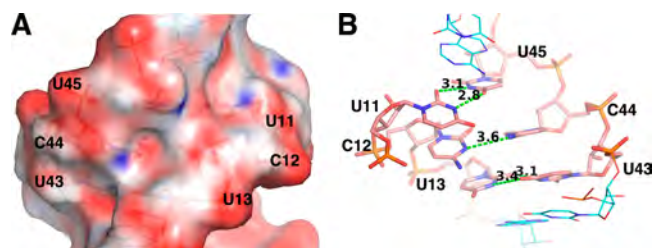


Figure 2. Overview of ^{5'}UCU^{3'}/^{3'}UCU^{5'} loop in L2 of chain A. (A) Charge distribution of the minor groove at the loop shown as a transparent surface model. The electrostatic potential was calculated using APBS and contoured at ±25 kT/e. (B) Loop represented as a stick model. Hydrogen bonds are represented as green dashed lines.

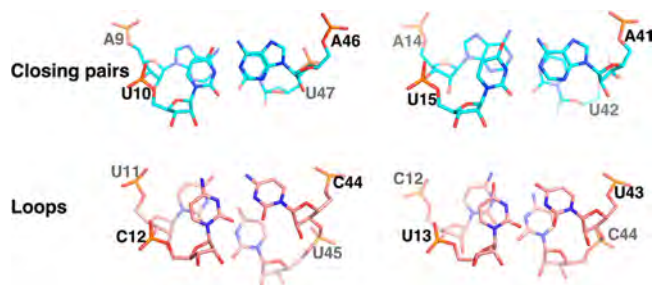


Figure 3. Base stacking of ^{5'}UCU^{3'}/^{3'}UCU^{5'} loops and ^{5'}AU^{3'}/^{3'}UA^{5'} closing pairs in chain A.

environment is different, the two RNA chains are essentially identical as judged by the rmsd between them (1 and 0.65 Å for the whole molecule and SCA10 repeat region, respectively). The *B* factor for both RNAs is lower at the tetraloop region and increases throughout the repeat and additional base pairs, which suggests that the end of the hairpin stem is labile (Figure 1D). G1–C7 and A51–C55 of both chains were not modeled because of a lack of traceable electron density (Figure 1D). The disorder observed at the stem ends is due to multiple factors, including inherent instability of the SCA10 loop and crystal contacts (Figure 1D).

The two asymmetric RNA molecules are antiparallel to each other (Figure 1C), and the interaction is stabilized by multiple hydrogen bonds (Figure S-1 of the Supporting Information). The tetraloop/tetraloop receptor interaction, which formed between crystallographic symmetry mates, is the most significant packing interaction and is invariant compared to other tetraloop/tetraloop receptor structures. Although there is a coaxial “end to end” alignment of symmetry-related molecules, no base stacking between them is observed because of disorder in the stem ends (Figure 1B).

Although 3DNA analyses display the average base pair step and local helical parameters within the range of A-form RNA (Table S-2 of the Supporting Information), individual steps have wide variations indicating irregularities. Not surprisingly, the largest deviations occur within the loops. Helical twist values at the ^{5'}UCU^{3'}/^{3'}UCU^{5'} loops show increased average values (39°) relative to those of A-form RNA. Large variations in inclination and tip indicate disrupted base stacking interactions within the loop.

Internal ^{5'}UCU^{3'}/^{3'}UCU^{5'} Loop Structures. A total of six potential U-U base pairs and two potential C-C base pairs are found in the two RNA molecules. The six carboxyl oxygens from the ^{5'}UCU^{3'}/^{3'}UCU^{5'} loops are all concentrated in the minor groove. This makes the minor groove of the SCA10

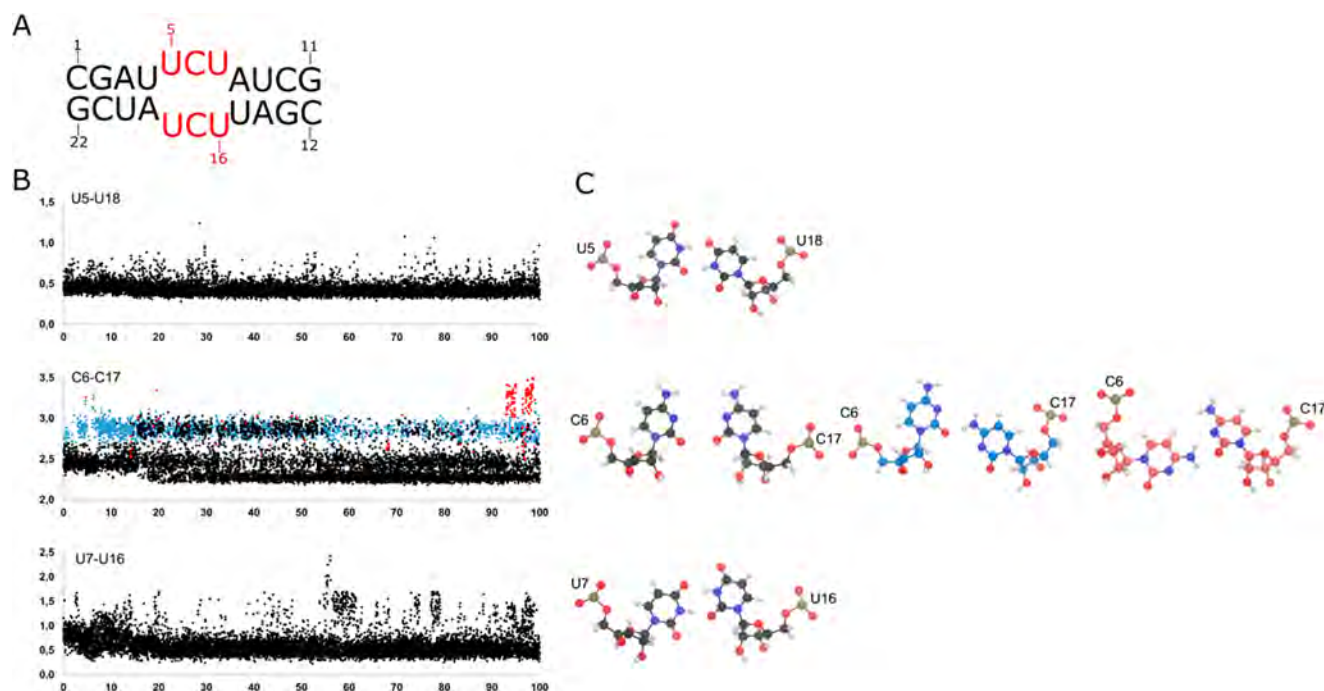


Figure 4. (A) Sequence of the model $r(\text{AUUCU})$ used in the computational studies. (B) Symmetry-corrected rmsd analysis of the U5-U18, C6-C17, and U7-U16 pairs. (C) Clustered conformations of the internal loop base pairs. Each data point in panel B corresponds to a cluster that is represented by a color code displayed in panel C. Only conformations with a >10% residence time were selected for the analysis.

repeats the most charge dense region among repeat expansion RNAs, which may have implications for the “druggability” of this RNA (Figure 2A).

The $5'\text{AU}^{3'}/3'\text{UA}^{5'}$ loop closing pairs have standard Watson–Crick geometries. In both chains, the U-U base pairs in L2 have a two-hydrogen bond geometry while the U8-U48 pair defined in L1 has a one-hydrogen bond geometry (Figure 2B and Figure S-2 of the Supporting Information). Interestingly, U-U internal loops in $r(\text{CUG})$ repeats have significant differences compared to the U-U pair geometries observed in the $r(\text{AUUCU})$ repeats. In particular, the U-U pairs in $r(\text{AUUCU})$ have larger propeller and buckle deviations, and the repeat overall has variations of roll and tilt larger than those of $r(\text{CUG})$ (Tables S-2 and S-3 of the Supporting Information).⁶

The central C-C base pair of chain A shows one hydrogen bond geometry, whereas that of chain B appears to be dynamic with poorly defined electron densities for both bases as well as parts of the phosphate backbone in C44 (Figures S-2 and S-3 of the Supporting Information). One hydrogen bond geometry of the C-C pair is similar to that previously observed in $r(\text{CCG})$ repeats.⁷ Local base pair parameters propeller and buckle of the C-C pair are also comparable to those of $r(\text{CCG})$ repeats. However, the opening (36°) is larger than the one observed in $r(\text{CCG})$ repeats because C44 opens toward the major groove (Table S-3 of the Supporting Information).⁷ Unusually short C1'–C1' distances, ranging from 7.8 to 9 Å, found in $5'\text{UCU}^{3'}/3'\text{UCU}^{5'}$ of L2 appear to contribute conformational deviation of U-U and C-C pairs from those found in $r(\text{CUG})$ and $r(\text{CCG})$ repeats (Figure S-2 of the Supporting Information).

As evidenced by increased helical twist values, stacking interactions among the bases throughout $5'\text{UCU}^{3'}/3'\text{UCU}^{5'}$ were marginal to nonexistent. In contrast, the loop closing base pairs and their nearest neighbors ($5'\text{AU}^{3'}/3'\text{UA}^{5'}$) show

extensive overlap, which is expected for adjacent Watson–Crick pairs (Figure 3). Some stacking is observed between the A of the closing pair and the neighboring U in the loop. However, this stacking interaction appears to be suboptimal because of a lack of planarity between the bases (Figure S-4 of the Supporting Information).

Dynamics of the $5'\text{UCU}^{3'}/3'\text{UCU}^{5'}$ Model System. Understanding the dynamics of biomolecules is often a computationally challenging process. Moreover, classical MD protocols are subject to sampling limitations, which are hard to overcome. The REMD method simulates several replicas at different temperatures with a certain probability of swapping periodically from one temperature replica to another to improve thermodynamic sampling in a simple and effective manner.³³ To achieve thorough sampling of the 3×3 internal loop conformational space, we employed REMD simulations with the hope that higher temperatures would destabilize the loop and provide possibilities of analyzing the dynamic features of this 3×3 RNA internal loop. A total of 40 replicas spanning a temperature range close to 300 K were used to study the intrinsic dynamics of the 3×3 internal loop. The initial coordinate of the $5'\text{UCU}^{3'}/3'\text{UCU}^{5'}$ model system is identical to the central fragment of the L1 crystal structure except for additional C-G and G-C pairs flanking the internal loop to enhance the structural stability.

The rmsd analysis was completed for each U-U and C-C pair contained in the 300 K replica individually (Figure 4). The rmsd values were plotted with respect to the average conformation, and each color represents an individual clustered conformation with a residence time that was >10% of the total simulation (Figure 4C). In good agreement with the crystal structure, the two U-U pairs in the model stay in a two-hydrogen bond conformation along the 300 K trajectory. However, some differences are observed in terms of stacking. For instance, the stacking surface area of $5'\text{U4U}^{3'}/3'\text{A19U18}^{5'}$

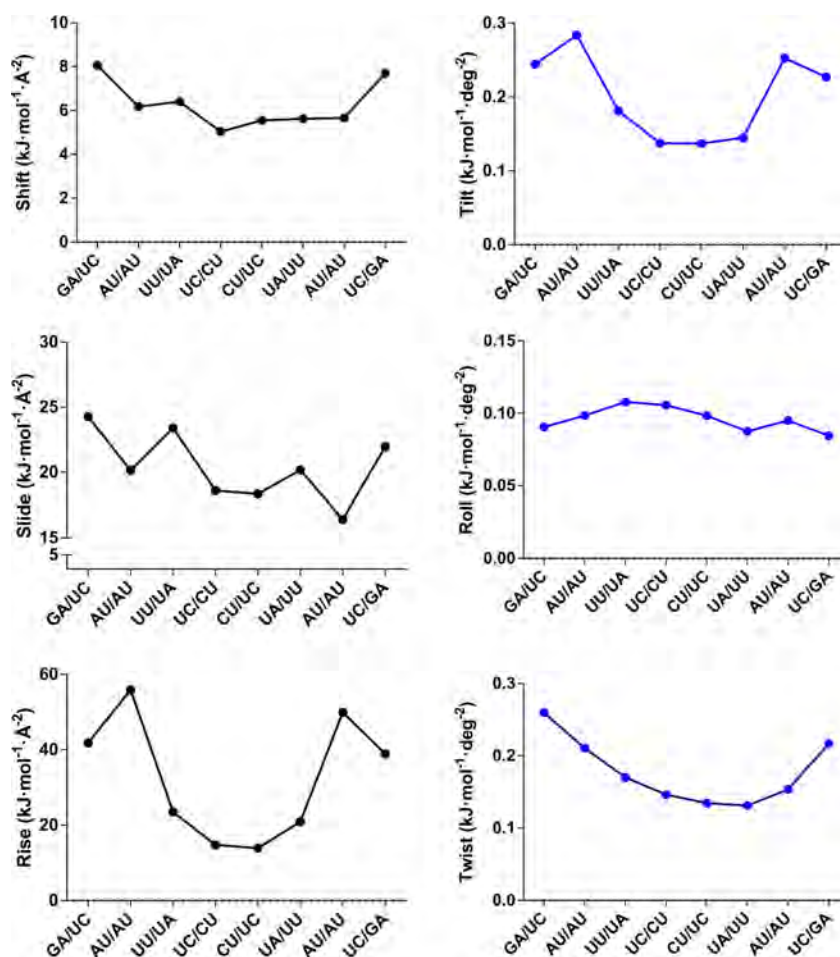


Figure 5. Stiffness force constants of the helical space (shift, slide, rise, tilt, roll, and twist) for the different nucleotides contained in the [GAUUCUAU] fragment. The most flexible regions correspond to those with associated low stiffness force constants. Overall, flexibility is especially pronounced in the [UCU] region as observed in the slide, rise, and tilt stiffness profiles.

is 0.56 Å² larger than that of 5'U7A8^{3'}/3'U16U15^{5'}, so the stacking contribution is greatly diminished in the latter case.

The C-C pair is clearly the most dynamic mismatch, whereby a large number of conformations are equally distributed during the simulation. Of these, a total of three different relevant conformations were found for the C-C pair while the U5-U18 and U7-U16 pairs visited only one relevant conformation (Figure 4B). Interestingly, the C-C pair showed a zero-hydrogen bond conformation during 34% of the simulation (Figure 4C), while two different one-hydrogen bond conformations were present during 64% of the total time (Figure 4C). This dynamic behavior of the C-C pair during the simulation is in line with the mainly one-hydrogen bond state in the crystal structure as well as the poorly defined electron density. Previous studies of r(CUG) repeats defined the dynamic nature of U-U pairs; thus, it is not surprising the poor stacking of the C-C pairs with the flanking uridines and the concurrent C-C dynamic behavior. Several transitions between *cis* and *trans* Watson-Crick conformations of the C-C pair are observed along the simulations that altogether clearly affect the stiffness of the entire internal loop. The C1'-C1' distances of the loop nucleotides remain in the range of 8.8–8.9 Å on average, as observed in the crystal structure, so no major backbone deviations were produced during the simulation.

The influence of the overall flexibility of r(AUUCU) by the 5'UCU^{3'}/3'UCU^{5'} internal loop was analyzed using both

stiffness constants of a helical space defined by Lankas et al. and the helical parameters of 3DNA.²⁷ Most of the base pair and helical parameters are close to those in the crystal structure except for buckle, propeller, and slide, which showed significant deviations (Table S-6 of the Supporting Information). However, the fluctuations are considered to be small, so the effect of the Jacobian factor was neglected in the stiffness analysis. In general terms, the internal loop presents the lowest stiffness constant values along the structure (Figure 5), especially rise and tilt, which can be explained by the lack of optimal stacking into this region. Altogether, the labile stacking capability of the internal loop and its large charge density may confer to this RNA a unique binding region for small molecules or other ligands such as ions or proteins.

Stabilization of C-C Pairs by Water-Mediated Hydrogen Bonds. Although the C-C pair remains mainly in a one-hydrogen bond conformation, the MD results suggest the presence of a stable zero-hydrogen bond state. Moreover, two-dimensional potential of mean force analysis (2D-PMF) revealed that the zero-hydrogen bond state (C1 in Figure 4C) is the most stable among the clustered conformations (Figure 6). The PMF map was constructed using slide and helical twist as variables because of the high variability of these parameters in the crystal structure and during the simulation. The C1 coordinates are close to local minima (−1.9, −23) and deviated from the coordinates of the crystal structure in chain A

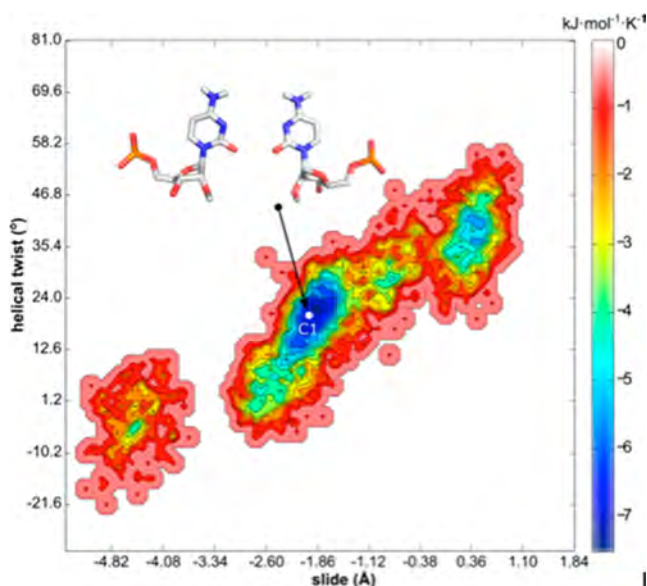


Figure 6. 2D-PMF surface showing the exploration of the C-C pair conformations along the MD simulation. The x - and y -axes represent the slide and helical twist values of the noncanonical CC pair, respectively. The zero-hydrogen bond CC conformation (C1 in Figure 4C) is displayed as a white dot in the figure that corresponds to a local minimum conformation.

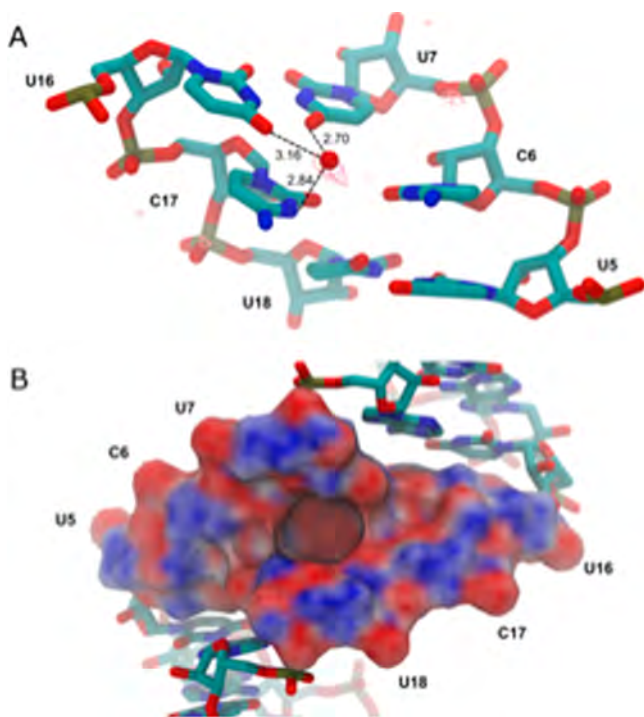


Figure 7. (A) Stick model representation of the internal loop and potential hydrogen bond with an explicit water molecule. GIST-computed water probability contours are represented as a grid. (B) Charge distribution of the minor groove at the loop shown as a surface model. The electrostatic potential was calculated using APBS and contoured at ± 10 kT/e.

($-1.1, 36.2$) and in chain B ($-0.9, 40.1$). This indicates an additional force should stabilize C1 to keep the zero-hydrogen bond conformation as a preferred conformation among all the sampled phase space. A closer inspection of this state suggests

that the hydration pattern of the noncanonical pair plays an essential role in its stabilization. Effects of the structure and thermodynamics of the solvent were investigated using GIST (grid inhomogeneous solvation theory), which analyzes the three-dimensional water density around a certain area termed voxels with associated solvent properties.²⁹ Indeed, a highly occupied volume appears to be close to the C-C pair. Surprisingly, the water molecule occupies a space among C17, U7, and U16 and coordinates a potential hydrogen bond between atom O4 of U7 and atom N3 of C17 (Figure 7A). At the same time, the U7-U16 pair inclines toward the C-C pair and breaks the hydrogen bonding pattern, yielding a highly negatively charged cavity in the RNA minor groove (Figure 7B). This phenomenon is produced repeatedly and approximately one-third of the time in the MD simulation.

Inherent Instability of the $5'UCU^{3'}/\beta'UCU^{5'}$ Internal Loop.

Our crystallographic and computational results presented here corroborate the findings of Handa et al. that the $5'UCU^{3'}/\beta'UCU^{5'}$ internal loop is metastable.⁸ Therefore, it has been hypothesized that RNA unwinding can start through internal loop destabilization. Under this premise, we investigated the stability of each pair individually within the 3×3 internal loop context using steered molecular dynamics (SMD). SMD is a powerful technique used to gain insights into several mechanisms, such as unfolding pathways of macromolecules. This procedure exploits nonequilibrium sampling by applying time-dependent biasing forces to guide the system transformation. During SMD experiments, several pulls are simulated in one direction. Previous SMD studies have been proven to be successful in computing free energy profiles on realistic biomolecular systems.^{34,35}

Herein, we conducted 25 successive SMD pullings per pair with a constant velocity of 0.14 \AA/ns and a spring constant of $10 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ and computed force of pulling and cumulative work profile for each system (Figure 8A). Each plot represents the work required to pull each pair from the bonded state with an $\sim 2.5\text{--}3.5 \text{ \AA}$ COM distance to 6.0 \AA where hydrogen binding interactions are no longer operative. Pulling force peaked at $\sim 4 \text{ \AA}$ for U7-U16 and U5-U18 and 5 \AA for C6-C17. The mean work performed in each pair clearly reflects the fact that the central C-C pair is the most probable starting point of unwinding ($W = 7.88 \text{ kcal/mol}$). Indeed, the U-U pulling experiments yield very close work values (12.21 and 12.13 kcal/mol for U5-U18 and U7-U16 pairs, respectively).

Interaction energies of noncanonical pairs vanished smoothly and had relatively marginal impact over the conformation of the neighboring pairs; thus, the overall RNA structure remained intact during all the pulling experiments. While both U-U pairs did not change their orientation during the steering process, the C-C pair experienced a rotation along the α torsion that flipped C6 out of the RNA duplex. The rotational movement of the C6 base was consistently observed in all SMD trials. Therefore, we concluded that this should be the lowest-energy pathway for breaking the C-C pair. The thermodynamic stability of an RNA structure depends not only on base pairing but also on stable stacking interactions. Therefore, the observed unstacking of a cytosine base can further destabilize the $5'UCU^{3'}/\beta'UCU^{5'}$ internal loop.

Biological Significance of the Structural Features. The crystallographic results herein confirm that the r(AUUCU) repeats indeed form an overall A-form geometry similar to that of other repeat expansion disease RNAs. However, unlike

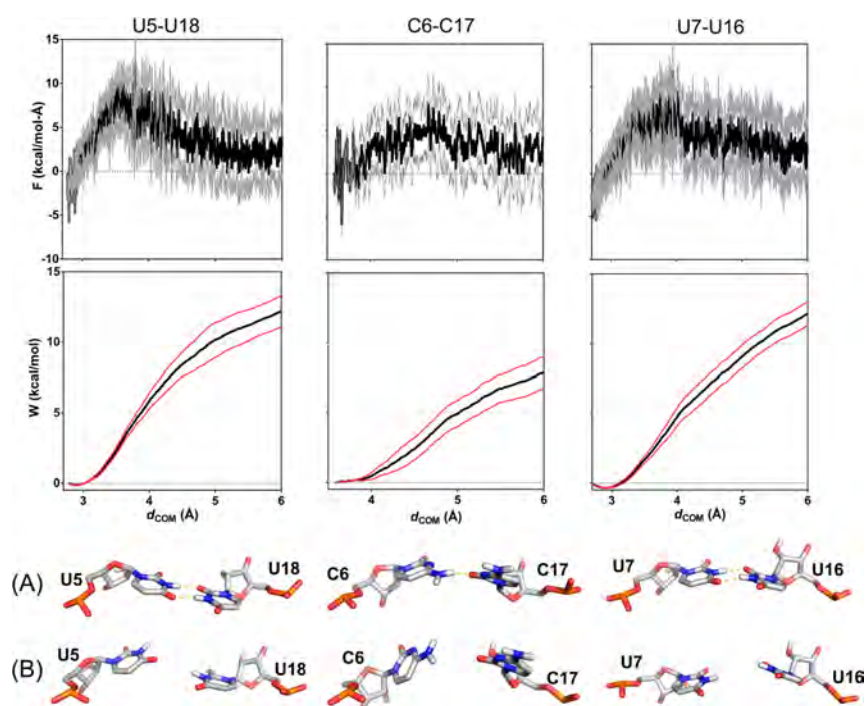


Figure 8. Pulling force and cumulative work distribution profiles vs distance from the center of mass (d_{COM}). For each base pair, initial (A) and final (B) representative conformations are depicted. The mean force and work are represented as black lines; the standard deviations of force and work profiles are depicted as gray and red lines, respectively.

trinucleotide repeats, which harbor two G-C closing pairs for every 1×1 nucleotide loop, the SCA10 repeat has two A-U pairs for every 3×3 nucleotide loop. Thus, the SCA10 repeat is likely less stable than $r(\text{CAG})$, $r(\text{CUG})$, or $r(\text{CGG})$ repeats because of (i) weaker closing base pairs (A-U vs G-C) and (ii) decreased loop stability. Thermodynamic studies of RNA duplexes containing two to four CNG repeats (where N represents any nucleotide) showed that C-C and U-U mismatches are less stable than A-A and G-G.³⁶ In SCA10, each $5' \text{UCU}^{3'}/3' \text{UCU}^{5'}$ forms two U-U mismatches and one C-C mismatch. Crystallographically, such instability is indirectly observed by the high temperature factors (and weak electron density map) (Figure S-3 of the Supporting Information). The accompanied computational simulations also explored the nature of the conformational flexibilities of the C-C and U-U mismatches. In addition, computational simulation observed that the central C-C mismatches spent significant simulation time in the zero-hydrogen bond state through stabilization by water-mediated hydrogen bonds near C17, U7, and U16.

Studies showed that SCA10 is caused when $r(\text{AUUCU})^{\text{exp}}$ binds and inactivates hnRNP K.² hnRNP K binds RNA through a KH domain, which is highly abundant particularly in proteins that regulate gene expression. Crystal structures have revealed that KH domains bind single-stranded RNA by forming favorable interactions with at least four nucleotides, N1–N4.³⁷ Interestingly, adenine and cytosine are preferred at N3 as they can most favorably hydrogen bond with a conserved hydrophobic residue in β -strand 2 (with the amino acid residue's carbonyl and a backbone amide oxygen).³⁷ In particular, structural studies of the KH3 domain of hnRNP K with single-stranded DNAs with cytosine at N3 determined the binding specificity through hydrogen bond interactions between O2 and N4 of cytosine and NH2 of Arg414 and the backbone carbonyl of Ile423, respectively. Additional bipartite

hydrogen interaction between N3 of cytosine and the backbone amide of Ile423/NH2 of Arg414 is mediated through a water molecule.^{38,39} It is conceivable that the other KH domains (KH1 and KH2) of hnRNP K have similar binding specificity for the single-stranded nucleic acids considering the high degree of sequence conservation among the three KH domains.³⁹ Our SMD study indicates the central C-C mismatch is particularly vulnerable to strand opening, which is likely done through rotational movement along the α torsion angle. Like base flipping of many DNA repair enzymes, hnRNP K may sample the weak C-C mismatch pair and (or) initiate unzipping of $r(\text{AUUCU})^{\text{exp}}$ by flipping out a base of the pair, i.e., C6 (Figure 4A). The unzipping event can be facilitated further by the disrupted base stacking as well as inherently weak interactions in the internal loop (as compared to a fully paired RNA). In such a scenario, the most likely register for $r(\text{AUUCU})$ binding is (N1–N4, where the fifth nucleotide of the repeat is denoted in parentheses) $(\text{A})^1\text{U}^2\text{U}^3\text{C}^4\text{U}$. Taken together, the structural information presented herein provides a framework to allow the design of small molecule therapeutic agents.

■ ASSOCIATED CONTENT

📄 Supporting Information

GIST input script, refinement and data collection statistics, and further analysis of the X-ray crystal structure. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.biochem.5b00551.

Accession Codes

The structure was deposited in the Protein Data Bank as entry SBTM.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: disney@scripps.edu. Phone: (561) 228-2203. Fax: (561) 228-2147.

*E-mail: hajpark@scripps.edu. Phone: (561) 228-2121. Fax: (561) 228-3067.

Funding

This work was funded by the National Institutes of Health (Grant R01-GM079235 to M.D.D.) and by The Scripps Research Institute.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Jessica Childs-Disney for discussions and critical review of the manuscript and the staff at the LS-CAT, APS, for synchrotron support.

■ ABBREVIATIONS

r(AUUCU)^{exp}, expanded r(AUUCU) repeat; DM, myotonic dystrophy; hnRNP K, heterogeneous nuclear ribonucleoprotein K; SCA10, spinocerebellar ataxia type 10.

■ REFERENCES

(1) Li, L. B., and Bonini, N. M. (2010) Roles of trinucleotide-repeat RNA in neurological disease and degeneration. *Trends Neurosci.* 33, 292–298.

(2) White, M. C., Gao, R., Xu, W., Mandal, S. M., Lim, J. G., Hazra, T. K., Wakamiya, M., Edwards, S. F., Raskin, S., Teive, H. A., Zoghbi, H. Y., Sarkar, P. S., and Ashizawa, T. (2010) Inactivation of hnRNP K by expanded intronic AUUCU repeat induces apoptosis via translocation of PKC δ to mitochondria in spinocerebellar ataxia 10. *PLoS Genet.* 6, e1000984.

(3) Zu, T., Gibbens, B., Doty, N. S., Gomes-Pereira, M., Huguet, A., Stone, M. D., Margolis, J., Peterson, M., Markowski, T. W., Ingram, M. A., Nan, Z., Forster, C., Low, W. C., Schoser, B., Somia, N. V., Clark, H. B., Schmechel, S., Bitterman, P. B., Gourdon, G., Swanson, M. S., Moseley, M., and Ranum, L. P. (2011) Non-ATG-initiated translation directed by microsatellite expansions. *Proc. Natl. Acad. Sci. U.S.A.* 108, 260–265.

(4) Ash, P. E., Bieniek, K. F., Gendron, T. F., Caulfield, T., Lin, W. L., DeJesus-Hernandez, M., van Blitterswijk, M. M., Jansen-West, K., Paul, J. W., III, Rademakers, R., Boylan, K. B., Dickson, D. W., and Petrucelli, L. (2013) Unconventional translation of C9ORF72 GGGGCC expansion generates insoluble polypeptides specific to c9FTD/ALS. *Neuron* 77, 639–646.

(5) Kumar, A., Fang, P., Park, H., Guo, M., Nettles, K. W., and Disney, M. D. (2011) A crystal structure of a model of the repeating r(CGG) transcript found in fragile X syndrome. *ChemBioChem* 12, 2140–2142.

(6) Kumar, A., Park, H., Fang, P., Parkesh, R., Guo, M., Nettles, K. W., and Disney, M. D. (2011) Myotonic dystrophy type 1 RNA crystal structures reveal heterogeneous 1 \times 1 nucleotide UU internal loop conformations. *Biochemistry* 50, 9928–9935.

(7) Kiliszek, A., Kierzek, R., Krzyzosiak, W. J., and Rypniewski, W. (2012) Crystallographic characterization of CCG repeats. *Nucleic Acids Res.* 40, 8155–8162.

(8) Handa, V., Yeh, H. J., McPhie, P., and Usdin, K. (2005) The AUUCU repeats responsible for spinocerebellar ataxia type 10 form unusual RNA hairpins. *J. Biol. Chem.* 280, 29340–29345.

(9) Disney, M. D., Labuda, L. P., Paul, D. J., Poplawski, S. G., Pushechnikov, A., Tran, T., Velagapudi, S. P., Wu, M., and Childs-Disney, J. L. (2008) Two-dimensional combinatorial screening identifies specific aminoglycoside-RNA internal loop partners. *J. Am. Chem. Soc.* 130, 11185–11194.

(10) Milligan, J. F., and Uhlenbeck, O. C. (1989) Synthesis of small RNAs using T7 RNA polymerase. *Methods Enzymol.* 180, 51–62.

(11) Batty, T. G., Kontogiannis, L., Johnson, O., Powell, H. R., and Leslie, A. G. (2011) iMOSFLM: A new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D* 67, 271–281.

(12) Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* 66, 213–221.

(13) Emsley, P., and Cowtan, K. (2004) Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D* 60, 2126–2132.

(14) DeLano, W. L. (2012) *The PyMOL Molecular Graphics System*, version 1.5.0.2, Schrödinger, LLC, Portland, OR.

(15) Lu, X. J., and Olson, W. K. (2003) 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* 31, 5108–5121.

(16) Rother, M., Rother, K., Puton, T., and Bujnicki, J. M. (2011) ModeRNA: A tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.* 39, 4007–4022.

(17) Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J., and Berman, H. M. (2008) RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA* 14, 465–481.

(18) Joung, I. S., and Cheatham, T. E. (2008) Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* 112, 9020–9041.

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

(20) Case, D. A., V. B. Berryman, J. T., Betz, R. M., Cai, Q., Cerutti, D. S., Cheatham, T. E., III, Darden, T. A., Duke, R. E., Gohlke, H., Goetz, A. W., Gusarov, S., Homeyer, N., Janowski, P., Kaus, J., Kolossváry, I., Kovalenko, A., Lee, T. S., LeGrand, S., Luchko, T., Luo, R. B., Wu, X., and Kollman, P. A. (2014) *AMBER 14*, University of California, San Francisco.

(21) Yildirim, I., Stern, H. A., Kennedy, S. D., Tubbs, J. D., and Turner, D. H. (2010) Reparameterization of RNA χ torsion parameters for the AMBER force field and comparison to NMR spectra for cytidine and uridine. *J. Chem. Theory Comput.* 6, 1520–1531.

(22) Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T. E., Laughton, C. A., and Orozco, M. (2007) Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers. *Biophys. J.* 92, 3817–3829.

(23) Patriksson, A., and van der Spoel, D. (2008) A temperature predictor for parallel tempering simulations. *Phys. Chem. Chem. Phys.* 10, 2073–2077.

(24) Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.* 23, 327–341.

(25) Sagui, C., Pedersen, L. G., and Darden, T. A. (2004) Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations. *J. Chem. Phys.* 120, 73–87.

(26) Toukmaji, A., Sagui, C., Board, J., and Darden, T. (2000) Efficient particle-mesh Ewald based approach to fixed and induced dipolar interactions. *J. Chem. Phys.* 113, 10913–10927.

(27) Lankas, F., Sponer, J., Langowski, J., and Cheatham, T. E. (2003) DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.* 85, 2872–2883.

(28) Paladino, A., and Zangi, R. (2013) Ribose 2'-hydroxyl groups stabilize RNA hairpin structures containing GCUAA pentaloop. *J. Chem. Theory Comput.* 9, 1214–1221.

- (29) Nguyen, C. N., Young, T. K., and Gilson, M. K. (2012) Erratum: Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril (The Journal of Chemical Physics (2012) 137 (044101)). *J. Chem. Phys.* 137, 044101.
- (30) Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD: Visual Molecular Dynamics. *J. Mol. Graphics* 14, 33–38.
- (31) Ferre-D'Amare, A. R., Zhou, K., and Doudna, J. A. (1998) A general module for RNA crystallization. *J. Mol. Biol.* 279, 621–631.
- (32) Coonrod, L. A., Lohman, J. R., and Berglund, J. A. (2012) Utilizing the GAAA tetraloop/receptor to facilitate crystal packing and determination of the structure of a CUG RNA helix. *Biochemistry* 51, 8330–8337.
- (33) Sugita, Y., and Okamoto, Y. (1999) Replica exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314, 141–151.
- (34) Patel, J. S., Berteotti, A., Ronsisvalle, S., Rocchia, W., and Cavalli, A. (2014) Steered molecular dynamics simulations for studying protein-ligand interaction in cyclin-dependent kinase 5. *J. Chem. Inf. Model.* 54, 470–480.
- (35) Tekpinar, M., and Zheng, W. (2014) Unzipping of neuronal snare protein with steered molecular dynamics occurs in three steps. *J. Mol. Model.* 20, 2381.
- (36) Broda, M., Kierzek, E., Gdaniec, Z., Kulinski, T., and Kierzek, R. (2005) Thermodynamic stability of RNA structures formed by CNG trinucleotide repeats. Implication for prediction of RNA structure. *Biochemistry* 44, 10873–10882.
- (37) Auweter, S. D., Oberstrass, F. C., and Allain, F. H. (2006) Sequence-specific binding of single-stranded RNA: Is there a code for recognition? *Nucleic Acids Res.* 34, 4943–4959.
- (38) Backe, P. H., Messias, A. C., Ravelli, R. B., Sattler, M., and Cusack, S. (2005) X-ray crystallographic and NMR studies of the third KH domain of hnRNP K in complex with single-stranded nucleic acids. *Structure* 13, 1055–1067.
- (39) Braddock, D. T., Baber, J. L., Levens, D., and Clore, G. M. (2002) Molecular basis of sequence-specific single-stranded DNA recognition by KH domains: Solution structure of a complex between hnRNP K KH3 and single-stranded DNA. *EMBO J.* 21, 3476–3485.

Publisher: Taylor & Francis

Journal: *Journal of Biomolecular Structure and Dynamics*

DOI: <http://dx.doi.org/10.1080/07391102.2015.1114971>

Deciphering structure-activity relationships in a series of Tat/TAR inhibitors

Lise Pascale,^[a] Alejandro López González,^[b] Audrey Di Giorgio,^[a] Marc Gaysinski,^[a] Jordi Teixido Closa,^[b] Roger Estrada Tejedor,^[b] Stéphane Azoulay,^[a] and Nadia Patino*^[a]

[a] Institut de Chimie de Nice UMR7272, Université Nice Sophia Antipolis, 06108 Nice Cedex, France. [b] Molecular Design Lab., IQS School of Engineering, Universitat Ramon Llull, 08017 Barcelona, Spain.

***Corresponding author:** patino@unice.fr; Tel: +33 (0)4 92 07 61 46; Fax: +33 (0)4 92 07 61 51

Acknowledgements. We thank Jean Marie Guignonis for mass analyses.

Fundings. This work was supported by SIDACTION ; the Agence Nationale de Recherche sur le SIDA ; and the Caisse Primaire d'Assurance Maladie des Professions Libérales. L. Pascale was recipient of a MENRT Ph.D. fellowship.

Deciphering structure-activity relationships in a series of Tat/TAR inhibitors

A series of pentameric "Polyamide Amino Acids" (PAAs) compounds derived from the same trimeric precursor have been synthesized and investigated as HIV TAR RNA ligands, in the absence and in the presence of a Tat fragment. All PAAs bind TAR with similar sub-micromolar affinities but their ability to compete efficiently with the Tat fragment strongly differs, IC_{50} ranging from 35 nM to $> 2 \mu\text{M}$. While NMR and CD studies reveal that all PAA interact with TAR at the same site and induce globally the same RNA conformational change upon binding, a comparative thermodynamic study of PAA/TAR equilibria highlights distinct TAR binding modes for Tat competitor and non-competitor PAAs. This led us to suggest two distinct interaction modes that have been further validated by molecular modeling studies. While the binding of Tat competitor PAAs induces a contraction at the TAR bulge region, the binding of non-competitor ones widens it. This could account for the distinct PAA ability to compete with Tat fragment. Our work illustrates how comparative thermodynamic studies of a series of RNA ligands of same chemical family are of value for understanding their binding modes and for rationalizing structure-activity relationships.

Keywords: RNA ligand interactions, thermodynamics, HIV TAR RNA, Tat inhibitors

1. Introduction

Today, it is becoming increasingly clear that non-coding RNAs play essential roles in a variety of fundamental cellular and pathological functions. For a multitude of diseases, targeting such RNAs constitute an alternative strategy that complements traditional protein-based targeting. These RNAs adopt folded hairpin structures leading to binding pockets suitable for the formation of specific interactions with their natural counterparts (RNAs, proteins, metabolites...). Small molecules able to selectively inhibit these interactions are of considerable interest both as therapeutic agents and as chemical

probes. The ability of some antibiotics to exert their activity by binding to defined regions of bacterial ribosomal RNAs provides the proof of concept that it is possible to target specifically some RNA structures with small molecules (Poehlsgaard & Douthwaite, 2005). However, even if numerous RNA-directed small ligands, mainly identified by standard or virtual screening approaches, have been reported in the literature so far (see reviews (Blond et al., 2014; Disney et al., 2014; Guan & Disney, 2012)), none of them has yet led to commercially approved drugs, except antibiotics. Efforts for the discovery of small synthetic molecules that combine high affinity and selectivity for a particular RNA target should be pursued. A better knowledge of the factors that govern small molecule-RNA recognition is therefore of great importance to achieve this goal. For several years, an extensive research has been dedicated to the highly conserved HIV-1 TAR RNA hairpin fragment (Blond et al., 2014; Massari et al., 2013; Yang, 2005), which plays a crucial role in the viral transcription step of HIV *via* its complexation with the transactivating transcription (Tat) protein and cellular factors (Stevens et al., 2006). Even if up to now these efforts have been unsuccessful in leading to an antiviral drug onto the market, the TAR RNA fragment remains nevertheless an excellent model to better understand the main molecular forces that drive RNA-ligand associations (Kumar & Maiti, 2013; Suryawanshi et al., 2010).

Previously, we reported a comparative TAR interaction study with three series of original ligands called PAAs “Polyamide Amino Acids” (α -PAA, β -PAA and C- α -PAA). These compounds were trimeric structures, each unit comprising an amide backbone (2-aminoethylglycyl or 2-aminoethyl β -alanyl) onto which an amino acid residue (L- α or L- β Phenylalanine, Arginine and Lysine) is condensed (Figure 1A). In addition to the identification of some lead structures, this study also revealed that the nature of the PAA (α -, β - or C- α -PAA) and to a lesser extent, its sequence, had a large

impact both on the TAR binding mode and on the ability to displace a preformed Tat/TAR complex (Pascale et al., 2013). In an attempt to increase both TAR affinity and Tat competitor behavior, we have designed a series of pentameric C- α -PAA deriving from one tri-C- α -PAA lead, identified in our previous work. In this paper, we describe the penta-PAA preparation as well as the TAR interaction studies in the absence and in the presence of a Tat fragment, by fluorescence and FRET spectroscopies. TAR-PAA interactions have been further characterized by UV-melting and Circular Dichroism experiments and the RNA recognition site confirmed by NMR. A detailed thermodynamic analysis has allowed us to rationalize structure-activity relationships by discriminating two distinct groups of PAA ligands that interact with TAR by two different interaction modes, which have been further supported by molecular modeling studies.

2. Materials and Methods

2.1 Experimental procedures

Unless otherwise stated, all reagents and solvents were of analytical grade and from Sigma-Aldrich (St Louis, MO, USA). HEPES [4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid] and all inorganic salts for buffers were purchased from Calbiochem-Merck Millipore (Fontenay sous Bois, France) (molecular biology grade). RNA and DNA oligonucleotides were purchased from IBA GmbH (Gottingen, Germany). Labeled Tat peptide was purchased from EZBiolab (Carmel, CA, USA). All buffers were filtered through 0.22- μ m Millipore filters (GPEXpressPLUS membrane).

2.2 Fluorescence binding assays

Ligand solutions were prepared as serial dilutions by an epMotion automated pipetting

system (Eppendorf) in buffer A (20 mM HEPES (pH 7.4 at 25 °C), 20 mM NaCl, 140 mM KCl and 3 mM MgCl₂) at a concentration twice higher than the desired final concentration to allow for the subsequent dilution during the addition of the RNA solution. The appropriate ligand solution (30 µL) was then added to a well of a non-treated black 384-well plate (Nunc 237105), in triplicate. Refolding of the RNA was performed using a thermocycler (ThermoStatPlus Eppendorf) as follows: the RNA, diluted in 1 mL of buffer A, was first denatured by heating to 90 °C for 2 min then cooled to 4 °C for 10 min followed by incubation at 20 °C for 15 min. After refolding, the RNA was diluted to a working concentration of 10 nM through addition of the appropriate amount of buffer A. The tube was mixed and 30 µL of the RNA solution was added to each well containing ligand. This subsequent dilution lowered the final RNA concentration to 5 nM. The fluorescence was measured on a GeniosPro (Tecan) with an excitation filter of 485 ± 10 nm and an emission filter of 535 ± 15 nm. Each point was measured 5 times with a 500 µs integration time and averaged. Binding was allowed to proceed at least 30 min at room temperature to achieve equilibrium.

To study the temperature dependence, the plates were incubated after 30 min equilibrium at different temperature ranging from 5°C to 35°C.

The salt dependence was studied in 20 mM HEPES (pH 7.4 at 25 °C), 20 mM NaCl, 3 mM MgCl₂ with the KCl concentration varied between 70 and 250 mM.

For competitive experiments in the presence of a dsDNA, a 15-mer sequence (5'-CGTTTTTATTTTGC-3') and its complement, annealed beforehand, were added to buffer A to obtain a 100-fold nucleotide excess over TAR RNA (900nM duplex; 5nMRNA). For competitive experiments in the presence of tRNA, a mixture of pre- and mature yeast tRNAs (containing >30 different species from baker's yeast

(*Saccharomyces cerevisiae*, Sigma, type X-SA) was added to buffer A to obtain a 100-fold nucleotide excess over TAR RNA.

2.3 Fluorescence resonance energy transfer (FRET) displacement assays

Ligand solutions and RNA (40 nM working solutions) were prepared as described above in buffer B [50 mM tris buffer (pH 7.4 at 25°C), 20 mM KCl and 0.005% Tween 20]. Labeled Tat peptide (40 nM in buffer B) was mixed to an equal volume of TAR RNA for 20 min at room temperature to form the Tat/TAR complex before adding the ligand. The appropriate ligand solution (30 μ L) was added to a well of a 384-well plate, in triplicate, followed by 30 μ L of the Tat/TAR solution. Fluorescence was measured as described above after 30 min of incubation at room temperature.

2.4 Temperature-dependent UV spectroscopy (UV melting)

Thermal denaturation scans were obtained using a Cary 300 (Varian) spectrophotometer equipped with an electrothermal multicell holder. Absorbance versus temperature profiles were recorded at 260 nm. After structuration of TAR RNA and incubation (1h) with the corresponding ligand, the temperature was raised from 20 to 90°C, with a heating rate of 0.5°C/min. Thermal denaturation studies were carried out at 2 μ M TAR RNA with 2 μ M PAA or without PAA (TAR alone). The experiments were performed in buffer C [10 mM sodium cacodylate, 10 mM NaCl (pH 7.5) and 0.1 mM EDTA]. The melting temperature (T_m) value was taken as the midpoint of the melting transition as determined by the maximum of the first-derivative plot with Prism software.

2.5 CD study

CD measurements were performed with a Jasco J-810 spectropolarimeter equipped with a Jasco PTC 423S Peltier temperature controller. Samples were prepared in buffer D [20

mM potassium phosphate buffer (pH 7.4 at 25 °C), 10 mM NaCl and 1 mM MgCl₂].

Spectra were obtained at 3 μM RNA or PAA (for individual spectra) or at a molar ratio of 1:1, 1:2 and 1:5 RNA:PAA for the complexes. Spectra were recorded at 20°C from 360 to 200 nm at 1-nm intervals, with an integration time of 4s and a 50-nm/min speed. CD scans were repeated five times and then averaged and corrected by the subtraction of the buffer background.

2.6 Data analysis

Binding data (K_D and FRET experiments) were analyzed using Prism 5 (GraphPad Software) by nonlinear regression following the equation:

$$Y = \text{Bottom} + (\text{Top} - \text{Bottom}) / (1 + 10^{((\text{LogIC}_{50} - X) * \text{HillSlope}))}$$

K_D values were converted to ΔG° values as $\Delta G^\circ = RT \ln K_D$.

Salt dependence of K_D was analyzed by the following equation:

$$\text{Log}(K_D) = \text{log}(K_{\text{nel}}) - Z\psi \text{log}([KCl]) \quad (\text{Equation 1})$$

where K_{nel} is the dissociation constant at the standard state in 1 M KCl, Z is the number of ions displaced from the nucleic acid and ψ is the fractional probability of a counterion being thermodynamically associated with each phosphate of the RNA number of cations. K_{nel} and $Z\psi$ were treated as fitting parameters.

For thermodynamic analysis, ΔG° values were plotted versus T . Nonlinear regression using the three-parameter fit in Prism 5 was used to fit the following equation to the data: $\Delta G^\circ_T = \Delta H^\circ_{Tr} + \Delta C_p(T - Tr) - T\Delta S^\circ_{Tr} - T\Delta C_p \ln(T/Tr)$ (Equation 2) where Tr is a constant reference temperature (in our study $Tr = 293.15$ K), and the three fit parameters are ΔH°_{Tr} the change in enthalpy upon binding at Tr ; ΔS°_{Tr} , the change in entropy upon binding at Tr ; and ΔC_p , the change in heat capacity. ΔC_p was assumed

to be independent of temperature; inclusion of a $\Delta C_p/\Delta T$ term in the analysis did not improve the quality of the fits and gave larger standard errors for the returned parameters.

ΔH°_T and ΔS°_T were calculated from the result derived from the fitting of the curve ΔG° values versus T by the equations 3 using:

$$\Delta H^\circ_T = \Delta H^\circ_{Tr} + \Delta C_p(T - Tr) \quad (\text{Equation 3.1})$$

$$\Delta S^\circ_T = \Delta S^\circ_{Tr} + \Delta C_p \ln(T/Tr) \quad (\text{Equation 3.2})$$

where ΔH°_T is the change in enthalpy upon binding at T (25°C) and ΔS°_T , the change in entropy upon binding at T.

For compensation analyses, graphs of ΔH° values versus $T\Delta S^\circ$ ones, ΔG° vs ΔH° and ΔH° vs ΔC_p were plotted. Linear regression in Prism 5 was used to fit the following equations 4 to the data:

$$\Delta H^\circ_T = aT\Delta S^\circ_T + b \quad (\text{Equation 4-1})$$

$$\Delta G^\circ_T = a\Delta H^\circ_T + b \quad (\text{Equation 4-2})$$

$$\Delta H^\circ_T = aT\Delta C_p^\circ_T + b \quad (\text{Equation 4-3})$$

2.7 Molecular dynamics simulations

AutoDock Vina (Trott & Olson, 2010) was used for generating 40 initial complex conformations for each PAA using PDB ID 2kx5 as the receptor structure. Starting from the best scored docking poses a molecular dynamics system was prepared using explicit solvent for its equilibration and free energy evaluation. AM1-bcc partial charges were assigned to each PAA using *antechamber* from the Amber suite. Then, the complexes were prepared with *tleap* using the standard ff10 Amber force-field (AMBER12, University of California, San Francisco). Each structure was minimized in two steps: first, all residues except solvent were held fixed with a restraint force of 100 kcal/mol-

\AA^2 . Steepest descent minimization followed by conjugate gradient was performed using 2,500 steps in both cases. The same minimization protocol was applied in 10,000 steps without positional restraints. After minimization, the temperature was raised from 0 to 300 K in 100 ps using constant volume dynamics. A restraint force of 10 kcal/mol- \AA^2 was applied to the complex and SHAKE was turned on for bonds involving hydrogen atoms. Then, 200 ps of constant pressure dynamics were applied for density equilibration. Finally, each production run was performed using Langevin dynamics with a collision frequency of 1 ps⁻¹. An atom-based long range cutoff of 9 \AA was applied during all the simulations.

3. Results and Discussion

The promising results that we previously obtained with trimeric-C- α -PAA structures as Tat/TAR inhibitors prompted us to design longer structures in order to increase their efficiency and their activity (Pascale et al., 2013). Starting from one of the most promising compound, namely « C- α -FRF », we arbitrarily decided to elongate it from its N-terminal extremity to form a pentameric PAA, using three C- α -PAA monomers (R, K and F). As already noticed (Bonnard et al., 2010) the two basic residues (K and R) were selected for their ability to form hydrogen bonds and/or electrostatic interactions. Moreover, it is well-known that arginine-rich peptides and peptidomimetics incorporating guanidinium motives recognize TAR at the major groove, between the bulge and the loop, as does the basic domain of the Tat protein (⁴⁹RKKRRQRRR⁵⁷) (Blond et al., 2014; Massari et al., 2013; Richter & Palu, 2006). Furthermore, the aromatic F residue was selected for increasing π -stacking and van der Waals interactions. Thus, 9 penta-C- α -PAAs « **YX-FRF** » (**X** and **Y** = “R” or “F” or “K” PAA monomer) were synthesized (**Ia-c**, **IIa-c** and **IIIa-c**, Figure 1).

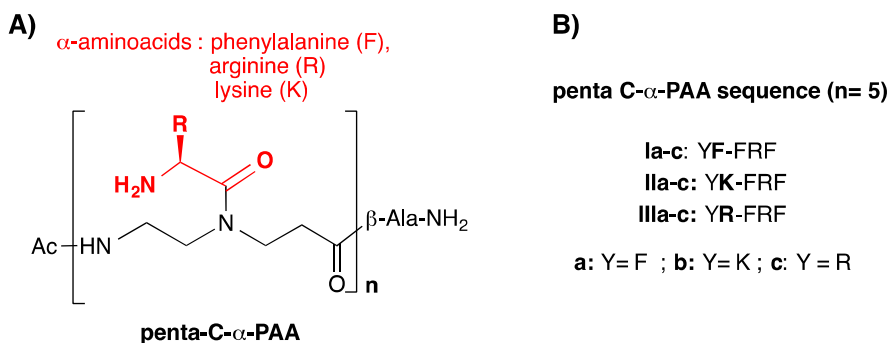
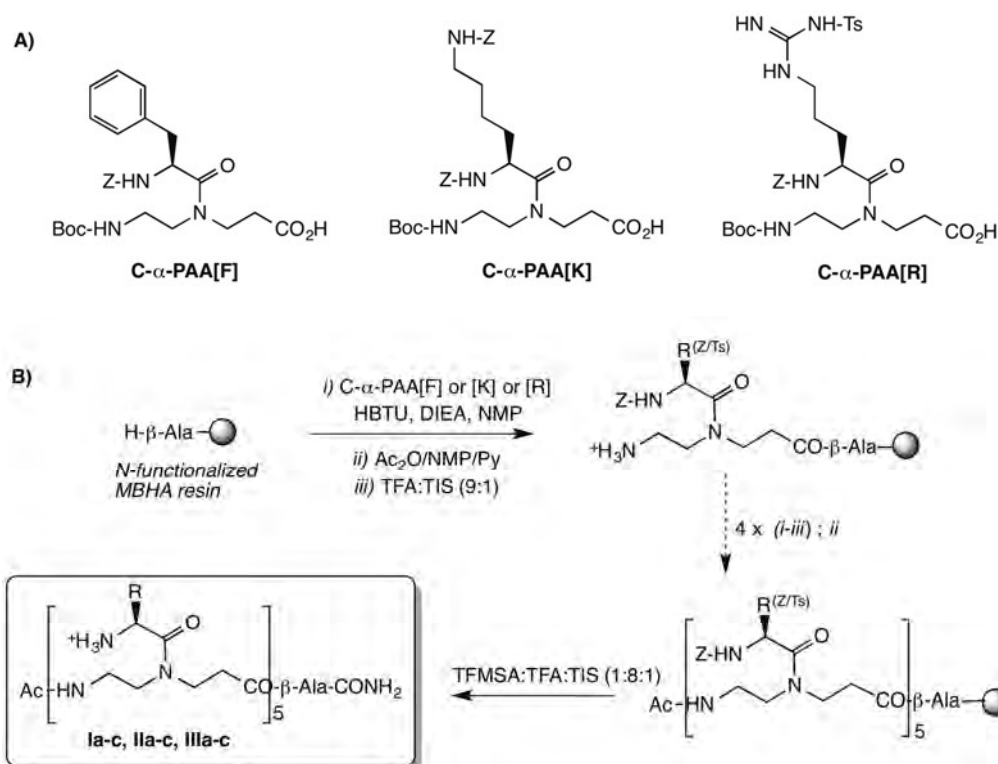


Figure 1. A) General structure of C- α -PAAs; B) penta-C- α -PAAs sequences

3.1 Chemistry

Penta-C- α -PAAs were prepared following a standard solid-phase strategy, using a β -alanine functionalized MBHA resin and starting from fully and orthogonally N-protected C- α -PAA monomers (Pascale et al., 2013) (Scheme 1A). This procedure includes: (i) successive elongation/deprotection steps involving selected monomers, (ii) acetylation of the last residue and (iii) acidic cleavage of the protecting groups and simultaneous release from the resin (Scheme 1B). Crude penta-C- α -PAAs were obtained in high HPLC yields (from 70 to 98%) and then purified by semi-preparative HPLC. Structures were confirmed by HRMS experiments (SI, Table S1).



Scheme 1. A) Protected C- α -PAA monomers used for B) the solid-phase synthesis of penta-C- α -PAAs

3.2 TAR affinity and FRET displacement assays

With the penta-C- α -PAAs in hand, we first evaluated TAR affinities (K_D) by monitoring the fluorescence change of a fluorescently labeled (Alexa 488) TAR fragment (18–44 nt), as previously described (Pascale et al., 2013) We also assessed their ability to displace a Tat fragment from a preformed TAR/Tat complex (IC_{50}) *via* a FRET assay, using a fluorescein-tagged Tat peptide fragment (amino acids 48–57) and a Dabcyl-labeled TAR fragment (18–44 nt) (Murchie et al., 2004) Dissociation constants (K_D) and inhibitory concentrations (IC_{50}) associated with each PAA are given in Table 1 and compared to the trimeric C- α -FRF precursor, used as a reference.

Table 1. TAR affinities (K_D), specificity (K_D' and K_D'') and inhibitory concentrations (IC_{50})^[a]

<i>C-α-PAA</i>	<i>n</i> ^o	<i>K_D</i> (nM)	<i>IC</i> ₅₀ (nM)	<i>K_D</i> ' (+ <i>tRNA mix</i> ^[c])	<i>K_D</i> '/ <i>K_D</i>	<i>K_D</i> '' (+ <i>dsDNA mix</i> ^[d])	<i>K_D</i> ''/ <i>K_D</i>
FFFRF	Ia	116.8	> 2000	190	1.7	176.2	1.5
KFFRF	Ib	79.6	> 2000	240	3	129.1	1.6
RFFRF	Ic	89.5	> 2000	320	3.6	169.3	1.9
FKFRF	IIa	78	1100				
KKFRF	IIb	77.8	593				
RKFRF	IIc	69.6	362				
FRFRF	IIIa	95.3	413				
KRFRF	IIIb	57.7	34.7	130	2.3	56	1
RRFRF	IIIc	48.6	101	100	2.1	64	1.3
FRF		260 ^[b]	> 2000 ^[b]		6.5 ^[b]		2.3 ^[b]

[a] All standard fluorescence measurements were performed in buffer A (20 mM HEPES (pH 7.4), 20 mM NaCl, 140 mM KCl and 3 mM MgCl₂) at 25°C. For clarity reason, incertitude values were not added but do not exceed 5%. [b] From ref (Pascale et al., 2013). [c] Measured in the presence of a 100-fold nucleotide excess of a mixture of natural tRNA (tRNAmix). [d] Measured in the presence of a 100-fold nucleotide excess of a 15-mer duplex DNA.

All penta-C-α-PAA strongly bind to TAR with similar affinities (48.6 nM < *K_D* < 116.8 nM) but they do not have the same ability to displace the Tat/TAR complex. Clearly, they can be divided into two distinct groups: the first one (series **I**: YF-FRF), « F-rich », has no ability to displace the Tat/TAR complex up to 2 μM, while the second one (series **II**: YR-FRF and series **III**: YK-FRF), « R/K-rich », displays low *IC*₅₀ values, except **IIa** which reveals an intermediate behavior. Thus, the nature of the penultimate residue **X** in YX-FRF pentamers has a major influence on *IC*₅₀ values. Similarly, the nature of the last PAA **Y** has also an impact on *IC*₅₀ values (see table 1, series **II** and **III**). Therefore, it seems that the N-terminal extremity of penta-PAA mainly governs their mode of interaction with TAR which drives or not the displacement of Tat protein. In addition, as previously reported with tri-PAA, there is no correlation between *K_D* and *IC*₅₀ values. It is also worth noting that even though the

increase in affinity is moderate (2 to 5 fold) compared to the original trimeric structure C- α -PAA (see table 1), the impact on IC₅₀ values is far more spectacular, from 67 to > 2000 times (in the case of the most active compound **IIIb**). This underscores the fact that the determination of a K_D value is not sufficient to predict the ability of a ligand to inhibit the interaction between an RNA fragment and its cognate partner.

To assess the specificity of penta-C- α -PAAs, K_D values of some representative compounds of series **I** and **II** were measured in the presence of a large excess of tRNA (K_D') and dsDNA (K_D'') (Table 1). All penta-PAA retain a specificity for TAR in the presence of both tRNA and DNA, this specificity being higher in the latter case. Moreover, it is noteworthy that penta-C- α -PAA are more specific than tri-C- α -PAA ($1.7 < K_D' / K_D < 6.5$; $1.8 < K_D'' / K_D < 2.4$) (Pascale et al., 2013), whatever their sequence. But as for trimers, penta-PAA containing F-rich sequences are less specific than PAA containing more cationic K/R-rich sequences (see PAA YFFRF *vs* PAA YRFRF **II** in table 1), likely indicating that cationic residues are preferentially involved in specific interactions rather than in ionic ones.

For better understanding the factors that account for the distinct behavior of penta-PAAs as Tat competitors, structural and thermodynamic studies were undertaken.

3.3 NMR experiments

Since tri-C- α -PAAs interact with TAR at the bulge level, it is likely that longer PAAs also bind TAR at the same site. However, we conducted NMR studies not only for verifying this hypothesis but also for comparing the mode of interaction for two PAAs, i.e. **Ic** and **IIIb**, representative of the two groups of PAAs (« F-rich » and « R/K-rich ») as defined in Tat competition assays.

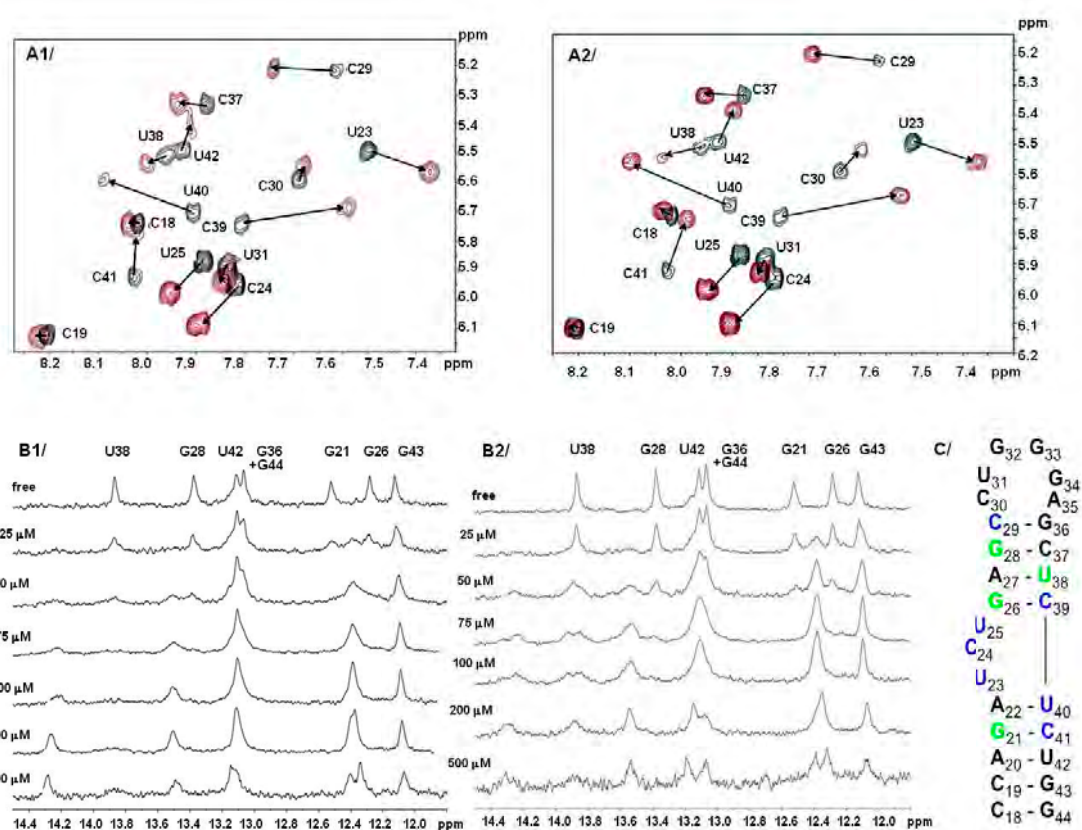


Figure 2.A/, *top*: 2D-TOCSY spectra showing pyrimidine H5–H6 cross-peaks for TAR. Black: free TAR (100 μ M); red: PAA/TAR complex with C- α -RFFRF **Ic** (A1/) or C- α -KRFRF **IIb** (A2/) at a ratio 5:1. Arrows indicate chemical shift changes on PAA binding. The spectra were acquired at 35 $^{\circ}$ C in a D₂O buffer (50 mM NaCl, 20 mM phosphate, pH 7.4). B/, *bottom*: stacked plot of 1D NMR spectra of the imino region of 50 μ M TAR RNA with increasing concentration of **Ic** (B1/) or **IIb** (B2/). Spectra were collected at 286K in a H₂O/D₂O (90/10) buffer (20 mM phosphate and 50 mM NaCl, pH 7.4). C/, *bottom*: secondary structure of the 27-mer TAR RNA fragment. Residues shown in blue are those exhibiting a pyrimidine H5–H6 proton chemical shift change (>0.1 ppm) upon addition of **Ic** or **IIb** (B2/). Residues in green are those experiencing an imino resonance change (>0.1 ppm).

The titration of TAR RNA with PAA **Ic** or **IIb** by 2D TOCSY experiments corroborated the specific interaction around the bulge for both compounds. Indeed, the overlaid TOCSY spectra with increasing amounts of **Ic** (Figure 2, A1/) or **IIb** (Figure 2, A2/) are very similar as they both show large chemical shift changes ($\Delta\delta \geq 0.1$ ppm)

for U23, C24, U25, C29, C39, U40 and C41, smaller changes ($\Delta d \leq 0.1$ ppm) for residues C30, C37, U38 and U42, and no significant changes for residues C18, C19 and U31 more remote from the bulge. The slight differences between the two ligands concern residues C30, U38, C41 and U42 which undergo larger chemical shift changes in **IIIb** than in **Ic** probably due to the higher affinity of **IIIb** for TAR. Comparison of the TOCSY experiments of penta-PAA with those previously reported for a tri-PAA sharing the same FRF sequence (**4b** in ref (Pascale et al., 2013)) did not show any significant differences except that the large chemical shift changes ($\Delta d \geq 0.1$ ppm) occur earlier (from 1eq.) for penta-PAA, probably reflecting once again the higher affinity of penta-PAA vs tri-PAA for TAR RNA.

Stacked 1D NMR spectra of the imino resonance region of TAR in the presence of increasing amount of **Ic** or **IIIb** (Figure 2, B1/ and B2/) are very similar and confirm the interaction of both compounds around the bulge as the imino protons of residues G21, G26, G28 and U38 undergo changes of chemical shifts ($\Delta d \geq 0.1$ ppm) upon addition of ligands. At 25 μ M, imino protons of these residues are in a slow rate of exchange at the NMR time scale. As the concentration of penta-PAA (**IIIb** and **Ic**) increases (75 μ M) the rate of exchange remains slow for G28 and U38 while imino signals of G26 and G21 become coalescent. If we compare the RNA imino resonances spectra obtained by the penta-PAA (**Ic** or **IIIb**) titration with the ones previously reported with the tri-PAA **4b** titration (Pascale et al., 2013) we can notice that penta-PAA binding seems to slow down the rate of exchange of the interactive imino protons. This could possibly reflect the higher affinity of the penta-PAA for TAR (≈ 10 -fold) as compared to the tri-PAA.

Overall, these NMR data unambiguously reveal that the binding site of both penta-PAA compounds (**Ic** and **IIIb**) is centered on the bulge region of TAR. Although

TOCSY experiments seem to indicate a stronger interaction for **IIIb** than **Ic** (some residues of **IIIb** display a larger chemical shift), these differences are not enough marked to explain the distinct ability of **Ic** and **IIIb** to compete with Tat.

3.4 - Circular Dichroism and UV melting studies

Even if penta-PAA's interact with TAR in the same region, their binding could induce distinct TAR conformational changes that could be related to their different ability to compete with Tat. Circular dichroism studies were undertaken to compare the TAR structural changes occurring upon binding of the two representative penta-PAA **Ic** and **IIIb** (SI Figure S1). For PAA alone, only negligible CD intensity in the 200-350 nm region was observed, suggesting the lack of structuration or secondary shape. By contrast, the CD spectrum of TAR alone is characteristic of an A-form double helix, with a strong positive band at 260 nm, very sensitive to base-stacking, and a strong negative band at 210 nm, related to the A-form of the TAR RNA helical structure (Loret et al., 1992). Addition of increasing concentration of PAA **Ic** or **IIIb** induces a strong reduction in ellipticity at 260 nm. This reflects that upon binding of **Ic** or **IIIb**, C24 and U25 residues of the bulge are exteriorized and concomitantly destacked as in the case of argininamide, Tat basic domain and Tat derived peptides (Tan & Frankel, 1992; Kumar & Maiti, 2013; Murchie et al., 2004). Nevertheless, the stronger decrease in intensity associated with a red shift displacement for the 210 nm band observed with **IIIb** compared to **Ic** could suggest a more pronounced effect of the latter on the TAR helicity (Davidson et al., 2011).

On the other hand, melting temperature studies conducted in the presence of **Ic** or **IIIb** at a 1:1 PAA/TAR molar ratio, shows that the binding of **IIIb** induces a slightly higher stabilization of the TAR structure than compound **Ic** (temperature increased by 4°C and 2°C upon binding, respectively (SI Figure S2). However, it is unlikely that this

small variation could account for the distinct ability of the two PAA to displace the Tat/TAR complex.

3.5 -Thermodynamic studies

Thermodynamic characterization is essential for understanding molecular interactions and their impact in biological processes (Chodera & Mobley, 2013; Lane & Jenkins, 2000; Pilch et al., 2003). To gain further insights on the TAR binding modes of penta-C- α -PAAs, thermodynamic binding profiles associated with each PAA/TAR equilibrium were determined. Enthalpy (ΔH°) and entropy changes (ΔS°) were calculated from equations 1 and 2 (see experimental section) after determination of ΔG°_T at several temperatures (278-308 K). We also determined the electrostatic (ΔG°_{el}) and non-electrostatic (ΔG°_{nel}) components of the Gibbs energy according to the polyelectrolyte theory (Record, Jr. et al., 1998). In the case of the two representative compounds **Ic** and **IIIb**, the K_D dependency on the ionic strength of the solution was studied over a range of KCl concentration from 70 to 250 mM. The results of thermodynamic analyses are summarized in table 2.

Table 2. Thermodynamic parameters for penta-C- α -PAA/TAR complexes^[a] at 25°C.

C- α PAA	<i>n</i> ^o	ΔG°	ΔH°	$T\Delta S^\circ$	ΔC_p	ΔG°_{el}	ΔG°_{nel}	$\Delta G^\circ_{nel}/\Delta G^\circ$
FFFRF	Ia	-39.6	-11.9	27.7	-0.75			
KFFRF	Ib	-40.4	-16.9	23.6	-1.33			
RFFRF	Ic	-40.2	-14.4	25.8	-1.10	-1.5	-38.6	96%
FKFRF	IIa	-40.5	-13.1	27.5	-0.99			
KKFRF	IIb	-40.6	-14.0	26.6	-1.1			
RKFRF	IIc	-40.8	-13.1	27.7	-1.03			
FRFRF	IIIa	-40.0	-17.2	22.8	-1.33			
KRFRF	IIIb	-41.3	-13.3	28.0	-1.07	-2.3	-39	95%
RRFRF	IIIc	-41.7	-11.3	30.4	-0.94			

a) ΔG° , ΔG°_{el} , ΔG°_{nel} , ΔH° and $T\Delta S^\circ$ are expressed in kJ/mol. ΔC_p is expressed in kJ/mol/K. For clarity reason, incertitude values were not added but do not exceed 5%.

At first glance, all penta-PAA/TAR equilibria display very similar thermodynamic signatures, with ΔG° varying from -39.6 to 41.7 kJ/mol. As for tri-C- α -PAA, non-electrostatic interactions dominate the binding by contributing to about 95% of the total binding free energy, whatever the PAA sequence (cf $\Delta G^\circ_{\text{nel}}/\Delta G^\circ$ ratio, table 2). This strongly suggests that the majority of ammonium and guanidinium groups of penta-PAA interact with TAR RNA *via* specific hydrogen bonding and/or π -cation interactions rather than *via* ionic interactions with the phosphate backbone.

Equilibria are entropically driven and enthalpically favored, ΔH° and $T\Delta S^\circ$ values lying in a small interval, from -17.2 to -11.3 kJ/mol and from 22.8 to 30.4 kJ/mole, respectively. ΔH° and $T\Delta S^\circ$ are quite well correlated ($R^2 = 0.93$, Figure 3), highlighting an entropy-enthalpy compensation (EEC) phenomenon (Chodera & Mobley, 2013) which results in Gibbs energies (ΔG°) included in a small interval (from -41.7 to -39.6 kJ/mol) (see supporting information for comment on statistical relevance). ΔC_p values for all PAA/TAR complexes range from -1.33 to -0.75 kJ/mol/K, falling into the range of -2.3 to -0.4 kJ/mol/K (100-150 kcal/mol/K), typically observed in the case of ligand–nucleic acid interactions (Islam et al., 2009; Pilch et al., 2003; Stolarski, 2003).

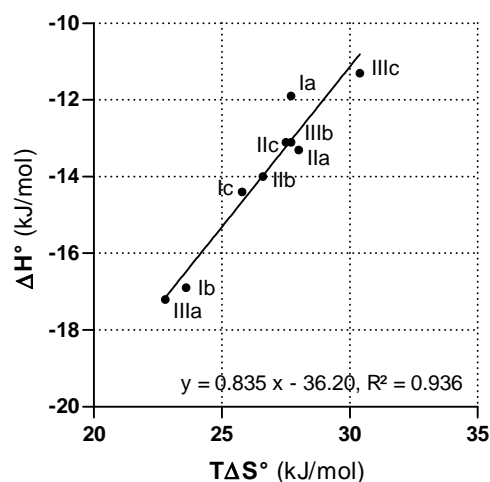


Figure 3. Entropy/enthalpy compensation in penta-PAA series. For clarity reason, error bars were omitted, but incertitude values do not exceed 5%.

A closer analysis reveals that each series **I-III** constitutes in fact a three ligands-set distinct from the others. Indeed, examination of the following relationships: (ΔH° vs $T\Delta S^\circ$), (ΔG° vs ΔH°) and (ΔH° vs ΔC_p) within each series (Figures 4A-C) points out strong correlations for two of them (**I** and **III**). Undoubtedly, the thermodynamic signatures of series **I** and **III** differ, proving that their interaction mode is distinct. Thus for example, in series **I**, the ligand of highest affinity (**Ib**) corresponds to the tighter one (the lower ΔH° and $T\Delta S^\circ$ of the series) whereas in series **III**, the looser ligand (**IIIc**, the higher ΔH° and $T\Delta S^\circ$ of the series) displays the best affinity. In these two series, the correlation between ΔH° and ΔC_p demonstrates that a part of the heat capacity change reflects the degree of tightness of the complex. Thus, a loosening in the structure complex (ΔH° increase) is associated with a ΔC_p increase.

Concerning series **II**, ΔH° and $T\Delta S^\circ$ values range in small intervals, as compared with the two other series. No linear correlation could be deduced, maybe because in all cases, compound **IIc** caused deviations. However, from a thermodynamic point of view, series **II** is closer to series **III** than to **I**.

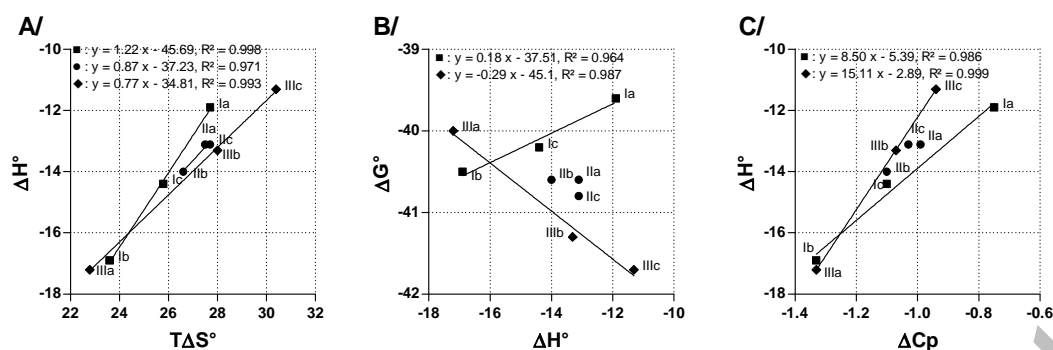


Figure 4. Variations of the enthalpy of binding (ΔH°) with A) the entropy of binding ($T\Delta S^\circ$); B) the Gibbs energy (ΔG°); C) the heat capacity change (ΔC_p) for the three PAA series (I, II and III). For clarity reason, error bars were omitted, but incertitude values do not exceed 5%.

As all penta-C- α -PAAs derive from the same C-terminal FRF sequence, these observations clearly demonstrate that the nature of the N-terminal residues has a strong impact on their TAR recognition mode and, consequently, on their ability to displace the Tat/TAR complex. Thus, compounds of series I, unable to displace the Tat/TAR complex ($IC_{50} > 2000$ nM), have a thermodynamic behavior which clearly differs from R/K-rich compounds of series II and III that are effective Tat competitors ($IC_{50} < 400$ nM). On the other hand, in the case of I and III, the nature of the last PAA residue Y has an impact on ΔH° and $T\Delta S^\circ$ values, whereas it has only little effect in II.

For a better comprehension on the comparative binding modes of the two PAA models, i.e. “RFFRF” **Ic** and “KRFRF” **IIIb**, we synthesized two truncated derivatives devoid of the F- β -alaninamide moiety, namely tetra-C- α -PAA “RFFR” and “KRFR”, and compared their thermodynamic profiles with **Ic** and **IIIb**, respectively (Table 3).

Table 3. Thermodynamic parameters for tetra-C- α -PAA/TAR complexes^[a] and comparison between penta and tetra C- α -PAA

tetra-C- α -PAA	ΔG°	ΔH°	$T\Delta S^\circ$	ΔC_p
------------------------	------------------	------------------	-------------------	--------------

RFFR	-39.8	-20.6	19.1	-1.57
KRFR	-37.0	-56.7	-19.8	-3.54
PAA vs PAA	$\Delta(\Delta G^\circ)$	$\Delta(\Delta H^\circ)$	$\Delta(T\Delta S^\circ)$	$\Delta(\Delta C_p)$
KRFR/RFFR	-2.7	36.1	38.9	1.97
RFFR/RFFRF Ic	-0.4	6.2	6.7	0.47
KRFR/KRFRF IIIb	-4.3	43.5	47.8	2.48

[a] ΔG° , ΔH° and $T\Delta S^\circ$ are expressed in kJ/mol. ΔC_p is expressed in kJ/mol/K. For clarity reason, incertitude values were not added but do not exceed 5%.

Even if the TAR affinity of RFFR (F-rich) is higher than that of KRFR (K/R rich), its RNA complex is considerably looser than the one formed with KRFR, as demonstrated by ΔH° , $T\Delta S^\circ$ and ΔC_p values. This is in line with our previous results concerning F-rich and K/R-rich tri-PAA and likely means that for F-rich PAAs, interactions are fewer and/or less optimized than for K/-rich PAA. Lengthening the loose tetra-RFFR binder by adding the F- β -alaninamide moiety (leading to RFFRF **Ic**) fails to maximize interactions and slightly destabilizes the complex, as demonstrated by the enthalpy loss and the entropy gain of 6.2 and 6.7 kJ/mol, respectively. By contrast, adding the F- β -alaninamide moiety to the tight tetra-KRFR ligand (leading to KRFRF **IIIb**) induced a considerable enthalpy loss of 43.4 kJ/mol, overbalanced by an entropy gain of 47.6 kJ/mol. We rationalized this thermodynamic behavior by hypothesizing that only the N-terminal tetrameric sequence KRFR tightly interacts with TAR in the interaction site, whereas the C-terminal moiety (i.e. F- β -alaninamide) remains free and exposed to the solvent. In such case, the mobility of this unbound fragment would result in the destabilization of the whole interaction network, leading to a decrease in tightness at the ligand/RNA interface. This phenomena would result in less unfavorable $T\Delta S^\circ$ (due to greater mobility) and less favorable ΔH° (due to weaker interactions) than for the corresponding tetra-PAA, as observed here. To support these hypotheses, we

Accepted Manuscript

performed a molecular modeling study concerning the TAR interaction of these two penta-PAAAs.

3.5 - TAR molecular recognition of penta-C- α -PAA **Ic** and **IIIb**

Molecular dynamics (MD) is a powerful tool to explore molecular complexes formation at an atomistic level and analyze the conformational energy landscape accessible to these molecules. Such protocols have been applied in this study in order to predict the interaction modes between C- α -PAA and TAR that may complement experimental data. Visual inspection of MD results is in good agreement with NMR data, which suggested a preferred binding site in the bulge region of TAR for both **Ic** and **IIIb** compounds. However, the orientation differs from both complexes: **IIIb** points towards the inter-helical junction of the bulged TAR, whereas **Ic** provides a more packed conformation and buries into the TAR major groove similarly to previously reported structures (Davidson et al., 2009; Davidson et al., 2011). The dynamic behavior of **Ic** and **IIIb** complexes along the MD is noteworthy. Analyses of the full trajectory suggest that **Ic** widens the bulge region up to 15.2 Å while **IIIb** produces a contraction of this region up to 9.0 Å with subsequent RNA helicity loss (Figure 5A). According to these results, both compounds induce non-negligible TAR conformational changes, as pointed out previously by CD spectra. In addition to major structural changes of TAR upon ligand binding, MD confirms the positive influence of the N-terminal sequence since F- β -alaninamide terminal moiety remains exposed to solvent during almost all the trajectory. Particularly, molecular modeling results suggest a remarkably difference in the exposition to solvent of the F- β -alaninamide moiety in PAA **Ic** and **IIIb**, due to the conformational changes in the RNA previously described. Moreover, no specific interactions have been characterized for this unbound fragment that might contribute to

ΔH° .

A closer inspection of the complex trajectories suggests that specificity is clearly achieved by side-chain interactions based on π stacking, cation- π stacking and charge-based interactions as well as the orientation of the compound along the major groove (Figure 5B). Compound **Ic** buries into the major groove of TAR by means of intimate charge-based contacts. Interestingly, U25 is flipped out from the RNA complex and drowns down the -hole UCU loop. This is counterbalanced by C24 stabilization into the bulge region through a F2 π -stacking interaction and R4 charge-based interaction with the phosphate group of U23. R4 also yields intramolecular contacts with F3 via a cation- π stacking interaction. This interaction is reproduced by R1 with the stack of the guanidinium group on top of C30, which significantly contributes to complex stabilization. Despite the sequence dissimilarity between compounds **Ic** and **IIIb**, both provide similar interactions with TAR. For instance, C24 stacks with R2 via a cation- π interaction, while the acetyl moiety buries into the major groove of the apical loop and interacts with the 2'OH group of U23 *via* hydrogen bonding. The same interaction is hypothesized for R4 through the sugar moiety of G33. Moreover, carbonyl and ammonium groups of penta-PAA backbone significantly contribute to binding through electrostatic interactions. For instance, K1 provides a charge-based interaction with the RNA backbone and F3 anchors to the C30/U31/G34/G36 cluster.

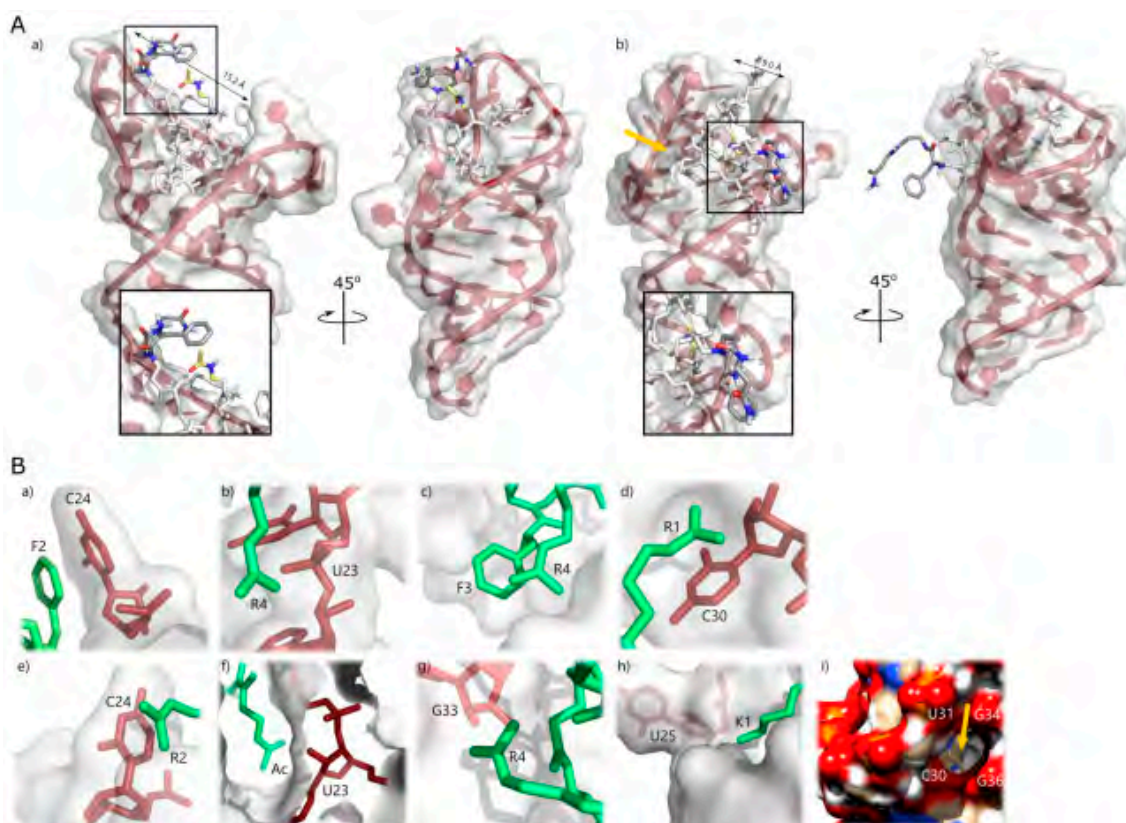


Figure 5. (A) Final MD frame representations of **Ic** complex (a) and **IIIb** (b). F- β -alaninamide and Acetyl terminus (in yellow) are framed. In the two cases, F- β -alaninamide is free of interaction. For **IIIb**, it remains exposed to the solvent while the rest of the penta-PAA buries into the TAR major groove. The bulge closure distance has been calculated as N2(G34)-N4(C24) distance for **Ic** and N2(G32)-N4(C24) distance for **IIIb**. (B) Main interactions provided by MD simulations for **Ic** (a to d) and **IIIb** (e to i). Henceforth, PAA are numbered as Ac-R1-F2-F3-R4-F5 and Ac-K1-R2-F3-R4-F5. HIV-1 TAR is represented as a white surface with red nucleotides and PAA are represented in green. On the one side, **Ic** interacts (a) with C24 into the bulge region through a F2 π -stacking interaction. R4 interacts at the same time with (b) phosphate U23 group and (c) F3 via intra- residue cation- π stacking interaction. (d) Arginine R1 stacks with C30 providing another cation- π interaction which stabilizes the complex. On the other side, **IIIb** (e) stacks with C24 via a R2 cation- π interaction. (f) Ac group buries into the major groove and interacts with the 2'OH group of U23. (g) G33 also provides a hydrogen bond interaction with R4 through the 2'OH group. (h) K1 provides charge-based interactions with the RNA backbone. (i) F3-NH₃⁺ group (yellow arrow) anchors to the TAR opening, which contains a high negative charge density.

To conclude, these molecular modeling studies validate our assumptions based on experimental data. Thus, contrary to neomycine B, which binds in the minor groove at the junction between the bulge and the lower stem of TAR and allosterically inhibits Tat binding (Lu et al., 2011), both penta-PAA **Ic** and **IIIb** recognize TAR at the major groove, between the bulge and the loop, as does the arginine-rich domain of the Tat protein (⁴⁹RKKRRQRRR⁵⁷). This is not surprising since **Ic** and **IIIb** contain at least one arginine residue and it is well-known that even a single argininamide residue has a specificity for this binding pocket (Frankel, 1992; Long & Crothers, 1995; Olsen et al., 2005). Upon binding of both PAA **Ic** and **IIIb**, the bulge nucleotides become unstacked, as observed in the case of arginamide, Tat basic domain and Tat derived peptides (Bardaro, Jr. et al., 2009; Davidson et al., 2011). However, while the binding of the tight ligand **IIIb** induces a strong contraction at the bulge region, the loose one, **Ic**, widens it. This may account for their distinct ability to inhibit the Tat /TAR complex. Indeed, **IIIb** is a strong antagonist of Tat ($IC_{50} = 35$ nM) whereas **Ic** has no ability to displace Tat from a preformed Tat/TAR complex until 2 μ M. Our results also demonstrate that in the case of penta-PAA **IIIb**, only the tetrameric N-terminal moiety interacts tightly with RNA, while the C-terminal part, *i.e.* F- β -alaninamide, is excluded from the interaction site and stays exposed to the solvent. Therefore, it appears that R-rich PAA tetramers would be of optimum length to allow a tight interaction with TAR and to compete efficiently with the Tat fragment for TAR binding.

4. Conclusion

Many RNA ligands have been developed in view of inhibiting the interaction of a target RNA with its cognate partner. Studies related to RNA/ligand interactions are often

limited to the determination of ligand affinity and/or of ligand ability to inhibit the natural complex. However, these studies do not provide insights into structure-activity relationships, which are important for a rational design of specific RNA ligands.

In the present work, increasing the length of previously studied trimeric C- α -PAAs to pentameric structures led in all cases to an increase in TAR affinity as one would expect but surprisingly the ability to compete with Tat was strongly improved only in the case of R/K rich-PAA. While NMR, CD and stability studies could not reveal any significant difference in the TAR binding of penta-PAAs, only a comparative analysis of thermodynamic profiles allowed us to highlight different binding features in the PAA series. This led us to propose two distinct interaction modes for F-rich and K/R-rich penta-PAA that were validated by molecular modeling studies. According to these computational models, distinct conformational changes would occur upon binding of the two kinds of ligands that could result in different ability to displace Tat fragment. Our PAA study illustrates how comparative thermodynamic and structural studies of a series of RNA ligands of same chemical family are of value for understanding their binding modes and for rationalizing structure-activity relationships. Such studies should be expended in this challenging research area to improve both the design of new ligands and the knowledge of their interactions.

On the other hand, although the conformational flexibility of PAA is well-suited for the induced-fit mechanism by which RNA recognition occurs, it makes structure prediction difficult and may induce a lack of specificity for the target RNA. So, based on all these new results, it would be interesting to reduce the flexibility of the PAA polyamide backbone and to study the influence on the TAR binding of this conformational restriction. Works are in progress and results will be reported in due course.

Supplementary data. Supplementary Tables (Table S1), Supplementary Figures (figures S1-S2) and Supplementary Methods (NMR experiments, molecular modelling protocols).

Reference List

- Bardaro, M. F., Jr., Shajani, Z., Patora-Komisarska, K., Robinson, J. A., & Varani, G. (2009). How binding of small molecule and peptide ligands to HIV-1 TAR alters the RNA motional landscape. *Nucleic Acids Res*, 37(5), 1529-1540.
- Blond, A., Ennifar, E., Tisne, C., & Micouin, L. (2014). The design of RNA binders: targeting the HIV replication cycle as a case study. *ChemMedChem.*, 9(9), 1982-1996.
- Bonnard, V., Pascale, L., Azoulay, S., Di, G. A., Rogez-Kreuz, C., Storck, K. et al. (2010). Polyamide amino acids trimers as TAR RNA ligands and anti-HIV agents. *Bioorg.Med.Chem.*, 18(21), 7432-7438.
- Chodera, J. D., & Mobley, D. L. (2013). Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annu.Rev.Biophys.*, 42, 121-142.
- Davidson, A., Leeper, T. C., Athanassiou, Z., Patora-Komisarska, K., Karn, J., Robinson, J. A. et al. (2009). Simultaneous recognition of HIV-1 TAR RNA bulge and loop sequences by cyclic peptide mimics of Tat protein. *Proc.Natl.Acad.Sci.U.S.A*, 106(29), 11931-11936.

Davidson, A., Patora-Komisarska, K., Robinson, J. A., & Varani, G. (2011). Essential structural requirements for specific recognition of HIV TAR RNA by peptide mimetics of Tat protein. *Nucleic Acids Res.*, 39(1), 248-256.

Disney, M. D., Yildirim, I., & Childs-Disney, J. L. (2014). Methods to enable the design of bioactive small molecules targeting RNA. *Org.Biomol.Chem.*, 12(7), 1029-1039.

Frankel, A. D. (1992). Peptide models of the Tat-TAR protein-RNA interaction. *Protein Sci.*, 1(12), 1539-1542.

Guan, L., & Disney, M. D. (2012). Recent advances in developing small molecules targeting RNA. *ACS Chem.Biol.*, 7(1), 73-86.

Islam, M. M., Pandya, P., Kumar, S., & Kumar, G. S. (2009). RNA targeting through binding of small molecules: Studies on t-RNA binding by the cytotoxic protoberberine alkaloid coralyne. *Mol.Biosyst.*, 5(3), 244-254.

Kumar, S., & Maiti, S. (2013). Effect of different arginine methylations on the thermodynamics of Tat peptide binding to HIV-1 TAR RNA. *Biochimie*, 95(7), 1422-1431.

Lane, A. N., & Jenkins, T. C. (2000). Thermodynamics of nucleic acids and their interactions with ligands. *Q.Rev.Biophys.*, 33(3), 255-306.

Long, K. S., & Crothers, D. M. (1995). Interaction of human immunodeficiency virus type 1 Tat-derived peptides with TAR RNA. *Biochemistry*, 34(27), 8885-8895.

Loret, E. P., Georgel, P., Johnson, W. C., Jr., & Ho, P. S. (1992). Circular dichroism and molecular modeling yield a structure for the complex of human

Accepted Manuscript

immunodeficiency virus type 1 trans-activation response RNA and the binding region of Tat, the trans-acting transcriptional activator.

Proc.Natl.Acad.Sci.U.S.A, 89(20), 9734-9738.

Lu, J., Kadakkuzha, B. M., Zhao, L., Fan, M., Qi, X., & Xia, T. (2011). Dynamic ensemble view of the conformational landscape of HIV-1 TAR RNA and allosteric recognition. *Biochemistry*, 50(22), 5042-5057.

Massari, S., Sabatini, S., & Tabarrini, O. (2013). Blocking HIV-1 replication by targeting the Tat-hijacked transcriptional machinery. *Curr.Pharm.Des*, 19(10), 1860-1879.

Murchie, A. I., Davis, B., Isel, C., Afshar, M., Drysdale, M. J., Bower, J. et al. (2004). Structure-based drug design targeting an inactive RNA conformation: exploiting the flexibility of HIV-1 TAR RNA. *J.Mol.Biol.*, 336(3), 625-638.

Olsen, G. L., Edwards, T. E., Deka, P., Varani, G., Sigurdsson, S. T., & Drobny, G. P. (2005). Monitoring tat peptide binding to TAR RNA by solid-state ³¹P-19F REDOR NMR. *Nucleic Acids Res*, 33(11), 3447-3454.

Pascale, L., Azoulay, S., Di Giorgio, A., Zenacker, L., Gaysinski, M., Clayette, P. et al. (2013). Thermodynamic studies of a series of homologous HIV-1 TAR RNA ligands reveal that loose binders are stronger Tat competitors than tight ones. *Nucleic Acids Res.*, 41(11), 5851-5863.

Pilch, D. S., Kaul, M., Barbieri, C. M., & Kerrigan, J. E. (2003). Thermodynamics of aminoglycoside-rRNA recognition. *Biopolymers*, 70(1), 58-79.

- Poehlsgaard, J., & Douthwaite, S. (2005). The bacterial ribosome as a target for antibiotics. *Nat.Rev.Microbiol.*, 3(11), 870-881.
- Record, M. T., Jr., Zhang, W., & Anderson, C. F. (1998). Analysis of effects of salts and uncharged solutes on protein and nucleic acid equilibria and processes: a practical guide to recognizing and interpreting polyelectrolyte effects, Hofmeister effects, and osmotic effects of salts. *Adv.Protein Chem.*, 51, 281-353.
- Richter, S. N., & Palu, G. (2006). Inhibitors of HIV-1 Tat-mediated transactivation. *Curr.Med.Chem.*, 13(11), 1305-1315.
- Stevens, M., De, C. E., & Balzarini, J. (2006). The regulation of HIV-1 transcription: molecular targets for chemotherapeutic intervention. *Med.Res.Rev.*, 26(5), 595-625.
- Stolarski, R. (2003). Thermodynamics of specific protein-RNA interactions. *Acta Biochim.Pol.*, 50(2), 297-318.
- Suryawanshi, H., Sabharwal, H., & Maiti, S. (2010). Thermodynamics of peptide-RNA recognition: the binding of a Tat peptide to TAR RNA. *J.Phys.Chem.B*, 114(34), 11155-11163.
- Tan, R., & Frankel, A. D. (1992). Circular dichroism studies suggest that TAR RNA changes conformation upon specific binding of arginine or guanidine. *Biochemistry*, 31(42), 10288-10294.

Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J.Comput.Chem.*, 31(2), 455-461.

Yang, M. (2005). Discoveries of Tat-TAR interaction inhibitors for HIV-1. *Curr.Drug Targets.Infect.Disord.*, 5(4), 433-444.

Accepted Manuscript