UNIVERSIDAD AUTÓNOMA DE SAN LUIS POTOSÍ

DOCTORADO INSTITUCIONAL EN INGENIERÍA Y CIENCIA DE MATERIALES

"Analysis of the separation between the ends of mRNA molecules from different organisms by using computational algorithms and smFRET"

Tesis que presenta:

M. en C. Nancy Anabel Gerling Cervantes

Para obtener el grado de Doctor en Ingeniería y Ciencia de Materiales

Directores de tesis:

Dr. Jaime Ruiz García y Dr. Eduardo Gómez García

San Luis Potosí, México.

2024

Table of Contents

Abstract

Considering that the innate proximity of RNA ends might have important unknown biological implications, we aimed to determine whether the close proximity of the ends of mRNA molecules is a conserved feature across organisms and gain further insights into the functional effects of the proximity of RNA ends.

We present two projects in this thesis; the first one comprises the study of the secondary structure of 274 full native mRNA molecules from 17 different organisms to calculate the contour length (C_L) of the external loop as an index of their end-to-end separation. Our computational predictions show bigger variations than previously reported and also than those observed in random sequences. From this project, we found that our results suggest that separations larger than 18.5 nm are not favored, whereas short separations could be related to phenotypical stability. Overall, the results obtained implies the existence of a biological mechanism responsible for the increase in the observed variability, suggesting that the C_L features of the exterior loop could be relevant for the initiation of translation, and that a short C_L could contribute to the stability of phenotypes. The second one, comprise the single molecule Fluorescence Resonance Energy Transfer (smFRET) system design and calibration to perform the experimental *in vitro* measurements of the distance between mRNAs ends from 4 organisms from the Eukarya domain. From the second project, we obtained only preliminary results, and some experiments are pending to be performed.

Keywords: Contour length, DNA, mRNA, phenotypic stability, RNA external loop, RNA secondary structure, smFRET.

Acknowledgements

First, I would like to thank my main supervisors Dr. Jaime Ruíz García and Dr. Eduardo Gómez García for giving me the opportunity to collaborate in their research group and for all the support they gave me during this study. I also want to give thanks to Dr. José Alfredo Méndez Cabañas for all the valuable advice and for helping me with my research study.

I am grateful to Dr. Elizabeth Reynaga Hernández for teaching me molecular biology techniques.

Thanks to Emmanuel Vazquez Martínez for his help with technical support, and also thanks to my lab co-workers for their help and motivation. I likewise give thanks to Dr. Pablo Luis Hernández Adame and Dr. Joan Anto for their help during the FRET optical system design and calibration.

I would like to thank CONACYT for support with the PhD fellowship.

Finally, I would also express my gratitude to my family, specially to my mom Elena and my mother-in-law Araceli who help me to carry my daughter Giselle while I was doing my research work. To my loved husband Andrés for all the support that he gives me over these years.

Dedication

I lovingly dedicate this thesis to the memory of my beloved father, Edmundo Gerling, who always encouraged me to do more and to believe in myself.

Although I cannot see you anymore, you will be always in my thoughts.

Until we met again,

Your daughter who misses you so much

Nancy Gerling.

CHAPTER 1. Introduction

Chapter 1 introduces some basic concepts and background about RNA molecules and previous studies performed about the close proximity between RNA ends. Finally, the objectives of the present thesis are mentioned.

1.1. Structure and function of RNA

Ribonucleic acid (RNA) is an essential biopolymer present in all forms of life. The RNA is an elastic chain composed by nucleotides (nt) consisting of a nitrogenous base, a pentose (ribose) and a phosphate group. In the RNA molecules are present four nitrogenous bases, two purines: adenine (A) and guanine (G), and two pyrimidines: uracil (U) and cytosine (C). The chemical components of the nitrogenous bases determine the interaction between them and between the RNA backbone. The RNA backbone comprises the phosphate group and the sugar and is always synthetized in the 5′ → 3′ direction. The directionality of ssRNA molecules is referred to the end to end chemical orientation, in which the 5′ end of the chain carried one free phosphate group attached to the 5′ carbon atom of the ribose sugar and the 3′ end of the chain carried one free hydroxyl group at the 3′ carbon atom of the sugar [1] (see Fig.1.1.1).

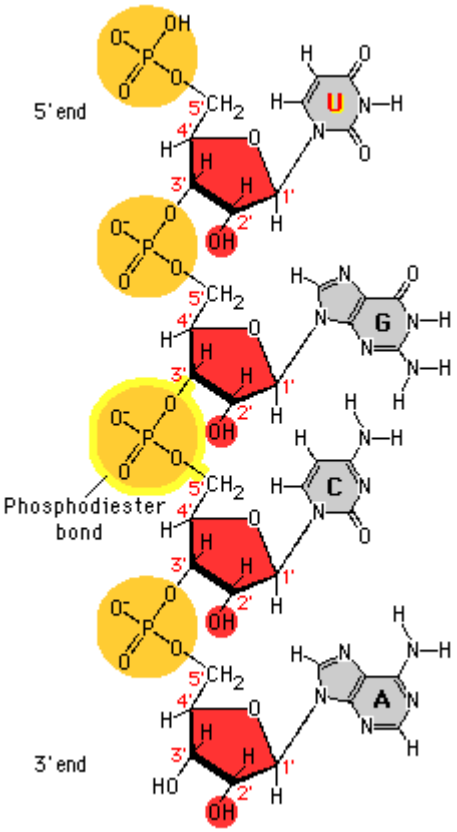

Figure 1.1.1. Structure of the RNA polynucleotide chain. Figure adapted from [2].

The RNA molecules participate in several key cellular functions [3], such as catalysis, splicing, regulation of both transcription and translation [3, 4]. They also help to maintain the telomers and protect against viruses [5, 6]. To perform all these functions, RNA molecules need to fold into complex secondary and tertiary structures (Fig. 1.1.2), and base paired regions are formed to increase the conformational stability of single-stranded RNA (ssRNA) molecules. The simplest secondary structures are formed by base pairing between distant complementary segments in ssRNA, such as hairpins and stem-loops (Fig. 1.1.2 A). The single stranded loop formed between the base pair helical hairpin is much shorter than the one formed in stem-loop. These simple secondary structures can fold into more complicated tertiary structures leaving the 5′

and 3′ ends of the molecule loose [7]. For example, the pseudoknots are one type of tertiary structure formed by the interaction of secondary loops through base pairing between complementary bases (see Fig. 1.1.2 B green and blue) [8-10].

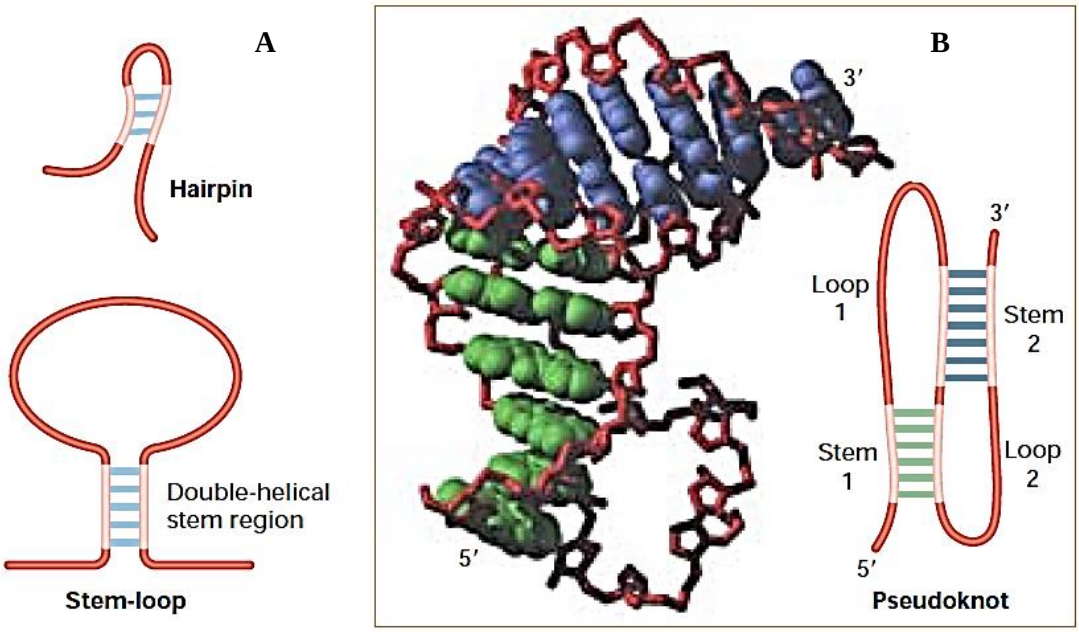

Figure 1.1.2. ssRNA folding conformations. (A) schematic representation of a simple secondary structures and (B) tertiary structures. Figure adapted from [10].

1.2. The ends of RNA molecules are close to each other

As it was mentioned in the previous section, the ssRNA molecules tend to fold into secondary and tertiary structures providing thus an effective circularization, where the ends of the molecules are in close proximity. In previous studies, Yoffe *et al.* [7] performed a theoretical analysis of long random ssRNA sequences, complemented by mfold and Vienna RNA computational algorithms, to show that the contour length of the exterior loop (the unpaired loop that contains both ends of the RNA molecules) remains

small (~ 12 nt links or ~ 7 nm) regardless of the overall nucleotide sequence or length [7]. It is known that there is a degeneration in base pair prediction accuracy that increase with increasing the sequence length [11-13]. Despite of this, single sequence secondary structure prediction is reasonably accurate with this RNA folding programs [14]. Different probabilistic models give similar results [15-17]. More recently, Leija-Martínez *et al.* [18], using single molecule Fluorescence (Förster) Resonance Energy Transfer (smFRET) *in vitro*, measured the end to end distance in the range of 5 to 9 nm between the 5′ and 3′ ends of several mRNA molecules from a fungus and two viruses. This range corresponds to an exterior loop contour length of 9 to 16 nt with a weak dependence on the molecule length but independent of its origin and secondary or tertiary structures [18]. Whether this closeness between the ends of an ssRNA molecule is a general feature in all forms of life remains to be proved. Nonetheless, the proximity of the ends in mRNA works as an effective circularization, which is thought to be important for proper regulation and translation of RNA into proteins [7, 18]. Such effective circularization has already been shown to be important on the interaction of RNA-binding proteins within cis-motifs on the 5′- and 3′- untranslated regions (UTRs) of mRNA molecules [19-22], and on the ability of the 5′ UTR to regulate translation initiation by recruiting translation factors [21-23]. Examples include the effective circularization of mRNA in yeast, where the poly-A bound PABP (poly-A binding protein) which interacts with the eukaryotic translation initiation factor eIF4G and then interacts with the translation initiation factor eIF4E (see Fig. 1.2.1) [20].

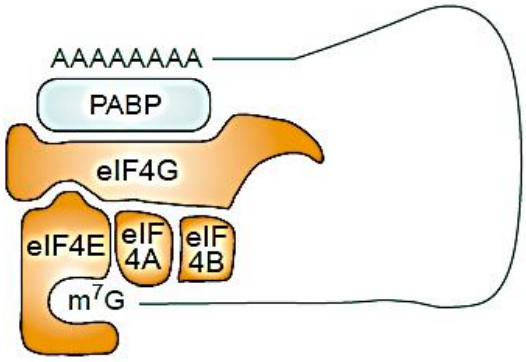

Figure 1.2.1. Schematic representation of mRNA translation protein interactions. The protein PABP mediate the interaction of the eukaryotic translation initiation factor (eIF4G) and the poly-A tail in the capped 5′ end. Figure adapted from [20].

In neurons, the protein CPEB1 mediate the translation repression and activation by binding to a specific cis-element in the 3′-UTR (see Fig. 1.2.2). Through a specific binding of translation initiation factors (eIFs), CPEB1 forms a complex with the 5′-cap maintaining the mRNA translationally latent. When the CPEB1 is phosphorylated, one gets the recruitment of CPSF to the hexanucleotide sequence (HEX) and the polyadenylation by Gld2. This polyadenylation helps to override the MASKIN inhibition allowing the eIF4G bind to the eIF4E and thus, activate the initiation of translation [24].

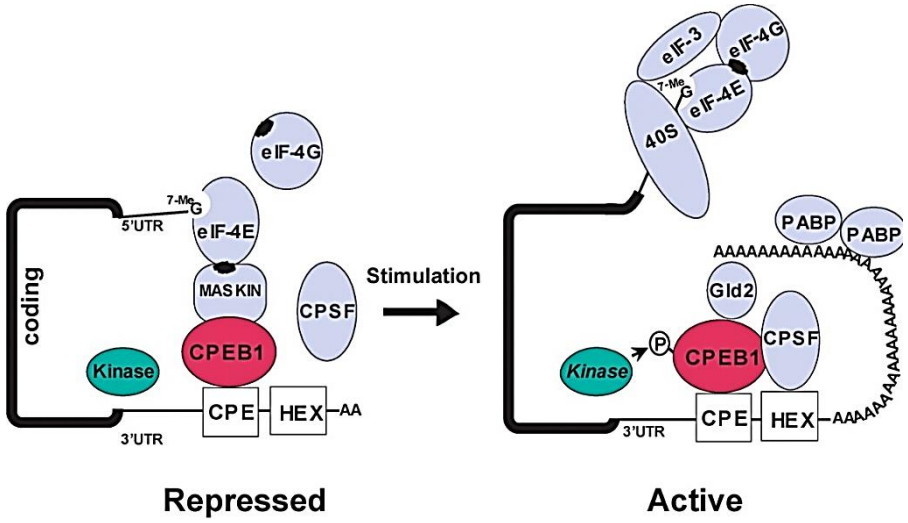

Figure 1.2.2. Schematic representation of mRNA translation repression and activation mediated by CPEB1. Repressed state is achieved by the MASKIN interaction with eIF4E, leading the mRNA in a translationally latent state. The CPEBI phosphorylation activate the recruitment of CPSF that allow the activation of the initiation of translation. Figure adapted from [24].

1.3. Objectives of the thesis

Considering that the innate proximity of RNA ends might have important unknown biological implications, the main objective of the thesis was to determine whether the close proximity of the ends of native mRNA molecules is a conserved feature and if has a constant value regardless of the organism. Also, investigate about if there is a biological role contributing to the close proximity between the mRNA ends.

To achieve the main objective, we set the following particular objectives:

1. Perform *in silico* measurements of the distance between native mRNA ends from 17 different organisms by using computational programs (mfold and Vienna RNA) and compare them with those obtained by random-generated sequences.
2. Perform *in silico* and *in vitro* measurements of the distance between mRNA ends from 4 organisms from the Eukarya domain by using computational programs (mfold and Vienna RNA) and smFRET.

CHAPTER 2. Study of the separation between the ends of native mRNA molecules by using bioinformatic analysis and computational algorithms

In this chapter, it is described the methodology used to select and to predict the distance between ends of native mRNA sequences. Effective circularization of mRNA molecules is a key step for the efficient initiation of translation. Research has shown that the intrinsic separation of the ends of mRNA molecules is rather small, suggesting that intramolecular arrangements could provide this effective circularization. Considering that the innate proximity of RNA ends might have important unknown biological implications, we aimed to determine whether the close proximity of the ends of mRNA molecules is a conserved feature across organisms and gain further insights into the functional effects of the proximity of RNA ends. Using both mfold [25] and Vienna RNA [26] algorithms, we obtained the minimum free energy (MFE) secondary structures of 274 full native mRNA molecules from 17 model organisms in order to calculate the contour length (C_L) of the external loop as an index of their end-to-end separation. The complete mRNA sequences were selected randomly, with the requirement of having the complete sequence of the 5′ and 3′ UTRs. It is known that for most of the species there is no information available about the complete mRNA sequences, so this added an extra effort in the selection of the mRNA sequences. Also,

we generate random RNA sequences to obtain the distance between their ends and compare them with those obtained by using native sequences. Our computational predictions show bigger variations (from 0.59 to 31.8 nm) than previously reported, and moreover, also than those observed in random sequences. Our results suggest that separations larger than 18.5 nm are not favored, whereas short separations could be related to phenotypical stability. Overall, our work implies the existence of a biological mechanism responsible for the increase in the observed variability, suggesting that the C_L features of the exterior loop could be relevant for the initiation of translation, and that a short C_L could contribute to the stability of phenotypes.

2.1. Methods

2.1.1. mRNA sequences

All the analysis was performed using native full transcriptional units reported to the National Center for Biotechnology Information Genome Database (GenBank (https://www.ncbi.nlm.nih.gov/genbank/)). All mRNA sequences were selected only if: a) the reported sequence have the presence of both 5′- and 3′-UTRs; b) the 3′-UTRs contained the polyadenylation signal and start of the poly-A tail; and c) the length falls between 200 and 7000 nt including both UTRs and coding sequence (CDS). A total of 3000 mRNA sequences were analyzed but 274 mRNA sequences complied with all of the requirements. The selected organisms and the data for their native full mRNA sequences are shown in Table 2.1.1.

It is important to note that, although 274 mRNA sequences were included in this study, our work is a prospective study limited by the availability of full native mRNA sequences

(containing both 5´- and 3´-UTR). In this regard, because full mRNA sequences for homologous genes are very scarce, we decided to construct the complete mRNA sequences for the homologous genes by determining the transcription initiation site as well as the polyadenylation site. Initial identification of homologous genes (shown in Table 2.1.2) was performed by means of bioinformatic analysis (from the GenBank) at the protein level and by using the BLASTp alignment program (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to verify if the same protein is present in their related specie. Then, an annotation method was used following the Inr context rule based on the specific consensus sequences for each species. The complete transcript sequence was determined by annotation of the polyadenylation site [27-30]. For comparison, we also performed the analysis for a set of 50 random-generated sequences (see Appendix) that were generated using a macro in the Excel program. These sequences contained 1600 nt and an equal probability for the four nucleotides.

Table 2.1.1. mRNA molecules studied from 17 model organisms. The mRNA names are placed in order of the temporal range of their first ancestors' appearance.

Organism/Geological era	mRNA name	Total length (nt)	GenBank
H. salinarum *(HbS)*	blp bacterioopsin-linked protein blp[a]	503	NC_010364.1
	Sod2 superoxide dismutase 2[a]	644	NC_010364.1
	thrC1 threonine synthase[a]	1302	NC_010364.1
	mcmA1 methylmalonyl-CoA mutase subunit A[a]	1766	NC_010364.1
C. reinhardtii *(CR)*	Putative copper chaperone Atx1 mRNA	391	AF280056.1
	LciA mRNA for low-CO2 inducible protein LCIA	1917	AB168092.1
	Atp2 (atpB) mRNA	2599	X61624.1
	Lcr1 mRNA for low-CO2 inducible Myb transcription factor LCR1	3195	AB168090.1
	Protein S5 precursor (Prps5) mRNA	2303	AY093615.1
	Thioredoxin f1 (TRXf1) mRNA	1879	AY184800.1
	Halo-acid dehalogenase-like hydrolase (HDH1) mRNA	1351	AY672644.1
	Nitrite transporter NAR1 mRNA	2044	AF149737.1
	Chloroplast ATP-binding protein (Sabc) mRNA	2253	AY536252.1
	Chloroplast sulfate permease (SulP2) mRNA	1863	AY536251.1

	Chloroplast beta carbonic anhydrase (Cah6) mRNA	2452	AY463239.1
	Possible membrane protein, low CO2-induced mRNA	1294	U31976.1
	delta-aminolevulinic acid dehydratase (alad) mRNA	1717	U19876.1
	gliding motility related CaM kinase mRNA	2002	AY348297.1
	Lci6 mRNA for low-CO2 inducible protein LCI6	1931	AB168091.1
	Thioredoxin x (TRXx) mRNA	1161	AY184799.1
	thioredoxin o (TRXo) mRNA	1176	AY184798.1
	cytosolic thioredoxin h2 (TRXh2) mRNA	1126	AY184797.1
	thioredoxin y (TRXy) mRNA, complete cds	1313	AY184796.1
	Chloroplast sulfate-binding protein (Sbp) mRNA	1853	AY536253.1
V. carteri	Small cysteine-rich extracellular protein VCRP1 precursor mRNA	1243	DQ521274.1
(VC)	Channelrhodopsin-2 mRNA	2411	EU285660.1
	GDP dissociation inhibitor protein GDIV1p (gdiV1) mRNA	2519	U62866.1
	Retinoblastoma-related protein 1 mRNA	3656	EU366288.1
	channelrhodopsin-1 mRNA	2694	EU285658.1
	small cysteine-rich extracellular protein VCRP2 precursor mRNA	1027	EU143653.1
	somatic regenerator RegA (regA)	6725	AF106963.1
L. camtschaticum	CNP mRNA for C-type natriuretic peptide	990	AB205156.1
(LC)	LjMA1 mRNA for muscle actin	1466	AB076674.1
	Chitinase (chit) mRNA	2797	EU741679.1
	ubiquitin-conjugating enzyme UBE2A mRNA	1289	KP203887.2
	wingless-type MMTV integration site family member 1 (Wnt1) mRNA	1313	KT897931.1
	nk2-1 family homeobox c (Nkx2-1/2-4C) mRNA	2925	KT897927.1
	LjMA2 mRNA for muscle actin	2472	AB052654.2
	mRNA for aldolase (EJM8)	2221	D38620.1
	mRNA for aldolase (EJL3U)	1761	D38619.1
	NF-kappaB mRNA	4764	KY652748.1
	I kappa B-epsilon (IkBe)	3625	KC335304.1
	CD29-like protein mRNA	3698	GU013762.1
P. imperator	mRNA for hemocyanin subunit 6 (hc6 gene)	1995	FN424083.1
(PI)	mRNA for hemocyanin subunit 3b	2063	FN424082.1
	mRNA for hemocyanin subunit 5b (hc5b gene)	2189	FN424086.1
	mRNA for hemocyanin subunit 3a (hc3a gene)	2494	FN424079.1
	mRNA for hemocyanin subunit 5a (hc5a gene)	2079	FN424084.1
	mRNA for hemocyanin subunit 3c (hc3c gene)	2246	FN424081.1
B. germanica	Prepro-hypertrehalosemic hormone mRNA	464	FJ943774.1
(BG)	Hypertrehalosemic hormone receptor mRNA	1745	GU591493.1
	1,4-alpha-D-glucan glucanohydrolase precursor (bgtg-1) mRNA	2035	AY945930.1
	ace2 type acetylcholinesterase mRNA	2430	DQ288847.1
	triosephosphate isomerase (tpi) mRNA	1233	DQ885469.1
	mRNA for Yorkie-L (yki gene)	1700	HF969252.1
	mRNA for Hippo (hpo gene)	2079	HF969251.1
	mRNA for FoxO protein	2266	HE648216.1
	mRNA for Na/K-ATPase subunit beta 1 (nrv1 gene)	1145	HE795995.1
	receptor for activated protein kinase C-like (RACK1) mRNA	1147	DQ885470.1
	mRNA for ecdysone inducible protein 75 isoform B (e75 gene)	3102	AM238654.1
	mRNA for ecdysone inducible protein 75 isoform A (e75 gene)	3232	AM238653.1
	mRNA for fruitless (fru gene)	1158	FN429764.1
	mRNA for squid, variant G (sqd gene)	1445	FM875794.1
	ace1 type acetylcholinesterase mRNA	2683	DQ288249.1

	mRNA for glutathione S-transferase (gstd1 gene)	956	AM778448.1
	mRNA for Yorkie-S (yki gene)	1622	HF969253.1
P. sylvestris	glyceraldehyde-phosphate dehydrogenase mRNA	1454	L26923.1
(PS)	CCAAT-box binding factor HAP3-like protein (HAP3A) mRNA	812	JF280795.1
	mRNA for ornithine aminotransferase (dOAT gene)	1967	AM228955.1
	NAD+-dependent glyceraldehyde-3-phosphate dehydrogenase	1615	L32560.1
	glyceraldehyde-3-phosphate dehydrogenase (GapC1) mRNA	1304	L07501.1
G. biloba	Defensin precursor mRNA	506	AY695796.1
(GB)	Nuclear-encoded chloroplast chlorophyll a/b binding protein mRNA	997	L23107.1
	WD40-repeat protein (WD40) mRNA	1421	KJ630503.1
	3-hydroxy-3-methylglutaryl coenzyme A reductase mRNA	2206	AY741133.1
	glyceraldehyde-phosphate dehydrogenase mRNA	1192	L26924.1
	ginkbilobin-2 precursor mRNA	712	DQ496113.1
	mRNA for putative auxin response factor 6/8 (arf6/8 gene)	3450	FN433179.1
	lipid transfer protein precursor, mRNA	700	DQ836633.1
G. gallus	Preproghrelin mRNA	843	AY299454.1
(GG)	Heme oxygenase 1 (hmox1) mRNA	1565	HM237181.1
	Vesicular glutamate transporter 2 (VGLUT2) mRNA	1760	JF320001.1
	EGF/TGF-alpha receptor (c-erbB) mRNA	2243	M77637.1
	paraoxonase-2 (PON2) mRNA	1262	L47573.1
	PGK mRNA	1453	L37101.1
	kinase related protein mRNA	2535	M88283.1
	(17.5) mRNA	2046	M88072.1
	p94 mRNA for n-calpain-1 large subunit	3454	D38028.1
	YB-1 protein mRNA	1507	L13032.1
	alpha-3 type IX collagen mRNA	2416	M83179.1
	liver ribonuclease A precursor, mRNA	577	DQ395277.1
	WDR1 protein mRNA	3280	AF020054.1
	chS-Rex-b mRNA	3187	U17606.1
	chS-Rex-s mRNA	1572	U17605.1
	protein arginine methyltransferase 4 (PRMT4) mRNA	1788	KY655811.1
	neural retina growth hormone mRNA	690	AY373631.1
	TRF2-interacting telomeric RAP1 protein (RAP1) mRNA	2399	AY083608.1
M. domestica	Early lactation protein precursor (ELP) mRNA	477	JN191340.1
(MD)	Sperm protein Sp17 mRNA	1087	AF054290.1
	Endogenous retrovirus ERV syncytin-Opo1 mRNA	2145	KM235357.1
	Endogenous retrovirus Opo-Env3-ERV Env3 mRNA	2481	KM235359.1
	mRNA for Interleukin-6 receptor alpha (IL6R gene)	2131	LT596680.1
	domestica mRNA for Interleukin-6 (IL6 gene)	1255	LT596676.1
	anterior pituitary glycoprotein hormone common alpha subunit mRNA	755	AY048590.1
	thyroid stimulating hormone beta subunit precursor mRNA	517	AY048589.1
	follicle-stimulating hormone beta precursor mRNA	450	AF406610.1
	somatotropin precursor mRNA	813	AF312023.1
	p21-ras mRNA	796	Z12125.1
	beta (2) microglobulin mRNA	1105	AY125947.1
Z. mays	Nuclear-encoded mitochondrial F1F0 ATP synthase epsilon subunit	488	L39120.1
(ZM)	mudrB mRNA	1030	U14598.1
	Beta-8 tubulin (tub8) mRNA	1625	L10636.1
	Sucrose transporter 2 (SUT2) mRNA	2149	AY581895.1

	O-methyltransferase mRNA	1268	L14063.1
	beta-7 tubulin (tub7) mRNA	1583	L10634.1
	beta-6 tubulin (tub6) gene and mRNA	1730	L10633.1
	cytochrome P-450 (cyp78) mRNA	2087	L23209.1
	putative bifunctional nuclease (nuc gene)	1095	AM710418.1
	mRNA for putative inositol-3-phosphate synthase (mips2 gene)	1848	AM295187.1
	mRNA for transcription factor MYB42 (myb42 gene)	1129	AM156908.1
	mRNA for transcription factor MYB39 (myb39 gene)	1078	AM156907.1
	mRNA for transcription factor MYB31 (myb31 gene)	1264	AM156906.1
	mRNA for transcription factor MYB8 (myb8 gene)	1082	AM156905.1
	mRNA for transcription factor MYB2 (myb2 gene)	1339	AM156904.1
	harpin binding protein 1 (HrBP1) mRNA	1218	AY388616.1
H. brasiliensis	Copper transport protein ATOX1 (CCH) mRNA	543	GU550955.1
(HB)	JAZ11 mRNA	914	KJ001648.1
	MYB transcription factor (MYB) mRNA	970	DQ323739.1
	JAZ9 mRNA	1391	KJ001646.1
	subtilisin-like serine protease C (SPC) mRNA	2389	KU845304.1
	subtilisin-like serine protease A (SPA) mRNA	2444	KU845302.1
	JAZ10 mRNA	804	KJ001647.2
	JAZ8 mRNA	526	KJ001645.1
	JAZ7 mRNA	644	KJ001644.1
	mRNA for latex allergen	1419	AJ223038.1
A. thaliana	Thionin (Thi2.2) mRNA	718	L41245.1
(AT)	rac GTP binding protein mRNA	985	AF079485.1
	Heat shock mRNA	3105	U13949.1
	cystathionine beta-lyase mRNA	1644	L40511.1
	molybdenum cofactor biosynthesis enzyme (cnx1) mRNA	2237	L47323.1
	serine acetyltransferase (SAT1) mRNA	1079	L42212.1
	GTP-binding protein mRNA	2234	L38614.1
	lipoxygenase mRNA	2790	L04637.1
	RNA polymerase subunit (isoform B) mRNA	1406	L34773.1
	RNA polymerase subunit (isoform A) mRNA	1282	L34772.1
	thionin (Thi2.1) mRNA	611	L41244.1
	recombination and DNA-damage resistance protein (DRT112) mRNA	705	M98456.1
	thaliana rac GTP binding protein Arac10 (Arac10) mRNA	794	AF079485.1
	PAC3 mRNA	1200	L35241.1
A. cerana	Glutaredoxin 1 (Grx1) mRNA	436	JX844656.2
(AC)	1-cys thioredoxin peroxidase (Tpx4) mRNA	931	KJ551847.1
	Arginine kinase (AK) mRNA	1650	KF772855.1
	Phenoloxidase subunit A3 (PPO) mRNA	2293	JX844653.1
	superoxide dismutase 2 (SOD2) mRNA	1003	JN637476.1
	CTL5 (CTL5) mRNA	1197	KT808468.1
	cuticular protein CPF1 (CPF1) mRNA	1021	KJ634544.1
	cuticle protein 2 (CPR2) mRNA	1134	KJ502287.1
	caspase 1 mRNA	1328	KF955542.1
	mitogen-activated protein kinase kinase 4 (MKK4) mRNA	1634	KF017207.1
	triosephosphate isomerase (Tpi) mRNA	1894	KP994676.1
	ERR (ERR) mRNA	1671	KP398511.1
	decapentaplegic (Dpp) mRNA	1652	KT750952.1
	glutathione S-transferase (GSTO1) mRNA	1040	KF496073.1

	thioredoxin 1 (Trx1) mRNA	649	JX844651.2
	CAT (CAT) mRNA	1905	KF765424.1
	glutathione S-transferase omega 2 (GSTO2) mRNA	1365	JX434029.1
	thioredoxin 2 (Trx2) mRNA	407	JX844649.1
	cytochrome P450 4G11 (CYP4G11) mRNA	2041	KC243984.1
	delta-class glutathione S-transferase (GSTD) mRNA	1009	JF798573.1
D. rerio	flii mRNA for flightless I	3878	AB355792.1
(DR)	fibrosin protein (fbrs)	5142	KY492383.1
	heat shock cognate (hsc70)	2323	L77146.2
	contactin-associated protein-like 2b (cntnap2b)	4751	HQ880438.1
	contactin-associated protein-like 2a beta isoform (cntnap2a)	888	HQ880437.1
	contactin-associated protein-like 2a alpha isoform (cntnap2a) mRNA	4558	HQ880436.1
	growth hormone receptor a mRNA	2332	EU649774.1
	myogenin mRNA	1364	AF202639.1
	QM protein (QM) mRNA	769	AY763500.1
	neuropeptide FF-related PQRF precursor (PQRF)	827	AY092774.1
	NADPH-cytochrome P450 oxidoreductase	2941	AY949986.1
	CL2 mRNA	2926	EU269066.1
	connexin 41.8 (cx41.8) mRNA	2801	DQ177156.1
P. troglodytes	Hepcidin (HAMP) mRNA	391	EU076436.1
(PT)	Glycolipid transfer protein (GLTP) mRNA	722	EF688398.1
	Gene for non-A non-B hepatitis-associated microtubular mRNA	1641	D90034.1
	Dusty protein kinase mRNA	3601	AY641092.1
	mRNA for Killer-cell Ig-like receptor KIR2DL8 (KIR2DL8 gene)	1064	AM279149.2
	mRNA for beta1,4-galactosyltransferase 7 (b4Gal-T7 gene)	1074	AM231264.1
	beta-defensin 104 (DEFB104) mRNA	285	EU126867.1
H. sapiens	mRNA for Ubiquitin protein ligase	2850	AB056663.2
(HS)	prostasin mRNA	1834	L41351.1
	K+ channel beta-subunit (Kvb1.3) mRNA	3103	L47665.1
	neuroendocrine-specific protein C (NSP) mRNA	1416	L10335.1
	STAT4 mRNA	2588	L78440.1
	ERK3 protein kinase mRNA	3324	L77964.1
	FRG1 mRNA	1042	L76159.1
	cyclin G2 mRNA	1410	L49506.1
	interleukin 8 receptor alpha (IL8RA) mRNA	2007	L19591.1
	phosphatase 2A B56-epsilon (PP2A) mRNA	3270	L76703.1
	cyclin G1 mRNA	1602	L49504.1
	5-HT6 serotonin receptor mRNA	1984	L41147.1
	pyruvate dehydrogenase kinase isoenzyme 3 (PDK3) mRNA	1599	L42452.1
	pyruvate dehydrogenase kinase isoenzyme 2 (PDK2) mRNA	1422	L42451.1
	casein kinase I epsilon mRNA	1331	L37043.1

Table 2.1.2. Homologous mRNA molecules studied from related organisms. The green algae *C. reinhardtii* with *V. carteri*, the fishes *L. camtschaticum* with *D. rerio*, the eudicotyledones *H. brasiliensis* with *A. thaliana* and the hominids *P. troglodytes* with *H. sapiens*.

Homolog mRNA name	Organisms	Total length (nt)	CDS protein identity (%)	GenBank
ubiquitin conjugating enzyme E2	*C. reinhardtii (CR)*	2282	25	XP_001690015.1
	V. carteri (VC)	1638		XP_002953636.1
thioredoxin-like protein (TRX10)	*C. reinhardtii (CR)*	1009	27	XP_001690017.1
	V. carteri (VC)	1456		XP_002958831.1
glyceraldehyde-3-phosphate dehydrogenase (GAP3)	*C. reinhardtii (CR)*	1930	45	XP_001689871.1
	V. carteri (VC)	1627		XP_002956882.1
p53-induced protein 8	*C. reinhardtii (CR)*	2684	69	XP_001690067.1
	V. carteri (VC)	1507		XP_002955375.1
vacuolar ATP synthase subunit H (ATPvH)	*C. reinhardtii (CR)*	2603	82	XP_001689562.1
	V. carteri (VC)	1808		XP_002955475.1
peroxiredoxin, type II (PRX5)	*C. reinhardtii (CR)*	1180	83	XP_001689455.1
	V. carteri (VC)	1336		XP_002956367.1
20S proteasome alpha subunit D (POA4)	*C. reinhardtii (CR)*	1267	91	XP_001689587.1
	V. carteri (VC)	1025		XP_002955451.1
translation initiation factor 4E (eif4E)	*C. reinhardtii (CR)*	2032	91	XP_001693235.1
	V. carteri (VC)	1696		XP_002958375.1
26S proteasome regulatory subunit (RPN11)	*C. reinhardtii (CR)*	1964	93	XP_001689423.1
	V. carteri (VC)	1419		XP_002955275.1
ribosomal protein L18a (RPL18a)	*C. reinhardtii (CR)*	1086	94	XP_001689743.1
	V. carteri (VC)	1121		XP_002948508.1
histone H2B	*C. reinhardtii (CR)*	571	96.77	XP_001691693.1
	V. carteri (VC)	717		XP_002955481.1
histone H4 (HFO24)	*C. reinhardtii (CR)*	1457	99.03	XP_001690685.1
	V. carteri (VC)	653		XP_002955420.1
ATP synthase F0 subunit 8	*L. camtschaticum (LC)*	794	40.74	YP_007517126.1
	D. rerio (DR)	630		NP_059335.1
NADH dehydrogenase subunit 6	*L. camtschaticum (LC)*	580	41.14	YP_007517133.1
	D. rerio (DR)	1868		NP_059342.1
NADH dehydrogenase subunit 5	*L. camtschaticum (LC)*	1971	54.48	YP_007517132.1
	D. rerio (DR)	2112		NP_059341.1
NADH dehydrogenase subunit 1	*L. camtschaticum (LC)*	1204	65.82	YP_007517122.1
	D. rerio (DR)	1396		NP_059331.1
cytochrome c oxidase subunit III	*L. camtschaticum (LC)*	1627	80.08	YP_007517128.1
	D. rerio (DR)	1694		NP_059337.1
cytoplasmic actin	*L. camtschaticum (LC)*	1921	98.4	BAB41207.1
	D. rerio (DR)	1664		NP_571106.2
E3 SUMO-protein ligase SIZ1-like	*H. brasiliensis (HB)*	2903	65.91	XP_021644935.1
	A. thaliana (AT)	3331		AAU00414.1
Two pore calcium channel protein 1	*H. brasiliensis (HB)*	3075	70.65	XP_021685070.1
	A. thaliana (AT)	2500		BAB55460.1
ribosomal protein L20	*H. brasiliensis (HB)*	731	80.34	YP_004327685.1
	A. thaliana (AT)	1664		NP_051082.1
ATP-dependent Clp protease proteolytic subunit	*H. brasiliensis (HB)*	642	83.59	YP_004327687.1
	A. thaliana (AT)	1516		NP_051083.1

ribosomal protein S4 (chloroplast)	*H. brasiliensis (HB)* *A. thaliana (AT)*	906 798	88.56	YP_004327664.1 NP_051061.1
NADH dehydrogenase subunit 3	*H. brasiliensis (HB)* *A. thaliana (AT)*	1272 428	90.83	YP_004327667.1 NP_051064.1
ATP synthase CF1 epsilon subunit	*H. brasiliensis (HB)* *A. thaliana (AT)*	725 610	91.67	YP_004327668.1 NP_051065.1
photosystem I subunit IX	*H. brasiliensis (HB)* *A. thaliana (AT)*	768 454	95.45	YP_004327682.1 NP_051079.1
photosystem II protein M	*H. brasiliensis (HB)* *A. thaliana (AT)*	813 181	97.06	YP_004327647.1 NP_051053.1
V-type proton ATPase 16 kDa proteolipid subunit	*H. brasiliensis (HB)* *A. thaliana (AT)*	736 665	98.18	XP_021659332.1 AAA99937.1
photosystem II protein D2	*H. brasiliensis (HB)* *A. thaliana (AT)*	1583 1613	98.58	YP_004327657.1 NP_051054.1
dopamine receptor D4	*P. troglodytes (PT)* *H. sapiens (HS)*	1639 1589	92.64	XP_016775504.1 NP_000788.2
hemoglobin subunit delta	*P. troglodytes (PT)* *H. sapiens (HS)*	927 927	99.32	XP_001162045.2 NP_000510.1
Arginine vasopressin receptor 1A	*P. troglodytes (PT)* *H. sapiens (HS*	2502 2493	99.52	XP_016778615.1 NP_000697.1
cytidine deaminase (CDA)	*P. troglodytes (PT)* *H. sapiens (HS)*	3056 892	100	XP_001161389.1 AAA57254.1
glutamate-cysteine ligase	*P. troglodytes (PT)* *H. sapiens (HS)*	1620 1610	100	XP_513572.3 AAA65028.1
hemoglobin subunit beta	*P. troglodytes (PT)* *H. sapiens (HS)*	627 626	100	XP_508242.1 NP_000509.1

2.1.2. Prediction of the distance between the ends of mRNA molecules

Prediction of the secondary structure of each mRNA was performed using both mfold [25] and Vienna RNA programs [26]. Based on energy models, these computational programs estimate the free energy (FE) of the secondary structures that are predicted for each ssRNA sequence. Both algorithms were used to calculate the secondary structures. To estimate the distance between the ends of each mRNA molecule, we calculate the contour length (C_L) of the exterior loop from each secondary structure. The C_L is obtained by counting the number of nucleotide links comprising the exterior

loop multiplied by the typical distance (d= 0.59 nm) between nucleotides in RNA molecules [31] (Fig. 2.1.2.1). Although a degeneration in base pair prediction accuracy that increases with the length of sequence has been proposed [11-13], single sequence secondary structure prediction is reasonably accurate with the RNA folding programs we used here [14]. Moreover, different probabilistic models give similar results [15-17]. Indeed, regardless of differences in the secondary structure, when compared, both algorithms gave similar results in the C_L values (Fig. 2.1.2.2 to 2.1.2.8). Furthermore, when the C_L of the 5 structures with the smaller FE were compared against the C_L of the minimum free energy (MFE) secondary structures, this is, the FE C_L value divided by the MFE C_L value, no significant differences were found despite the differences in the secondary structure (p= 0.1753) (Fig. 2.1.2.9). For this reason, we decided to use the MFE secondary structure in our study, which corresponds to the thermodynamically most stable structure with the lowest energy possible (ΔG).

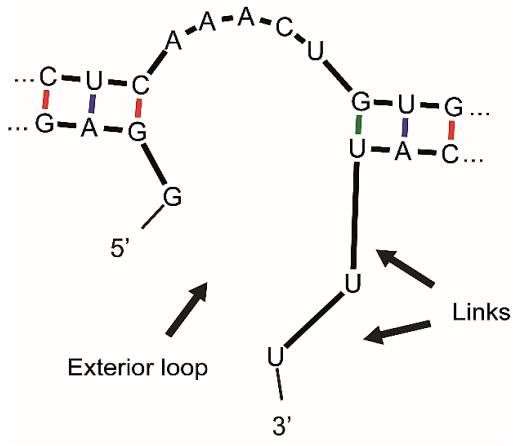

Figure 2.1.2.1. Exterior loop of the minimum free energy mRNA secondary structure. The structure corresponds to *HS* mRNA for ubiquitin protein ligase predicted by mfold. The contour length (C_L) is given by the total number of links in the exterior loop (L = 11) multiplied by the distance between nucleotides (d = 0.59 nm) giving 6.49 nm in this case.

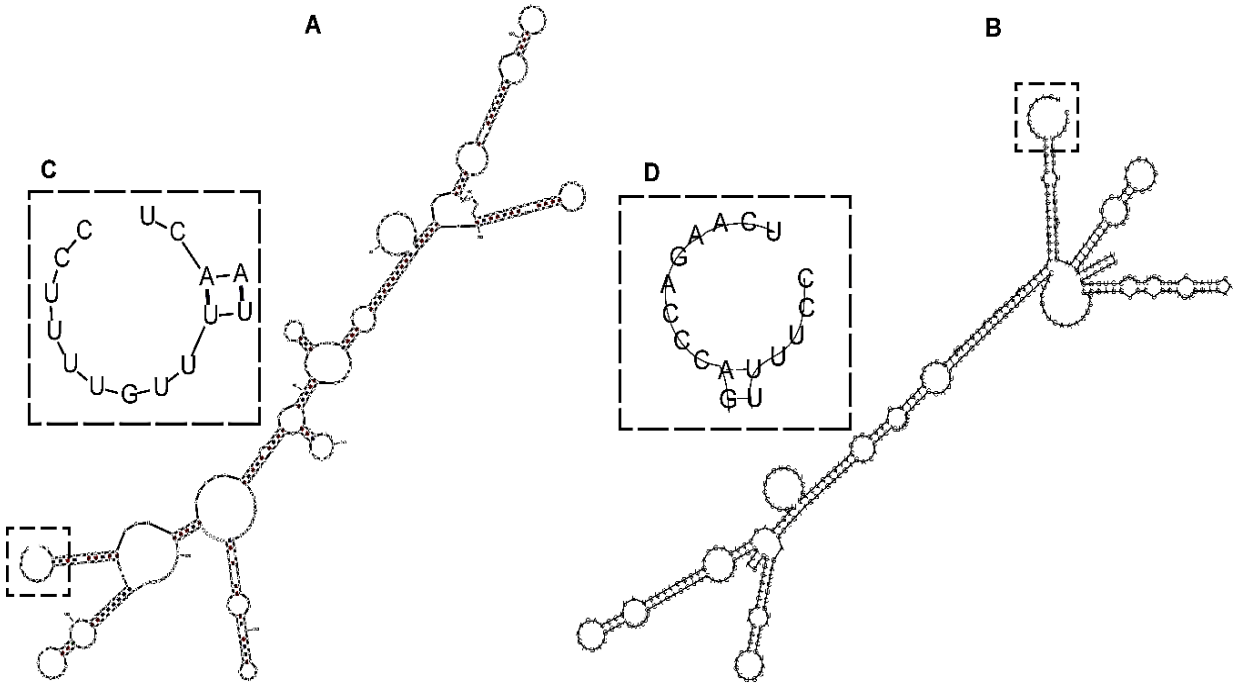

Figure 2.1.2.2. Minimum free energy secondary structure for 391 nt mRNA of *PT hepcidin (HAMP)*. Obtained by (A) mfold and (B) Vienna RNA algorithms. The exterior loop is quite similar despite differences on their secondary structure. (C) and (D) zooms of the exterior loop of (A) and (B), respectively.

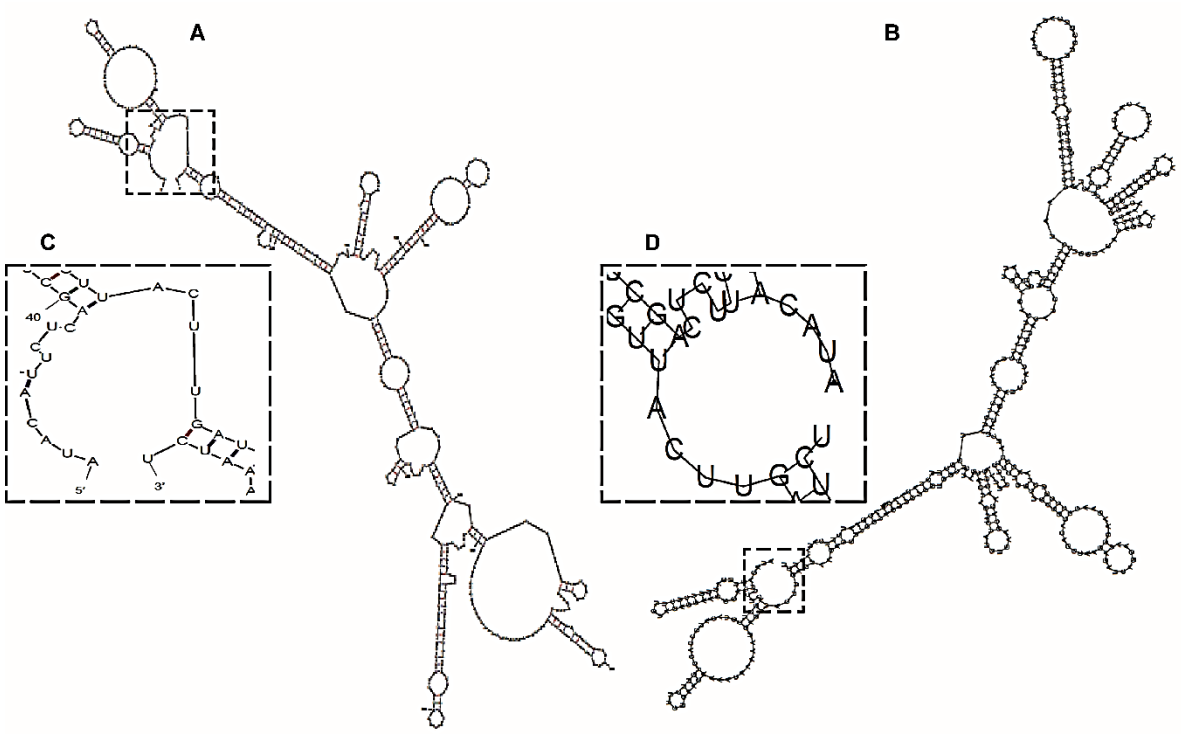

Figure 2.1.2.3. Minimum free energy secondary structure for 464 nt mRNA of *BG Prepro-hypertrehalosemic hormone*. Obtained by (A) mfold and (B) Vienna RNA algorithms. The exterior loop is quite similar despite differences on their secondary structure. (C) and (D) zooms of the exterior loop of (A) and (B), respectively.

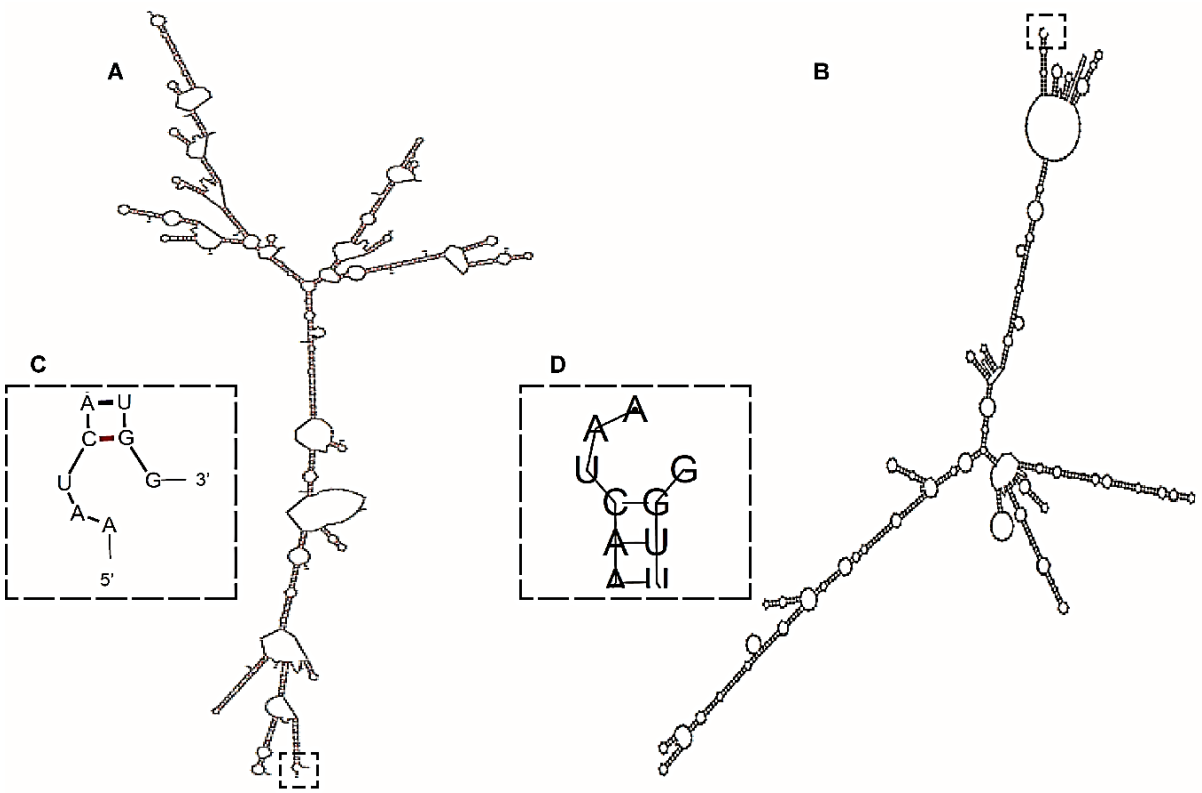

Figure 2.1.2.4. Minimum free energy secondary structure for 997 nt mRNA of *GB Nuclear-encoded chloroplast chlorophyll a/b binding protein.* Obtained by (A) mfold and (B) Vienna RNA algorithms. The exterior loop is quite similar despite differences on their secondary structure. (C) and (D) zooms of the exterior loop of (A) and (B), respectively.

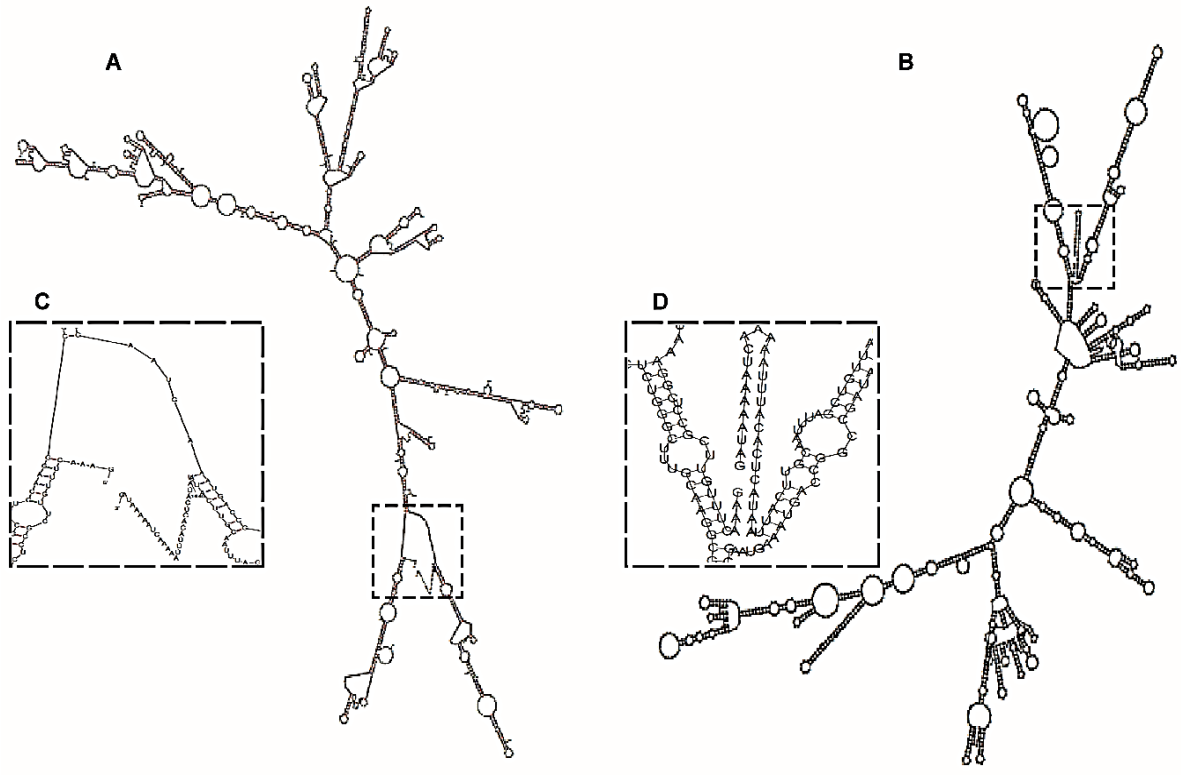

Figure 2.1.2.5. Minimum free energy secondary structure for 1466 nt mRNA of *LC LjMA1 muscle actin.* Obtained by (A) mfold and (B) Vienna RNA algorithms. The exterior loop is quite similar despite differences on their secondary structure. (C) and (D) zooms of the exterior loop of (A) and (B), respectively.

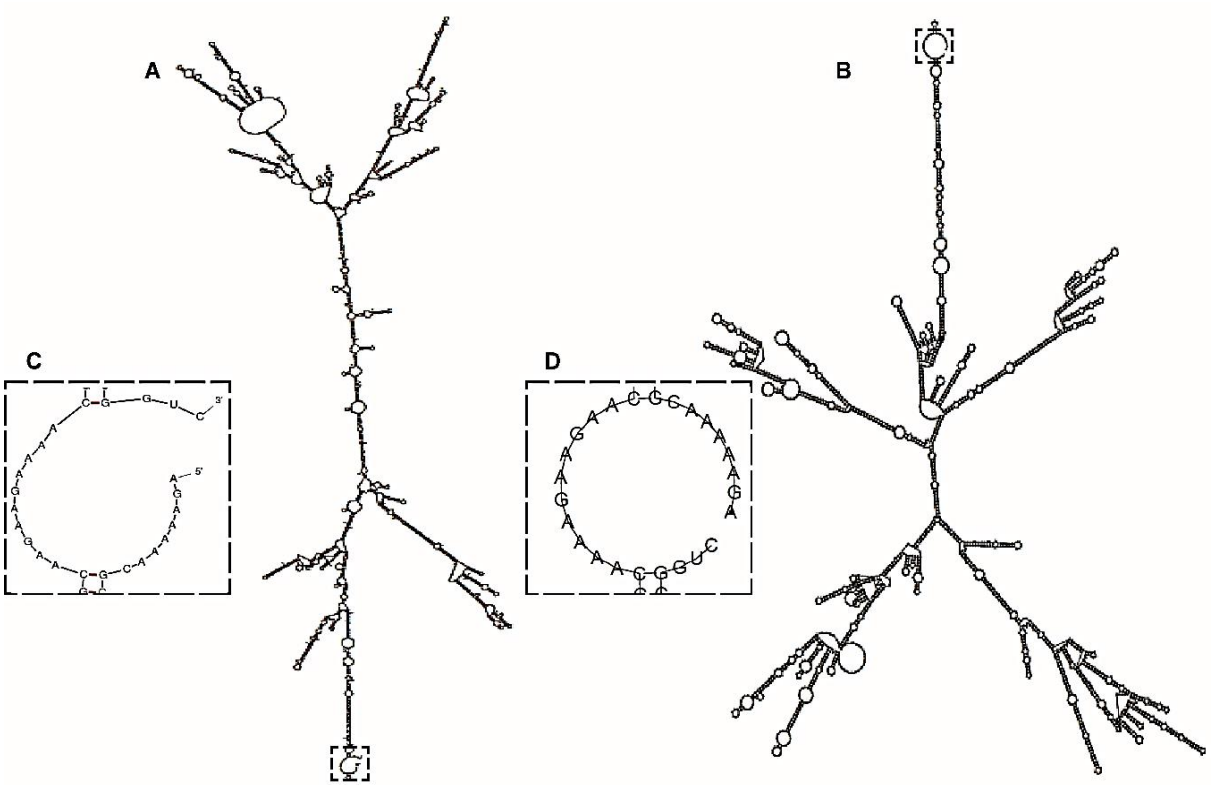

Figure 2.1.2.6. Minimum free energy secondary structure for 2411 nt mRNA of *VC Channelrhodopsin-2.* Obtained by (A) mfold and (B) Vienna RNA algorithms. The exterior loop is quite similar despite differences on their secondary structure. (C) and (D) zooms of the exterior loop of (A) and (B), respectively.

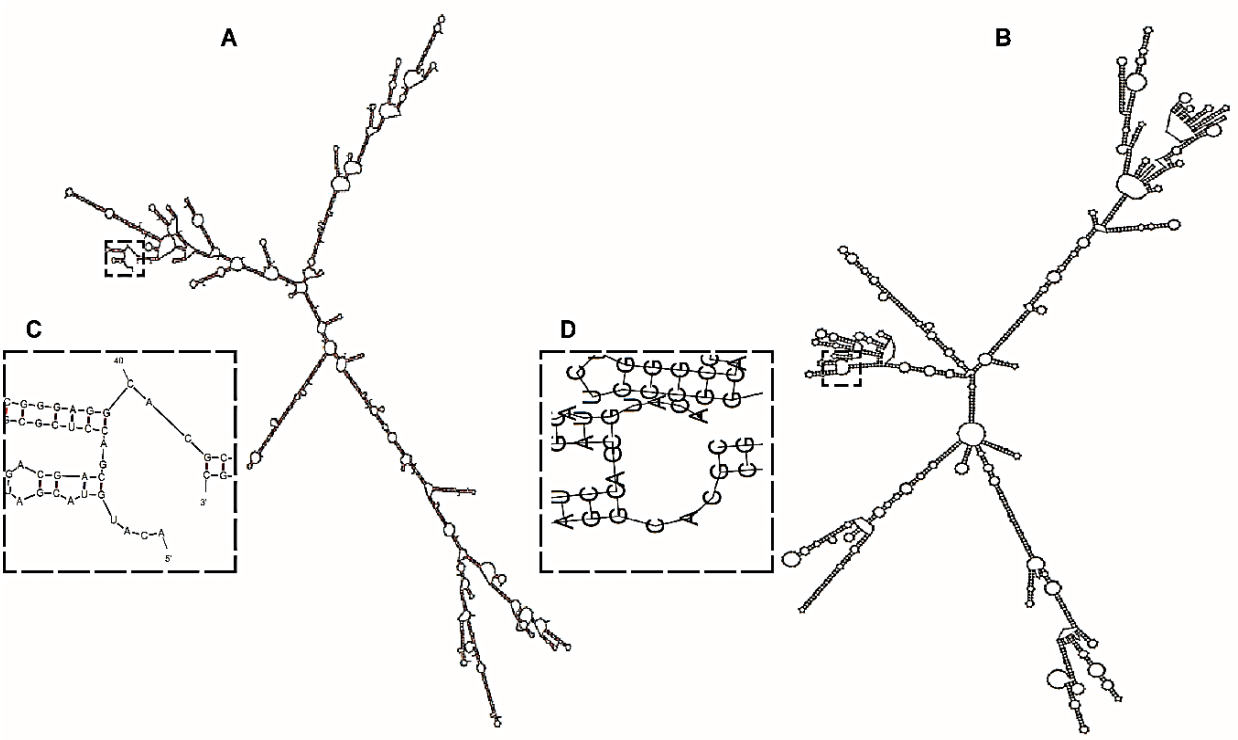

Figure 2.1.2.7. Minimum free energy secondary structure for 1766 nt mRNA of *HbS mcmA1 methylmalonyl-CoA mutase subunit A.* Obtained by (A) mfold and (B) Vienna RNA algorithms. The exterior loop is quite similar despite differences on their secondary structure. (C) and (D) zooms of the exterior loop of (A) and (B), respectively.

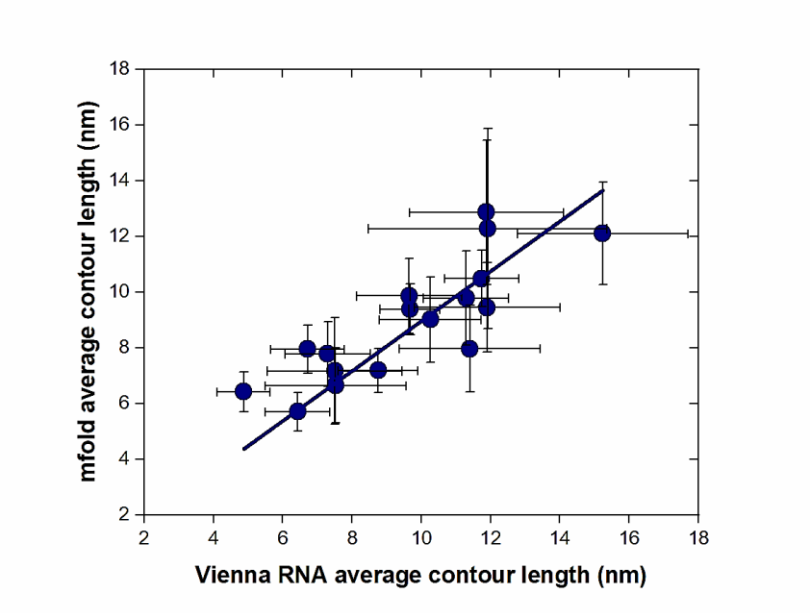

Figure 2.1.2.8. mfold vs Vienna RNA contour length values. Both algorithms give similar contour length values. The plot includes the linear fit ($y = bx$) with $b = 0.895 \pm 0.002$. The Pearson correlation coefficient is r (16) = 0.84 and $p = 0.000025$, consistent with significant correlation. The data presented are ± SD.

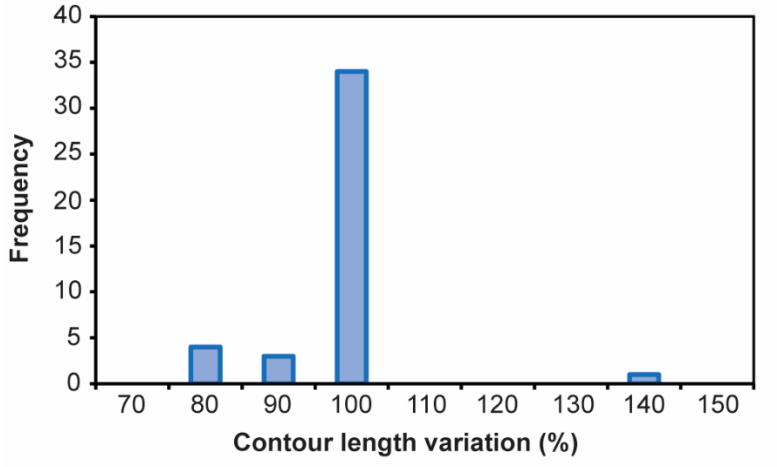

Figure 2.1.2.9. Contour length ratio between the minimum free energy and free energy structures. No significant differences appear between the MFE structure (100%) and the subsequent 4 FE structures for the same mRNA from 9 species, $p = 0.1753$ and total sample size of $N = 45$.

2.1.3. Statistical Analysis

All data presented are mean ± SEM (standard error of the mean) unless otherwise mentioned. Statistical differences were analyzed by the two tailed Welch´s test for unequal variances. Differences were significant at $p < 0.05$. For the linear fits in our data, we used linear regression and two tailed test of significance at $p < 0.05$.

2.2. Results

It has been demonstrated computational and experimentally that the ends of RNA molecules have an innate proximity, however using only a low number of RNA molecules [7, 18, 32]. We therefore decided to extend these observations by computationally determining the end-to-end distance of 274 full native mRNA molecules from 17 species reported to the GenBank. The species used for this study are: the halophilic archaeon *Halobacterium salinarum (HbS)*; the single-cell green alga *Chlamydomonas reinhardtii (CR);* the colonial green alga *Volvox carteri (VC)*; the jawless fish *Lethenteron camtschaticum (LC);* the scorpion *Pandinus imperator (PI)*; the cockroach *Blatella germanica (BG)*; the scot pine *Pinus sylvestris (PS)*; the maidenhair tree *Ginkgo biloba (GB)*; the fowl *Gallus gallus (GG)*; the opossum *Monodelphis domestica (MD)*; the monocotyledon *Zea mays (ZM)*; the dicotyledons rubber tree *Hevea brasiliensis* (*HB*); the flowering plant *Arabidopsis thaliana (AT);* the honey bee *Apis cerana (AC)*; the fresh water jawed fish *Danio rerio (DR)* as well as the hominids *Pan troglodytes (PT)* and *Homo sapiens (HS)* (Fig. 2.2.1).

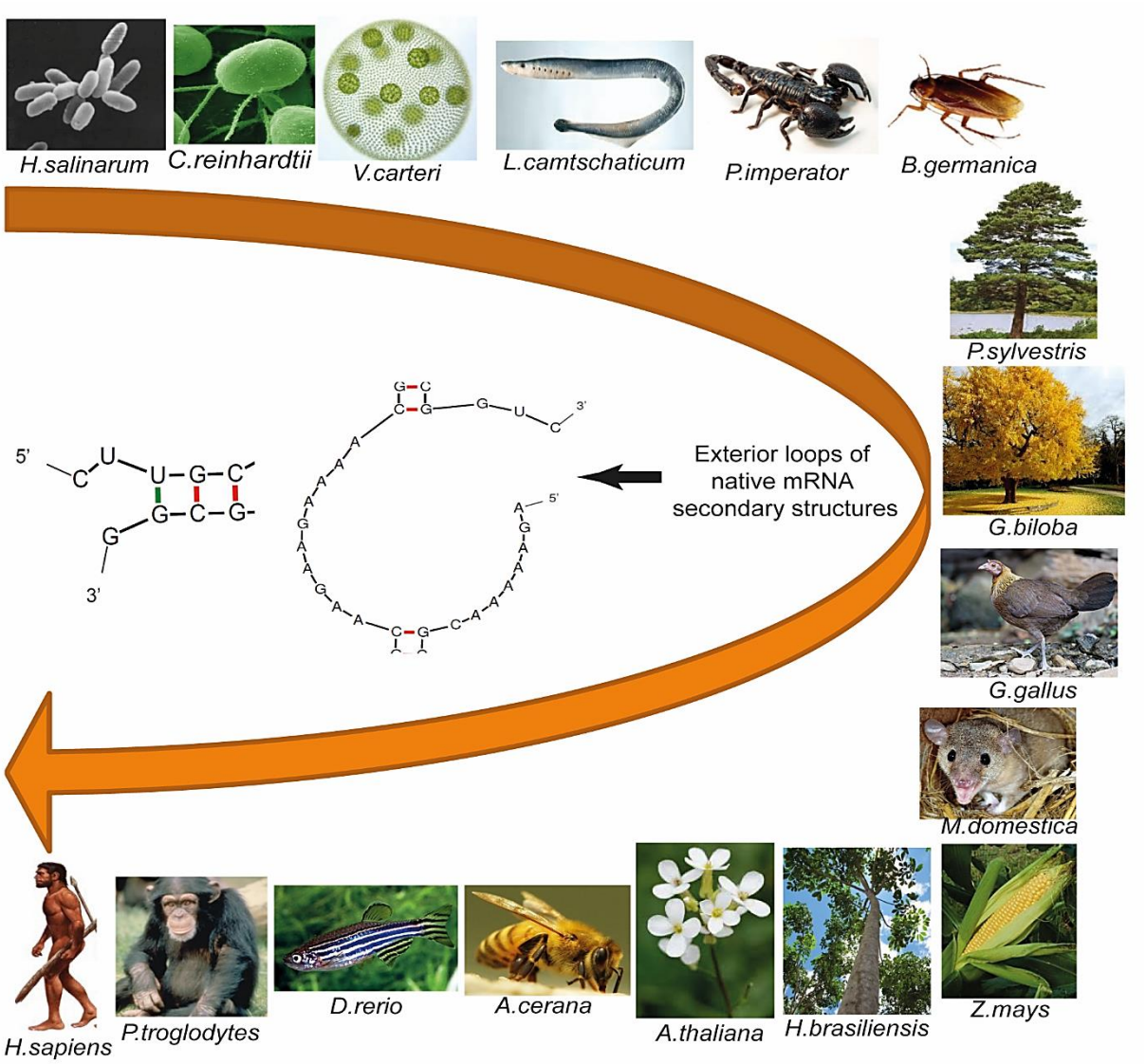

Figure 2.2.1. Seventeen model organisms from where the native full mRNAs sequences were selected. From each selected mRNA we obtained the MFE secondary structure to estimate the distance between the mRNA ends from the exterior loop.

Both mfold and Vienna RNA programs were used to obtain the MFE secondary structures of full-length native mRNAs randomly selected from these species. The end-to-end distance was determined by calculating the C_L of the exterior loop. (Fig. 2.1.2.1).

Remarkably, although clear differences in the whole secondary structures can be found with each program, both programs gave quite similar results for C_L values (Fig. 2.1.2.2 to 2.1.2.7), thus, we chose to report here the values obtained with the Vienna RNA algorithm.

The average contour length of the randomly selected full native mRNA sequences reported to the GenBank is shown in Fig. 2.2.2. On average, the C_L varies from 4.8 (*GB*) up to 15.2 nm (*PI*), which is wider than previously reported [7, 18, 32]. Moreover, as can be noted in Fig. 2.2.2, the separation between the ends of mRNA molecules do not remain constant. Indeed, when individually evaluated, the separation between mRNA ends varies from 0.59 up to 31.8 nm (Fig. 2.2.2. B). The Gaussian fit is centered on x= 9.0 ± 0.8 nm and has a width of w= 6.03 ± 0.86 nm (± Standard Deviation (SD)). The statistics give a C_L smaller than 17.5 ± 2 nm (± SD) (95 % confidence level) thus larger C_L values are not favored.

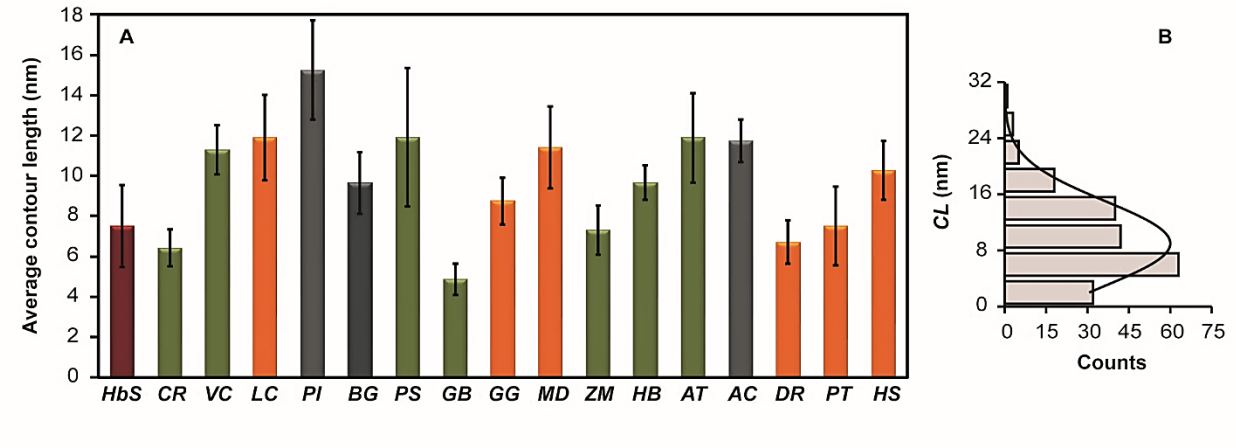

Figure 2.2.2. Contour length distributions from the predicted mRNA secondary structures. (A) Average contour length from the predicted mRNA secondary structures obtained by Vienna RNA predictions. *H. salinarum* (*HbS*), *C. reinhardtii* (*CR*), *V. carteri* (*VC*), *L. camtschaticum* (*LC*), *P. imperator* (*PI*), *B. germanica* (*BG*), *P. sylvestris* (*PS*), *G. biloba* (*GB*),

G. gallus (*GG*), *M. domestica* (*MD*), *Z. mays* (*ZM*), *H. brasiliensis* (*HB*), *A. thaliana* (*AT*), *A. cerana* (*AC*), *D. rerio* (*DR*), *P. troglodytes* (*PT*) and *H. sapiens* (*HS*). The bars correspond to the average value of all mRNA used for each species. Bars with the same color represent organism of a same clade: halobacteria (red bars); Viridiplantae (green bars); invertebrates (gray bars) and vertebrates (orange bars). We include the standard error of the average of the mRNA molecules included per species with a total sample size of N= 204. (B) We show the histogram for all C_L values obtained with Gaussian fit (black line). We can observe that the mRNA molecules have a C_L smaller than 17.5 ± 2 nm (± SD) with 95 % confidence level.

Therefore, as a control, we performed a similar analysis to emphasize the difference, using only randomly generated sequences (Fig. 2.2.3). The 50 sequences had 1600 nt with an equal probability for the four different nucleotides. The resulting histogram (Fig. 2.2.3 A) shows that the C_L values varies from 1.77 up to 21.2 nm and the statistics give a C_L smaller than 11.4 ± 1.4 nm (± SD) with 95 % confidence level. The Gaussian fit is centered on x_r = 7.7 ± 0.4 nm and has a width of w_r = 2.6 ± 0.7 nm (± SD). Comparing the width of the native (w_n = 6.03 ± 0.86 nm) (Fig. 2.2.3 B) and random-generated sequences (w_r = 2.6 ± 0.7 nm) we see that the native RNA sequences show considerably higher variations (4.8 σ confidence). To reinforce this difference, we consider two effects that could biased this number. The first effect is that the average values for the C_L are different between native and random-generated sequences, and we can think that this is the reason to observe higher variations when the width distributions are compared. To account for this, we can scale w_r as if we had the same center values. Considering that the center values of a distribution usually grow as the square of the distribution width, we can scale w_r by $\sqrt{\frac{x_n}{x_r}}$ giving w_s = 2.8 ± 0.8 nm (± SD)) (4.2 σ confidence). The second effect is the fact that the GC content and the C_L values

are highly correlated (as we will explain later with more detail (Fig. 2.2.7)) with a linear dependence with slope b = -0.18 ± 0.04 nm / % (± SD). This means that the variations in the GC content impact directly over the C_L values. Therefore, we scale w_s due to the variations in the GC content. The histogram of the GC content for native sequences would give a Gaussian fit centered on x_{nGC} = 46.7 ± 1.4 % and a width of w_{nGC} = 7.0 ± 1.4 %, whereas the random sequences have x_{rGC} = 50 % and w_{rGC} = 0.86 ± 0.03 % (± SD). The variations of GC of the native sequences given by w_{nGC} translate into variations of the C_L value of $\Delta C_L = (b)(w_{nGC})$ = -1.26 nm. This corresponds to a 14 % variation over the C_L value at 50 % GC content. Since the width goes as the square root of the C_L value, that corresponds to a scaling factor of $\sqrt{(1.15)(w_s)}$, giving a final scaled value of w_f = 3.0 ± 0.8 nm (± SD) (Fig. 2.2.3 A and B). Finally, the probability to have these variations due to a statistical fluctuation is smaller than 0.01 % (3.7 σ confidence (Fig. 2.2.3 C)). The above result indicates that this difference cannot be accounted for by the different value of the Gaussian center or by the variations in GC content. Indeed, there must be an underlying biological mechanism contributing to the selection of a particular contour lengths, because the variations cannot be explained by thermodynamically reasons alone.

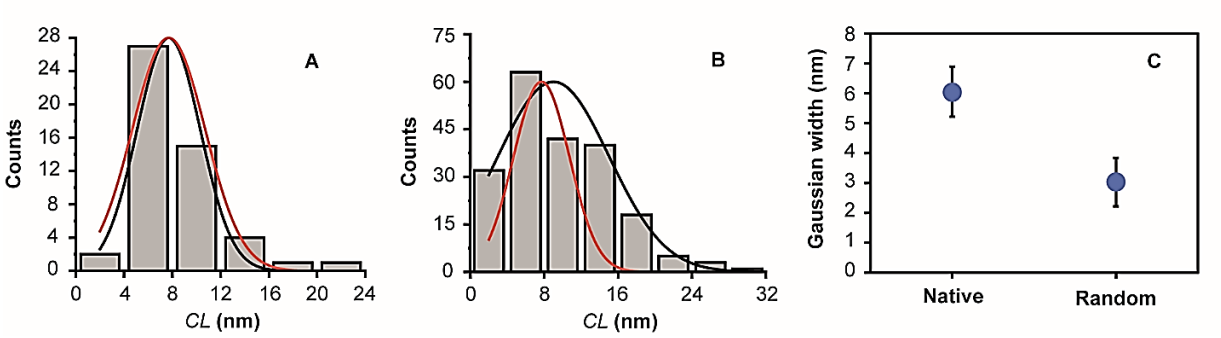

Figure 2.2.3. Difference between the contour length of random-generated and native sequences. A) Histogram obtained for the C_L values of random-generated sequences with N = 50. We can observe that the sequences have a C_L smaller than 11.4 ± 1.4 nm with 95 % confidence level. We show the Gaussian fit for not scaled (w_r = 2.6 ± 0.7 nm) and scaled values (w_f = 3.0 ± 0.8 nm) (black line and red line respectively) with the same amplitude for the comparison of both. B) Histogram obtained for the C_L values of native sequences. We show the Gaussian fit for native (w_n = 6.03 ± 0.86 nm) and the scaled value for random-generated sequences (w_f = 3.0 ± 0.8 nm) (black line and red line respectively), again with the same amplitude for the comparison. C) We observe a higher variation between the native and the scaled value for random-generated distribution widths with a 3.7 σ confidence. The data presented are ± SD.

The biological implications of the end-to-end proximity have not yet been explored, and because our results point toward an underlying biological mechanism contributing to the selection of a particular contour lengths, we decided to explore further some possible biological mechanisms.

The efficiency during translation initiation is particularly dependent on the effective circularization of mRNAs, so any increase in C_L would have serious implications on the rate of translation of any given mRNA. Intriguingly, the length of 3′-UTRs shows an increase as the level of complexity of organisms increases, suggesting that 3′-UTRs may have increased with evolution [20], and thus modifying the efficiency of translation. Moreover, phenotypical stability, which is the ability to maintain the same phenotype in

response to environmental changes, depends on the efficiency of proteins synthesis; conversely, low translation efficiencies could increase the opportunity for variability. As can be seen in Fig. 2.2.2, when the C_L values are plotted in order of the temporal range of their first ancestors' appearance [33-51], although no evolutionary trend could be found, the separation between the ends of mRNA molecules do not remain constantly small. For instance, insects (*BG* and *AC*) as well as mammals (*MD*, *PT* and *HS*), angiosperms (*ZM, HB* and *AT)* and fishes (*LC* and *DR*) have similar C_L among them, this is, no variability could be found when related species are evaluated. However, the multicellular green alga *VC* has a significantly longer C_L value than its closest evolutionary predecessor *CR* (p= 0.007). In contrast, the eudicotyledones *HB* and *AT*, which have a similar level phylogenetic divergence, present similar C_L values (9.7 ± 0.8 nm for *HB* versus 11.9 ± 2 nm for *AT*, p= 0.36), and somewhat similar with the hominids *PT* and *HS* (7.5 ± 2 versus 10.3 ± 1.5 nm, respectively, p= 0.27).

To have a better insight of any impact of the end-to-end separation of mRNA molecules on translation, the separation between the ends of mRNA molecules was analyzed in homologous (see Table 2.1.2), housekeeping and highly expressed genes (see Table 2.1.1).

As would be expected, when compared, homologous genes from related species (green algae *CR* and *VC*; fishes *LC* and *DR*; eudicots *HB* and *AT* and the hominids *PT* and *HS*), showed no significant differences (Fig. 2.2.4). However, when compared against heterologous genes, the C_L values of homologous, housekeeping and highly expressed genes, have lower C_L values, suggesting that constant level of expression at the protein level of these three types of genes could be related to a small end-to-end distance of their respective mRNA.

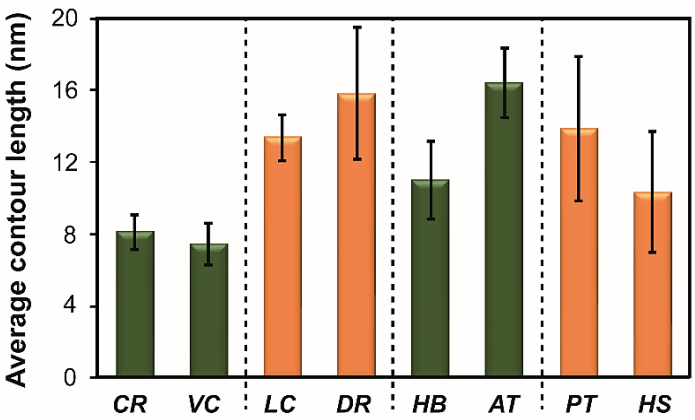

Figure 2.2.4. Contour length distributions from the predicted mRNA secondary structures in related species. Average contour length from the homologous genes in related species *C. reinhardtii* (*CR*), *V. carteri* (*VC*), *L. camtschaticum* (*LC*), *D. rerio* (*DR*), *H. brasiliensis* (*HB*), *A. thaliana* (*AT*), *P. troglodytes* (*PT*) and *H. sapiens* (*HS*). Black dashed lines divide the related species. Bars with the same color represent organism of a same clade: Viridiplantae (green bars) and vertebrates (orange bars). We include the standard error of the average of the mRNA molecules included per specie with a total sample size of N = 70. Statistical differences between the CL values of related species were analyzed by the two tailed Welch´s test for unequal variances.

Since both the length of the 3′-UTRs and their whole GC content could impact on the overall structure of RNA molecules, we decided to investigate whether these two parameters have any correlation with the separation between both ends of mRNA molecules. As can be found in Fig. 2.2.5 and 2.2.6, the 3′-UTR average length is not correlated with their C_L values. Thus, the 3′-UTR length no impact on the C_L value. However, the GC content negatively correlates with a lower C_L value. Fig. 2.2.7 shows that higher GC content negatively correlates with a lower C_L value (Pearson correlation coefficient $r(202) = -0.30$, $p < 0.01$; linear fit to the data a= 18.6 ± 2 nm and b= -0.18 ± 0.04 nm / % GC) (± SD). Likewise, when only the homologous genes were analyzed, a

higher GC content negatively correlates with a lower C_L value (Fig. 2.2.8) ($r(68) = -0.42$, $p < 0.01$; a= 23.7 ± 3.2 nm and b= - 0.25 ± 0.06 nm / % GC) (± SD).

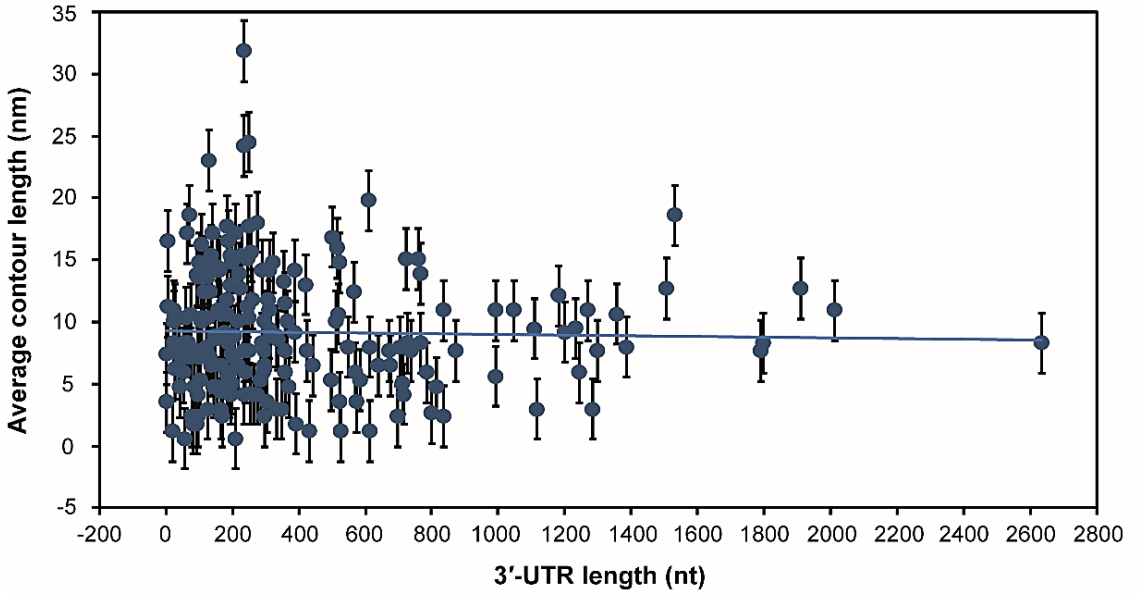

Figure 2.2.5. Average contour length vs 3′-UTR length. The error bars represent the typical variations obtained with mfold and Vienna RNA. The solid line is a linear fit ($y = a + bx$) with a = 9.26 ± 0.47 nm and b = - 0.0002 ± 0.0008 nm/nt. The Pearson correlation coefficient is r (202) = -0.02, p = 0.74, consistent with no correlation. The data presented are ± SD.

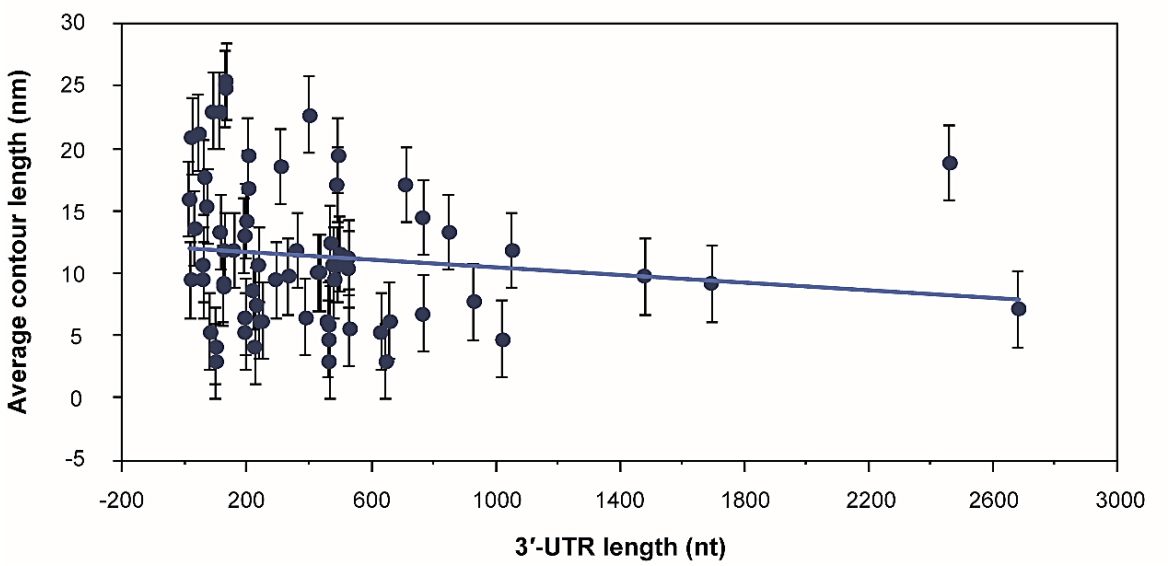

Figure 2.2.6. Average contour length vs 3′-UTR length from homologous genes in related species. The error bars represent the typical variations obtained with mfold and Vienna RNA. The solid line is a linear fit ($y = a + bx$) with a = 11.9 ± 0.9 nm and b = - 0.001 ± 0.001 nm/nt. The Pearson correlation coefficient is r (68) = -0.12, p = 0.28, consistent with no correlation. The data presented are ± SD.

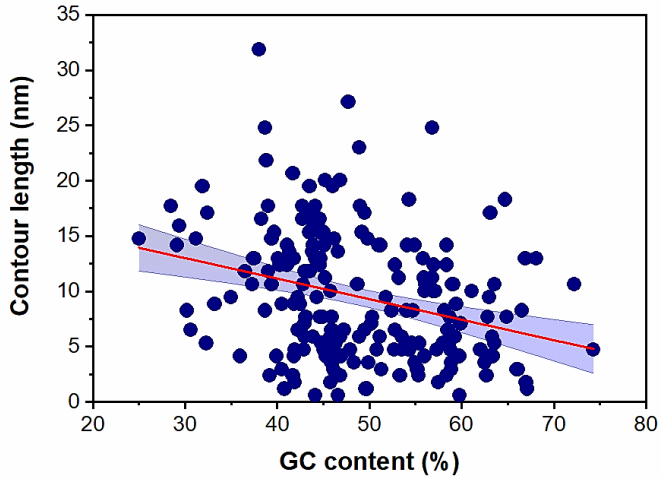

Figure 2.2.7. Contour length vs GC content from the predicted mRNA secondary structures. Blue circles correspond to values obtained using Vienna RNA, *N= 204*. The plot includes the linear fit (red line) ($y= a + bx$) with a= 18.6 ± 2 nm and b= -0.18 ± 0.04 nm / % GC. The Pearson correlation coefficient is r(202) = -0.30, p < 0.01, consistent with significant correlation. The data presented are ± SD. We include the 95% confidence band.

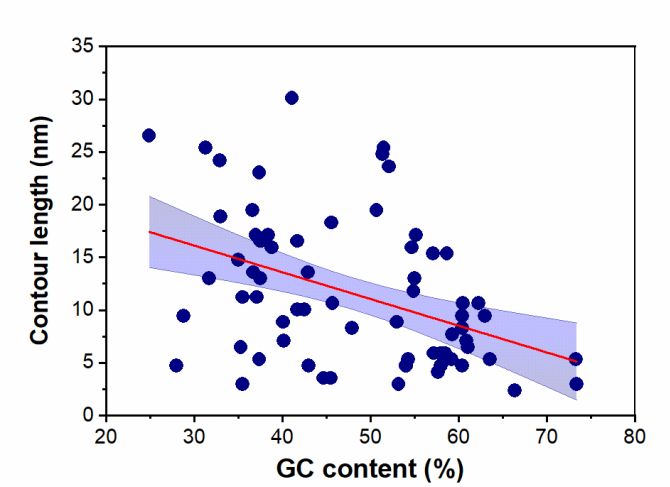

Figure 2.2.8. Contour length vs GC content from the predicted mRNA secondary structures in related species. Blue circles correspond to values obtained from homologous genes using Vienna RNA, *N= 70*. The plot includes the linear fit (red line) (*y= a + bx*) with *a*= 23.7 ± 3.2 nm and *b*= - 0.25 ± 0.06 nm / % GC. The Pearson correlation coefficient is *r*(68) = -0.42, $p < 0.01$, consistent with significant correlation. The data presented are ± SD. We include the 95% confidence band.

2.3. Discussion

There is no doubt about the necessity of an effective circularization of the ends of mRNA molecules to increase their translation efficiency [52]. However, our working hypothesis is that the intrinsic thermodynamic properties given by the sequence of RNA molecules determines the extend of that circularization (C_L size), and therefore, their rate of translation initiation efficiency. In our hypothesis, instead of a mRNA circularization induced by RNA binding proteins [53, 54], the inherent proximity of both ends in the mRNA molecules promotes the initiation of cap-dependent protein synthesis by favoring the recognition of the 5′-cap and the 3′ poly(A) tail by eIF4E and

PABP, respectively [52, 55]. Similarly, in uncapped, non-polyadenylated positive-single RNA stranded plant viruses, the initiation of protein synthesis is driven by 3´ cap-independent translation enhancers (3´CITEs) through base-pairing with complementary sequences in the 5´ UTR inducing an effective circularization that favors the recruiting the initiation factor eIF4F and PABP [56]. Moreover, positive-single-stranded RNA genomes of Flaviviruses (like Dengue and Zika viruses, ~ 11 kb) require an active self-induced circularization for full replication efficiency [57]. Therefore, small C_L values of mRNA molecules would favor their translation efficiency, whereas large separations would decrease it. Strikingly, we found larger C_L values than previously reported [18, 32] (Fig. 2.2.2). Furthermore, full native mRNA sequences show much larger variations than those of random sequences that could not be explained just by statistical variations, and native RNA sequences present a lower MFE values than their correspondent random sequence [58, 59]. Therefore, we reasoned that there must be a biological impact related to the variability in C_L values we observed. In other words, statistical variations are not big enough to explain the variations observed in native sequences, with remarkably high confidence.

In this regard, the efficiency of proteins synthesis is one of the important internal features that allows organisms to adapt and survive to changes in the external conditions. Therefore, the stability of phenotypes could be a feature that might depend on small distances between the ends of mRNA molecules. In line with this, *GB* and *CR*, the two species with the smallest C_L values (4.8 ± 0.7 nm and 6.4 ± 0.9 nm, respectively), have an increased ability to cope with adverse conditions or environmental changes. For example, *GB,* has an impressive capacity to resist serious pests and diseases, as well as a high tolerance to city smoke and industrial fumes [60,

61], whereas *CR* shows a strong phenotypic stability after exposure to either high CO_2 concentrations during 1000 generations [62] or to CO_2 limiting conditions [63]. Moreover, viroids of the Avsunviroidae and Pospiviroidae families, whose genomes are composed of circular ssRNA [64], show a lack of divergence and diversification [65]. Conversely, larger C_L values could favor variability upon stress or pressure, leading to phenotypical instability, and perhaps, divergency. Although the ability to survive or diverge depends on a combination of several characteristics under the appropriate environmental and intrinsic conditions, we expected to obtain some insight by considering the length of the exterior loop as one of the intrinsic conditions to take into account. In this regard, if there is a biological impact for the extent of C_L values, an upper limit in the distance between ends of a mRNA molecule should exist. Using statistical analysis, we determined that C_L values larger than 17.5 ± 2 nm (± SD) are not favored. In addition, we expected that related species in a similar evolutive process would not show significant differences in their C_L values, whereas groups with few extant species would show the smaller C_L values. Supporting these ideas, the eudicotyledones (*HB*/*AT*) and the hominids (*PT/HS*) pairs, which are in similar phylogenetic divergence level, present similar C_L values (Fig. 2.2.4); and *GB*, the most ancient living tree [60, 61] and the only extant species of the Order Ginkgoales have the smallest C_L values we found. Interestingly, *CR* (C_L of 6.4 ± 0.9 nm) is the evolutionary predecessor of *VC* (C_L of 11.3 ± 1.2 nm), and when the C_L values of their homologous genes are compared, no statistical differences could be found (Fig. 2.2.4). However, when the C_L values of their heterologous genes are compared, they are clearly different. Indeed, homologous genes maintain similar C_L values (Fig. 2.2.4). Thus, our results might suggest that heterologous genes could be involved in the ability

of species to diverge, as their C_L values are increased in comparison to that of homologous genes.

Finally, it is important to note that all analyses were performed using full-length mRNA sequences, all of them contained their respective 5´-UTR and 3´-UTR. This means that all sequences used to calculate their C_L values started with their transcription starting nucleotide and included the typical polyadenylation signal (PAS) for that specie, located downstream of the coding sequence. No particular attention was put on alternative PAS, as we only wanted to analyze full-length mRNAs. Furthermore, it could be argued that the poly-A tail should increase the size of the C_L values, although in essence this is true, the cytoplasmic PABP (PABPc) has the ability to bind to a stretch of 12 As with high affinity [66, 67] while covering 25 nt [68]. Therefore, it is very likely that PABPc could bind to the initial segment of the poly-A tail while interacting with the 3´end portion of the exterior loop, generating the appropriate distance to interact with eIF4F, located in the 5´end portion of the exterior loop, thus making the entire size of the poly-A tail irrelevant, at least for the initiation of translation.

CHAPTER 3. smFRET system design and calibration for *in vitro* measurements

In order to perform the physical measurements of the distance between the ends of mRNA molecules from species from the Eukarya domain, we had to work on the smFRET system design and its calibration based on previously reported studies [69]. First, in this chapter, we mention a brief introduction about the smFRET technique. Also, we described the optical setup design used to make some adaptations to an epifluorescence microscope with the purpose to be used as a smFRET for the *in vitro* experiments. Finally, the methodology used, and the main results obtained for the calibration of the equipment are described.

3.1. FRET

Fluorescence resonance energy transfer (FRET) is a powerful spectroscopic technique for characterization of biomolecular structure and dynamics. This technique is used to measure the distance between molecules in the range of 1 to 10 nm and is based on the distance dependent energy transfer between two fluorophores attached to the biomolecule of interest. These fluorophore pair must show a spectral overlap between

the emission spectrum of the donor and the absorption spectrum of the acceptor molecule. By using the appropriate excitation light for the donor, their excitation energy is transferred to a nearby suitable acceptor via an induced dipole interaction. This process can be explained by the Jablonski diagram shown in Fig. 3.1.1 [70-72].

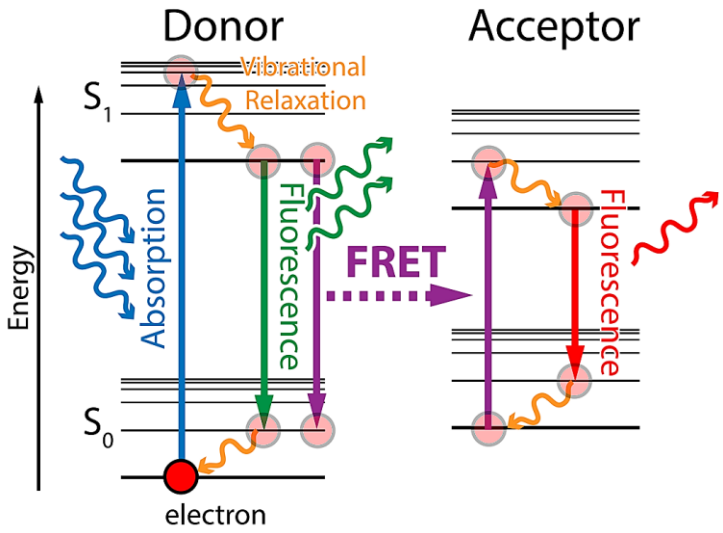

Figure 3.1.1. FRET effect described by Jablonski diagram. The stimulation of a donor fluorophore by the appropriate photon, excite an electron from the ground state S_0 into a higher energy state S_1. Part of that energy is lost by vibrational relaxation. Afterwards, the electron falls back to S_0 and can either emit a photon or the energy can be transferred to an electron of a closer acceptor fluorophore, which is then exited to a higher state S_1 resulting in photon emission of the acceptor. Figure modified from [73].

The efficiency of energy transfer (E), is given by

$$E = \frac{1}{1 + \left(\frac{R}{R_0}\right)^6}$$

where R is the distance between the donor and acceptor and R_0 is the Förster radius at which 50% of the energy is transferred and is a function of the properties of the

fluorophores. Figure 3.1.2 shows energy transfer efficiency as a function of the distance between the dyes. The R_0 value for the fluorescent dye pair is generally provided by the suppliers.

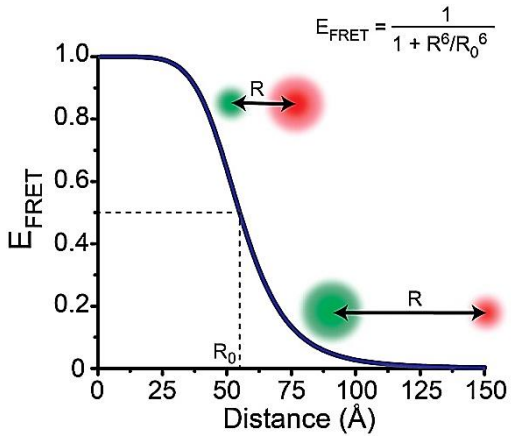

Figure 3.1.2. Energy transfer efficiency vs distance. The Energy transfer efficiency rapidly increases as the separation distance decreases below R_0, and conversely.

For FRET experiments, the selection of pair of fluorescent dyes should be considered depending on the separations involved in the sample under study. For example, the dyes pair chosen for our experiments were Alexa Fluor 546 (AF546) and Alexa Fluor 647 (AF647). In Fig. 3.1.3 it is shown the fluorescence spectra for the selected dyes pair which, has a reported value of R_0= 7.4 nm [74], therefore we will have information at distances larger than 7.4 nm.

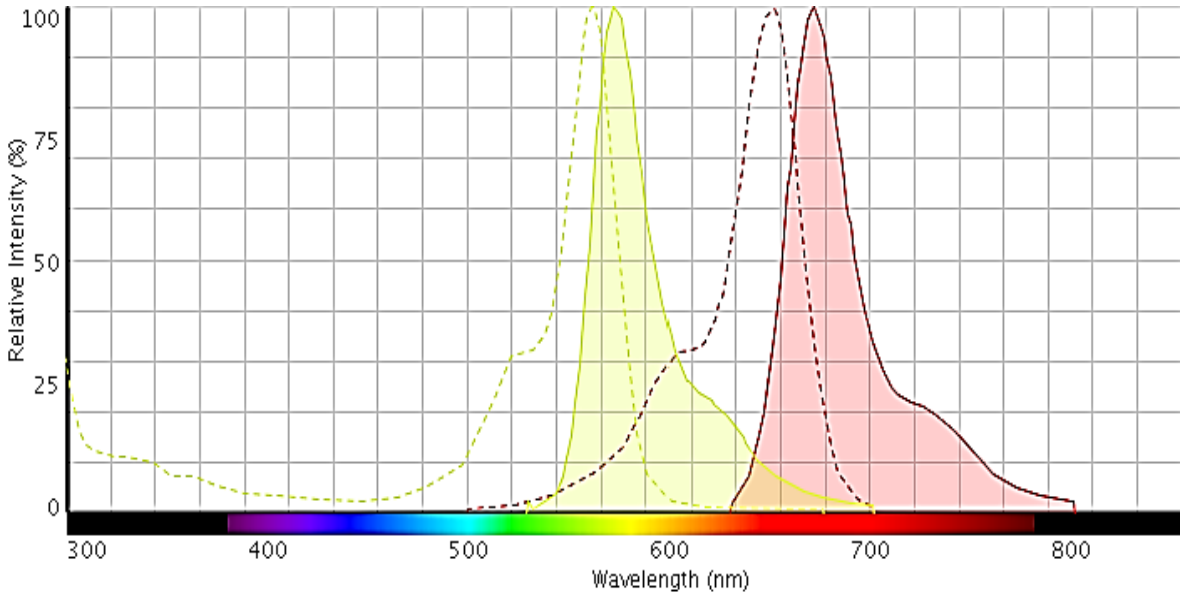

Figure 3.1.3. Fluorescence spectrum from AF546 and AF647. The dye pair shows an overlap between the emission spectrum of the donor (AF546, yellow solid line) and the excitation spectrum of the acceptor (AF647, red dashed line). Image modified from ThermoFisher Fluorescence SpectraViewer [75].

In vitro experiments with FRET using microscopy comes with a set of technical challenges to recover relevant information. FRET experiments measure the donor and acceptor intensity by passing the emission through a series of optical elements to avalanche photodiode detectors or a sensitive digital camera. Then, the apparent FRET efficiency (E_{app}), is given by

$$E_{app} = \frac{I_A}{I_A + I_D}$$

where I_A and I_D represent the acceptor and donor intensities (fluorescent signals), respectively. $Eapp$ provides only an approximate indicator of the inter-dye distance because of uncertainty in the orientation of the dipole moments (κ^2) between the two

fluorophores and the required instrumental corrections [76]. As a rule of thumb, if fluorescence anisotropy of both fluorophores is less than 0.2, κ^2 is close to 2/3 [76].

To recover the true FRET efficiency, is necessary to make some corrections to the intensity values as

$$E = \frac{I_A - \beta I_D}{(I_A - \beta I_D) + \gamma I_D}$$

where βI_D corrects for leakage of donor emission into the acceptor channel, the factor γ depends on the difference between the detection efficiencies of the donor η_D and the acceptor η_A as well as the quantum yields φ_A and φ_D respectively,

$$\gamma = \frac{\eta_A \varphi_A}{\eta_D \varphi_D}$$

Thus, the value of γ adjusts for differences between the donor and acceptor dyes in their probability of photon emission upon excitation and the probability that emitted photons will be detected. In single molecule FRET (smFRET) microscopy, methods for γ determination vary depending on experimental methodology.

Due to its strong distance dependence, smFRET efficiency can be used as a spectroscopic ruler. The principal advantage to use this technique is the possibility to resolve the signal of each individual molecule with a single fluorophore pair allowing for precise analysis of heterogeneous populations. For example, it is known that RNA folding goes through multiple interactions, folding pathways, and intermediates before reaching its native state. Thus, this technique is accurate to study RNA folding dynamics, and allow to capture data from different conformation transitions of the molecule [70].

3.2. smFRET Optical setup

To perform the measurements of the distance between the ends of the Eukarya mRNA molecules, we made some adaptations to an epi-fluorescence microscope Nikon Eclipse E800 (Figure 3.2.1) placed at the Biological Physics Laboratory in the Physics Institute from the Universidad Autónoma de San Luis Potosí (UASLP).

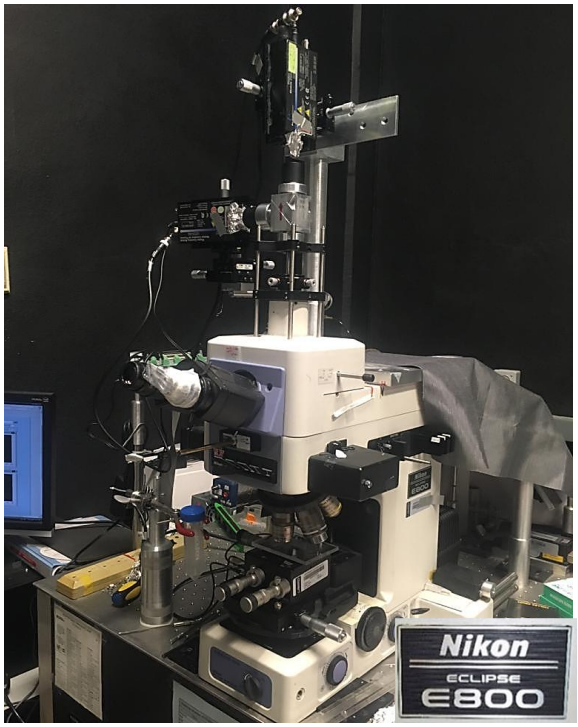

Figure 3.2.1. Picture of the epi-fluorescence microscope adapted as a smFRET. The microscope is placed in the Biological Physics Laboratory from UASLP.

For the light system illumination, the excitation laser light was chosen to match the properties of the donor dye (AF546) and to reduce direct excitation of the acceptor (AF647). A diode pumped solid state laser centered at 515 nm (Spectra-Physics

Excelsior-515-50) was adapted to the microscope by using two mirrors to direct the laser light into the microscope collector lens. To increase the incident laser beam diameter into the collector lens, a beam expander (10X objective CP-Achromatic 10X/0.25, Zeiss Optics), mounted on three-dimensional micrometer translation stage (NF15AP25, Thorlabs) to facilitate its alignment with the optical path, was placed between the optical path from M2 to the collector lens.

In order to achieve single molecule sensitivity, above the head of the microscope, we placed at the image plane position a mounted precision 100 µm pinhole (Thorlabs) to block the out of focus fluorescence signal. Then, we used a trinocular tube cube to place a dichroic mirror (DMLP605R Longpass Dichroic Mirror, 605 nm Cut-On, Thorlabs) mounted at 45° from the vertical plane (inside the cube) to separate the donor and acceptor emissions. Even when our dichroic mirror can separate efficiently the components of light from the donor and acceptor, we decided to add two band pass filters. For the horizontal component, we mounted a band pass filter centered at 580 nm with FWHM=30 (580DF30, Omega optical) and for the upper vertical component we mounted a band pass filter centered at 670 nm with FWHM=40 (670DF40, Omega optical). By using doubled achromatic lenses (AL-D with EFL = 5 mm and AL-A with EFL = 4 mm, Edmund Optics), each component (donor and acceptor fluorescence) was focused onto a single photon counting module (SPCM-AQR-14, Perkin-Elmer Optoelectronics) mounted on three-dimensional micrometer translation stages (NF15AP25, Thorlabs). Fig. 3.2.2 shows the configuration of our smFRET optical set up mounted in the laboratory. For each fluorescent photon burst detected by the single photon counting module (SPCM) in an integration time t = 1 ms, a stream of pulses was sent directly to a counter in a SCB-68 card (National Instruments) and stored in a

computer through a home-built Labview algorithm. FRET signals analysis was performed offline by using an algorithm written on MATLAB.

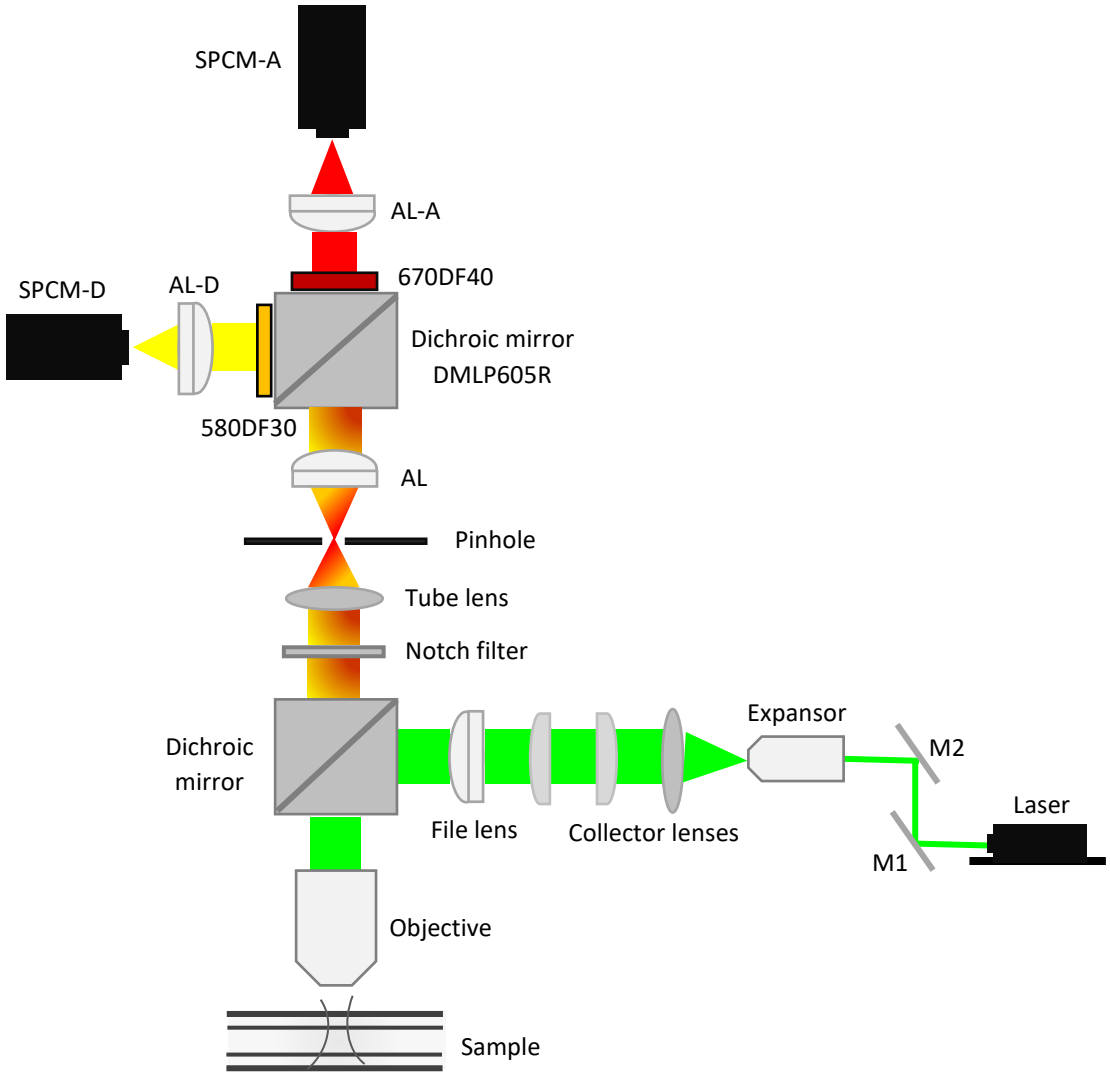

Figure 3.2.2. Schematic representation of the smFRET optical setup configuration mounted in the laboratory. Excitation laser light at 515 nm is shown in green. Fluorescence emission from donor and acceptor is shown in yellow and red respectively. The dichroic mirror separates the fluorescent light components from the donor and acceptor directing it toward the corresponding SPCM-D (for the donor) and to the SPCM-A (for the acceptor).

3.2.1. smFRET detector alignment

To be able to line up the SPCMs with the correspondent signal emission, SPCM-D was aligned by using 1 µM of AF546 in TE buffer to achieve signal in the donor channel (D-channel). SPCM-A was aligned by using a mixture of 500 nM of AF546 and 500 nM of AF647 in TE buffer to achieve FRET emission in the acceptor channel (A-channel). Both SPCMs were lined up with the corresponding light by hand through the micrometer translational stage, until a maximum fluorescent signal was detected in a bin time of 1 ms.

3.2.2. Background signal contribution

After having the detectors aligned, is necessary to measure the background signal from the electronic dark counts. First, we prepared a clean sample chamber by using a rectangle miniature hollow glass tubing (0.05 x 0.5 x 50 mm, VitroCom, cat. 5005-050). Each chamber was prepared as shown in Fig.3.2.2.1. A plastic tube was attached at both extremes of the glass tubing and fixed with silicon. Then, the prepared glass tubing was placed at the top of a half of microscope slide and then fixed with epoxy resin onto another microscope slide (both microscope slides previously cleaned with acetone) covering very well the junction between the extremes of the glass tubing and the plastic tube to avoid leakage at the moment to flow the sample. Before loading the sample, each chamber was cleaned by flowing through the plastic tube (using a micro syringe) autoclaved deionized water (18 $M\Omega cm^{-1}$) and incubated with 1 M of potassium hydroxide (KOH) for 3 min. After that, washed again with autoclaved deionized water

and with absolute ethanol (ETOH). Finally, washed 2 times with autoclaved deionized water.

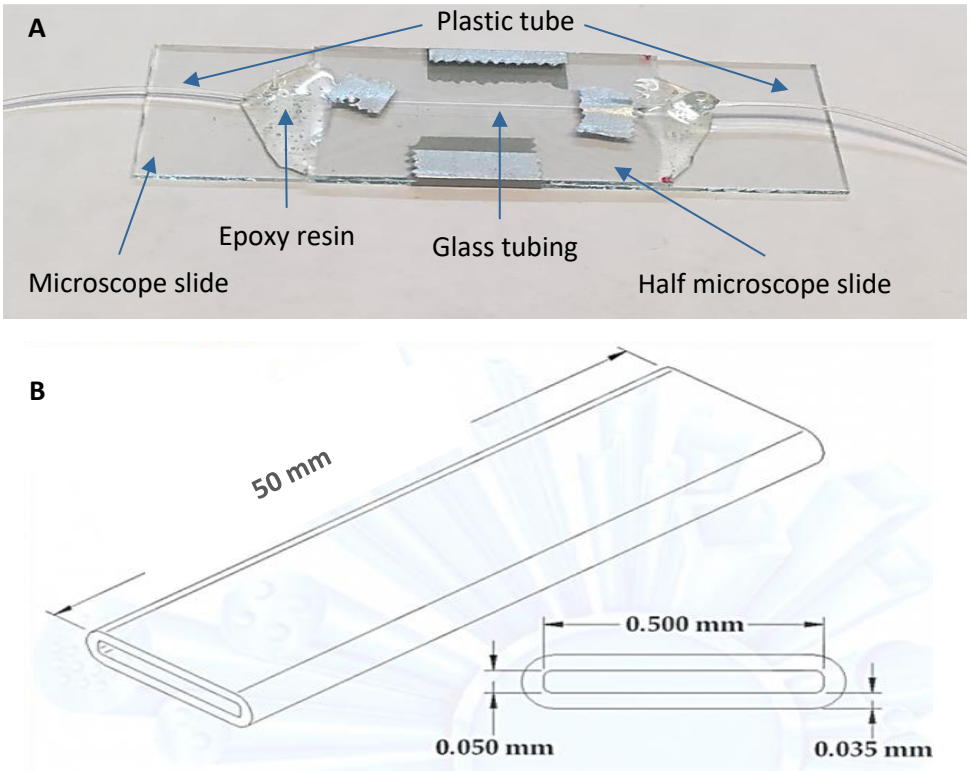

Figure 3.2.2.1. smFRET sample chamber. A) Sample chamber built in the lab for *in vitro* experiments. B) Large schematic representation of the glass tubing used to flow the sample (figure adapted from [77]).

The background counts were measured by flowing into the clean sample chamber 1X TE buffer, which is the solvent used in this thesis, and focused on different laser power with a 100X immersion oil objective. For all FRET experiments we used the immersion oil for microscopy (Merck) with a refractive index between 1.515 to 1.517 and fluorescence ≤ 1500 ppb. Fig. 3.2.2.2 shows that a higher excitation power corresponds to an increased level of background counts. Therefore, for the smFRET experiments it

is necessary to use low intensity power to focus the sample to reduce the background signal. For example, at 122.5 µW excitation power on 1X TE buffer, dark counts are low, allowing to differentiate between signal and noise in smFRET experiments (Fig. 3.2.2.3 A). Moreover, at this excitation power the mean counts in both channels are around 2 counts/ms (Fig. 3.2.2.3 B). Therefore, is convenient to work at maximum excitation power of 122.5 µW in the future smFRET experiments in order to distinguish signal (approximately 30 counts/ms [69]) that comes from the sample from the signal that comes from the background.

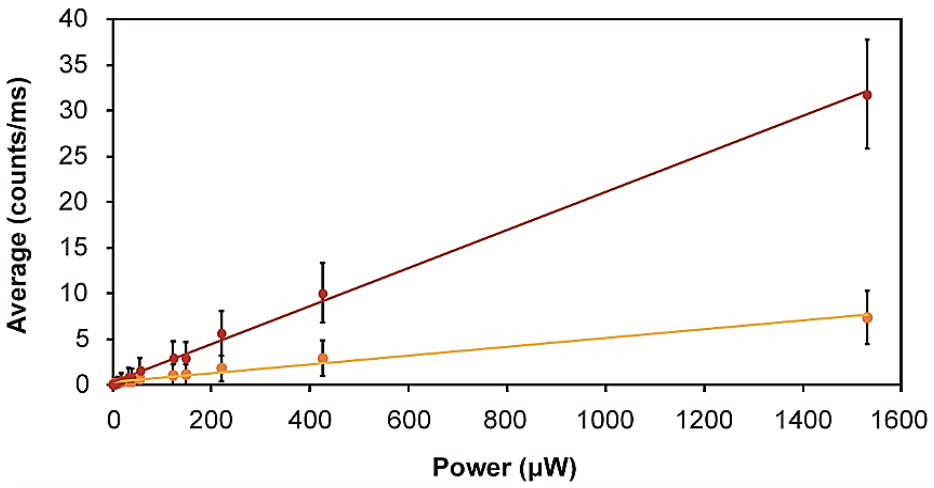

Figure 3.2.2.2. Average photon counts per millisecond vs power. Number of counts detected at different excitation power on 1X TE buffer. The red circles and yellow squares correspond to data obtained for the A-channel and D-channel, respectively. The solid line is a linear fit ($y = a + bx$) with a = 0.13 ± 0.17 counts/ms and b = 0.02 ± 0.003 counts/ms µW (A-channel) and a = 0.18 ± 0.15 counts/ms and b = 0.005 ± 0.001 counts/ms µW (D-channel). We include the standard deviation of the average with N = 15.

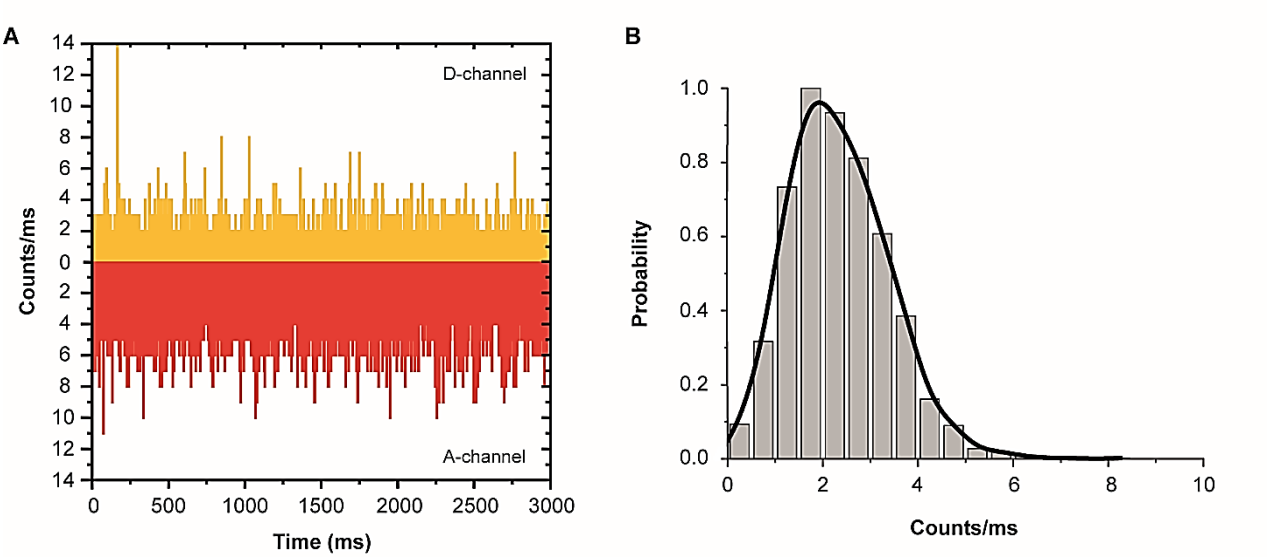

Figure 3.2.2.3. Background counts in both channels. A) Number of counts detected at 122.5 μW excitation power on 1X TE buffer solution in 1 ms. Yellow and red correspond to data obtained in the D-channel and A-channel respectively. B) Histogram to show the mean counts in both channels (the average counts from the sum in D-channel and A-channel) for 1X TE buffer using 122.5 μW excitation power with a Poisson fit with λ = 2.08 ± 0.03 (black line).

3.3. Optical system calibration

The methodology used to perform our optical smFRET system calibration was by determining the transfer efficiencies of short fragments of dsDNA with lengths of 10, 13, 22, 28 and 45 base pair (bp) labeled with fluorescent dyes at the ends of the molecule. Thus, our short fragments of dsDNA can be treated as a rigid rod since they are much smaller than the DNA persistence length (150 bp) [78]. In this way, we can use this short dsDNA molecules as a spectroscopic ruler for single-molecule fluorescent calibration. The DNA lengths were chosen considering the R_0 of the fluorophores pair (AF546 and AF647), which is 7.4 nm [74]. The fluorescent dyes used were Alexa Fluor 546-14-dUTP (AF546-dUTP, Invitrogen) as a donor and Alexa Fluor

647-aha-dCTP (AF647-dCTP, Invitrogen) as acceptor. These fluorescent molecules are nucleotides (deoxyuridine triphosphate and deoxycytidine triphosphate) attached to the dyes which are modified at the C-5 position of uridine and cytosine by an alkynyl amino flexible linker, thus providing a spacer between the nucleotide and the dye in order to reduce interaction between them. The length of this flexible amino linkers is about 14 atoms (given by the company), which correspond to the length of long linkers. This length of the linker can cause uncertainties in smFRET distance measurements that is necessary to consider [79]. Thus, is necessary to add the contribution to the effective length for long linkers, which is approx. 0.75 nm with a deviation of 0.22 nm, to the length of each DNA [69].

3.3.1. Annealing ssDNA oligos

To obtain dsDNA, we designed short fragments of ssDNA and its complement to be able to hybridize them and get dsDNA with 3' recessive ends. In this way, we can procced with the dsDNA labeling by using the Klenow enzyme and incorporate the fluorophores AF647-dCTP and AF546-dUTP by filling the protruding ends in the dsDNA.

ssDNA fragments and their complementary ones were purchased to ADN SINTETICO T4oligo. The oligonucleotide sequences forward (FOR) and reverse (REV) are shown in Table 3.3.1.

Table 3.3.1. ssDNA sequences and its complement for smFRET calibration.

FOR10	AGACGTGAG
REV10	GCTCACGTC
FOR13	AGACAAGGTGAG
REV13	GCTCACCTTGTC
FOR22	AGACGTGTTGTGAACCGTGAG
REV22	GCTCACGGTTCACAACACGTC
FOR28	AGACGTGTGACCGCATTTTTGAGTGAG
REV28	GCTCACTCAAAAATGCGGTCACACGTC
FOR45	AGACGCGCTTACTAGTGCAAATTGTGACCGATTTTGAGTGAG
REV45	GCTCACTCAAAATCGGTCACAATTTGCACTAGTAAGCGCGTC

To anneal FOR and REV complementary sequences, first, each oligo was resuspended with NFW (nuclease free water) to a final concentration of 100 µM. Then, 20 µM of each complementary oligo was mixed in a PCR tube (free of DNases) with 10X annealing buffer (100 mM Tris pH.8, and 500 mM NaCl, filtered with 0.22 µm Millipore membrane) and NFW to get a final reaction volume of 30 or 50 µl. Finally, the hybridization was started by using the MasterCycler Nexus Gradient 94°C x 3 min, 75 cycles (-1°C/cycle, 1min per cycle) and 4°C HOLD.

The annealing DNA molecules obtained were verified by running an 8% and 15% polyacrylamide gel (PAGE) at 60 Volts for 15 and 130 min respectively (see Fig. 3.3.1.1). Then, the dsDNA molecules were purified from PAGE by following the Crush and Soak Method [80]. Finally, the samples were stored at -20° C until needed.

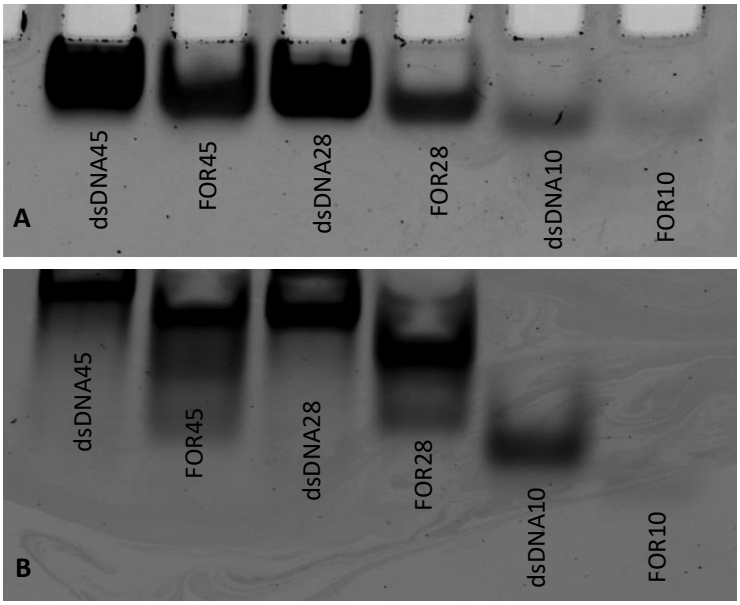

Figure 3.3.1.1. Polyacrylamide gels showing the hybridization of the DNA molecules. DNA hybridization was verified by loading samples on A) 8% PAGE and running at 60 V for 15 min and B) 15% PAGE and running at 72 V for 130 min. Molecules with a lower molecular weight (ssDNA (FOR)) run first than the higher molecular weight (dsDNA).

3.3.2. dsDNA labeling

To label the both ends of our 5' overhang dsDNAs with the fluorescent dyes (AF546-dUTP and AF647-dCTP), we used the Klenow Fragment (3' → 5' exo$^-$) from New England BioLabs. This enzyme can fill protruding ends of DNA molecules by its polymerase activity but has lost the 5' → 3' exonuclease activity and has mutations to abolish the 3' → 5' exonuclease activity. The labeling reaction was as follow: In a microcentrifuge tube add 5 µl of 10X NEBuffer, 0.6 µg of dsDNA, 1 nmol of AF546-dUTP, 1 nmol of AF647-dCTP and 5 U of Klenow enzyme. Adjust with NFW to a final reaction volume of 50 µl and incubate at 25°C for 2 h. After labelling reaction, the sample was purified 2 to 4 times by using Illustra MicroSpin G-25 Columns from GE

Healthcare or mini–Quick Spin Columns from Roche, following the manufacturer protocol. Figure 3.3.2.1 shows the absorption spectra obtained for the dsDNAs labeled with AF546-dUTP and AF647-dCTP after one purification by using the columns. From the absorption spectra is possible to estimate the labeling dye efficiency, by using

$$E_{dye} = \frac{\left(A_{dye} \times \epsilon_{base}\right)}{\left(A_{base} \times \epsilon_{dye}\right)}$$

where E_{dye} is the dye efficiency, A_{dye} and A_{base} are the absorbances for the dye and nucleic acid respectively, ϵ_{base} and ϵ_{dye} are the molar extinction coefficients for the nucleic acid and the dye respectively. Also, there is a correction factor in A_{base} that needs to be considered because many fluorophores absorb light at 260. The correction factor value for the AF546 is 0.21 and for AF647 is 0 [81].

The average values obtained for the estimation of the DNA labelling efficiency with the AF546-dUTP and AF647-dCTP was above 50 ± 12 % and 66 ± 24 % (± SD), respectively. Nonetheless, this result should be interpreted with care because the method used to know the labeling efficiencies is not a robust method. Indeed, with this method is not possible to differentiate between the dyes attached to the nucleic acids of the unincorporated dyes.

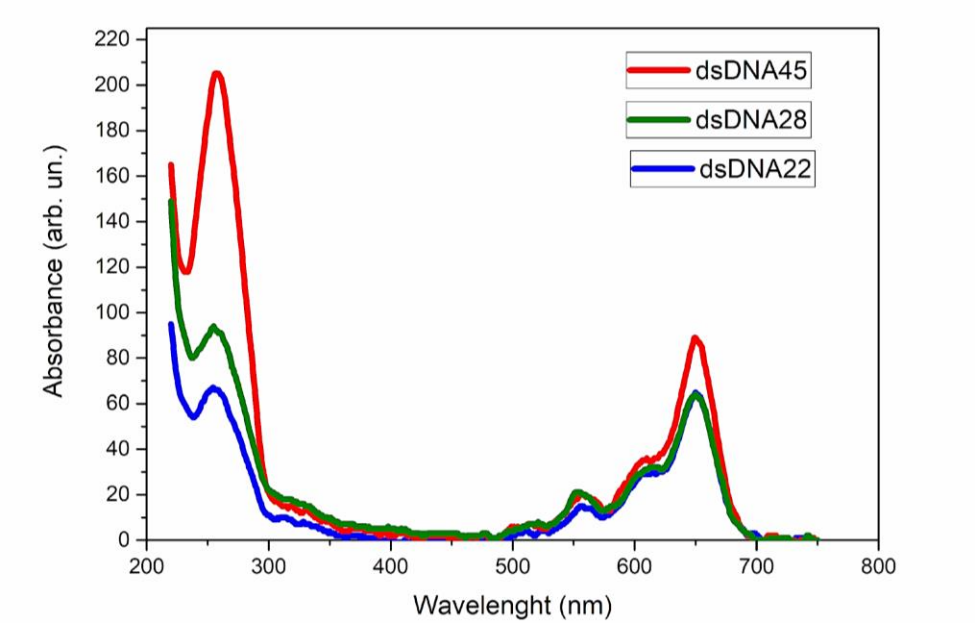

Figure 3.3.2.1. Labeled dsDNAs absorption spectra. The peak at 260 nm corresponds to the absorbance of the nucleic acids. The peaks at 546 nm and 647 nm correspond to the AF546-dUTP and to the AF647-dCTP respectively. Red, green and blue curves correspond to the labeled dsDNA of 45, 28 and 22 bp in length.

It is important to note that, these efficiencies have been achieved after one column purification. Surprisingly, after purifying the labeling reaction one more time, the labeling efficiency decreases. Thus, we had a remaining amount of unincorporated dyes on the labeling reaction. This feature led us to think, that by using this labeling protocol, we are not obtaining a high DNA labeling efficiency. At the end, after 4 purifications we get labeling efficiencies of 7.5 ± 6 % and 14 ± 12 % (± SD) for AF546-dUTP and AF647-dCTP respectively. However, contrary to the labeling efficiencies reported before [18], this final labeling efficiencies should be enough to achieve good signal for the smFRET calibration due to the fact that we are working at a single molecule level.

3.3.3. Measurements of the distance between the ends of labeled dsDNA molecules

Attempting to get smFRET calibration, 0.5 ml of labeled dsDNA (10 to 100 pM) of different lengths (10, 13, 22, 28 and 45 bp) in TE buffer (10 mM Tris, 1 mM EDTA, pH 8) was flowing into a clean sample chamber (see section 3.2.2 for further details) and focused with 100X immersion oil objective. The smFRET signals that we achieved were similar to the signals shown in Figs. 3.3.3.1 and 3.3.3.2.

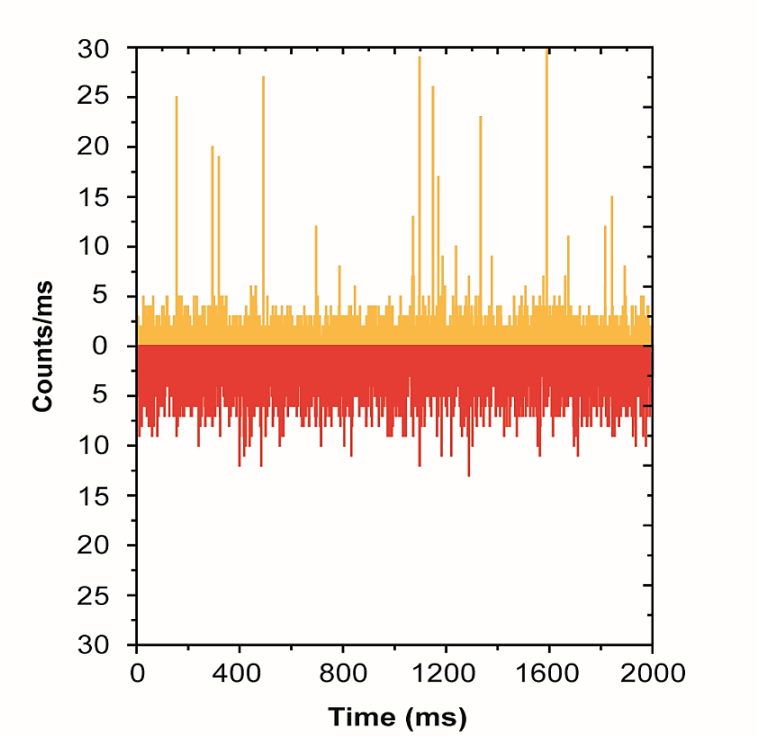

Figure 3.3.3.1. smFRET signals obtained from freely diffusing labeled dsDNA10 molecules in TE buffer. Sample concentration of 10 pM and excitation power of 220 µW. Signals detected in the donor and acceptor channels correspond to data in yellow and red respectively.

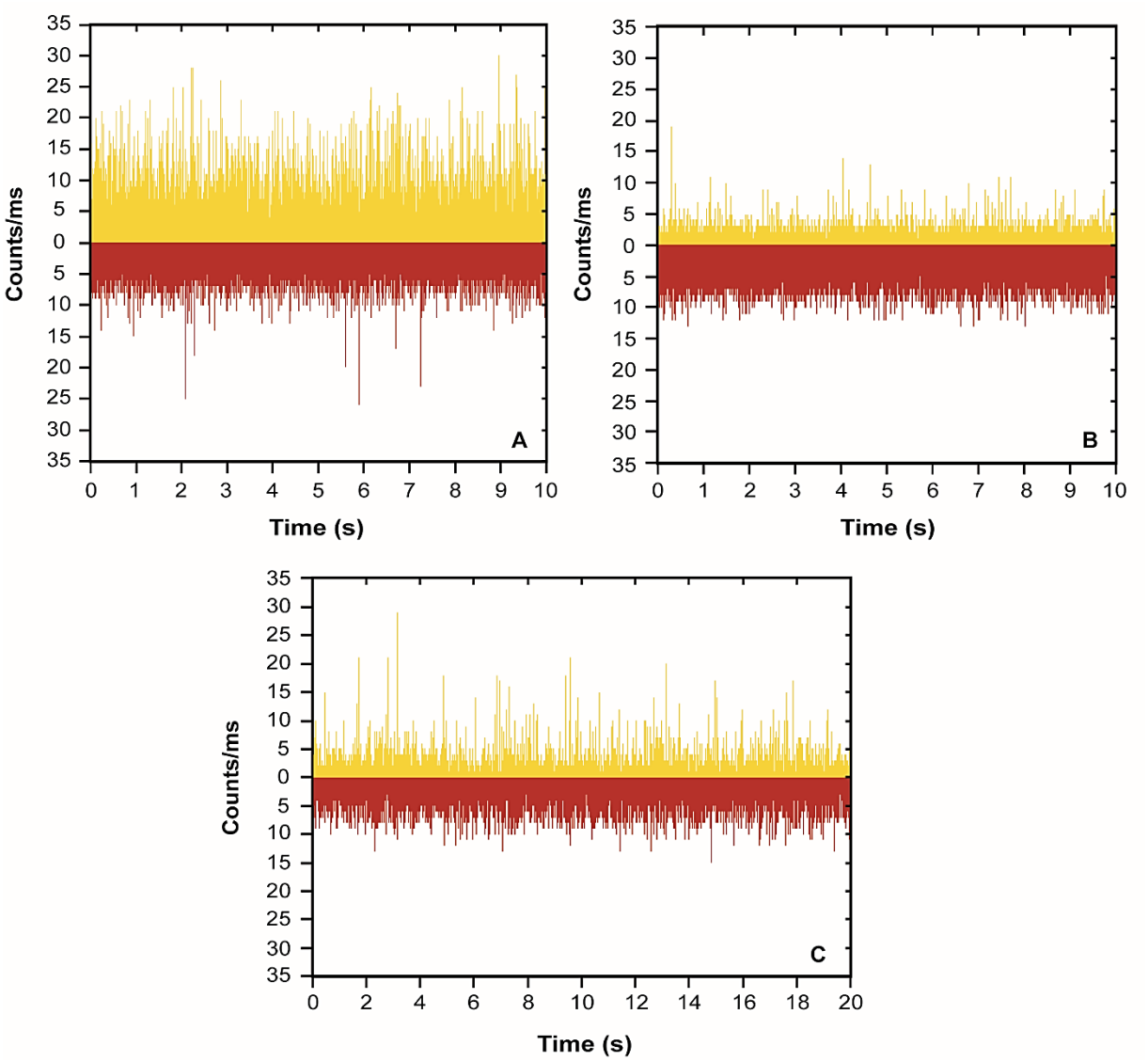

Figure 3.3.3.2. smFRET signals obtained from freely diffusing labeled dsDNA13 molecules in TE buffer. Signals detected in the donor and acceptor channels correspond to data in yellow and red, respectively. A) Sample concentration of 72 ng and excitation power of 122.5 µW. B) Sample concentration of 10 pM and excitation power of 0.22 µW. C) Sample concentration of 120 pM and excitation power of 0.22 µW.

After a lot of attempts, by using the labeled dsDNA13 (13 bp in length) we get something that seems to fit some of the expected smFRET signal features (Fig. 3.3.3.3). Each donor peak has its corresponding acceptor peak, and they show the

correct time widths of 3 to 5 ms. Although, this kind of signal was achieved only once and is important to note that the counts detected per millisecond are remarkably high, it is not clear that is a valid signal yet.

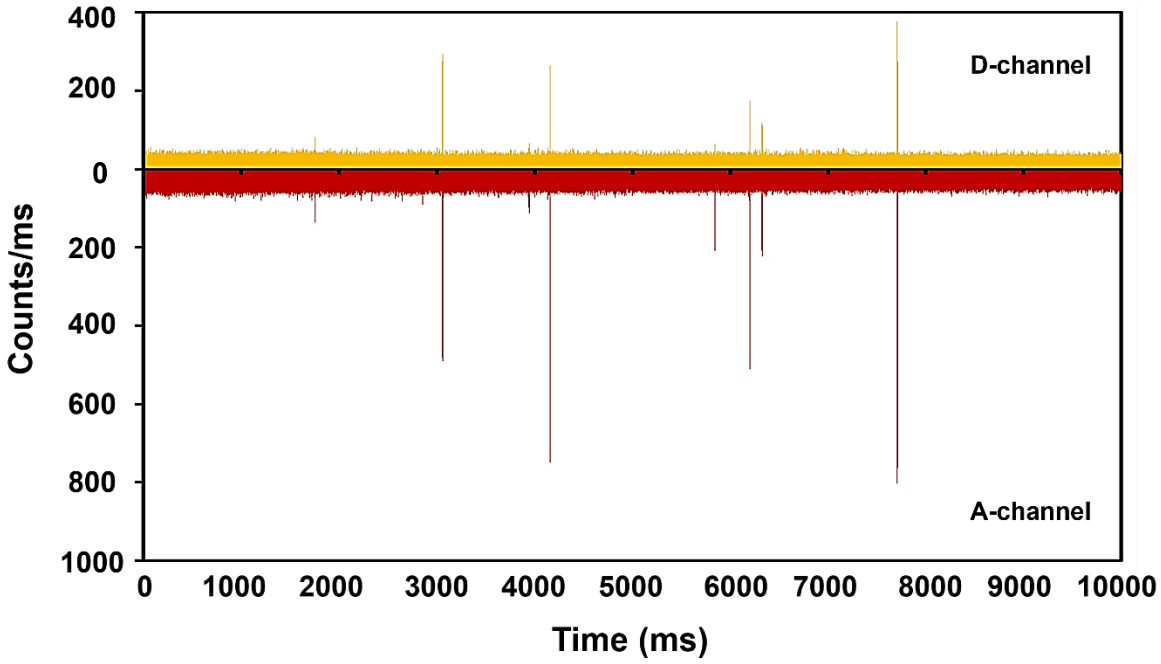

Figure 3.3.3.3. smFRET signals from freely diffusing labeled dsDNA13 molecules in TE buffer. Signals detected in the donor and acceptor channels correspond to data in blue and red respectively. Sample concentration of 1790 ng and excitation power of 122.5 µW. The width of each peak is 3 ms but we can find peaks up to 5 ms.

Many attempts were made to obtain smFRET signal without positive results. Also, changes in the DNA labeling protocol were implemented (by changing the amounts of the reagents) but the expected FRET signals never appeared. For this, we decided to check the DNA labeling products by PAGE. In Fig. 3.3.3.4 we can observe the differences between the non-labeled dsDNA25 (25 pb), the labeled dsDNA25, the AF546 and AF647. As we can observe from the gels, it seems that no remnants of

AF546 or AF647 are present in the labeled dsDNA25 (lane 2), suggesting that we have a complete labeling reaction. Is important to note that samples in lanes 1 and 2 were loaded by using bromophenol blue, which indeed has a fluorescence, as it can be confirmed for the appearance of one band which is marked with an arrow. Also, another PAGE experiment could be conducted, but this time by loading the samples by using glycerol and acquiring the images by using the preferred filter for each application. However, this kind of experiment does not ensure that we go in the right direction to finally obtain the real smFRET signals. For this reason and because of the uncertainty if the samples are correctly labeled or if the smFRET equipment is working properly, to save time and money, the best solution is to corroborate that the smFRET equipment is working correctly. To do this, a dsDNA12 (with 12 bp in length) labeled with AF546 NHS Ester and with AF647 NHS Ester was purchased from IDT. This labeled dsDNA12 probe was first ssDNA labeled at the 5' ends of each complementary sequence FOR (5' AGACAAGGTGAG 3') and REV (5' CTCACCTTGTCT 3') with AF546N and AF647N respectively and then hybridized to form a labeled dsDNA. Unfortunately, due to the lack of time to conclude the Doctoral project, these experiments were pending. However, this proposal should be effective to continue with the experiments, and the result will show us if another labeling strategy should be implemented or if the smFRET equipment needs some improvements.

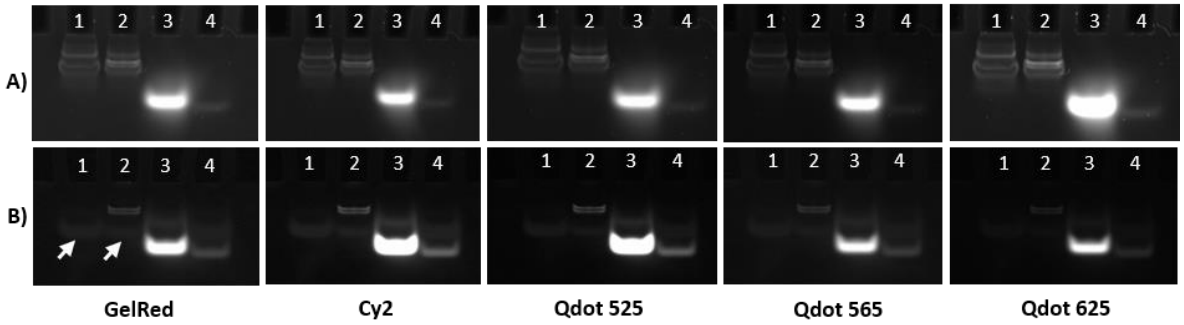

Figure 3.3.3.4. PAGE to corroborate the labeling dsDNA reaction. The dsDNA25 was loaded on 15 % PAGE and was runed at 75 Volts for 1 h. The rows correspond to A) post-stained gel with GelRed and B) not post-stained gel. The 5 columns correspond to different manner of image acquisition of the same gel (A) or B)) by using ChemiDoc XRS+ equipment from BIO-RAD (GelRed, Cy2, Qdot 525, Qdot 565, Qdot 625 applications), all of these images were acquired by using UV light source and the alternate standard filter. The lane 1, 2, 3 and 4 correspond to dsDNA25, labeled dsDNA25, AF546 and AF647 respectively.

CHAPTER 4. Study of the separation between the ends of mRNA molecules from organisms from the Eukarya domain by smFRET

In Chapter 2 it was shown, by using computational programs, that the distance between mRNA ends is not constant and varies among organisms. This distance implies the existence of a biological mechanism responsible for the increase in the observed variability, suggesting that the C_L features of the exterior loop could be relevant for the initiation of translation. In this chapter, we studied the distance between mRNA ends molecules from 4 organisms from the Eukarya domain by using computational programs and smFRET.

Using both mfold and RNA Vienna algorithms, we calculated the minimum free energy (MFE) secondary structures of 16 mRNA molecules from 4 organisms from the Eukarya domain to estimate the distance between their ends before we start with the smFRET *in vitro* experiments. The complete mRNA sequences were selected randomly, with the requirement of having the complete sequence of the 5' and 3' UTRs. Also in this chapter, it is described all the experiments performed for the *in vitro* measurements to determine the end-to-end distance between the mRNA ends by using smFRET.

4.1. mRNA sequences from Eukarya domain species

To select the mRNA sequences from Eukarya domain species (Table 4.1), we performed a search of complete sequences from the GenBank considering the same features mentioned in section 2.1, with the requirement that all mRNAs have a length that falls between 400 and 3500 nt including both UTRs and coding sequence (CDS). The four organisms selected from the Eukarya Domain to obtain the mRNA sequences are shown in Figure 4.1.1. Following a simple complex organism order, the organisms chosen were the obligate intracellular organism *Plasmodium falciparum* (*PF*), the unicellular fungi *Saccharomyces cerevisiae* (*SC*), the flowering plant *Arabidopsis thaliana* (*AT*) and the hominid *Homo sapiens* (*HS*).

Table 4.1. mRNA molecules studied from 4 organisms from the Eukarya domain.

Organism	mRNA name	Total length (nt)	GenBank
P.falciparum (PF)	17 kD sexual stage protein	1094	M64107.1
	cAMP dependent protein kinase catalytic subunit	1230	U78291.1
	gamma-tubulin	1888	X62393.1
	dynamin like protein	2996	AF336796.1
S.cerevisiae (SC)	13 kD vacuolar	434	U21240.1
	Ribosomal protein S21	539	D11386.1
	Sec61 beta-subunit homolog	829	L38891.1
	Y-helicase protein 1	3104	AB016599.1
A.thaliana (AT)	Thionin Thi2.2 mRNA	718	L41245.1
	rac GTP binding protein mRNA	985	AF079485.1
	Glutamate decarboxylase mRNA	1874	NM_121739.3
	Heat shock mRNA	3105	U13949.1
H.sapiens (HS)	Interleukin 5 mRNA	816	NM_000879.2
	Actin beta mRNA	1812	NM_001101.3
	mRNA for Ubiquitin protein ligase	2850	AB056663.2
	Cadherin 10 mRNA	3436	NM_006727.3

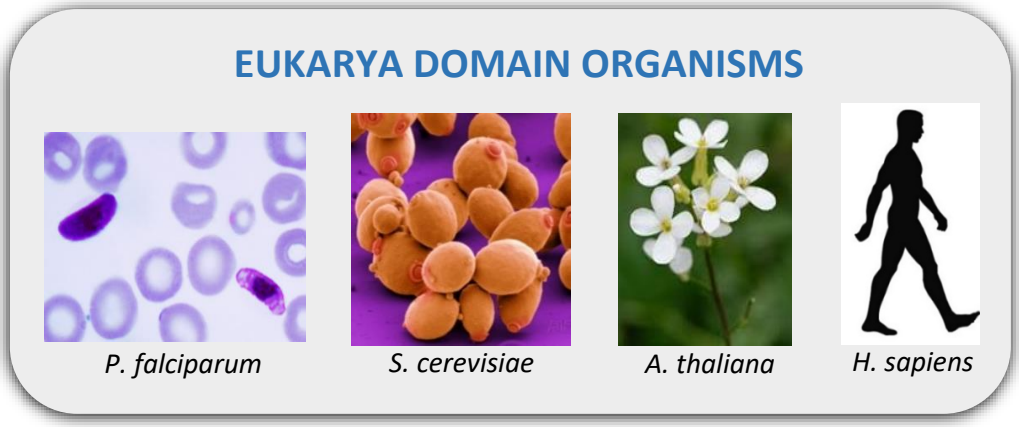

Figure 4.1.1. Model organisms from the Eukarya domain. From each species, four native mRNA sequences reported to the GenBank were selected.

4.2. Predicted distance between ends of mRNA molecules

Using both mfold and Vienna RNA algorithms we obtained the MFE secondary structures and calculated the contour length (C_L) of the exterior loop in the same way previously described in section 2.1.2. Figs. 4.2.1 and 4.2.2. shows that the C_L values varies from 0.59 (*SC*) up to 31.8 nm (*AT*). Thus, we can speculate that for the *in vitro* measurements of the mRNA end to end distance, there will be no FRET signal for some of the *AT* genes which have presented larger C_L values. However, it is important to consider that the contour length is flexible and indeed is possible to find a configuration with the ends closer [18, 82].

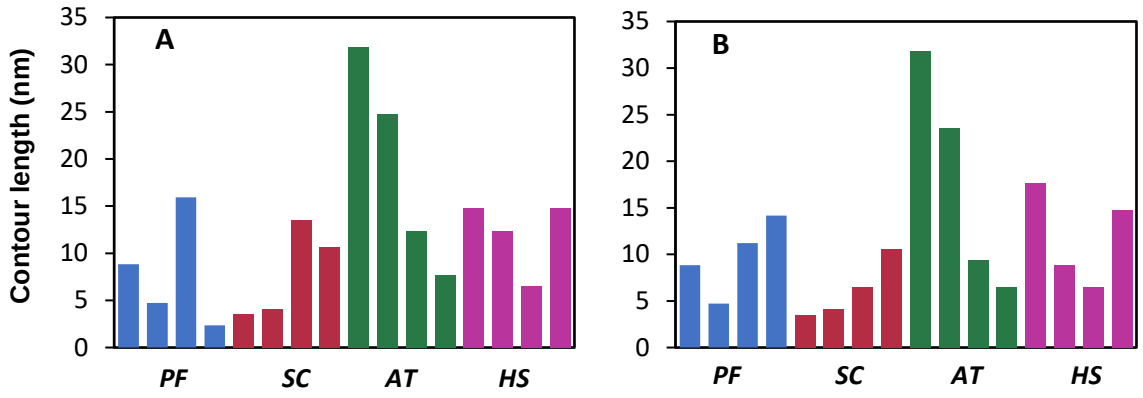

Figure 4.2.1. Contour length distributions from the predicted mRNA secondary structures. Obtained by (A) Vienna RNA and (B) mfold algorithms and by simplex complexity organism order, starting with *P. falciparum* (*PF*), *S. cerevisiae* (*SC*), *A. thaliana* (*AT*) and *H. sapiens* (*HS*). The bars correspond to the value of each mRNA used for each organism, with a total sample size of *N* = 16.

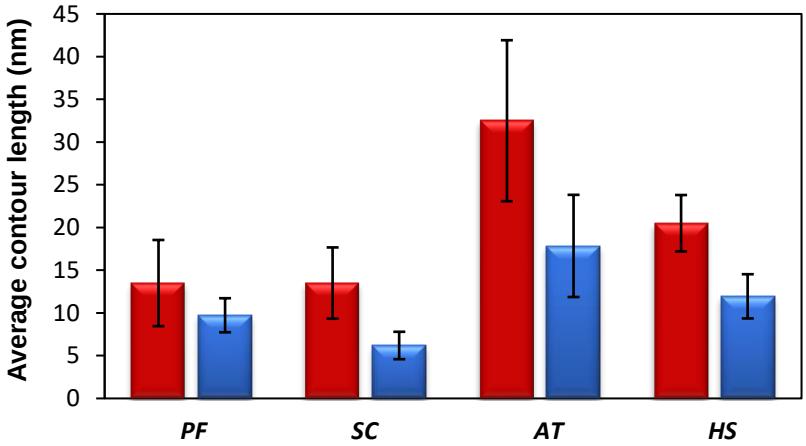

Figure 4.2.2. Contour length distributions from the predicted mRNA secondary structures. Average contour length from the predicted mRNA secondary structures obtained by Vienna RNA (red bars) and mfold (blue bars) algorithms and by simplex complexity organism order, starting with *Plasmodium falciparum* (*PF*), *Saccharomyces cerevisiae* (*SC*), *Arabidopsis thaliana* (*AT*) and *Homo sapiens* (*HS*). The bars correspond to the average value of all mRNA used for each organism. We include the standard error of the average of the mRNA molecules included per organism with a total sample size of *N* = 16.

4.3. Obtaining the mRNA molecules for *in vitro* measurements

To obtain the mRNA molecules with the specific required sequence (selected from the GenBank), we used a plasmid cloning vector to insert the gene of interest. Then, by following a plasmid amplification protocol we obtained a great amount of DNA material to perform *in vitro* transcription and finally get the mRNA molecules of interest. As we mention above, the native mRNA sequences play an important role in the variation of the C_L, for this reason is important to obtain only the specific mRNA sequence of interest to be able to continue with the smFRET experiments. To achieve this, is important to understand what is happening in the *in vitro* transcription process. In the next section we explain the principal disadvantage of this process and how we design a strategy to solve it.

4.3.1. *In vitro* transcription and its disadvantages

In vitro transcription is a simple standard procedure that allows for templated directed synthesis of short and long RNA molecules of any sequence from a linear DNA template. The basis of this process comes from the engineering of a template that includes a bacteriophage promoter sequence upstream of the sequence of interest. This sequence is recognized by the corresponding polymerase (such as T7, T3 and SP6 RNA polymerases) that catalyzes the formation of RNA from DNA. Therefore, *in vitro* transcription assays have been widely used as a molecular biology technique [83]. Despite being a widely used standard procedure, has a disadvantage, that it is not possible obtain the specific nucleotide sequence of interest. This is, the RNA polymerase starts to add nucleotides that comes from its own recognition sequence

(see Table 4.3.1) to the nascent RNA. This process added an extra nucleotide that does not correspond to the initial sequence required. The number of extra nucleotides added depends on how close the required sequence is to the bacteriophage promoter sequence. Moreover, when a plasmid cloning vectors are used, is necessary to linearize them to stop the transcription reaction. This procedure is achieved by using restriction enzymes as close as possible to the 3' end of the inserted sequence, but this inevitably adds extra nucleotides that are not a part of the required transcript and come from the chosen restriction enzymes.

Table 4.3.1. Most common bacteriophage promoter sequences in a cloning plasmid vector. T7, T3 and SP6 RNA polymerase starts transcription at the underlined G in the promoter sequence. Thus, the first base in the transcript will be a G.

PROMOTER	PROMOTER SEQUENCE
T7 Promoter	5' TAATACGACTCACTATAG 3'
T3 Promoter	5' AATTAACCCTCACTAAAG 3'
SP6 Promoter	5' ATTTAGGTGACACTATAG 3'

For example, in our case, the promoter sequence used in all our plasmids is the T7 promoter recognized by the T7 RNA polymerase. This polymerase transcribes from 5' → 3' direction. Then, the first base in the transcript will be a G, that corresponds to a base that is not part of our sequence of interest. Moreover, at the end of the transcription we are going to have extra sequences that, in the same manner, are not part of our sequence of interest. This comes from the linearization process, which needs the usage of a restriction enzyme.

In conclusion, with the usage of this standard procedure is not possible to get the required mRNA molecule with the specific native nucleotide sequence.

4.3.2. Implemented strategy to solve the disadvantages that comes from the *in vitro* transcription

As it was explained above, with this standard procedure is not possible to obtain the native specific sequence required for our experiments. Thus, we decided to implement a strategy to eliminate the extra nucleotides in order to have only the required sequence. For this purpose, we used the RNase H which is an enzyme that recognize DNA-RNA hybrids with at least 4 bp in length [84]. In this regard, we added extra nucleotides at the 5′ and 3′ ends of our 16 selected mRNA sequences. At the 5′ and 3′ position we added between 10 to 20 nt, at the 3’ position we included the nucleotides that will remain after we linearize the plasmid. The restriction enzymes used here were PmeI and XhoI (depending on the plasmid construction), which are the enzymes used to linearize all the cloning vectors. Then, each sequence was inserted in the clone vector pBluescript II SK (-) (see Fig. 4.3.2.1) opening at SmaI position. Thus, 16 cloning vectors were generated for each of our 16 sequences that were synthetized by ADN SINTÉTICO T4 Oligo. In this way, the strategy to obtain only the gene of interest is as follows: First, it is necessary to linearize the plasmid to apply the *in vitro* transcription protocol (using the T7 promoter) and get the mRNA. This generated mRNA is going to have the extra nucleotides that is possible to cut by hybridizing it with the complementary DNA sequences. To achieve this, it is required to design the complement DNA oligos with the correspondent extra nucleotides in the RNA sequence

and hybridize them. After that, we will be able to cut only the DNA-RNA hybrids by using the RNase H. Finally, the specific mRNA sequence would be obtained.

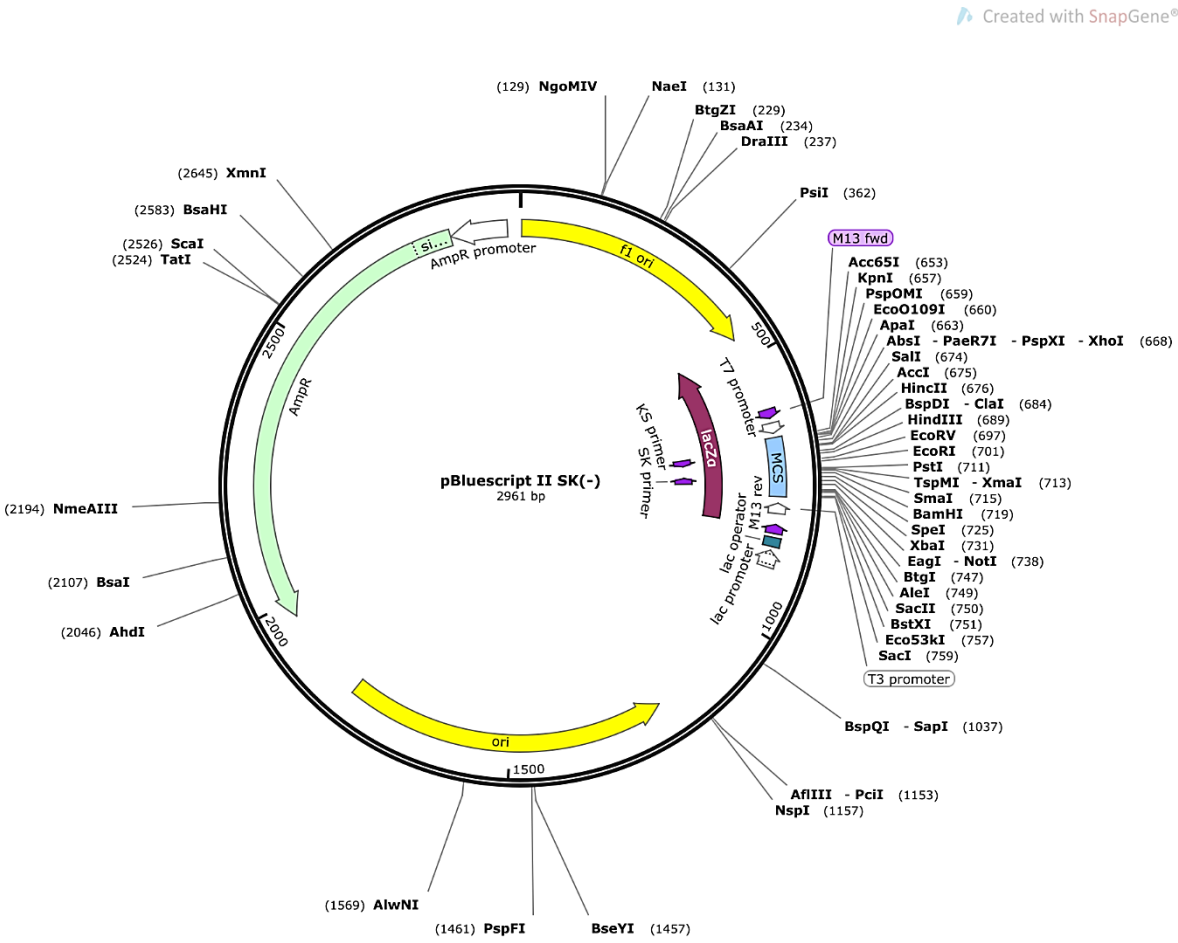

Figure 4.3.2.1. Standard cloning vector pBluescript SK (-) generated with SnapGen. Each of our 16 mRNA sequences were inserted at SmaI position.

4.3.3. DNA transformation

To obtain a great amount of genetic material for our experiments, we used the DNA transformation process. This process allows to transfer exogenous DNA into the host cell. In this thesis, the host cell used for the transformation was the 5-alpaha

Competent *E. coli* (High Efficiency) from NEB. Before we start with the transformation protocol, we first prepared LB (Luria Bertani) medium by pouring 17.5 gr Difco LB agar, Lenoxx (BD, Becton) into a 500 ml flask and filling with autoclaved ultrapure water to rise the 500 ml mark, we mixed well and autoclave for 20 min. Then, LB agar plates were prepared in a sterile environment as follows: 25 µl of ampicillin (from 100 µg/ml stocks) and 25 ml of warm LB medium (approximately 50°C) was poured in each petri dish, swirling carefully in a circular motion to mix. After pouring, the plates were let it sit to cool until the agar becomes solid. Then, the plates were placed at 37°C overnight (ON). If no contamination was present, the petri dishes can be stored at 4°C until needed.

For the transformation we followed the next protocol:

1. Placed the agar plates (stored at 4°C) at room temperature to let them warm up.
2. Thaw on ice the competent cells (50 µl stocks stored at -80°C) as well as the DNA cloning vectors (1 ng/µl).
3. Mix 5 µl of DNA into 50 µl of NEB cells. Carefully mix by flicking the bottom of the tube and place on ice for 30 min.
4. For the heat shock, place the samples at 42°C for 90 seconds in thermocycler.
5. In a sterile atmosphere, add to each sample 250 µl of LB medium and place in a shaker incubator at 37°C for 1 h.
6. In a sterile atmosphere, plate 250 µl of the sample transformation onto LB agar plate.
7. Incubate plates at 37°C ON to allow colonies to form.

8. Place the plates at room temperature to let them warm up. Finally, seal the plates with parafilm paper and place at 4°C until needed or continue with the next step.
9. In a sterile atmosphere, inoculate 7 ml of LB medium plus ampicillin (100 µg/ml) in a falcon 50 ml tube with a single picked colony. Grow in a shaker incubator at 37°C and 250 rpm for 4 to 6 h.
10. In a sterile atmosphere, add the 7 ml of the grow LB medium onto a 200 ml of LB medium in 1 Lt sterile Erlenmeyer flask containing 200 µl of ampicillin (100 µg/ml). Grow ON in a shaker incubator at 37°C and 250 rpm.
11. Isolate plasmid by following a maxi-prep protocol. We used the QUIAGEN Plasmid Maxi Kit following the manufacture protocol.
12. Dilute the obtained pellet with 50 µl of nuclease free water (NFW), quantify by UV-vis and store at -20°C until needed.

4.3.4. DNA digestion

In advance to the *in vitro* transcription, plasmid linearization was performed. The enzymes used to linearize the plasmids were PmeI and XhoI depending on the plasmid construction (see Table 4.3.4). The reaction was performed by digesting 60 µg of the plasmid, 180 U of 10 U/µl of PmeI or XhoI enzyme (from NEB) and 10 µl of 10X NEBuffer in a total reaction volume of 100 µl. The reaction was divided equally in 4 RNase free tubes, incubated at 37°C for 2 h and inactivated at 65°C for 20 min. The digested products were purified by using phenol-chloroform extraction. Finally, the obtained pellet was resuspended with NFW, quantified by using NanoDrop spectrophotometer and placed at -20°C until needed. The expected length (see Table

4.3.4) and the integrity of the digested DNAs were verified by running a TAE 8% agarose gel at 85 Volts for 90 min (see Fig. 4.3.4.1). The agarose gel in Fig.4.3.4.1 shows that except for the sample number 8 (in which it is possible to see the appearing of two closer bands), we obtain full digested products and the expected DNA lengths for all the samples.

Table 4.3.4. mRNA molecules studied from 4 model organisms. The expected digested DNA length corresponds to the length of the inserted gene onto the cloning vector.

Organism	Gene number and name	Enzyme used to linearize the plasmid	Digested DNA length (bp)	Expected length of the mRNA (nt)
A.thaliana (*AT*)	1.AT.Thi2	PmeI	3599	718
	2.AT.GTP	PmeI	3866	985
	3.AT.Glutamate	PmeI	4745	1874
	4.AT.Heat shock	PmeI	5986	3105
H.sapiens (*HS*)	5.HS.IL5.	PmeI	3697	816
	6.HS.AB.	PmeI	4691	1812
	7.HS.Ubiquitin	PmeI	5731	2850
	8.HS.Cadherin	PmeI	6307	3436
P.falciparum (*PF*)	9.PF.17kD	PmeI	3965	1094
	10.PF.cAMP	PmeI	4101	1230
	11.PF.gtubulin	PmeI	4759	1888
	12.PF.dynamin	XhoI	5867	2996
S.cerevisiae (*SC*)	13.SC.13 kD	PmeI	3315	434
	14.SC.RS21.	PmeI	3420	539
	15.SC.Sec61	PmeI	3710	829
	16.SC.Y-helicase	PmeI	5975	3104

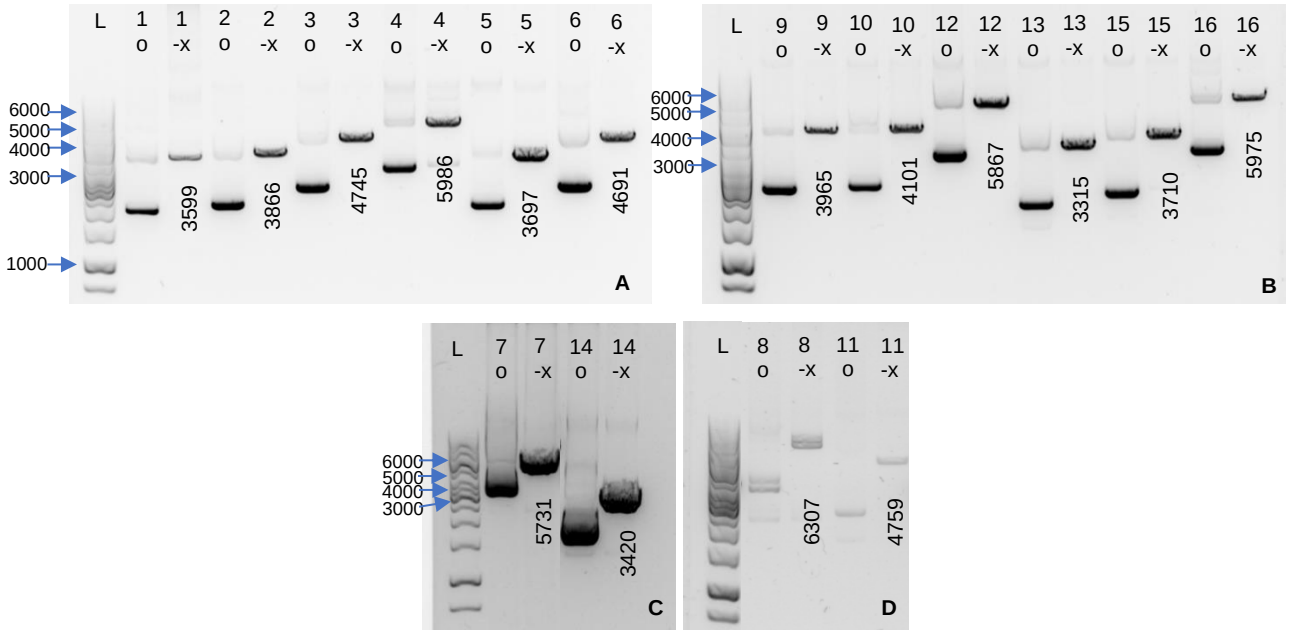

Figure 4.3.4.1. Circular DNA vs digested DNA in TAE agarose gel. The digested samples were verified by loading on a A) 0.8% agarose gel running for 75 min at 100 Volts, B and C) 0.8% agarose gel running for 90 min at 90 Volts and D) 1% agarose gel running for 80 min at 90 Volts. The sample line number corresponds to the gene number and name shown in table 4.3.3. Samples loaded in the gel correspond to circular DNA (o) and digested DNA (-x) to compare them and corroborate the complete digestion. The expected digested DNA lengths (nt) are shown below the sample bands.

4.3.5. *In vitro* transcription protocol

In vitro transcription reactions were performed according to the manufacturer instructions in the RiboMAX Large Scale RNA Production Systems- SP6 and T7 (Promega) as following:

In a 1.5 ml RNase free tube mix 4 µl T7 Transcription 5X Buffer, 1.5 µl of each rNTPs (25 mM ATP, CTP, GTP, UTP), 10 µg of linear DNA, 2 µl of enzyme mix (T7) and bring

to a final volume reaction of 20 µl with NFW. The reaction was incubated at 37°C for 4 h. After that we add 1 U/µg of RQ1 RNase-free DNase to the reaction and incubate at 37°C for 15 min. The transcript product was purified by using phenol-chloroform extraction. Finally, the obtained pellet was resuspended with 15 to 20 µl of NFW, quantified by using NanoDrop spectrophotometer and placed at -80°C until needed. The transcripts integrity and length were verified by loading samples on a MOPS agarose gels (1 % and 2 % of agarose) and running at 85 Volts for 90 min (see Fig. 4.3.5.1).

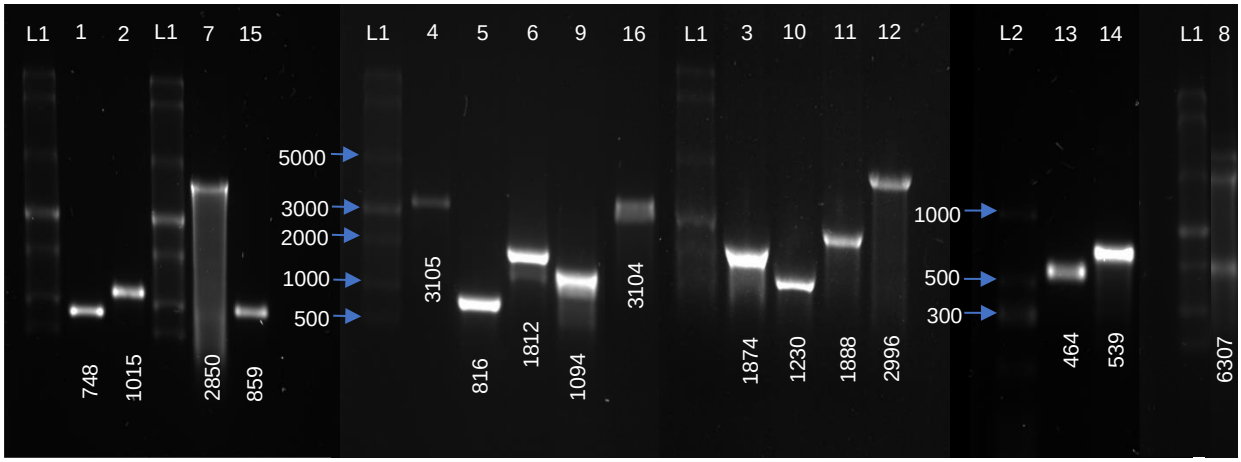

Figure 4.3.5.1. Integrity and expected length of the transcript products. The samples (1 to 16, see Table 4.3.4) were loaded on a MOPS agarose gel electrophoresis. L1= ssRNA Ladder (NEB) on a 1 % agarose gel and L2= Low Range ssRNA Ladder (NEB) on a 2 % agarose gel. The expected nucleotide lengths for each mRNA are shown below each band.

4.3.6. Cutting extra nucleotides by using RNase H

The mRNAs obtained from the *in vitro* transcription were hybridized with their complementary DNA oligos at the ends of the molecule to obtain a DNA-RNA hybrid.

RNase H (from NEB) was used to eliminate the extra nucleotides present in the sequence as it was explained in section 4.3.2. This enzyme specifically recognizes and cleaves DNA-RNA hybrids (Fig. 4.3.6.1).

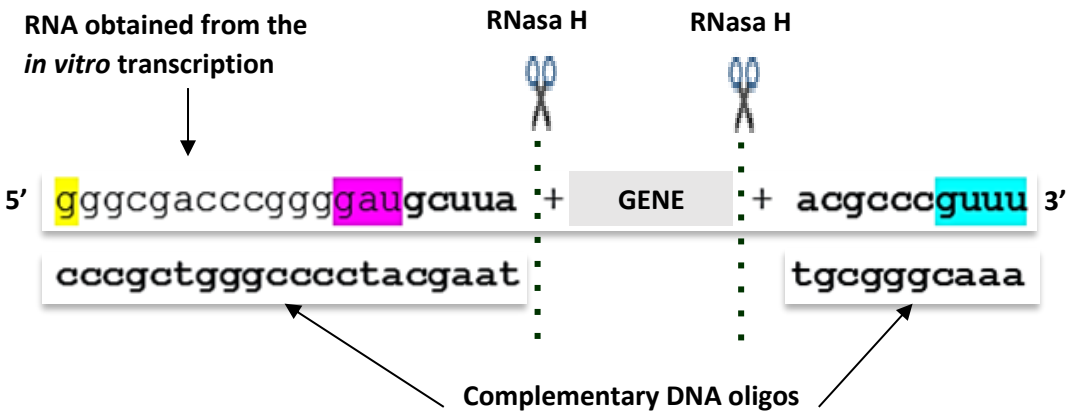

Figure 4.3.6.1. Strategy to obtain only the specific sequence required from the *in vitro* transcription products. The nucleotide highlighted in yellow corresponds to the first base incorporated by the T7 pol. The sequence highlighted in magenta comes from one part of the enzyme sequence used to open the plasmid vector. The sequence highlighted in blue comes from one part of the enzyme sequence used to linearize the plasmid vector (in this example corresponds to PmeI). After hybridizing the complementary oligos with the RNA, the RNase H was used to cut the hybrids and get only the sequence of interest (GENE).

The cutting reaction to eliminate the extra nucleotides in the transcript products was performed as follows: for a final reaction volume of 10 µl, mix in an RNase free tube 1 µg of RNA from the *in vitro* transcription, 1 µg of each of the two complementary DNA oligos, 1 µl of 10X RNase reaction buffer adjusting the reaction volume with NFW. Heat the reaction at 65°C for 5 min and immediately place it on ice for 5 min. Finally, add 5 U/µg of RNase H and heat the reaction at 32°C for 9 h. The digested mRNA products were verified by loading samples on a 15% UREA-PAGE (previously pre-running at 25

W for 25 min) and running at 25 W for 135 min (Fig. 4.3.6.2). Due to the minimal difference in length between the digested and the non-digested mRNAs, is not possible to obtain a more separated band. The gel in Fig. 4.3.6.2 shows that at 1 h of digestion with the RNase H we still had two bands, and an incomplete digestion was obtained. At 9 hours of digestion, the reaction was complete, as confirmed from the appearance of a unique band (sample 1.RNA.9h) in the gel. At this point, it seems that this protocol could serve in an efficient way to digest the other mRNA products used in this thesis, but this protocol was proved only with one sample (1.RNA, which correspond to 1.AT.Thi2 (see table 4.3.4) then, further experiments are needed to prove the functionality of this method with the others mRNA samples. After digesting the mRNAs, we have two options, one is to use the entire reaction obtained from the sample 1.mRNA-x.9h (see Fig. 4.3.6.2) and procced with the purification by using oligo clean and concentrator kit (Zymo Research) in order to start with the mRNA labelling protocol. Otherwise, if the digested reaction was not totally complete is possible to purify the mRNA by cutting from the gel only the band that contains the digested mRNA and purify it by using the Crush & soak method and placed at -80°C until needed.

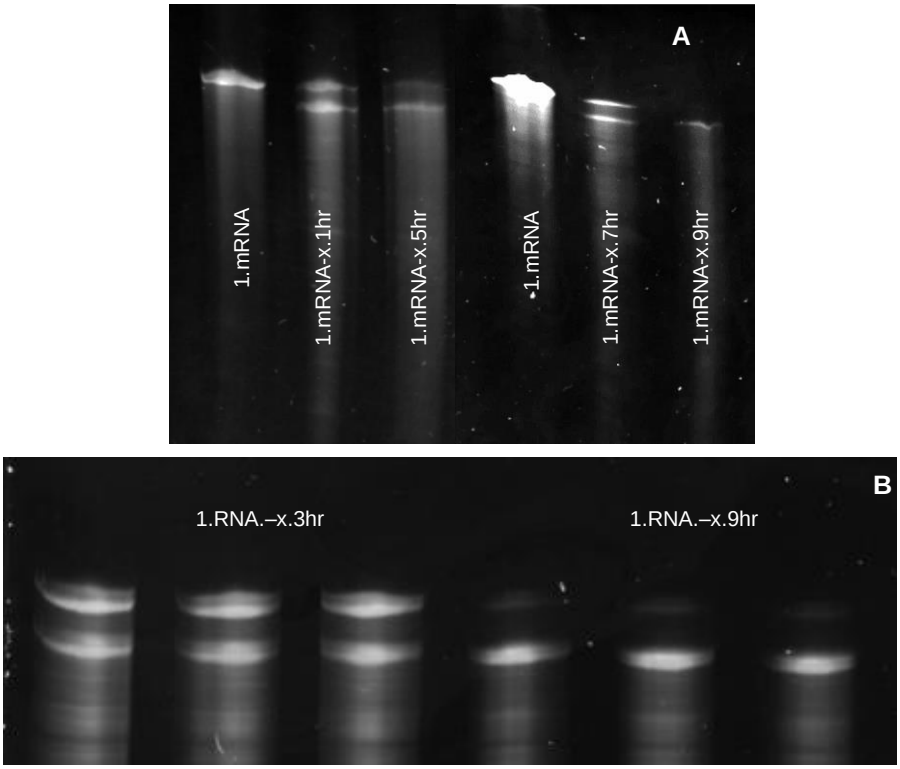

Figure 4.3.6.2. mRNA products after cutting with RNase H on a 15% UREA-PAGE. A) Using a 20 cm x 22 cm gel plates and running at 25 W for 4 hours. B) Using a Mini-PROTEAN glass plates (BIO-RAD) and running at 25 W for 135 min.

4.4. RNA labeling protocol

To measure the physical distance between mRNA ends by smFRET, the mRNAs were labeled at the 3'-end with a custom-made r-Adenosine-3',5'-bis-phosphate-8-[(6-Amino) hexyl]-amino-Alexa Fluor 546 (Jena Bioscience, Germany) and at the 5'-end with Alexa Fluor 647 C2-maleimide (Thermofisher). A schematic representation of the strategy used to label the mRNAs at the 5′ and 3′ ends of the molecules is shown in Fig. 4.4.1. The labeling reaction was performed first at the 3′ position and then at the 5′ position of the mRNA as follow:

- ***3'-end labeling***

The labeling reaction consists of attaching a single 3',5'-bis-phosphate nucleotide to the 3′ hydroxyl group of an RNA strand by using T4 RNA ligase. The protocol steps were applied avoiding RNA contamination as follow:

1. Thaw the Dimethyl sulfoxide (DMSO, Pierce™) at room temperature and warm the 50% polyethylene glycol 8000 (PEG, NEB) at 37°C for 5-10 minutes until volume is fluid.
2. Adjust a heating block to 85°C.
3. Transfer 1 µl of 1 µg of linear mRNA (diluted in NFW) in a microcentrifuge tube (RNase free). Add 25% DMSO and heat the mRNA for 3-5 minutes at 85°C. Place the mRNA immediately on ice.
4. In a microcentrifuge tube add 3 µl of 10X T4 RNA ligase Reaction Buffer (NEB), 1 µl of 40 U/µl RNase Inhibitor (Pierce™), the mRNA from step 3, 1.5 nmol of r-Adenosine-3',5'-bis-phosphate-8-[(6-Amino)hexyl]-amino-Alexa Fluor-546 (A-AF546) diluted in DMSO, 40 U of T4 RNA ligase (NEB), 4 mM ATP (NEB) and 20% PEG from step 1. Use a new pipette tip to mix ligation reaction after the PEG addition. Bring to a final reaction volume of 30 µl with NFW.
5. Incubate the reaction at 37°C ON (overnight).
6. Add 70 µl of nuclease-free water to the ligation reaction.
7. Add 100 µl of phenol-chloroform to extract the RNA ligase. Vortex the mixture briefly, then centrifuge 2 to 3 minutes at high speed in a microcentrifuge to separate the phases. Carefully remove the top aqueous phase and transfer to a nuclease-free tube.

8. Tagged 3′-mRNAs were purified two times using Zymo columns following manufacturer protocol and stored at -20°C until needed.

- ***5'-end labeling***

Purified 3′-end labeling mRNAs were labeled by using the 5′ EndTag Nucleic Acid Labeling System from Vector Labs and the Alexa Fluor 647 C2-maleimide (AF647-M). The labeling reaction consists in transferring a thiophosphate from ATPγS to the 5′ hydroxyl group of the nucleic acid by T4 PNK (polynucleotide kinase), then the maleimide label fluorophore is chemically coupled to the thiol group attaching covalently. Avoiding RNA contamination, the protocol steps were applied as follow:

1. In a microcentrifuge tube mix 1 µl of universal reaction buffer, mRNA (up to 0.6 nmols of 5′ ends in ≤ 8 µl) and 1 µl of CIAP (calf intestinal alkaline phosphatase). Bring the total reaction volume to 10 µl with NFW. Incubate for 30 minutes at 37 °C.
2. Combine the entire reaction mixture from Step 1 with 2 µl of universal reaction buffer, 1 µl of ATPγS and 2 µl of T4 PNK. Bring the total reaction volume to 20 µl with NFW and mix. Incubate for 30 minutes at 37 °C.
3. Add 10 µl of 0.005 mg/µl of AF647-M (dissolved in DMSO). Mix and incubate for 30 minutes at 65 °C.
4. Add 70 µl of NFW and 100 µl of phenol and vortex briefly. Remove upper aqueous layer to a clean microcentrifuge tube.
5. To this aqueous fraction add 5 µl of precipitant and 270 µl of 95% ethanol. Mix. Pellet the precipitated nucleic acid by centrifugation at 13,000 x g in a microcentrifuge for 30

minutes. Wash the pellet briefly with 70% ethanol and centrifuge at 13,000 x g for 3 minutes. Dry the pellet and resuspend in TE buffer (10 mM Tris, 1 mM EDTA, pH 8).

6. To remove the trace amounts of unincorporated fluorophore, use size exclusion chromatography. For this purpose, we used the oligo clean and concentrator kit (Zymo Research) following the fabricant protocol.

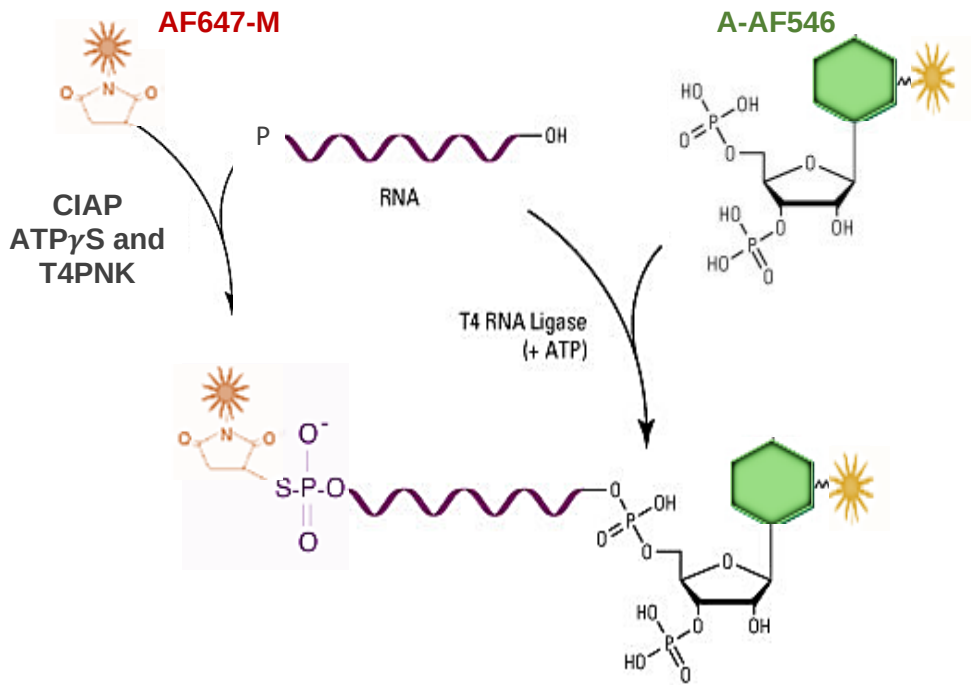

Figure 4.4.1. Schematic representation of the strategy used to label the mRNA molecules. The molecules were labeled at the 3′ and 5′ ends with A-AF546 and AF647-M respectively.

Because of the time to conclude with the doctoral project, only one attempt was performed using the 13.mRNA sample (which correspond to 13.SC.13kD without cutting the extra nucleotides, see Table 4.3.4). The labeling efficiencies obtained for the donor and acceptor molecules for this sample were about 59 ± 2 % and 76.8 ± 2 %

respectively (± correspond to the detection limit for pre-defined dyes reported in http://tools.thermofisher.com/content/sfs/manuals/3091-NanoDrop-One-Help-UG-en.pdf, page 11). The labeling efficiencies obtained here were after one column purification. Therefore, it is important to keep in mind that these values could be reduced after using more column purifications (for more information see section 3.3.2).

4.5. Physical distance between mRNAs ends

To measure the physical distance between the mRNA ends, approximately 0.5 ml of the labeled mRNA diluted at final concentration of 90 pM with TE buffer was flowed into a clean sample chamber. Then, FRET measurements were carried out by using the 13.mRNA labeling sample. Unfortunately, measurements of the distance between the mRNA ends were not achieved. Then, because the uncertainties and unreproducible results obtained from the smFRET calibration, we decide to leave pending the measurements of the distance between the mRNAs until we corroborate if the smFRET equipment is working properly (as was previously mentioned in section 3.3.3) to be able to continue with these experiments.

CHAPTER 5. Final remarks

5.1. Conclusions

In the first project of this thesis, we determined the exterior loop contour length of full native mRNA molecules from different species of different clades.

A fundamental question in biology is to predict how changes in genotypes could result in changes in phenotypes. In the first project, we show that the variations in the distance between ends of native mRNA molecules, represented here by the C_L of the exterior loop, are somewhat larger than previously reported using housekeeping and highly expressed genes. The variations observed are bigger than those obtained for random sequences. In other words, statistical or thermodynamic variations are not big enough to explain the variations observed in native sequences, with very high confidence. This result indicates that there is a biological mechanism responsible for the observed variations. Considering that end-to-end separation of mRNA molecules could impact the initiation of transcription, our results suggest that the variability in C_L could be related to phenotypical stability.

In this regard, it is important to note that phenotypical stability, which could modulate the ability of a species to diverge or survive, does not depend on just one characteristic. Instead, it depends on a combination of several characteristics under the appropriate

environmental and intrinsic conditions. Here we propose that the length of the exterior loop is one of the intrinsic conditions to be considered.

In the second project, we implemented the strategies and carried out the molecular biology experiments that are going to give us at the end the 16 different mRNA sequences of interest from organisms of the Eukarya domain. This was of utmost importance before we were ready to continue with the fluorescence mRNA labeling and the posterior analysis of the distance between their ends by smFRET. Along with the development of the DNA and mRNA labeling experiments we had some troubles that resulted in a small labeling efficiency, even with this problem, the efficiencies obtained here should be enough to be detected by the smFRET equipment. We found difficulties obtaining the correct smFRET signals by using DNA or RNA labeling samples. This led us to think that perhaps we need to change the labeling protocols, or we are missing something in the smFRET optical set up. To check this, we thought that the best way is corroborate first if the smFRET equipment works properly. To this end we purchased DNA labeled probes with a desired known length to measure the distance between their ends with the smFRET equipment. We believe that the result of this experiment will define the course of the next pending experiments, because now we should be able to know for sure if is needed to change the protocols in the nucleic acid labeling or if we can simply start with a new configuration or improvements on the smFRET equipment.

5.2. Future work

The studies presented here to analyze the distance between ends of full native mRNA molecules comprise both, theoretical and experimental work. From the theoretical work we show that the separation between mRNA ends is wider than previously reported in the literature either by theory, computer simulations and even experiments. Also, we propose that the effective circularization given by the distance between mRNA ends should play a role in the initial recognition by different translation initiation proteins by improving their interaction. Our results show that this distance differs from random sequences and varies depending on the organism, something that could be linked to the divergence and phenotypical stability of species. As a future work, this distance between ends could be corroborated by means of smFRET to compare experimental versus theoretical data. Also, it could be interesting to apply the same methodology used here to predict the distance between ends for long non-coding RNAs (lncRNAs), which are a type of RNA molecules that are not translated into proteins. The lncRNAs are involved in gene expression regulation at transcriptional and post-transcriptional level. Although there are a few well characterized lncRNAs that have an important role in the progression of diseases, the understanding of their function, expression regulation and secondary and tertiary structures is very limited [85]. Therefore, a comparison of the structure and distances between the ends in those molecules (mRNAs and lncRNAs) could provide insights about how the lncRNAs perform their functions in the cell. In this regard, we are currently working on developing this analysis by means of bioinformatics and computational algorithms for structure prediction.

From the experimental work, further experiments should be carried out in the future to conclude with the smFRET calibration and to perform the measurements of the end-to-end distance between the mRNAs from the Eukarya domain.

Finally, altogether these results could contribute to a better understanding of the role that plays the RNA molecules in the evolution of the species, in which the length of the exterior loop could be one important characteristic to be consider.

Bibliography

1. Lodish H BA, Z.S., et al., *Molecular Cell Biology (W. H. Freeman, New York) 4th edition Ed.* 2000.
2. http://www.phschool.com/science/biology_place/biocoach/bioprop/chemrna.html.
3. Watson JD, B.T.A., Bell S.P, Gann A, Levine M, Losick R (2014) Molecular Biology of the Gene (Pearson) 7th Ed.
4. Geisler, S. and J. Coller, *RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts.* Nat Rev Mol Cell Biol, 2013. **14**(11): p. 699-712.
5. Voinnet, O., *Induction and suppression of RNA silencing: insights from viral infections.* Nat Rev Genet, 2005. **6**(3): p. 206-220.
6. Montero, J.J., et al., *Telomeric RNAs are essential to maintain telomeres.* 2016. **7**: p. 12534.
7. Yoffe, A.M., et al., *The ends of a large RNA molecule are necessarily close.* Nucleic Acids Research, 2011. **39**(1): p. 292-299.
8. Cox MM (2012) Molecular Biology. Molecular Biology, P.a.P., ed Company WHFa (Kate Ahr Parker).
9. Hermann, T. and D.J. Patel, *Stitching together RNA tertiary architectures1.* Journal of Molecular Biology, 1999. **294**(4): p. 829-849.
10. Ed., L.H.M.C.B.N.Y.W.H.F.a.C.t.
11. Reuter, J.S. and D.H. Mathews, *RNAstructure: software for RNA secondary structure prediction and analysis.* BMC Bioinformatics, 2010. **11**: p. 129-129.
12. Dowell, R.D. and S.R. Eddy, *Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction.* BMC Bioinformatics, 2004. **5**(1): p. 71.
13. Doshi, K.J., et al., *Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction.* BMC Bioinformatics, 2004. **5**(1): p. 105.
14. Mathews, D.H., et al., *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.* Proc Natl Acad Sci U S A, 2004. **101**(19): p. 7287-92.
15. Clote, P., Y. Ponty, and J.M. Steyaert, *Expected distance between terminal nucleotides of RNA secondary structures.* J Math Biol, 2012. **65**(3): p. 581-99.
16. Fang, L.T., *The end-to-end distance of RNA as a randomly self-paired polymer.* J Theor Biol, 2011. **280**(1): p. 101-7.
17. Han, H.S. and C.M. Reidys, *The 5'-3' distance of RNA secondary structures.* J Comput Biol, 2012. **19**(7): p. 867-78.
18. Leija-Martínez, N., et al., *The separation between the 5'-3' ends in long RNA molecules is short and nearly constant.* Nucleic Acids Research, 2014. **42**(22): p. 13963-13968.
19. Pesole, G., et al., *Structural and functional features of eukaryotic mRNA untranslated regions.* Gene, 2001. **276**(1–2): p. 73-81.
20. Mazumder, B., V. Seshadri, and P.L. Fox, *Translational control by the 3'-UTR: the ends specify the means.* Trends in Biochemical Sciences, 2003. **28**(2): p. 91-98.
21. van der Velden, A.W. and A.A.M. Thomas, *The role of the 5' untranslated region of an mRNA in translation regulation during development.* The International Journal of Biochemistry & Cell Biology, 1999. **31**(1): p. 87-106.
22. Hughes, T.A., *Regulation of gene expression by alternative untranslated regions.* Trends in Genetics, 2006. **22**(3): p. 119-122.

23. Pichon, X., et al., *RNA Binding Protein/RNA Element Interactions and the Control of Translation.* Current Protein and Peptide Science, 2012. **13**(4): p. 294-304.
24. Wells, D.G., *RNA-binding proteins: a lesson in repression.* J Neurosci, 2006. **26**(27): p. 7135-8.
25. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Research, 2003. **31**(13): p. 3406-3415.
26. Gruber, A.R., et al., *The Vienna RNA Websuite.* Nucleic Acids Research, 2008. **36**(suppl_2): p. W70-W74.
27. Wu, X., G. Ji, and Y. Zeng, *In silico prediction of mRNA poly(A) sites in Chlamydomonas reinhardtii.* Molecular Genetics and Genomics, 2012. **287**(11): p. 895-907.
28. Yamamoto, Y.Y., et al., *Identification of plant promoter constituents by analysis of local distribution of short sequences.* BMC Genomics, 2007. **8**(1): p. 67.
29. Zhao, Z., et al., *Bioinformatics analysis of alternative polyadenylation in green alga Chlamydomonas reinhardtii using transcriptome sequences from three different sequencing platforms.* G3 (Bethesda, Md.), 2014. **4**(5): p. 871-883.
30. Tian, B. and J.H. Graber, *Signals for pre-mRNA cleavage and polyadenylation.* Wiley Interdisciplinary Reviews: RNA, 2012. **3**(3): p. 385-396.
31. Hyeon, C. and D. Thirumalai, *Mechanical unfolding of RNA: from hairpins to structures with internal multiloops.* Biophys J, 2007. **92**(3): p. 731-43.
32. Lai, W.-J.C., et al., *mRNAs and lncRNAs intrinsically form secondary structures with short end-to-end distances.* Nature Communications, 2018. **9**(1): p. 4328.
33. Noffke, N., et al., *Microbially Induced Sedimentary Structures Recording an Ancient Ecosystem in the ca. 3.48 Billion-Year-Old Dresser Formation, Pilbara, Western Australia.* Astrobiology, 2013. **13**(12): p. 1103-1124.
34. Ruhfel, B.R., et al., *From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes.* BMC Evolutionary Biology, 2014. **14**(1): p. 1-27.
35. Stanley, S.M., *An Ecological Theory for the Sudden Origin of Multicellular Life in the Late Precambrian.* Proceedings of the National Academy of Sciences, 1973. **70**(5): p. 1486-1489.
36. Renz, A.J., A. Meyer, and S. Kuraku, *Revealing Less Derived Nature of Cartilaginous Fish Genomes with Their Evolutionary Time Scale Inferred with Nuclear Genes.* PLOS ONE, 2013. **8**(6): p. e66400.
37. Mallatt, J. and J. Sullivan, *28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes.* Molecular Biology and Evolution, 1998. **15**(12): p. 1706-1718.
38. Shu, D.G., et al., *Lower Cambrian vertebrates from south China.* Nature, 1999. **402**(6757): p. 42-46.
39. Roeding, F., et al., *A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (Pandinus imperator).* Molecular Phylogenetics and Evolution, 2009. **53**(3): p. 826-834.
40. Zhou, X., et al., *De Novo Transcriptome of the Hemimetabolous German Cockroach (Blattella germanica).* PLoS ONE, 2014. **9**(9): p. e106932.
41. Rothwell, G.W., et al., *The seed cone Eathiestrobus gen. nov.: Fossil evidence for a Jurassic origin of Pinaceae.* American Journal of Botany, 2012. **99**(4): p. 708-720.
42. Looy, C.V., et al., *The delayed resurgence of equatorial forests after the Permian–Triassic ecologic crisis.* Proceedings of the National Academy of Sciences, 1999. **96**(24): p. 13857-13862.
43. Poort, R.J., H. Visscher, and D.L. Dilcher, *Zoidogamy in fossil gymnosperms: The centenary of a concept, with special reference to prepollen of late Paleozoic conifers.* Proceedings of the National Academy of Sciences, 1996. **93**(21): p. 11713-11717.

44. Suh, A., et al., *Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds.* Nature Communications, 2011. **2**(1): p. 443.
45. Luo, Z.-X., et al., *A Jurassic eutherian mammal and divergence of marsupials and placentals.* Nature, 2011. **476**: p. 442.
46. Chanderbali, A.S., et al., *Evolving Ideas on the Origin and Evolution of Flowers: New Perspectives in the Genomic Era.* Genetics, 2016. **202**(4): p. 1255-1265.
47. Magallón, S., et al., *A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity.* New Phytologist, 2015. **207**(2): p. 437-453.
48. Michener, C.D. and D.A. Grimaldi, *The oldest fossil bee: Apoid history, evolutionary stasis, and antiquity of social behavior.* Proceedings of the National Academy of Sciences, 1988. **85**(17): p. 6424-6426.
49. Smith, G.R., *Late Cenozoic Freshwater Fishes of North America.* Annual Review of Ecology, Evolution, and Systematics, 1981. **12**(Volume 12): p. 163-193.
50. Stauffer, R.L., et al., *Human and Ape Molecular Clocks and Constraints on Paleontological Hypotheses.* Journal of Heredity, 2001. **92**(6): p. 469-474.
51. White, T.D., et al., *Pleistocene Homo sapiens from Middle Awash, Ethiopia.* Nature, 2003. **423**(6941): p. 742-747.
52. Kim, B., et al., *Single-molecule visualization of mRNA circularization during translation.* Experimental & Molecular Medicine, 2023. **55**(2): p. 283-289.
53. Alekhina, O.M., et al., *Functional Cyclization of Eukaryotic mRNAs.* Int J Mol Sci, 2020. **21**(5).
54. Jackson, R.J., C.U. Hellen, and T.V. Pestova, *The mechanism of eukaryotic translation initiation and principles of its regulation.* Nat Rev Mol Cell Biol, 2010. **11**(2): p. 113-27.
55. Richter, J.D. and N. Sonenberg, *Regulation of cap-dependent translation by eIF4E inhibitory proteins.* Nature, 2005. **433**(7025): p. 477-480.
56. Nicholson, B.L. and K.A. White, *3' Cap-independent translation enhancers of positive-strand RNA plant viruses.* Current Opinion in Virology, 2011. **1**(5): p. 373-380.
57. De Falco, L., et al., *The Pseudo-Circular Genomes of Flaviviruses: Structures, Mechanisms, and Functions of Circularization.* Cells, 2021. **10**(3).
58. Bonnet, E., et al., *Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences.* Bioinformatics, 2004. **20**(17): p. 2911-7.
59. Seffens, W. and D. Digby, *mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences.* Nucleic Acids Research, 1999. **27**(7): p. 1578-1584.
60. Major, R.T., *The Ginkgo, the Most Ancient Living Tree.* Science, 1967. **157**(3794): p. 1270.
61. Gertz, H.J. and M. Kiefer, *Review About Ginkgo Biloba Special Extract EGb 761 (Ginkgo).* Current Pharmaceutical Design, 2004. **10**(3): p. 261-264.
62. Collins, S. and G. Bell, *Phenotypic consequences of 1,000 generations of selection at elevated CO2 in a green alga.* Nature, 2004. **431**(7008): p. 566-569.
63. Brueggeman, A.J., et al., *Activation of the carbon concentrating mechanism by CO2 deprivation coincides with massive transcriptional restructuring in Chlamydomonas reinhardtii.* Plant Cell, 2012. **24**(5): p. 1860-75.
64. Sanger, H.L., et al., *Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures.* Proceedings of the National Academy of Sciences of the United States of America, 1976. **73**(11): p. 3852-3856.
65. Di Serio, F., et al., *Current status of viroid taxonomy.* Archives of Virology, 2014. **159**(12): p. 3467-3478.
66. Kühn, U. and T. Pieler, *Xenopus poly(A) binding protein: functional domains in RNA binding and protein-protein interaction.* J Mol Biol, 1996. **256**(1): p. 20-30.

67. Sachs, A.B., R.W. Davis, and R.D. Kornberg, *A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability.* Mol Cell Biol, 1987. **7**(9): p. 3268-76.
68. Baer, B.W. and R.D. Kornberg, *The protein responsible for the repeating structure of cytoplasmic poly(A)-ribonucleoprotein.* J Cell Biol, 1983. **96**(3): p. 717-21.
69. NL, M., *Single Molecule FRET Studies of Short and Long RNA Samples.* , in *(Universidad Autónoma de San Luis Potosí).* (2014).
70. Ha, T., *Single-Molecule Fluorescence Resonance Energy Transfer.* Methods, 2001. **25**(1): p. 78-86.
71. Sasmal, D.K., et al., *Single-molecule fluorescence resonance energy transfer in molecular biology.* Nanoscale, 2016. **8**(48): p. 19928-19944.
72. https://www.olympus-lifescience.com/en/microscope-resource/primer/techniques/fluorescence/fret/fretintro/.
73. Hochreiter, B., et al., *Advanced FRET normalization allows quantitative analysis of protein interactions including stoichiometries and relative affinities in living cells.* Scientific Reports, 2019. **9**(1): p. 8233.
74. https://www.thermofisher.com/mx/es/home/references/molecular-probes-the-handbook/tables/r0-values-for-some-alexa-fluor-dyes.html.
75. https://www.thermofisher.com/order/fluorescence-spectraviewer/#!/.
76. McCann, J.J., et al., *Optimizing methods to recover absolute FRET efficiency from immobilized single molecules.* Biophysical journal, 2010. **99**(3): p. 961-970.
77. https://www.vitrocom.com/products/view/5005.
78. Yan, J. and J.F. Marko, *Localized Single-Stranded Bubble Mechanism for Cyclization of Short Double Helix DNA.* Physical Review Letters, 2004. **93**(10): p. 108108.
79. Sindbert, S., et al., *Accurate distance determination of nucleic acids via Förster resonance energy transfer: implications of dye linker length and rigidity.* J Am Chem Soc, 2011. **133**(8): p. 2463-80.
80. Ed., R.S.M.C.A.I.M.T.E.
81. https://assets.thermofisher.com/TFS-Assets/LSG/manuals/td07604.pdf.
82. Hyeon, C., R.I. Dima, and D. Thirumalai, *Size, shape, and flexibility of RNA structures.* J Chem Phys, 2006. **125**(19): p. 194905.
83. Beckert, B. and B. Masquida, *Synthesis of RNA by in vitro transcription.* Methods Mol Biol, 2011. **703**: p. 29-41.
84. Donis-Keller, H., *Site specific enzymatic cleavage of RNA.* Nucleic Acids Res, 1979. **7**(1): p. 179-92.
85. Losko, M., J. Kotlinowski, and J. Jura, *Long Noncoding RNAs in Metabolic Syndrome Related Disorders.* Mediators Inflamm, 2016. **2016**: p. 5365209.

Appendix

We provide 10 of the 50 random-generated sequences as example (all the sequences are saved in a computer at the Biological Physics Laboratory from UASLP with a name file of SECUENCIAS ALEATORIAS).

AL1.

CTTACGGAATCTGACGCCTAGCCGAGTCTGCCTAAGATCGACGGGGTTCTGTGTAGGTGCAGGA
GGGCACGGAAAGACACGACACCAGGTCGTATCAGATGGCAGTGAGAAGAATGAGTATGTCTGAA
ACGAAGACATTGCATAACTCATGTTATTGGGAATATATTGGTGTTCTAGCGCTCTAATTAGAACTGT
AGTTGGGCGGATGACGGATGAATGCTCCTATCGAAAAGTACCGCATTACGAGACATGCTAAACAT
AATGAGTCCCGCTACTAGTTCTTCTATAGCCCATGTCCAAGTAATCCTTTTTAAACTGGAGCTAATA
GATAACCCCCCACACCAATTAAACCTCTTTACTAAGCCTTAGTCAAAGTAATGTGGGTCTTCCCGT
ATGTCCTTACTCCAACTGGGTATTTATGAGACCTCGGCAGCACGAGCTTTCGCGCTCCACACCTG
ACACTTTCCCACTTAGGTTGGGTAAGTATCCACGAATTCGAAAGCCGCGCAGATATTATAACAATC
GTTCAGCCTTAGGATTACACGAGTCTCGGGAAGAGCAATTATAAGGTCGAACCGAGATCCAGATG
GTGCTTCTAGAAAGCAGCATCCGTCTAGAAGGGACGGCGAATCGTAGAATTATAAAGGATCCTCT
GTGATGGAAAGAACTAACAGATCGATTTGAGGCCAACTTTGAAGTAAGGAACAGTGTAGTGAGGG
GTTGCTGGAATGAATGGGGTCCCCGCACTATCTCCCGGCTCTTTTACTTAAAGTGTTTGAGATTCC
CTATGGAACTAGCAAATTGACCATAGTTAATGAGACCGCCACATAACAGTAATTATTGCTCGAATG
GCGAGTGTGCGAGACGTCTTCAGGTGAGGCTCAGAACAGCTCATAAGTCTTCAGTTAGAATATGG
TCCCCAGCACACACGAAGGCAATGGAATGGTAAGACCGTCTCGAACAACCATATGCAAGCCTAGC
GATTGCCTGCTTAAGTGTGATTCGCGGCTAGTGGCTCAACTTATTTCTCCCCTCCCAAACAGTTCA
TGCTACAAGTAGACTTGAGTTTTGCCTCCCAGTCTCATTTCATGAAATTGTGGTAATGGCCAATTG
ACCCTTTGCAAGTGTTGATCACATTAGACAATCTTCAAAAGCCAACTAATACAAAAGAATCGGGGC
AGCTTCGTGCGATTTTACTCGGCCTGGAGATATCCGCTACCGGACTGCTATCCAAACTCGGCAGG
CGGACATGCTAAAATCATTACACGAGGACGTCGCCTACTGGCGAATTACAGCCATCGTGCCCTAC
TCTGTTTAAATTACGATTGCTGACTGGTTATGACGTGTTTCCAAACGTCGAGTCTGTGTGCCTCTTT
TAGGTCGCTCAAGCATGAGCGTAACCTCCCACTCCATATTTACCCCTACCGGCCGATCCAGAATA
CACCGTGCTGCTGTTTTTTCCTCCGTAGCAGCGGTATGTTCTACATAGGAACGTAATCTATTCTCG
GGCTGGGATCTGACCTAACGCGGTTCTAAGGCGCCATTGAGAGAAAGCGGTCCACCCATACAGC
GCAGAGATCAGCCGGAGAGCGTTCGAGAA

AL2.

AATTCTGCCCCTGTTAGGCACTCCCTTGTTCTTTGTAATCAGGCAGATGTTTGCGACGACTGTGGT
TTGGCCTATGTTAACGGGCTTTGGAGGGACTGCGCCCCCGGCCGTGGAGTTCGTTTGTCTGCAA
GAAGCGATAATTTAGGCCTGTCGGTCGTACCTGCAGATGCTCGTTATGTATTCCGTCGGATTCACT
AGAACGAGGCCATCTCGTGATCTAAAGGAATAAGGTGCGATATCCTGAAGGCTGACTTTTGGTAA
GATGTCACCCCCACAGAATTTACGGTTCGTTCAACGACACGTAATTAGCATAATGGCGATTAGTGC
CCTTTAACCAGCTCGCTCTTGTGAGCAGAAGTAAGGGCTTGGGCGTGAAATCGTAGGGTCAGCG
GACCCTGACCTTAGGTGGCGATAGACTCCCAATCACAAGAAGTTTAAAACTATGCCCGTGCCTGA
TAGCACGTGCGAATCTGATATCATAACATTATCTTAACATCATTCCCTCAGTGGCGACGTAATCATT

GAGTGGGTAAGTCTGACGAACTTTATCGCCCATTCTTAAATACCGAAATGACAGCATCGGCCTCA
CTGAAAGGTCTATGGTTATAGGTCCGCCGTTTACCCCAGTAGGCGTGGCACTACCGGATCAAGTA
TATGAATAGGAACCTGGATAATTAGACTATAAATATTCACGGTAGAAGACAGCGCACACAAAGCCG
TCGGCTTTTAGAGGAAACCGCTCAATTAGAGGTAACTACGACACACGAAAAGCCTGTCCGTAGGA
TGTCCGTGGCAATAGCCTAGTTGCAGGGGCTCCCCGGGGCAATGAGATGAATTCCAAGGTAATC
AGGGGGGCAGAGCCGTTCATTGGATGATACCGGGCCTTTGACGACGATAATACAATGCGTGGTA
GCCGCCATGAACTATACTACAATTCGCACGAGTACAACGGTAGAAGAAATCCGCCTCGGTAGCAA
CTTAACAAATAAGCCCAGTCACGTCCTCAGATTATAGGTAAAAAGCTTAATGCCCCTCATGGCTGA
TGGAACTTCAGAAGCCATTCGGTGCACGCGTCTGGAGCCACAACCCATGAGTCTCTTTTGAAATT
AATCTAAAAAGCTGTCGAGCCCGGACCTATTGCTTGATGTCTAATCTTCCTAGCATTTGCCCAGAT
ATCGAGGGCATCTGATGGTCTCGTTAGTCTTTCCGGTGCGCTTGAGACCACTTGGATTTCAACTTA
TTTCGCCGATCCATCTTGTACTTCATGCATCAAGAGTGCCTTGTATCCACAGCGCCATCAGCAGG
GTTGCTGTTCTGGAAGAATACTAGGGGAACTCGGTATTCGATGACCGGGGTGTATAGATCCCGGC
AGTGCGCCGCCCGTCCAGACGGGTAGCTCACAGGTTGTGCACACAAACGTAACTATTCGAGAAT
GCCGGAGCGCCCGTCATGGTGCGCGCTTGTCACTCACGACCCATTAAAGTGGGGCGGATGTCTT
GTTATACTTGGGCCGCACGTAATACCTCCCATAAGGAGAATGGGCACCGTATTGGTGCCCATGCA
CCAAATCTGGATCGCGGCGGCGATTTGTTGCGTTGCG

AL3.

CTAGGTCATCTCACCGATCACAAGAACCAGCAAGGAGAGCATATTAACCGAGCTGTAACGCCTCG
ATGTATGGCCCCGTTTGGGGCCTGACCTTGTTGCGTGGCTCCTATATCCGGATCACTATAGGGCC
CCACCCCCGGTCCTCTTGAATAGCGAGCAGCTCTTCGAACTTCGGTCATACTCATCAAAATTTTAC
AGGTTATGACTGACCGTCACTAGATTAGCGTTTATAGGCGCCCTCCCTGTAGCACGCTCATGGTA
ACTGGAACCCCCCGGAAGCCGTGCCGGCAAGCAATCTAGACTGAACAAGCCCCCTTACTCCACA
AGAGGACATCGTTAGACTACCAAACAGGGCCGACCTCACTAATCGAATAAAACCCGCCATTGCAG
AGAACTACTGTAGCAGCTACTCTGGGGCGATCCAGAAAGGCTCGAGCGTGTACAGCAGTCTTAG
GGCGCCGACCCTTGCTTGAAAGGATTTCCACGTCTTCCGCGTACTCCAGATTTTTTTAACCCGGT
CCTGCAATATCGAATATCTTGGAGCAGCCTTGGGACCCACATACTTGGTTCTTGATAAAGCGGGT
ACTTGAACAAATGACATGCCCCCCTTCGAACGGTGAAGAAAATAATTGGTAAAACTGGGTGCCGA
CTGTTCAGAATGTGGGAACAAACCGTCTCCAGATCGGTTTAGAGGTGCTAAATAAAGCTATCACG
GCAATGGACTGCCGGAGGCTTCACATCTGTTCTACGGTAGATTCAGCCTAAGATTCCAACTCTCAT
CTCATCGTTCGACGGACACCTGTGTTGTAATCCGCATCTAACAGACGAGCTCATAAACTCGAAGA
GCCATTTACTGGACTACACTCGATTGCCACGCAGTGGGATCGCTGGTTGCCCGGTCATCTCTAAG
TCGGTACGTCTCCGAATCCGGTCGCTGCCTGGATATCGCGAAATGTTCAATATCATGGTACTCGG
GAGTAAGCTTGATCCCGAGGGTCGGTCGACCTACGGCTTACATGCTGTAATACGCTGAAAAAATC
TGGTTGTGCTGAGACCTTCGTATTGATGACTGAAACGTTACAGGGAAATTGGATTCCTTGCCAGG
GTGACAGATCTATAGCAATTTAGGGTACATCATCGTCGTATAGTCATGTCTCCCCATACCTGACTT
CTTTCGGACTGTATGGCGGATGGCGGCTATGAAGTGGGATGCTTGGGTGGATCAATAATTCTCGG
CTTAGCTATTATCTGGGCTGTCCGGGGGAGGGTTTTCATAAAGCTTTAGCTCAGTTTTGGTGACTC
GCCTATGATCTATAAAGCTCCCAGCTCCGATGCCCAGTGGAATATGCTCGGAGAGATGTTAAAGT
GTTTCTGTTATCCGTAATGGAAAATCAGCGATAATTCTGTGACCTTCCTGCAAACGATAACTTTAG
GAGCCCACCGGCCTACCATCTGAGCGCGGGAAGCGGTTTTTCTCCCTGTTATTCGCCAGCAGCA
GCGCCCGAGTCCCACTTACCTAACTCACTACCCTGCTCGGGCCTTTCGACTGCCGATGATCGGCA
CCTCTAACGTTCCGATGGACGCTTGGTTCGGAATATAC

AL4.

ATTTTGTATGATTTTAGCGTGTCACGGCCGCATATAAAGCAGGAAGGTTTTGTGCCTCAACTCGAC
ATCTCCTGTGATGCTAGCAAGCTTAAAGACCGCCAAGACCCCTGAGTCTCCTTTGTCTGAGACCG

CCTCTGTTGCGAGTAAAATATGTGTCATCCGAACATATATCGTGCAACCCAACTTCGGTTATGTCT
CCGGAATAGATTCATTCCGGCTGAGGATAGGCACTTGCAAGGACAAACGTGTGTAGTTACTCATG
GAACCTCGCGGCGGGGGTAGTCGTGTAATACAAGAGAACGCTTTAGGAATTGGTGGACTGGGCA
ATCGGGATTAGGTACGTGCGCATCACACGGGCGGTAGGATAACTGCTAACGATGCTAAGACAATC
CGCCTACAGACAATCACAGTGTTTAGTCAATTATGTAGCTTGGATCTCAAGGATCTAGGGCCGAAA
GATGTTAGACCAGTAAAATACTGTTTGCAGAATCCTCCACGATAAAGCCCCGCAGGATTTGTTTGT
GAATTTGGATGGGCCAAGGCAACGGTGGCGCTCAGTGGGGTTTGCCAAAAATACACTGTAGCAT
CAGAGCGGTTTTCCTCAAACTTTCAATTAAACTGGAGAACATGATCCTCTGGTGCTACCTGATAAA
ATCGCAATACTCATCGGTATGATCGGTGGATTGGTTTCAGTCACCTGTGCCGTCAGTATCGCGCT
TCAGTATAGACCGGGTGGTACTTCGGACATGGTACTACGAATTATTCAGCCGGAAGACTTTCCCA
TTGACGGGTCTGCGACCATAATATCAGTGTCATCCAAGCTGTGAGTATATTCATAAAAATGCCATC
CTATCAAGATTTCCAGCATTTTTAGGGTCAGCGCGAACCACGGTCCTATTCTCTGGCCAATGGTCT
GTCACCCTCGGCACAATACCCATCGGCGTAATTGCCGATTTAACTCTGTATTGTGGTCCATCACAT
CGCGTTCCGCGCGGTGCTGTGTCGCTATACGTCGCCGCGCTAAATGGATCGTCGCATGTGCGCA
TCAGTAATGGCTGCCCATGAAGATGCGGGAGCAAGGATTGTCCTACCCAGTCGTCTAGCAGCCA
GCTATCCGCGTCTCGATCCATACACGTATCCCAGTTGCGGCGTGTACACCATGCCCGCAGGCCT
GAGGAGGTAGTGCACACAATTCAGAGATTGCAAGCGGGGCGAGAAGTAACGCGCCACGCCTAAT
TAGAGCACACGCTAGACTGCCAGGTCAACTAGTTGTCCTGAAGCTGTACCCACTCCACCTCACGA
GCGAAGCCTCTCGCGTACGGAGTAATCGCCCATTTACTAGCGCCGTATCTCGTCTCGATACAAGA
CAATACTCGTTGCTCTATTCTGGTTCTGTGGGAGCGTCCTGCATGGATGTTAGCTGCACGGATATA
CTTCAGCTACGGAGTTTGTTCTAGCCGATTCGCCGCTACCTGAGTGCTTTATAAATGGGTATGTTC
GCCTCGCCGTAACAAAATGTAATCGTTGTACAACAAGCCCAATTTGTGGCCTCAAGGCATATCTCG
TTGCGTTTTGATTAGCTCATAAATGATCAACACTG

AL5.

TGTGTAGAGTTCCGCTCGTACGGAGTAGCGGCTCAATAGTGCGGAGCCCAAGTATAGGTTTATGC
TCGTGCCAGCCAAGTGTTAAACGCAAAAGTCATCCGGACATGACAGTTCTATTACCTCCGAGGAC
GTAATCCGCTAACAATGCATCACACAAATAAGCCAATCATACAGAGGGCTTAGGGATATTCCCTTT
GCCCTTACCGGCCGTTCTCAAGATTAGGCTCCTATACGTAAGAAGTACGCCGCAGCAAGCGCCG
AGATCCTCTGAGTTGCGTGCAATGGCTGATGGTCACGTCGCATGGAGTACGAGTGGCAAGACCG
GCTGCCCAACTCAATGCCCTGTGCGTTCGTACGATTGCATGGAGTCGATAAGTGGTGGTATGGAT
CACGGCAGCGAACTCACTCAGCGGTGCGACCGTACTAAACCATAGGCCCTCCAGATGGCGCGTT
ATAATCATAACGTCAAGACCGGCACGACTCATAGGTAGCTACTACATTCCGAAAATAGGTGCGGC
CAACCTACCGATAGACGGATTAATTCTAACTATCACCGGGAAACCAGAATACCCCGTAGGAAGAT
AACCGAGCGATCATTCAAGCGCACTAACGGGGTACCAGACCCATCATAATTGATCCCCCATCGGG
CTAGTTGTGTTCGGCTCTTTGCATAAAGACTTCTCCTCATTCTTGCTGAGGGGGTTTGTGCCTAAC
TGCCGATTCCGTTCGGTCAATGACACTTTCTGGCGTCGCAGTGCGACGGCTCCAGTCCATCGCG
GGCTTCTTGCAGGGGCCACGGTACTTCAAAAATTGCCGTTTTCATCGTCAAGTCAACCAATCTAAA
TAAATTCCATGTCAGGGGTGTTATTCTGCACGTTACATTGGGACGTCCTATAAGAACGTACTCCGC
AGACGTACACAATGGATCCATTCGAGCTATGTAGCCTCTTTTCGGGATTGTTTCACAGCTGAGCG
GCCGTAGCTTTAAGATCAACGAGTTAGTATACCGCTACCTAAAACGGGGCTCTAACGGTGCGTTA
TTGACGCGGTTTTATTAACGGAACCGGGCGACTCCGCCGCCACGCTCTCTTGGTCTGAGGGACT
AGTGGCACGGAGGGCCTTTGAGTTCGCAGAAACGTCCCGGACCTACAATAGGCTTTCTCATAAAA
GCAAGCTGACTGTGGTGTGCTCATCTGAACACCACCTATGGTCCTACTTTCGTCCGCGAATTTCG
AACCCAGCCGGCGCCCTCGCTACGGTCAACGGATCTGCCCTTGTGAGGTGCTCCTGCAAGCCGG
ATTCTCGGCGGCATAGAGTTAGGGAACACCATTTCTAGAGCCGCCAATATGAGCGCCGAGACCCA
GTCCCGGACCAACCCGGTTCGGTAACTCCTTTCTGTACAAACGGTGATGTGTTAGCGTTGTCATC
ACCTTTACAGAGTGCTGGAAATGGAGGTCTGCATAGGTAAAGAGGAGTATTACTCGATTGATGGG
GCAGAGGACAACGCCGAGCCCCTTTCGTTACACTTTATATGACTTCCCTGCATTCAGCACCAAGT
GCAACCGCTGAGCTACGAGTTAATCATCTAAACACCATCGAG

AL6.

ATTGTTTAGCGGTGACACTAACCTAAAGAATTTAATGATCACGTGCCCCGAGCACCATGCTGCGC
GTATACCGGAGTCTAACCGCGGTCACACATCAGTCTTATCTAGGCGCGTCGCCTCCATCCGGTCG
TCAAGTCTGCGTCGGAGACGTTCGACAGATACGCTTACAACCTCATTACATCACGGTGGCGTGAA
AATCCCCGCGCACCGAATGAGGTCTAGAAGCTGGCGCTCGAATGGCTCAAGGTGTGCAGACCCT
ACCCGTCAGTAAGCGATAACTATTCCGAAATATTGTGCCGCTGTTAATTAATTGGCGCCTGCAGTG
AAGTCTATGGTAGCGCATTAACAGGTTCCCAAGCTAACGATATCTTGCTCGCGGCACTCTTTCAAC
GGATTGTTAACGCACCTTGCTAGGTGAAGTTGAGATAGATGTAGGCTCATTTAGAAGTTGGCGCTT
AAGGTAGGGTAAAAGCAGAGAGTAGCTTAGAAGCGCGATCCCGACTCAAAGAGGGTCTCTAAGC
ACCTTGATCTTTGTCAGAAGTTATATTCGACTTTGCAGAACAAGTGAGCGCCGGTATCAATCCATC
TTCCCTTCGGCAAACGAATGATGCAACAAGCGGCACCGGAGGTTGGCTTCTGAAGTGAGCGTGT
GCAATATGTATCTGACTCGGCATGCCTACCACGCTTGTACACGCCATATGCTCACGGACTGTGAC
ATATATCCGCCGAGGAGATAGTTGTATTGTGACCGGGGTTCCACCATATACTCTGTAAATGACGG
CGTATGTCGGGGAAATCAACGAGACGATACAGGTCGCAAGCTGCTGAGTGCGTGGTTTCAGAGC
AGTGTATGAAAACTCCATGGGAATTTGTATCGTTTGGTACGGTACTACTCGCACGATATGCAAAAT
ACCGTGGTTCCATTCCAAACGCCATTGATATCCTAAATAAGGAAGCCTCTGTTGAGGGTCGACAC
AGGAAAAATTCGCCACGTGCCCTACTGTGTACGGGTAATTTAACCGATAATCGCTTATATGAGGAG
GAGCCAAATCAAACCTCACGCCCTTTCAGCCTAGCATGTCGTAGATATTGGAAACGCTTATGACG
ACAAAGAGGTCCACCAAAACCAAGCGTCGATCGTCTCCGGCACACTCTCATTGTATATTAATCACT
ACGCTCAAACGCAAAACCCAGGCCAACTTTTACTGGGGGTGTTGCTGTAGGTCGTTGCACCTACA
TGTGAGTTATTGACAGAAAGCTGAGGAGTACCCCCAGAGCAAACCTACGTAGTGACGCCTGTGCG
AGTGTGCCCCCATACGAGCTGAATCGACAGTCATCATAAAACGTACATATCATTTAGCCAGACATG
ACGCGTAAAGGCGAGAGGAAACTCCCTAGCAGGCCAGTCACTCTGATGGCCCTTGATAGGTCGT
AACCCGCCAAACGCGCATCAGGAGGAGAGCCAAGCGCAGCAGGTGGCAGCTATCGAGGTTAGAT
GCGACTGTACCACGATGCCGAGTGCAAATCGCAAACAGAACTTCATTTGTCCCGGACTCAGGGCA
GTGGCTATACTAGGAGGGATAATCATATCCTTATTTGC

AL7.

TCTCTTTCATGCGCAGTGCTCTCGCAAGAGCCTTCTGAACGCTCGCAAGTGTATGCGTTCATATTT
GATCTCGCTTGCAGGAACAGTCGATCACTTTAGACCTAGCATAATGACCCAGTCATACCAATACTA
CGAATGTACTGGGCCTGAGCAACATGCTGGCAGGGATCTGACATGCAACGAGGCATTAACTCGAT
GTGGGCGCTGTCATGCGTATATCCCGATTTTATGCGATTCCTGACGTGCTACCCTATGCTAGGTA
GTGAGTTAGCTTTTCACCGACGACCGTTAGCATTCGCGATGTTTTTCTTACCGTCTTCCTCGGGTT
AAGAAGGCACAGTGTTTGAGGATAATTTTCCATGTGTTTGATCTTTCAATTCCGGTGAAATGCCATT
ATGTCAGGAATCCGTGATATTATTCGAGCATGGCGGAGGTGGTACCTAATTCGGGGAATCTCCCC
CTAATCTACATTGGAGGACAGGCTTTATATTTGTGGTCCCATCAGAGGCCCCAACGGAGCCACCA
CATAACCTCAAATAATACACCTTCCCTATGTATTATCACTGTGACAGAACCCGGGGTAACATACTC
TAGCTTTTATTCAACGTGCGATTACTGGAGGCCAGCTCGCTTAGATCATTAACTTCGTCGACATCC
CTGCTGGTCTTCGGGTGGCTCAAATCGGCGATCGGTCCGATAATGTGACAATATGGCGGGAACC
ATATATGGCGTTCGCTCGTTGAACGTAAGATTCGAGGCGAAAAAGCTCACGAAACTTCTATATACC
AAGCTACTGCTGTCTAGCCGCTCCCGTGAACTCACTTTGTTATGACTTCAGGCCGCTGTTGCGGC
ACGACGCCGGGTAGATGAAGAAGTCGATTGCGAGATACTGATCCACTAGTGAGGAGCCTCATAC
GGGAGAGATACGGAAACATTGCTACACTTCCTTTACGTGATGTATAAGACAGTCACTGATCTTACC
GCCGCTGAGATATTAACAACCTATGACGGCAATGCTACCCACAGGGATGACGGCTATTCCGGCTT
GAACACCCGCCTGGGGTGTTTGGGTTTTGAATGCACGTTGGGACGCTATCCTTGCAACGTATTGG
ACCTCCGACTAGTACAATCTCACTCCGTAAGCAGGCTTCCTGCAACCACACTAATATTTCCGTTGC
CTCCCCTTAGTAGAAACAGAAGCTATCGATTCTAATGATTAACTATAAACCATCATTCTTTAATCAA
GGTGACGATTTCCATCCCTGCTAAATTGAGCAGCCGCGCGGGGGATTTGACATATTGGATATTAG
TAACCCCTTGGATAACTGATATGCCGATCAATTGGGAGCTGCAGGTGAACCTAATCAGCCGATCT

CCGTATGCGGTATAGGTCGGCTGGGCTAAATGGTCATCGTCCGCCTCTAGCGTAGCTGAGCGGT
CATAGTGTCCAATGCGCCGGTTATTCACTGACTAGGTCGCCCCCCCAGAGGCTTGTGGTTCCTAC
TGCATTGCGGTCGCTGCCCCGTTAGTTTTGAGACCCCTATCTTACGCACCATAAAACAGATACAG
GGGGGAGTCATTTCTGGTGCGGACTCTCGAA

AL8.

TCTCATCTAGGCGAACCCTGAAATCGCCGTTTATCGTCAGCGATGTTTACGCTCCGTAGATAGTCT
CGCGCTTAGGAGCACGGGCAGACGGATATCCTCGGAGGCTAATACTTGTTGACTATCTCGATGAT
AAAGAAGATTGGAAACCGCACGTTGCTGACCATTATTTAGATCATCTCCTTTCCACGTTGCCTATA
CAACGCGCCAACCGCTCCCACGGACCATGCGGTTCAGTCCTCGCCCACTTTCTAGTAATGTACAT
AAACGCGAAACACTTGATCTCGTTAGTCTGAAACCTATGGTTCTGGACTCTCGATGGGTTTGTTTT
GTTACCGTTATAAATTCAACCACAACTCGTCGCATTCTAAACTCGCAAGGTGTAGGGAACCTGTCT
TCCAGGAGGCTGGTTTGTTAGAACCAGTTGCGATGGAGGACTCGGGCGTAGTCCTAGGGTAACG
GAATGAACCTGGAATGACCTTGATCATAGTAAACTCCGCGGATTTCCCGTGCTTCAGCGGCTACA
AGTTTGCGCCTGGATTTATGTGTTCAATTGGAACTCGATACCGAGATTTTACCCTACTGCGTTCAC
AATCAGCTAAGGTTAACACCCGGACAACTGGCAATCACAGCAGCCCTGAGAAGAGAAGACGTCGT
CACAAAACTCCATGCCCAACGCGGCCCGTAATCGACTATTATACAGTATCGCTTAGCTTCCCACG
CACTGGGCATCTAGGACGTGCCAGTACCTATTGCCTCCAGTGATATTGCTACATAGGGCGGGTGG
ACGGAGAGGAGTTTTATCTTGCATCTAAAGGTTGTTATTTAGCCCTTATGCGGCACCCCGAGACAA
CAGACAGTTAGGGTCGCCGAAAAGGGAGGTTAAAGTTGTTACCTCATTGGTGGCAATGTTAGATC
CTTCAATCCTCTAATGCCCTACGATAGTCACTCCGCCACTGGCATCATCGTGGTCTGGGTTTACCA
TGCCCCAGCAGATCCTAGCTTGCTACTCGGTACTGGCAACTGAAGGCATCCGCTGTTGGGTGGC
TCTCCGCACCATATAGTTCACCGCGTTGGCCCTTCCACATGTTACCCTTTCCAATTGCTAAATGGT
GCTGGCATGCTCATCCAATTTTGTTAATTTCTAGAGTTTAAGCGGGGTTTATGAGCTTAATTCCCG
GACCTCGAAGATGTCTAGAACCAACACTAACATGCAGGTCAGGCGTGTAACTGTTTTATGATTCCG
TCTCTTCGAGCGAGTATGACCTCATACCAGGATCTTCTGCTACACTCGAAATCTAGCATCAACTAC
GTTGCCAACAGCGCTTGTAGGGCCGTGTCAGACAAGCATTTTTGTCACTAACGTGTGCCAGGAAA
TTCACAATGGGCGGCGACACAGAGGATAGATCCGGGCTAGGTTCACACTGGCGGCAACGTGAAG
GTAGTTACAGACCGTATTCGGATAGCACCTTCCGCATTACTTGGTGATAGGGCCGTCTTTTGTGAC
GTGGTAAGGCGTGGATCTTGTTTCATACCTAGCGTTGTTCCCCCCCCTACGCGTTCGCGTTCAGC
GAAGAAATCGATTACCCAAGCCGTGCCCGAC

AL9.

GTGCTCTCATGTGTCACACGAGGCAACTCTAGCTTGTGGGAGGTCGGTGCACAACGAAAGATTTG
ACCGAAATGGCGTAGATGACGAAAGGGGTACGGAGGTCTACCGCTTGCTTTCTACAGGTAGATG
CGCGGAAGCCTGATGCGAAGGATGTGAGTGGACCTAGACGAGCCAATCATTACAGTCTAGTTTAT
GGAAGTTTTTCTTAGACATATCTATTCGGCCGGAACTACGTAGTCCCCTTTAAACATTGAATATTTG
CACCGTCTCTCCGCAGGGTGGTTTCGTCCACGTTCATGGTATTCGAGGTACGGTACGACTAGGAC
ATATGGAAAGCCCATTTGACCCTACCCTCTTTTTCGGAGGAAAACTGATCGACGGCTAAGGCGCC
TGCGATGGTGAAGCTACTGCAAAGGTGACATACGCAGCGACCCTGGCAACGGAAAATGATTCATG
CCCGCTTCAGGCGCGACGAATCATATTTTCCACGCTCATTTTATTTCCTTGAGCACTCGAGTCTGT
TACCTACCCGCTAAAGGTAAGATGATCTCTACAACACTCCGCTTCCACGAGCTCTTCTCTTAGCCG
GTGACCCACAGCGCATCGGTCTCGAGTGCGGTCCTTGCATGAATGGGCGAGCTGCCGTATAACA
CCTTGCGCTCGATCACCCTGGCATGCCCTATATGGAATGGACTGGGACCATGAATCTCCTGTGGC
AGCTTACGCAGACTTCTAAAATTGAGCCCACGTAATAATACCTGTTTCCAGCGGTAGGCGGCACC
CTCTCGATCCGCGTGAGGTCGTACACCTATGATACCAACGATCTGGTGTCTGAACCCGCAACATG
AACCTGCAAATGGTGCGATAAGATTACCATGTTTGTTGCCCGGTACCCAGAACGGACTCATTGAA
GTAGACCGGTCGATCTCTTAGTGACGACGCGCTCGCTTATGGGATGACGCAACGACTGCATATTC

CTATACCTGAGCTACAAGCGCCAATCGCGGGCGTTCAAAATCTTTCAGTAGGCAAGAGTGCGAGG
ATAGTACTGGCTGCAGAAAGTATAGCTGCGAAATGGCTCATATCGTCCGACTCTGAGTATGCTCA
CCATAACTTTCATGTCTTGGGCTTACAAGAAACGGTGATTTCAGCATAGAAACGCTATCATGGTAG
ATAAGCAGGGATGCCGGATTTTAGTTAGCTGGCAAACTCCTTACAGGCCGTGTACTACACCGCGC
TTCCATCAGTTCGAGAGGGTCGGACTGGCGCCTGACCGTACACTATGAATTTCTTCGCCACTACA
AAGCGCCCCACATATCTTTCTTTACCTTGGGGGTGTTGAGCTGTAGCCGAACAGACTGGGGGACC
GCCAAGATGTAACAGTACTTTAATAGCGGGTGGGGACTCCCGCCGCGTTAAACAACTAACTAGAT
GAATCCGCGGCATAACAACAGTGCGGGCATGTTTTAGGGCGTCCGTCGCGTTTAATCCCCACCAC
GCTCTTTTCCCTCTCACCAAAGTTGATAACACCATACCAAACCTCAAGCCTAGCAGCTACCACGAG
CTAATCACTATCTCCATCGCTGTCTGAGCTGACTAG

AL10.

GGTCGCTGGACTGCTTGATCCATGTCCGTTCGAATACATTGGCGGCGTTATCTAAAAAGGCTTTA
GCTACCGTTTGTCTGCTCGAAAGCGTGCATTCGCCACGGCTAATCCTTTTCCGTGTTCACTGAGC
GTATCTGTCTTTAGACCCGAAATTACCAAATACAGGTAATTTTCAAGGATAATTCTCTCGATGACAG
TTCCCGCTGGCTCAGAAATCGTGGGTGTTCAAACTAACACGACCGGTTCACGAGCTTAGCACAAG
CGGGCTTATAGGCATAAGTAATATCGATTTCCGTTGTTAAGCTTAGATACCGCTGCACGGCTGATC
ATAATCTAAGTCAATGCTTAACGCTAATTCTTTCTAGGCGGGACTAGGGTGCTCAAGTTCTATGTA
AGCTACAACTTAAGCGTTTTTCCCAGACCTCAAGATCCACCCTTAACACTACCTCCCTTGTTCTTAA
ACGAGTTTTGTGTGAATCGAACCAGGACTGTAGTGAGAATTCTACCATAGAGCTGCGCTTGTTCTA
TGGTTTAAGGTCGGGGCCCGCCACCTGTCGGGGATTGGCCTTCGGTCTGGCTGTCTAAAGGGGC
GCAAGCATCTATACTGCCCGTTGGGTCACGGTTCGGACGTCGTATCGTCGTGGTATCAACGGCCT
CTAAGAGTGATTGGTTAAGGCTCAGTTAGTTTTGATGGCAGTGAGGTGGCTTAACCGTGACTCAG
TTAAGGTTAGGGACTCTAAGGTAAGCTTCATCCACCAAAATTAGCCCGGTTCCCGTGGAACAGAT
ATAACACAGCTTCTGAGCCCAACAGTCAATTTACTGCCGCCTGGATGCTGCCTAGGCTGTCGGGT
GTTTAATGCCTGCGGTACCGGGTTTCGTCTGGCGCCACTATTGCAATGAGTAGGACTACGAGGTT
TAGCACGAGAGGGTGCCGCACGTTTAAGCTGTACCAGGTGATCGTTTATAACATGTCTCTCGTGG
TGTAGCGTCCACTTTCTCCTTCAGAAAGTTGGGTCTCCTGCGCGGGAGGGCGTAACTGCCAGAG
GACTCCCTTGAACAGCTGTAGTTGCAATCTTCTGCTTCTTGGGCGGGGGTATTAGATAGAATGCT
GGTAAAGCGCCATATATCTTGCTAGCCATGTCTGCCACGGAAACGAATAATTCGGGAGAATTAAA
CCACGCATTGATCAACAACGGGAAATATAGGCACATGCGGGCACCTGGGGGGTCAAACTAATCT
GTGGGAGTCACTGGTGCTGCGAATTATCTTGTCCGTCTGACAGGATATGGAAGGGATTGAGCATC
CTACGTATTCTGGGGAATACCATCCGGGAATCGCACTGCCAGAAGCCTTGTTAGTCTTGTAATAC
GAGAGAGGACGAAGCCATCGCGCAAGGCTGTAAACAATTGTAACGTGGTGGTAATTTCAGTAGTA
GCACTATCTTGAATTAAGGATTTTTGGGCCCGGGGTCTGGTCGACGCAACCAAGTATCACTTGGG
ATCATTCTGACGTGGGGGTCCGGGGTTAAAAAGATGGGGAGAGTAGACGCCCTCTTCCTAACTTA
ATACCGCGTACAGACTGCACTCTAATGGTAGGGAGA