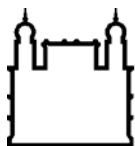MINISTÉRIO DA SAÚDE
FUNDAÇÃO OSWALDO CRUZ
INSTITUTO OSWALDO CRUZ

Doutorado em Programa de Pós-Graduação Biologia Computacional e Sistemas

SEARCHING FOR TPP RIBOSWITCH IN THE HUMAN GENOME AND COMPARISON WITH OTHERS FROM DIFFERENT SPECIES

DEBORAH ANTUNES DOS SANTOS

Rio de Janeiro
Junho de 2018

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ
Programa de Pós-Graduação em Biologia Computacional e Sistemas

DEBORAH ANTUNES DOS SANTOS

Searching for TPP riboswitch in the human genome and comparison with others from different species

Tese apresentada ao Instituto Oswaldo Cruz como parte dos requisitos para obtenção do título de Doutor em Biologia Computacional e Sistemas

Orientador (es): Dr. Ernesto Raúl Caffarena
Dr. Fabio Passetti

RIO DE JANEIRO
Junho de 2018
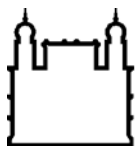
Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ
Programa de Pós-Graduação em Biologia Computacional e Sistemas

AUTOR: DEBORAH ANTUNES DOS SANTOS

SEARCHING FOR TPP RIBOSWITCH IN THE HUMAN GENOME AND COMPARISON WITH OTHERS FROM DIFFERENT SPECIES

ORIENTADORES: **Dr. Ernesto Raúl Caffarena**
Dr. Fabio Passetti

Aprovada em: 28/06/2018

EXAMINADORES:

Dra. Ana Carolina Paulo Vicente – Presidente (Fiocruz/RJ)
Dr. Leonardo Soares Bastos (Fiocruz/RJ)
Dr. Paulo Sergio Lopes de Oliveira (CNPEM/SP)
Dr. André Elias Rodrigues Soares – Suplente (LNCC/RJ)
Dra. Ana Carolina Ramos Guimarães – Suplente (Fiocruz/RJ)

Rio de Janeiro, 28 de junho de 2018
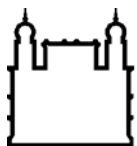
Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

Anexar a cópia da Ata que será entregue pela SEAC já assinada.

Dedico aos amigos
Aos que se tornaram familiares,
Aos que nasceram familiares
e aos que conheci antes de ontem.

Dedico tanto aos que me deixam louco,
Quanto aos que enlouqueço.

Aos que me criticam em tudo,
E a um ou outro que atura
Minha "chatura"

Aos amigos que correm,
Aos amigos que contemplam.

Aos que me consideram muito,
E aos que, com razão, fazem pouco.

Aos que conhecem o que penso,
E aos que só conhecem o que faço.

Aos que passam o dia todo comigo,
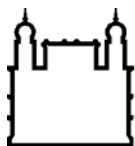*e aos que estão todo tempo em mim.*

Este trabalho é a soma de todos vocês.
E se ele não é melhor,
É por falta de memória,
Mas não por falta de amigos.

Adaptado de Efraim Rodrigues, Richard B. Primack

AGRADECIMENTOS

*"We build too many walls and not
enough bridges.'
- Isaac Newton*

Ministério da Saúde

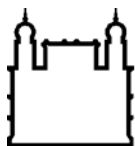**FIOCRUZ**
Fundação Oswaldo Cruz

INSTITUTO OSWALDO CRUZ

BUSCA DE TPP RIBOSWITCH NO GENOMA HUMANO E COMPARAÇÃO COM OUTROS DE DIFERENTES ESPÉCIES

RESUMO

TESE DE DOUTORADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

Deborah Antunes dos Santos

Riboswitches são sensores de RNA que afetam os processos pós-transcricionais através de sua capacidade de se conectar a metabólitos. A classe de tiamina pirofosfato (TPP) riboswitch é a mais difundida entre os riboswitches ocorrendo nos três reinos da vida. Mesmo controlando diferentes genes envolvidos na síntese ou transporte da tiamina e seus derivados fosforilados em bactérias, archaea, fungos e plantas, o aptâmero de TPP possui uma estrutura conservada. Riboswitches são considerados alvos potenciais para drogas antibióticas. A piritiamina, um análogo de tiamina, foi demonstrado ser tóxico para bactérias e fungos, tendo TPP riboswitches como alvos e, assim, reprimindo a biossíntese de tiamina. Portanto, torna-se essencial investigar a existência de riboswitches no genoma humano para evitar efeitos adversos. Neste estudo, visamos compreender o comportamento funcional de TPP riboswitches de bactérias e plantas, baseado em suas estruturas cristalográficas (TPPsw[ec] e TPPsw[at], respectivamente), nos estados *apo* e *holo* em solução aquosa. Adicionalmente, buscamos candidatos a TPP riboswitch no genoma humano, propusemos modelos 3D, e analisamos o comportamento estrutural dos candidatos humanos comparando com as estruturas cristalográficas disponíveis de outras espécies. Uma combinação de abordagens computacionais, envolvendo Bioinformática, Modelagem Comparativa, Simulações de Dinâmica Molecular e Análise de Redes, possibilitou encontrar diferenças no comportamento estrutural dos TPP riboswitches em espécies de bactérias e plantas, juntamente com a identificação de um potencial candidato para TPP riboswitch no genoma humano. Nossos resultados sugerem que diferentes interações no microambiente ao redor do nucleotídeo U36 de TPPsw[ec] (e U35 em TPPsw[at]) podem estar relacionadas a distintas respostas ao TPP. A análise de redes mostrou que pequenas diferenças estruturais no aptâmero permitem uma comunicação intramolecular aprimorada na presença de TPP em TPPsw[ec], mas não em TPPsw[at]. Os TPP riboswitches de plantas apresentam mecanismos de regulação mais sutis e lentos que as bactérias. Interessantemente, um potencial candidato a TPP riboswitch no genoma humano foi identificado no gene FBLN2. Dos três modelos construídos, dois mantiveram interações específicas do RNA com TPP: os modelos CANtrunc e CANcomp[2] mantiveram-se conectados ao TPP através de ligações essenciais específicas com o anel de aminopirimidina do TPP. No geral, o gene FBLN2 pode ser considerado como um possível candidato a TPP riboswitch.

INSTITUTO OSWALDO CRUZ

SEARCHING FOR TPP RIBOSWITCH IN THE HUMAN GENOME AND COMPARISON WITH OTHERS FROM DIFFERENT SPECIES

ABSTRACT

<u>**PHD THESIS IN COMPUTATIONAL SYSTEMS BIOLOGY**</u>

Deborah Antunes dos Santos

Riboswitches are RNA sensors that affect post-transcriptional processes through their ability to connect to small molecules. TPP riboswitch class is the most widespread riboswitch occurring in all three kingdoms of life. Even controlling different genes involved in the synthesis or transport of thiamine and its phosphorylated derivatives in bacteria, archaea, fungi, and plants, the TPP aptamer has a conserved structure. Riboswitches are considered potential targets for antibiotic drugs. The pyrithiamine, a thiamine analogue, has been determined to be toxic to bacteria and fungi, targeting TPP riboswitches and thereby repressing thiamine biosynthesis. Thus, it becomes essential to investigate the existence of riboswitches in the human genome to avoid adverse effects. In this study, we aimed at understanding the functional behavior of TPP riboswitches from bacteria and plant, based on their crystallographic structures (TPPsw^{ec} and TPPsw^{at}, respectively), in the *apo* and *holo* states, in aqueous solution. Additionally, we searched for candidates for TPP riboswitch in the human genome, proposed 3D models, and analyzed the structural behavior of human candidates and compared them to the available crystallographic structures from other species. A combination of computational approaches, involving Bioinformatics, Comparative Modeling, Molecular Dynamics Simulations and Network Analysis, made possible to find out slight differences in the structural behavior of TPP riboswitches in bacteria and plants species along with the identification of a potential candidate for TPP riboswitch in the human genome. Our results suggested that distinct interactions in the microenvironment surrounding nucleotide U36 of TPPsw^{ec} (and U35 in TPPsw^{at}) might be related to different responses to TPP. The networking analysis showed that minor structural differences in the aptamer enable enhanced intramolecular communication in the presence of TPP in TPPsw^{ec}, but not in TPPsw^{at}. TPP riboswitches of plants present subtler and slower regulation mechanisms than bacteria. Strikingly, a potential candidate for TPP riboswitch in the human genome was identified in FBLN2 gene. Out of the three models created, two maintained specific RNA interactions with TPP: CANtrunc and CANcomp^2 models maintained connected to TPP through specific essential bonds with the aminopyrimidine ring of TPP. Globally, the FBLN2 gene can be regarded as a possible candidate for TPP riboswitch.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

2D	Two-dimensional
3D	Three-dimensional
CANcomp[1]	Complete model 1 of putative human TPP riboswitch
CANcomp[2]	Complete model 1 of putative human TPP riboswitch
CANtrunc	Truncated model of putative human TPP riboswitch
CM	Covariance model
CASP	Critical Assessment of protein Structure Prediction
DCCM	Dynamic cross-correlation matrices
DNA	Deoxyribonucleic acid
FBLN2	Fibulin 2 gene
GAIT	Gamma-interferon-activated inhibitor of translation
HMM	Hidden Markov Model
hnRNPL	Heterogeneous nuclear ribonucleoprotein
Infernal	INFERence of RNA ALignment
IUPAC	International Union of Pure and Applied Chemistry
MD	Molecular dynamics
ncRNA	non-coding RNAs
NMR	Nuclear magnetic resonance
PCA	Principal component analysis
PDB	Protein databank
PTPP	Pyrithiamine pyrophosphate
RMSD	Root-mean-square deviation
RMSF	Root-mean-square fluctuations
RNA	Ribonucleic acid
SCFGs	Stochastic context-free grammar
SIP	Square inner product
TPP	Thiamine pyrophosphate
TPPsw[ec]	TPP riboswitch 3D-structures of *Escherichia coli*
TPPsw[at]	TPP riboswitch 3D-structures of *Arabidopsis thaliana*
UTR	Untranslated region
WC	Watson–Crick

1 INTRODUCTION

1.1 General Considerations

Nearly sixty years ago, the central dogma of molecular biology was proposed as the preferential flow of information, stating that DNA is transcribed into RNA, which in turn is translated into proteins with structural or catalytic functions (1,2). Since then, new findings have indicated that this theory was incomplete. For instance, in 2007, the ENCODE Project Consortium showed that, although most of the DNA is transcribed, only a fraction of the transcriptome is translated into proteins. RNA portions that do not encode proteins were then termed non-coding RNAs (ncRNA) (1–3).Those ncRNAs, belonging to the same class, share sequence and structural characteristics, which have been conserved throughout several evolutionary processes. The degree of sequence conservation is smaller than that observed for protein-coding genes but is crucial to explain the functional heterogeneity of the ncRNAs (4,5). One of the most significant examples of conserved functional RNAs are the riboswitches (6).

1.2 Riboswitches

Riboswitches are natural RNA sensors located in the untranslated regions (UTRs) or the introns within an mRNA sequence. These sensors are capable to be bound by a great variety of small molecules, such as vitamins, amino acids, and nucleotides (7,8) and controlling the transcription or translation of the host mRNA. Riboswitches can be classified into different classes according to their binding metabolite, being the most significant class the one including those capable of recognizing coenzymes, such as adenosylcobalamin (AdoCbl) (9), thiamine pyrophosphate (TPP) (10), flavin mononucleotide (FMN) (11), S-adenosylmethionine (SAM) (12), S-adenosylhomocysteine (SAH) (13), tetrahydrofolate (THF) (14) and molybdenum/tungsten cofactors (Moco/Tuco) (15). Riboswitches in the second largest group of purines with binding ability and some compounds derived from purines, such as adenine (16), guanine (17), pre-queuosine-1 ($preQ_1$) (18), deoxyguanosine (dG) (19), cyclic-di-GMP (c-di-GMP) (20), and cyclic-di-AMP (c-di-AMP) (21). They also recognize amino acids, including lysine (22), glycine (23), and glutamine (24). Other metabolites include metal cations such as Mg^{2+} (25), the halide anion F^- (26) and

glucosamine-6-phosphate (GlcN6P) (27). However, riboswitch classification could be more extensive as there are several putative structures to be validated and orphan riboswitches yet to be identified (reviewed by Peselis and Serganov (28)). Examples of known riboswitches are depicted in **Figure 1**, which is part of a review published by our research group early this year (29) (Appendix A).

Genes regulated by riboswitches are involved in the biosynthesis, catabolism, signaling or transport of its binding metabolite, which create a negative feedback regulatory mechanism to maintain the adequate levels of this molecule in metabolic processes (30). When metabolite levels increase, the binding to the riboswitch occurs, leading to down-regulation of the expression levels of the metabolite-related genes and, consequently, of the metabolite itself. This negative feedback mechanism can be regarded as a fast reaction to changes in the environmental metabolite concentration that does not require the assistance of other supporting molecules (31), minimizing consequently energy waste (32).

The structure of a riboswitch includes the aptamer and the expression platform, both of which are connected by the switching sequence. The aptamer region is evolutionarily conserved and responsible for metabolite recognition and binding (31,33,34). Metabolite binding induces a structural change in the expression platform, which is a highly variable region (31). This last modification controls gene expression (32). An example of this class of riboswitch is the guanine riboswitch, which is present in the *xpt-pbuX* operon of *Bacillus subtilis* (28,35). In some riboswitches, such as the SAM-II riboswitch in the *metX* transcript of the Sargasso Sea metagenome, both aptamer and expression platform are merged into a single region (36,37). In this particular case, SAM binding promotes the formation of a pseudoknot[1] structure, which includes the Shine-Dalgarno sequence, preventing its recognition by the ribosome.

---

[1] Pseudoknot is an RNA structure in which the loop of a hairpin pairs with either a stem or a loop outside the original hairpin (198).
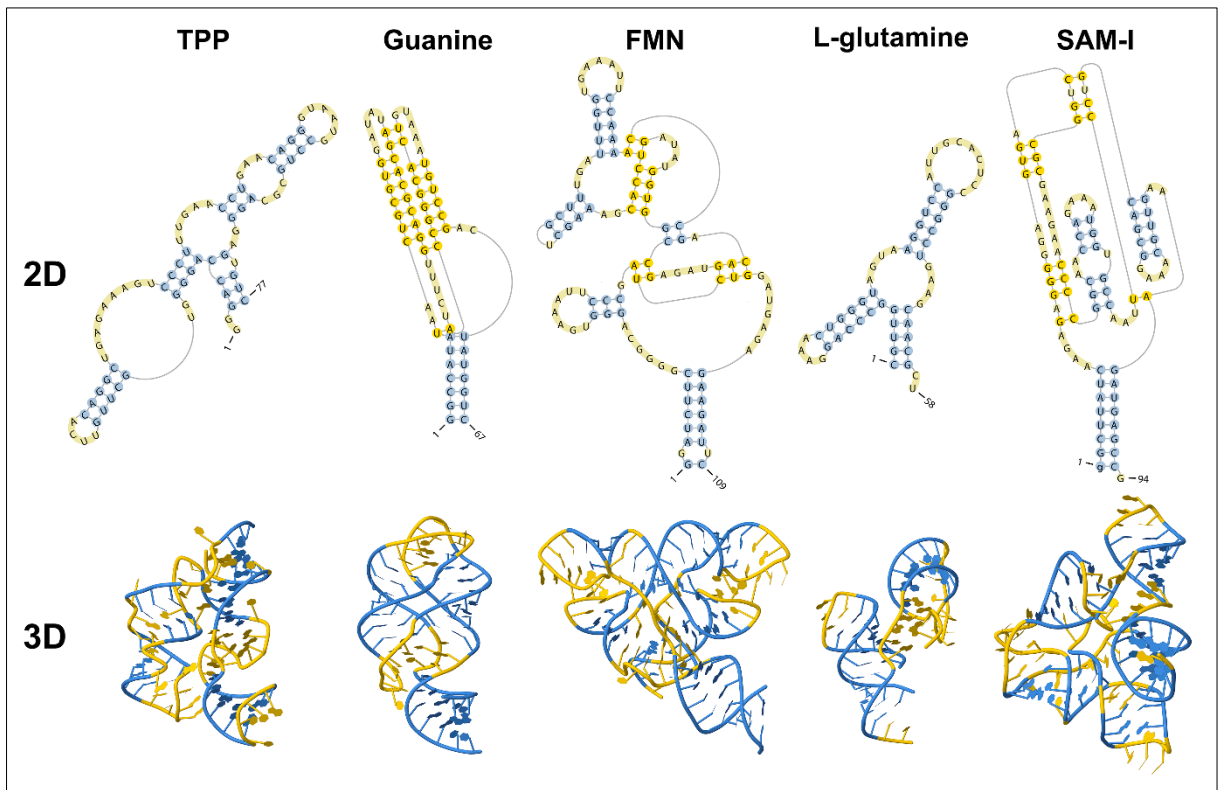
Figure 1: Secondary and tertiary structures of known riboswitches. **Source:** Antunes et al. 2018 (29).

Riboswitches have two stable conformations, "ON" and "OFF", which depend on metabolite binding (38). However, the adenine riboswitch in the thermophile *Vibrio vulnificus add* gene is the only known exception with three stable conformations (39). In one of them, the metabolite was inside the structure, and a free Shine-Dalgarno sequence allowed translation. On the other side, the metabolite was not inside the riboswitch in the two other conformations, and the Shine-Dalgarno sequence was not free. The difference between these two ligand-free conformations is that one of them, termed apoB by the authors, it is incapable of interacting with the metabolite. To adapt its 3D-structure to the other ligand-free conformation that is able to bind adenine, termed apoA, a change in the environmental temperature and metabolite concentration is needed.

The aptamer has an extremely high specificity to bind the metabolite, attributing the capacity to act in the presence of many related compounds (33). For instance, the AdoCbl riboswitch cannot bind to methylcobalamin or cyanocobalamin (40), and the TPP-binding riboswitch does not interact with thiamine or thiamine monophosphate (TMP) (41). This specificity is due to the evolutionary conservation of sequence together with other structural features. If mutations occur within metabolite-binding regions, the function of the riboswitch will be affected or even abolished (42).

Riboswitches can regulate gene expression in two different ways: prematurely terminating transcription (**Figure 2A**) or preventing the translation of its host mRNA (**Figure 2B**) (43,44).
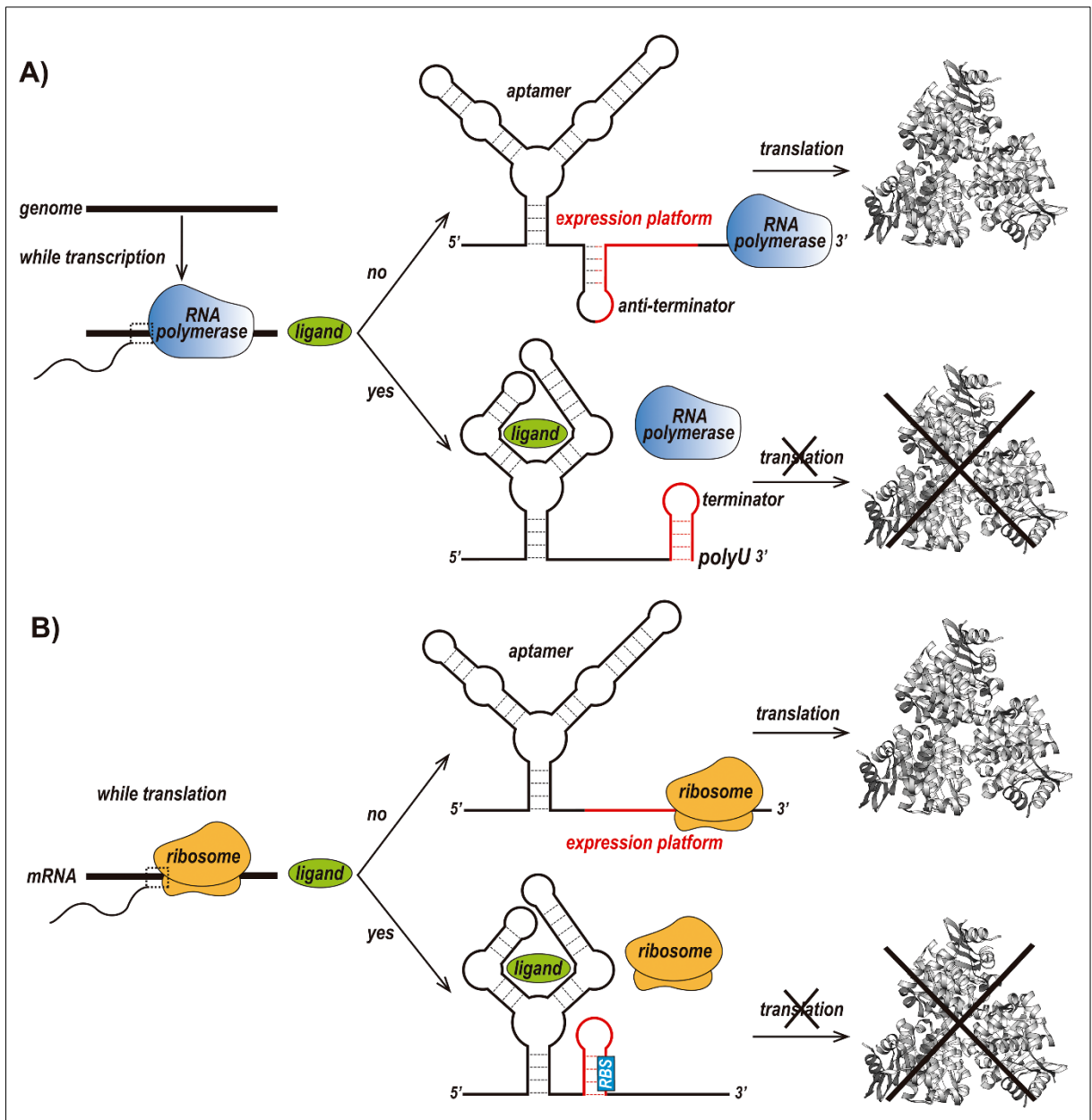
Figure 2: Two different forms of the riboswitch regulatory mechanism. **(A)** Premature termination of transcription. In the absence of a ligand, transcription of the downstream gene is permitted due to the formation of an anti-terminator stem. Upon binding of the ligand to the aptamer, a terminator stem is assembled the anti-terminator instead, and transcription is terminated. **(B)** Prevention of translation initiation. In the absence of a ligand, a ribosome binds to the ribosome-binding site (RBS) of an mRNA sequence and initiates translation. When the ligand is available, the RBS is sequestered and is not recognized by the ribosome, preventing translation from occurring. **Source:** Antunes et al. 2018 (29).

In case of premature transcription termination, the structure of the expression platform folds giving rise to either a terminator or an anti-terminator hairpin (31,45). For instance, in the above-mentioned example of guanine binding riboswitch from the *Bacillus subtilis xpt-pbuX* operon*,* binding to guanine leads to the formation of a Rho-independent transcription terminator, while the ligand-free conformation forms an anti-terminator hairpin. The Mg^{2+} and FMN riboswitches, which are found in the *mgtA* transcript from *Salmonella enterica* serovar Typhimurium and the *ribB* transcript from *Escherichia coli,* respectively, prevent transcription elongation by a Rho-dependent transcription termination mechanism (46). Upon riboswitch-metabolite binding, Rho binds to the transcribing mRNA, translocates up to the RNA–DNA helical region, and separates the transcribing mRNA from the template DNA thereby terminating transcription prematurely (45).

Prevention of translation initiation occurs due to the absence of the ribosome-binding site (RBS) (45). Examples of such riboswitches are the SAM-II riboswitch in the *metX* transcript of the Sargasso Sea metagenome (47), the adenine riboswitch within the *Add* mRNA from *Vibrio vulnificus* (39), and the lysine riboswitch in the *lysC* transcript from *E. coli.* In conditions of high lysine concentration, the expression platform of these riboswitches acquires a structure that prevents translation and exposes RNase E cleavage sites simultaneously (28).

1.3   Thiamine

Thiamine (vitamin B1 or aneurin) is a water-soluble vitamin that constitutes the B complex. It is composed of two rings: one of pyrimidine and one of thiazole and exists in four forms: alcohol, thiamine monophosphate, thiamine diphosphate and thiamine triphosphate (**Figure 3**) (48)*.* Its active form is the thiamine pyrophosphate (thiamine diphosphate or TPP), which acts as an essential coenzyme for the citric acid cycle enzymes: pyruvate dehydrogenase and α-ketoglutarate dehydrogenase, which respectively catalyze the oxidative decarboxylation of pyruvate for the formation of acetyl coenzyme A (CoA) and α-ketoglutarate to succinyl-CoA. TPP also acts as a coenzyme for α-ketose transketolase of the pentose phosphate pathway in animals and the Calvin cycle in plants. This enzyme catalyzes the transformation reaction of xylulose-5-phosphate and erythrose-4-phosphate into fructose-6- phosphate and glyceraldehyde-3-phosphate and the transformation of ribose-5-phosphate and fructose-6-phosphate into sedoheptulose-7-phosphate and glyceraldehyde-3-

phosphate. Also, it is involved in branched-chain ketoacid dehydrogenase related to the metabolism of leucine, valine, and isoleucine (49). Due to its role in these pathways, thiamine is vital for cells in all organisms.

Bacteria, fungi, and plants produce the vitamin in its active form, while humans rely on diet for the supply of thiamine. When ingested, thiamine pyrophosphate is hydrolyzed by intestinal phosphatases. The alcohol is actively transported by transporters ThTr1 and ThTr2 and phosphorylated again in the cytoplasm by thiamine pyrophosphokinase (TPPK) or by other kinases not characterized yet (48). The recommended daily supplement of thiamine is 1.2 mg for men and 1.1 mg for women. The vitamin is found in diets rich in brown rice, meat, liver, legumes (beans, pods, peas), vegetables, potatoes and whole grains (50).

Thiamine deficiency is a public health problem in many developing countries where polished rice is the main component of the diet due to the thiamine content being lost during rice processing (51). An alternative is the enrichment of refined flour products. Other conditions directly associated with thiamine deficiency are severe malnutrition, eating errors, hypermetabolic states and alcoholism, causing beriberi, a disease that can affect both cardiovascular and nervous systems (52).
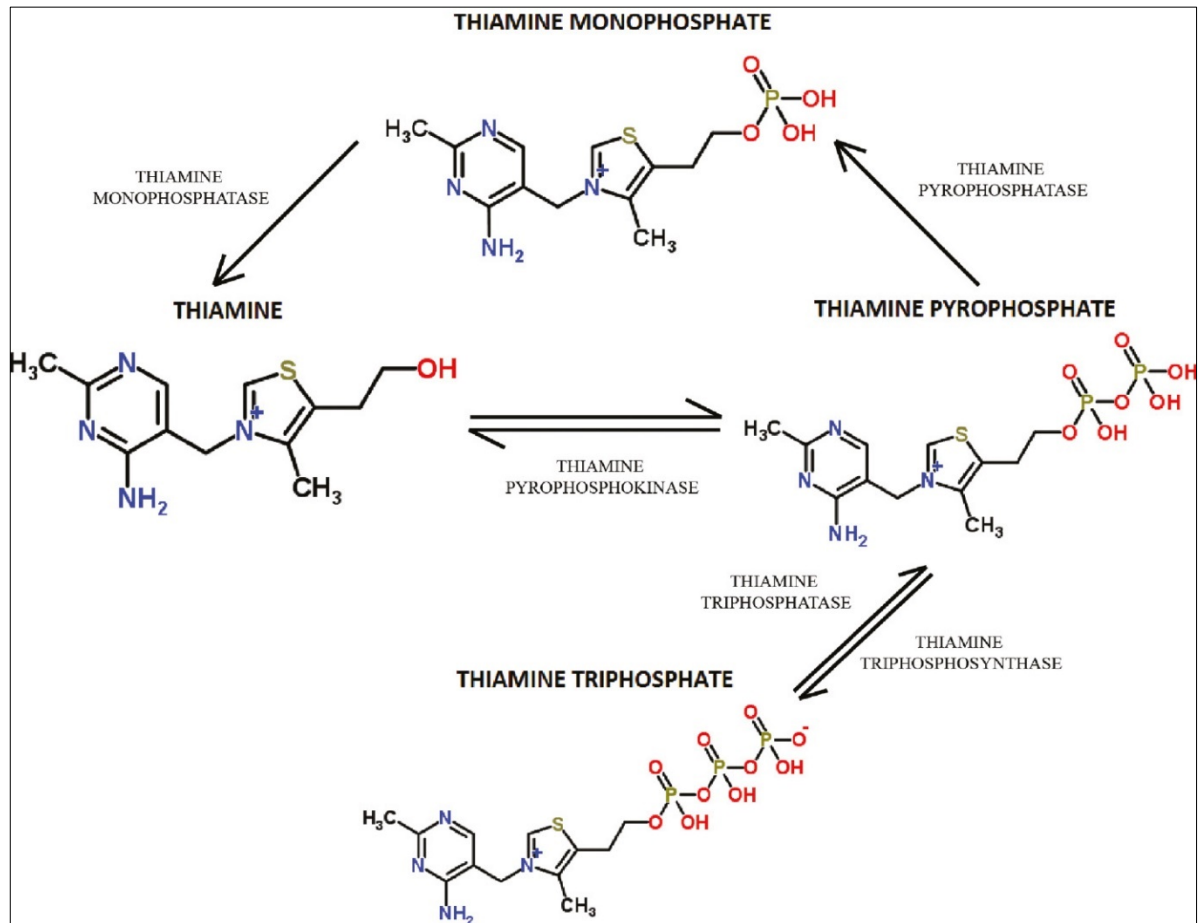
Figure 3: The chemical structure and currently understood biochemical pathway of thiamine, thiamine monophosphate, thiamine pyrophosphate and thiamine triphosphate. **Source:** Collie et al. 2017 (53).

1.3.1 *Thiamine and Cancer*

The determination of transketolase activity in erythrocytes is used to evaluate the nutritional status of vitamins in the body (49). In patients with cancer in advanced stages, this test detects vitamin deficiency, so it is common to prescribe a nutritional supplement. This metabolic situation occurs by the activation of the pentose phosphate pathway in tumor cells, generating a higher consumption of the molecule and reducing the absorption by the healthy cells. It was proposed that the activation of the pentose phosphate pathway would be the underlying reason why tumors can proliferate under hypoxic conditions (48). This fact has raised questions whether the advantages of the thiamine nutritional supplement are higher than the risks of tumor proliferation (54).

Ehrlich tumors implanted in rats showed that oxythyamine, a transketolase inhibitor, maintains tumor cells in the G1 state (48). Further research under the same conditions revealed that nutritional supplementation at doses 25 times the recommended daily value for rats stimulated tumor growth by 164%. Another relevant fact is associated with the fact that doses 2500 times the recommended daily amount for rats decreased tumor growth by 10% (54).

1.4 TPP riboswitch

Most of the studies on riboswitches have been carried out in prokaryotic organisms (55). In eukaryotes, to date, the only riboswitch described was the TPP riboswitch. In bacteria, such as *Escherichia coli*, two functionally different TPP riboswitches were reported. The first one is located in the 5' UTR region of the *thiM* gene, where it controls gene expression at the translation level (55,56). The second appears to occur both at the translation level and transcription of 5' UTR region of the *thiC* gene (55).

TPP riboswitches have been found in 5' UTRs regions of genes encoding thiamine biosynthetic enzymes in fungi (57,58), and algae, in which they promote alternative splicing of transcripts (59). Conversely, in all species of plants previously studied, the TPP riboswitch resides in the 3' UTRs region of the *THIC* gene. This difference in mRNA localization suggests a unique mode of action for plant riboswitches (60,61).

In fungi, TPP riboswitch is abundantly found in Ascomycota and Basidiomycota (62) such as *Aspergillus oryzae* (63) and *Neurospora crassa* (64). This class of

riboswitch is located within the 5' UTR region. In these organisms, when TPP levels are increased, metabolite binding to the riboswitch exposes an alternative splicing site while retaining part of its intron. This event changes the open reading frame and interrupts the biosynthesis of thiamine (65) (**Figure 4A**). The TPP riboswitch employs a similar mechanism in the transcription of the *THI4* and *THIC* genes from *algae* such as *Chlamydomonas reinhardtii* and *Volvox carteri* (66) (**Figure 4B**).

In plants, such as *Arabidopsis thaliana*, *Oryza sativa*, and *Poa secunda* (67,68), the 3' UTR region of the *THIC* gene is highly conserved and harbors a TPP riboswitch. In this type of mRNA, the start codon is followed by an intron, a small exon and a second intron that is tightly linked to the TPP riboswitch. This last intron may be kept or removed according to the intracellular TPP concentration. After binding to TPP, an alternative splice site is exposed, and the entire intron is removed along with its polyadenylation site, thus generating an unstable transcript with several polyadenylation sites (69) (**Figure 4C**).

Even controlling different genes involved in the synthesis or transport of thiamine and its phosphorylated derivatives in bacteria, archaea, fungi, and plants, the TPP aptamer has a conserved structure. The TPP aptamer consists of five stems. The P1 stem is responsible for linking the aptamer domain to the expression platform of the TPP riboswitch. Stems P2 and P3 are involved in binding of the TPP pyrimidine ring whereas stems P4 and P5 bind to the pyrophosphate group.

Rfam database (70) has a total of 9180 TPP riboswitch sequences (Rfam accession RF00059) from different organisms, and the consensus secondary structure of these sequences is highly conserved. P3 stem is the region with the highest variation among species. Bacteria and archaea commonly have a P3a stem (71), which is not observed in eukaryotic riboswitches. The eukaryotic P3 stem is significantly variable in length, sequence, and base pairings (72) (**Figure 5**).
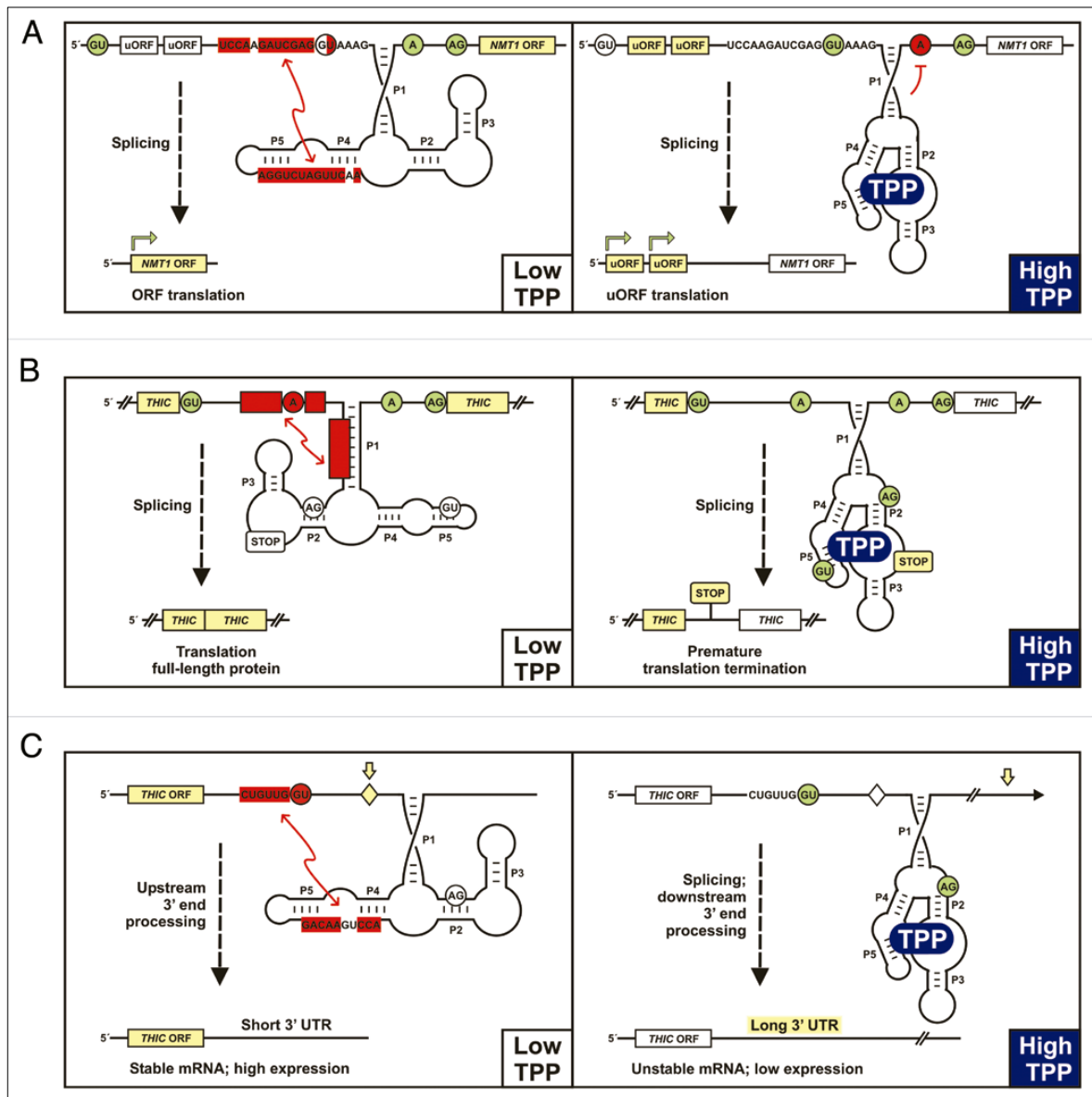
Figure 4: TPP riboswitch-mediated regulation of gene expression in filamentous fungi, green algae, and higher plants. **(a)** 5' region of *N. crassa* NMT1 pre-mRNA and prevalent alternative splicing products under conditions of low (left) and high (right) cellular TPP levels. The red boxes highlight base-pairing between P4/P5 stems of the aptamer and a complementary region adjacent to the proximal 5' splice site in the ligand-unbound state. Splice sites shown in green are available and utilized, whereas red color indicates diminished accessibility. TPP binding to the aptamer results in preferred usage of the proximal splice site and retention of uORF encoding sequences, which compete with the translation of the main ORF. Yellow boxes represent translated regions. **(b)** Alternative splicing of the TPP riboswitch-carrying intron positioned between coding region exons of the *THIC* pre-mRNA from *C. reinhardtii*. The red boxes indicate a proposed interaction of nucleobases in the P1 stem and the region surrounding the first branch site. Under low TPP conditions (left), splicing from the distal sites generates a mRNA encoding full-length protein. TPP binding to the aptamer (right) releases the upstream branch site resulting in cassette exon retention by use of additional 3' and 5' splice sites. The retained intron fragment introduces a premature termination codon ("STOP") preventing expression of the functional THIC protein. **(c)** TPP riboswitch-mediated regulation of splicing and alternative 3' end processing of *THIC* pre-mRNAs from higher plants. The TPP riboswitch is positioned within the 3' UTR of *THIC* pre-mRNAs and controls usage of a 5' splice site by exploiting base-pairing potential between the aptamer and the region adjacent to the respective splice site. Complementary regions are highlighted in red and depicted nucleotides represent the sequence example from *A. thaliana*. The diamond symbol upstream of the aptamer represents a 3' end processing site, which is used at low TPP levels and yields *THIC* mRNAs with short 3' UTR and high protein expression (left). TPP binding to the aptamer releases the 5' splice site, which results in splicing in the 3' UTR and removal of the major 3' end processing site (right). Subsequently, 3' end processing occurs at downstream sites generating transcripts with long 3' UTRs, which are less stable and yield lower amounts of THIC protein. **Source:** Wachter, 2010 (73).
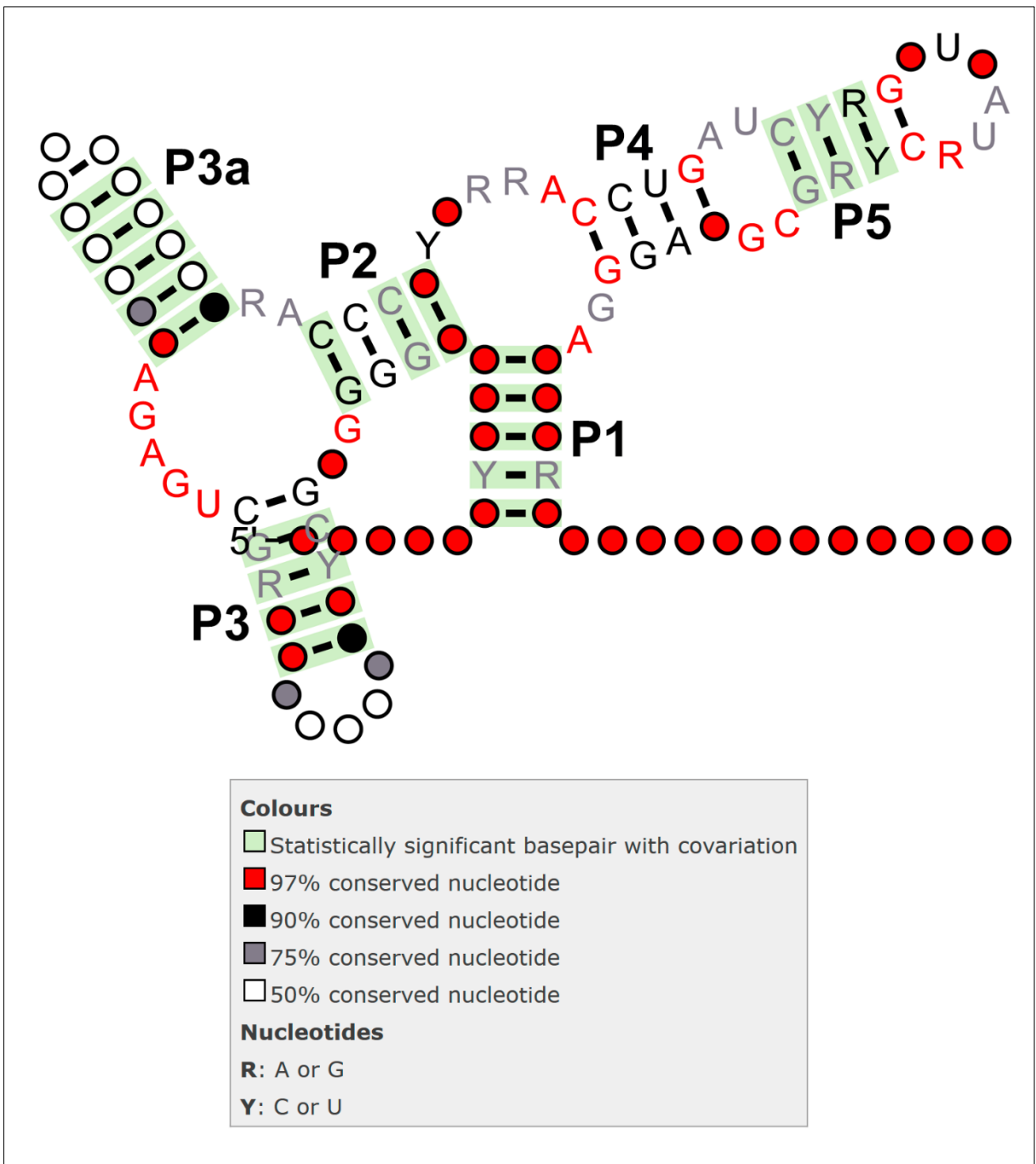
Figure 5: Consensus secondary structure of TPP riboswitch aptamer. Seed alignment and consensus secondary structure were provided by Rfam (http://rfam.xfam.org/family/RF00059). Colours and markup of the structure diagrams were identified according to the caption in the figure.

1.4.1 TPP riboswitches as targets for antibiotics

Riboswitches can be found in many genes, and participate in vital metabolic pathways of various organisms, including bacteria and pathogenic fungi. They also have characteristics that suggest they can be possible targets of useful antibacterial drugs. Due to the structural conservation of the aptamer riboswitch, it can be speculated that a similar drug to the metabolite would be able to bind to the riboswitch inactivating the expression of the vital gene.

Before the discovery of riboswitches, the researchers had synthesized numerous thiamine analogues to enable the identification and characterization of thiamin biosynthesis enzymes. An analogue, pyrithiamine, has been identified as toxic to bacteria and fungi, but the precise mechanism of pyrithiamine toxicity has remained a mystery for decades. More recently, the discovery of riboswitches led to the hypothesis that pyrithiamine could inhibit bacterial growth, targeting the TPP riboswitches and thereby repressing thiamine biosynthesis (for a review (74)).

After uptake by the cell, pyrithiamine is phosphorylated in pyrithiamine pyrophosphate (PTPP) (**Figure 6**), and this compound has been shown to bind (*in vitro*) to the TPP riboswitch with similar affinity to TPP and to repress the fused reporter gene with TPP riboswitch in bacteria. Besides, strains of *Bacillus subtilis*, *Escherichia coli* and *Aspergillus oryzae*, that were cultured to resist PTPP, showed a mutation in the conserved riboswitch region. Such evidence suggests the possibility that PTPP is capable of inhibiting bacteria and fungi growth (75).

New thiamine analogs have been investigated for inhibition of TPP riboswitches in bacteria. An example of this is the triazolethiamine, which has been shown to have superior activity compared to pyrithiamine. In contrast, the activation of triazolethiamine is dependent on proteins involved in the metabolic pathways of uptake and thiamine synthesis (76). These findings represent a promising starting point for the development of novel antibacterial compounds targeting TPP riboswitches.

Although discovered compounds that displayed inhibitory capacity against bacterial growth in vitro, there is still much to discover about these analogs and their interactions with different riboswitches. It is necessary to verify whether these compounds can cure infections and investigate potential toxicity in humans. This latter issue is of great importance because, to date, none of riboswitch has been described in the human genome. However, one cannot rule out the possible undesired interaction between the drug and host enzymes, such as the inhibition of thiamine-phosphate

pyrophosphorylase by PTPP in rats (75). For these reasons, it is crucial that researchers apply most up-to-date available technologies to design new efficient and safe drugs.
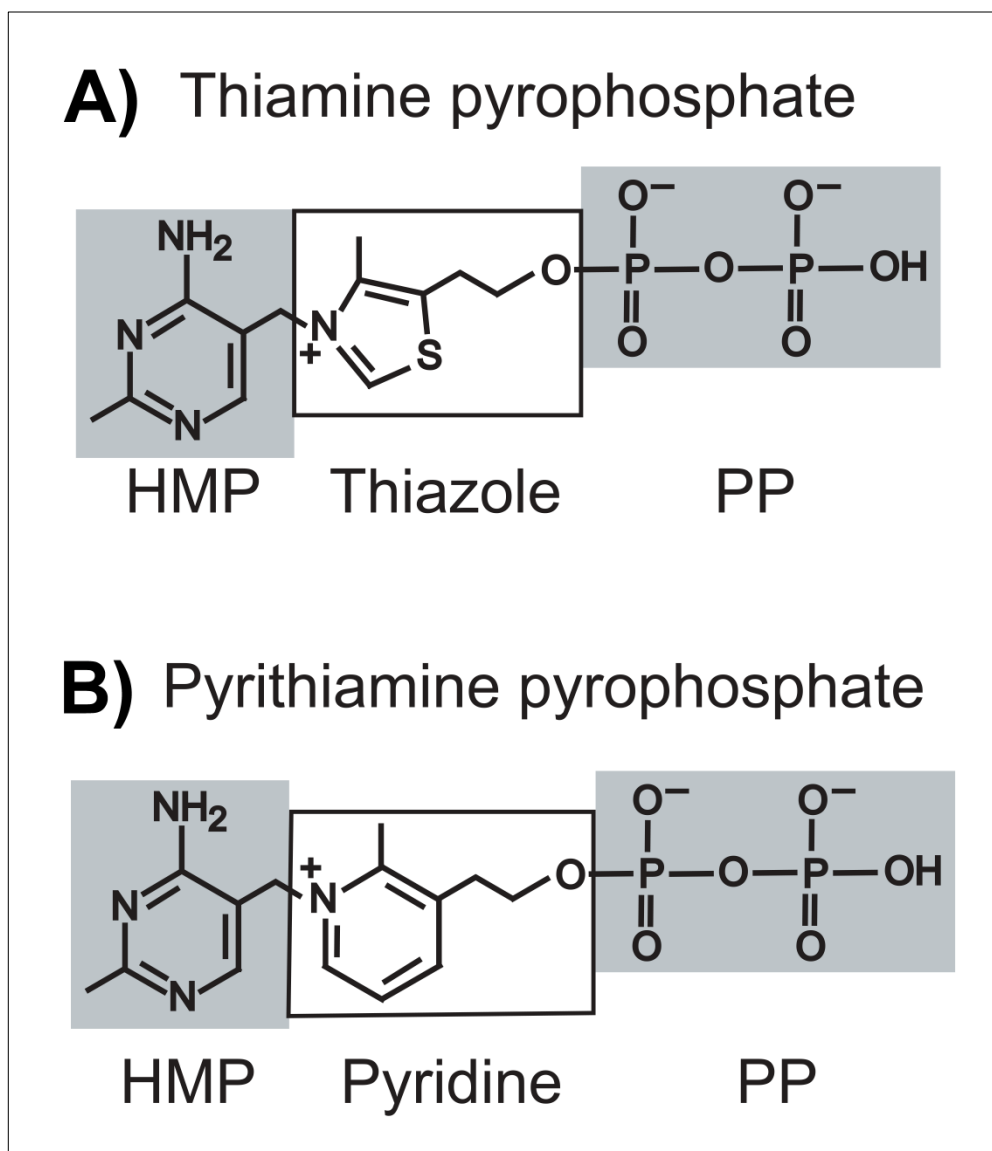
Figure 6: Chemical structures of the natural metabolite TPP **(a)** and the antimicrobial compound PTPP **(b)**. Rectangle filled in light gray highlights the 4-amino-5-hydroxymethyl-2-methylpyrimidine (HMP) ring and pyrophosphate group (PP). Adapted from Serganov et al. 2006 (77).

1.5   Human RNA switch

Despite the existence of riboswitches in some eukaryotes, these RNA structures were not identified in the animal kingdom. However, in humans, a structure functionally analogous to riboswitches has been determined, the RNA switch structure in the 3' UTR region of *VEGFA* gene. The RNA switch undergoes conformational changes in response to environmental signals to regulate gene expression. However, the detection mechanisms of riboswitches and RNA switch are different. While riboswitches respond to stimuli by direct binding to effector molecule, the RNA switch *VEGFA* gene is an independent metabolite and its conformational change is controlled by different binding proteins (78).

VEGF-A acts as both a growth factor and vascular endothelial cell survival factor contributing to development of new blood vessels (angiogenesis) and playing an essential role in physiological and pathological antigenic processes throughout embryonic development and during adult life (79,80).

In quiescent vasculature in adult organs, basal levels of VEGF-A are responsible for the protection of endothelial cells from apoptosis. However, the level increases in certain physiological situations, such as repair of wounds, adaptation to hypoxia, arthritis, psoriasis, proliferative retinopathies, among others (81,82). Besides, VEGF-A acts as a critical mediator of tumor angiogenesis by stimulating the growth of new blood vessels from neighboring capillaries, allowing tumor cells to acquire oxygen and nutrients for metastasis (83).

VEGF-A has critical importance during embryogenesis and in adults, resulting in strict control of its expression. In fact, this gene has two promoters, two polyadenylation signals and 14 alternative splicing events that allow the expression of up to 56 potential mRNAs. At the post-transcriptional level, the regulation includes alternative polyadenylation, regulation of mRNA stability, IRES-dependent translation, ORF and isoform expression, miRNA mediated regulation, G-quadruplex structure and RNA switch (**Figure 7A**) (for a review (84)).

15
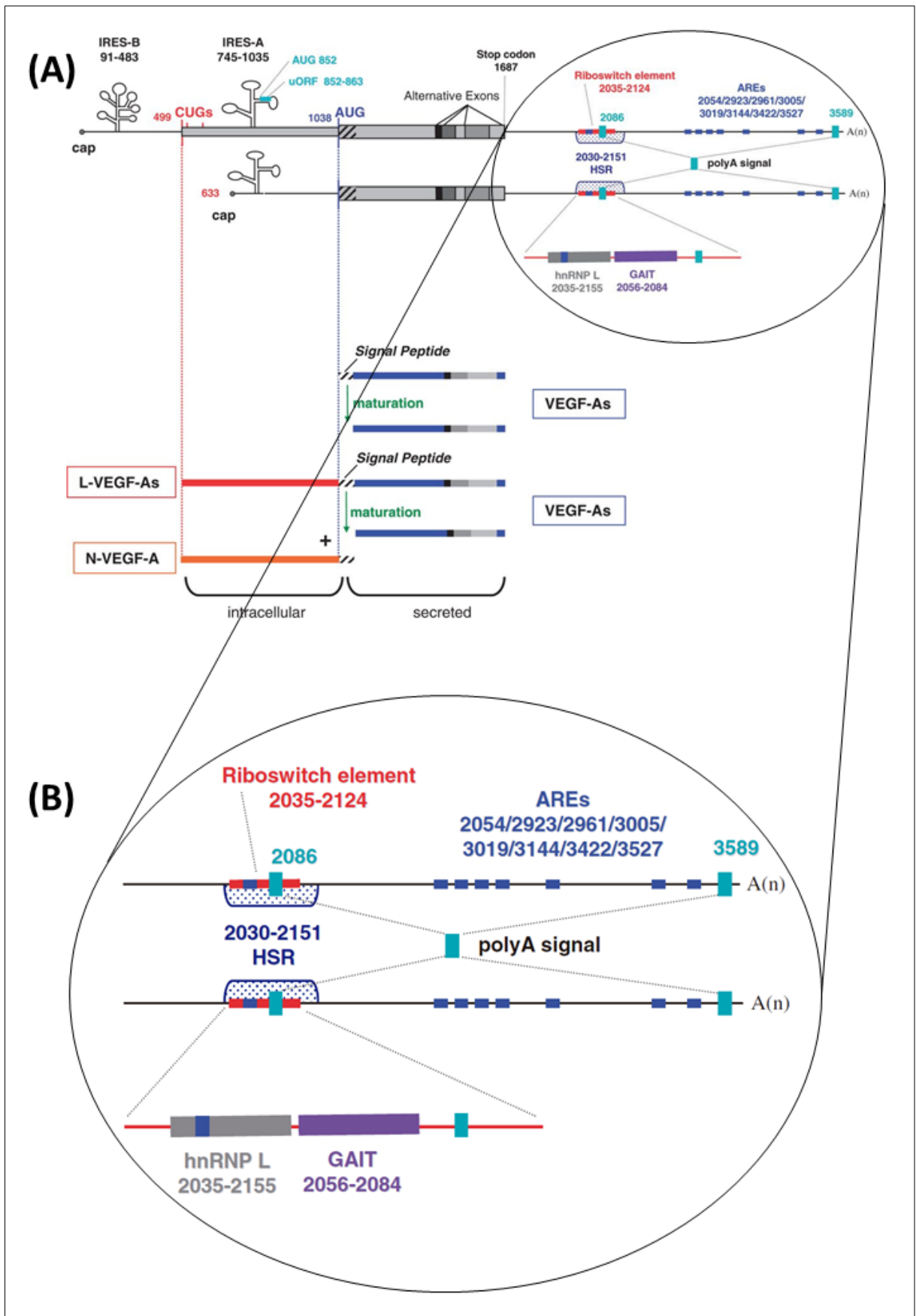
Figure 7: (A) Human *VEGFA* gene and its main regulatory elements. **(B)** Featured on the location of the RNA switch. Adapted from Arcondéguy (2013) (84).

In the RNA switch, a gamma-interferon-activated inhibitor of translation (GAIT) and heterogeneous nuclear ribonucleoprotein complex (hnRNPL) alter the production of VEGF-A during oxidative stress through translation repression or activation, respectively. The proximity between GAIT and hnRNPL binding sites (**Figure 7B**) suggests that their interactions with the RNA switch are mutually exclusive (78).

Hypoxia induces the nuclear hnRNPL translocation to the cytoplasm, which significantly increases the binding of the hnRNPL to the hypoxia stability region (HSR). This binding is required for stabilization of VEGF-A mRNA, preventing the binding of GAIT, contributing to the enhancement of VEGF-A expression, which reverses hypoxia by inducing blood vessel formation (85,86) (**Figure 8**).

Unlike riboswitches, the VEGF 3'UTR binds to two different protein elements to control gene expression. Nevertheless, the discovery of an RNA switch in human cells highlights the possibility of similar mechanisms playing essential roles in translation and transcription regulation in animal cells. Therefore, large-scale prediction of RNA motifs can serve as a tool to uncover these mechanisms and enhance our current knowledge of riboswitches.
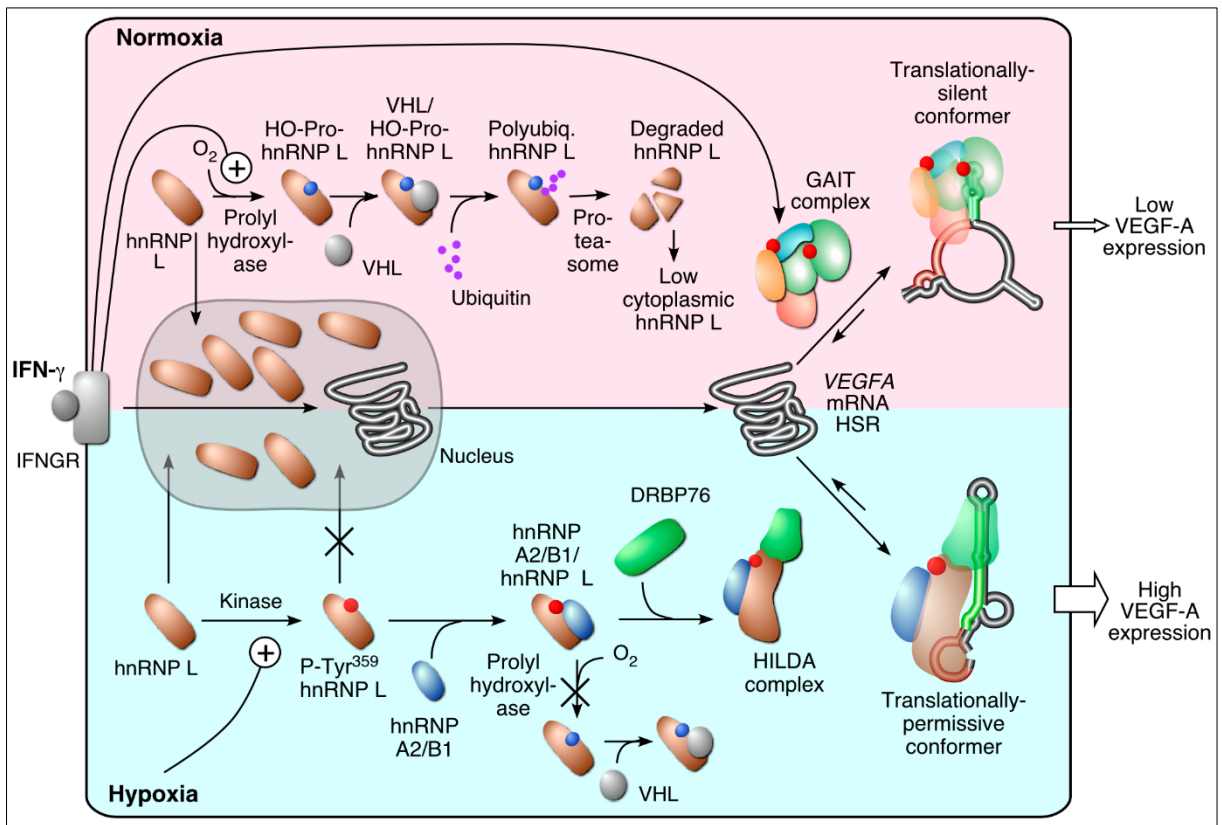
Figure 8: Regulation of hnRNP L expression by IFN- γ and hypoxia and the role of the HILDA complex in the VEGFA RNA switch (86).

1.6   Computational tools for riboswitch aptamer prediction based on RNA motif

There are several methods for predicting RNA motifs, such as using an algorithm for predicting the secondary structure and then comparing to the conserved stem-loops (like RiboSW (87)), searching for riboswitch particular sequence motifs followed by the comparison of the secondary structures (riboswitch Finder (88), RibEx (89), and DRD (90)) and the usage of probabilistic models such as Hidden Markov Model (HMM) and Covariance Model (CM) (HMMER (91), Infernal (92)). A review describing the current computational methods to detect riboswitch aptamer candidates was published by our research group recently (29).

Riboswitches control the expression of genes involved in the biosynthesis and transport of ligands, and also work as transcription factors (30). So, it is crucial to develop tools for the accurate identification of different riboswitch classes, given their significant regulatory role in bacteria and a few eukaryotic organisms. Several approaches have been used for the computational identification of riboswitch aptamers (**Table 1** and **Figure 9**). The current riboswitch search tools employ hidden Markov model algorithm, covariance model, and machine learning methods, which often use riboswitch aptamers, identified from seed alignments performed with sequences retrieved from the Rfam database.

Most of the tools mentioned here are web-based. These instruments often impose restraints on the input sequence length and number of riboswitches that can be detected at once. They also rely on sequence or structural conservation of the aptamer to perform the analysis. Therefore, the aptamer prediction affects the detection of more variable riboswitches, such as the TPP and the Cobalamin, or smaller ones, such as the guanine riboswitch.

Several computational methods have been created to identify novel riboswitches and to characterize those that are already known. Amongst the methods that use primary sequences, the HMMER and Infernal tools stand out due to their ability to run locally, with the advantage of not having upload limits. Both methods utilize similar approaches by applying probabilistic models to the sequence datasets to infer patterns.

DRD group (90) compared their server with RiboSW. The advantage of DRD compared to the other server is the ability to scan genome-scale files for riboswitches. In analyses of overall sequences obtained higher sensitivity (0.95) than RiboSW (0.85).

DRD server was able to detect 64 instances that RiboSW was not identified, and 12 instances in which the opposite was true.

Singh and collaborators compared the performance of HMM to other two CM web-based tools (Riboswitch finder and RibEx) in the search for ten riboswitches families on Rfam or RefSeq databases (93). Their results showed that HMM models run faster than CM and were more accurate than Riboswitch Finder and RibEx. The recently released version 1.1 of Infernal (92) was reported to be 100 times faster than earlier versions and has been used for the identification of functional RNA homologs in metagenomic data (94).

In a review on the computational prediction of riboswitches (95), Infernal was considered as the most valuable tool to predict riboswitch aptamers mainly because the relevant Rfam database relies on Infernal for maintenance and extension. Some studies have used Infernal to identify riboswitches and other ncRNAs in archaeal metagenomes (94,96), species in the phyla *Actinobacteria* (97) and Proteobacteria (98), *Methanobrevibacter ruminantium* (99), *Neisseria gonorrheae* (100), and *Brassica rapa* (101).

Table 1: Feature comparison among different software programs used for riboswitch identification. **Source:** Antunes et al. 2018 (29).

Feature	HMM	Infernal	Riboswitch Finder	RibEx	RiboSW	RNAConSLOpt	DRD	Riboswitch Scanner
Considers structural conformations	no	yes	yes	no	yes	yes	Yes	Yes
Considers conserved functional sequences	yes	no	yes	yes	yes	yes	Yes	No
Software package	yes	yes	no	no	yes	yes	No	No
Max input length	none	2 kb	3 Mb	40 kb	10 kb	none	None	None
# riboswitches	any	any	13	17+	12	any	13	24
New user definition	yes	yes	no	no	yes	yes	Yes	Yes
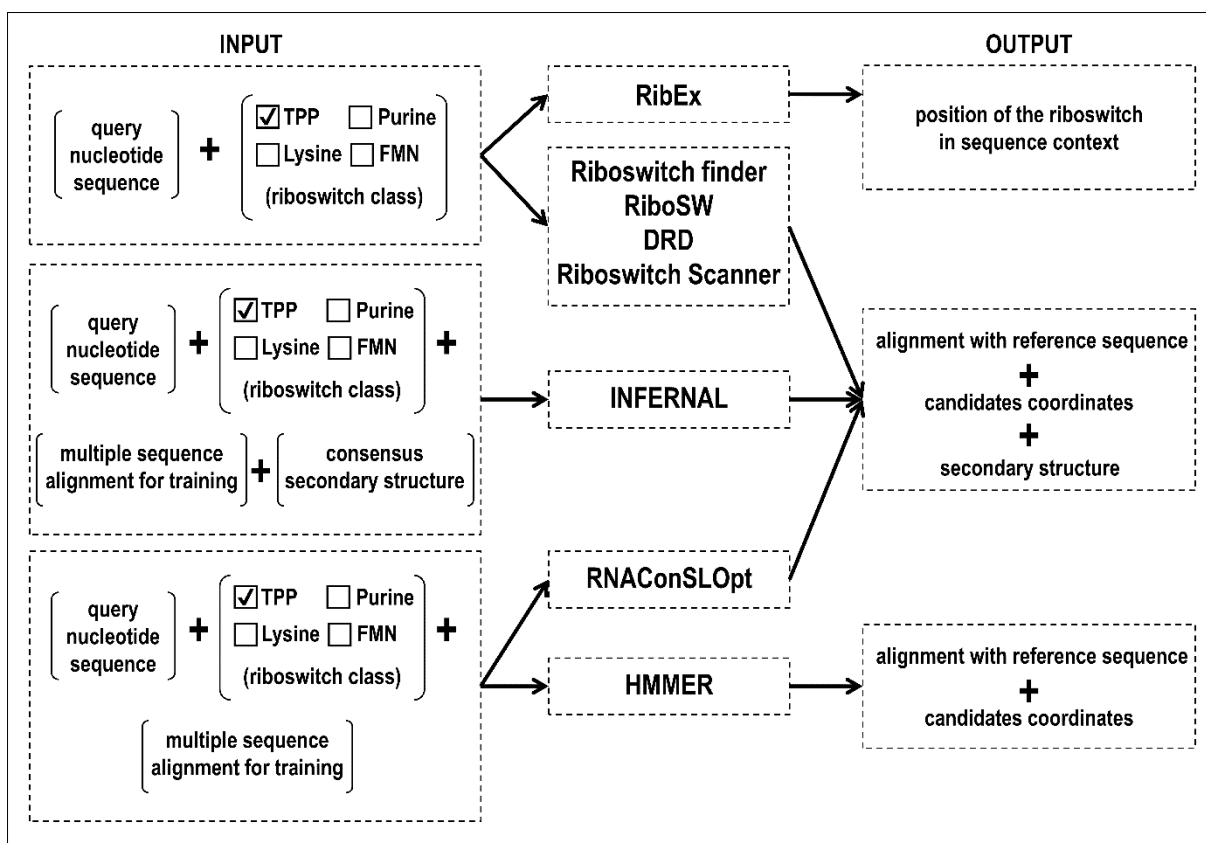
Figure 9: Input and output files of RNA motif prediction tools. **Source:** Antunes et al. 2018 (29).

1.7 Computational tools for riboswitch aptamer prediction and candidate evaluation based on RNA structure models

Likewise to what happens in proteins, RNA functions depend on its structure and dynamics, which are determined by its nucleotide sequence. The number of computational methods and algorithms to predict the 3D structure of proteins from its amino acid sequence is vast. Unfortunately, only a few are available for the prediction of RNA structure (102).

So far, knowledge-based approaches are the most suitable methods to determine a 3D configuration with the best possible accuracy. Comparative (or homology-) modeling, for instance, which is based on sequence similarity, works correctly when there is an experimentally elucidated structure to be used as a template (103). However, RNA templates are rarely available. In the RCSB PDB (104), only approximately 0.9% (1322 structures) of all deposited structures correspond to RNA structures (accessed July 2018).

Physics-based approaches are successful for the prediction of relatively small molecules. These tools are comparatively more appropriate for building models of RNA molecules with less than ~40 nt and display reasonable reliability for molecules up to ~80 nt. The prediction of larger molecules is possible, but the model reliability decreases as the length of the sequence increases (105).

The combination of knowledge- and physics-based approaches resulted in the development of the so-called *de novo* folding methods, which is the assembly of the target structure from small fragments derived from other known structures (106). Here, we compiled some programs using different approaches to predict RNA 3D modeling.

Riboswitches undergo conformational changes upon ligand binding and act as a switch at the transcriptional or translational levels. To understand the molecular mechanisms associated with their regulatory functions, regarding that riboswitches are functional entities that can undergo conformational changes, knowing their structures is of essential importance. Hence, predicting the structure of riboswitches can provide useful insight into the mechanism through which small molecules bind to RNAs, as well as shed light on how this process induces conformational changes in riboswitches.

The application of energy minimization methods for secondary structure prediction of the riboswitch expression platform domain is still limited as it involves

22

conformational changes. However, the prediction of this domain may be useful to support experimental assays. Barash and Gabdank (107) predicted a single point mutation positioned in the non-conserved TPP riboswitch region responsible for transforming the terminator to an anti-terminator state.

The recent developments in the secondary structure prediction allow to include probing data, like SHAPE and DMS, for restriction and prediction of a structure with high accuracy (reviewed in (108)). Among the programs listed by us, the RNAstructure includes the option of incorporating the probing data as restraints.

3D prediction of a single RNA sequence is still limited, especially when extended RNA sequences (reviewed in (109)) are involved. Comparative approaches using homologous sequence information increase the accuracy of a secondary structure prediction. In many circumstances, homologous RNA sequences of the target RNA sequence could be obtained, and it would be crucial to know the common secondary structure to those sequences (110).

The common secondary structure is a fundamental element in riboswitch aptamers prediction. Programs such as Infernal, Riboswitch Finder, RiboSW, DRD and Riboswitch Scanner use 2D structural conformations for homologous searching. Secondary structure information is also crucial for tertiary structure prediction. Template-based methods assist in modeling mutations or structural changes, whereas in *de novo* methods, base pair constraints when creating 3D models are allowed. For instance, the MC-sym tool was used to construct models of the SAM-I riboswitch RNA segment by incorporating elements of the expression platform allowing the formation of an antiterminator (AT) helix in the 3D structures (111).

RNAComposer uses 2D restraints to create models and has provided positive results regarding the structural prediction of riboswitches. The server has been tested using a set several riboswitches containing pseudoknots and extensive tertiary interaction (112). In this set, nine examples were characterized with high accuracy and acceptable recovery of canonical and non-canonical base pairing and stacking. Input and output files of tertiary structure tools are shown in **Figure 10**.

Prompted by the increasing number of 3D RNA prediction framework methods, the RNA-Puzzles started in 2012 (113). RNA-Puzzles is a CASP-like (114) event in which collective blind experiments for the evaluation of 3D RNA structure prediction are carried out (113,115,116). In the three rounds of RNA-Puzzles, predictions based on homology models already attained a high-level precision, providing useful insight into the understanding of the RNA structure (117). Moreover, ligand binding prediction

and subsequent conformational changes can also be described but cannot be reliably guaranteed.

Gong and collaborators (118) showed other approaches that aid to the investigation of the folding kinetics of aptamers and co-transcriptional folding kinetics using coarse-grained SOP model and, BarMap and helix-based computational approach, respectively. A new method StreAM-Tg (119) also allows analyzing structural transitions. This method gains insights into RNA dynamics based on a coarse-grained representation of RNA MD simulations.

Current modeling methods for template-based predictions have consistently reached a high accuracy level, *i.e.*, now it is possible to model nearly all the structural details, provided that a reliable homologous structure is identified. Also, the ligand binding sites were readily inferred via homology (117). Different classes of riboswitches can be found in the RCSB PDB (**Table 2**), facilitating the use of model-based approaches such as ModeRNA and MMB.

In the case of targets lacking sequence identities with previously experimentally resolved structures, modeling quality strongly depends on the size of the target. The third edition of RNA-Puzzles provided models for two small RNAs – the ZMP riboswitch (60-nt) and L-glutamine riboswitch (61-nt) – only approximately 6Å away of RMSD from the crystallographic structure. Although the tools are less accurate, they can correctly predict the overall global folding. Thus, the larger the targets without a template - ydaO riboswitch (108-nt) -, the less accurate the predictions will be (10Å best-case RMSDs).
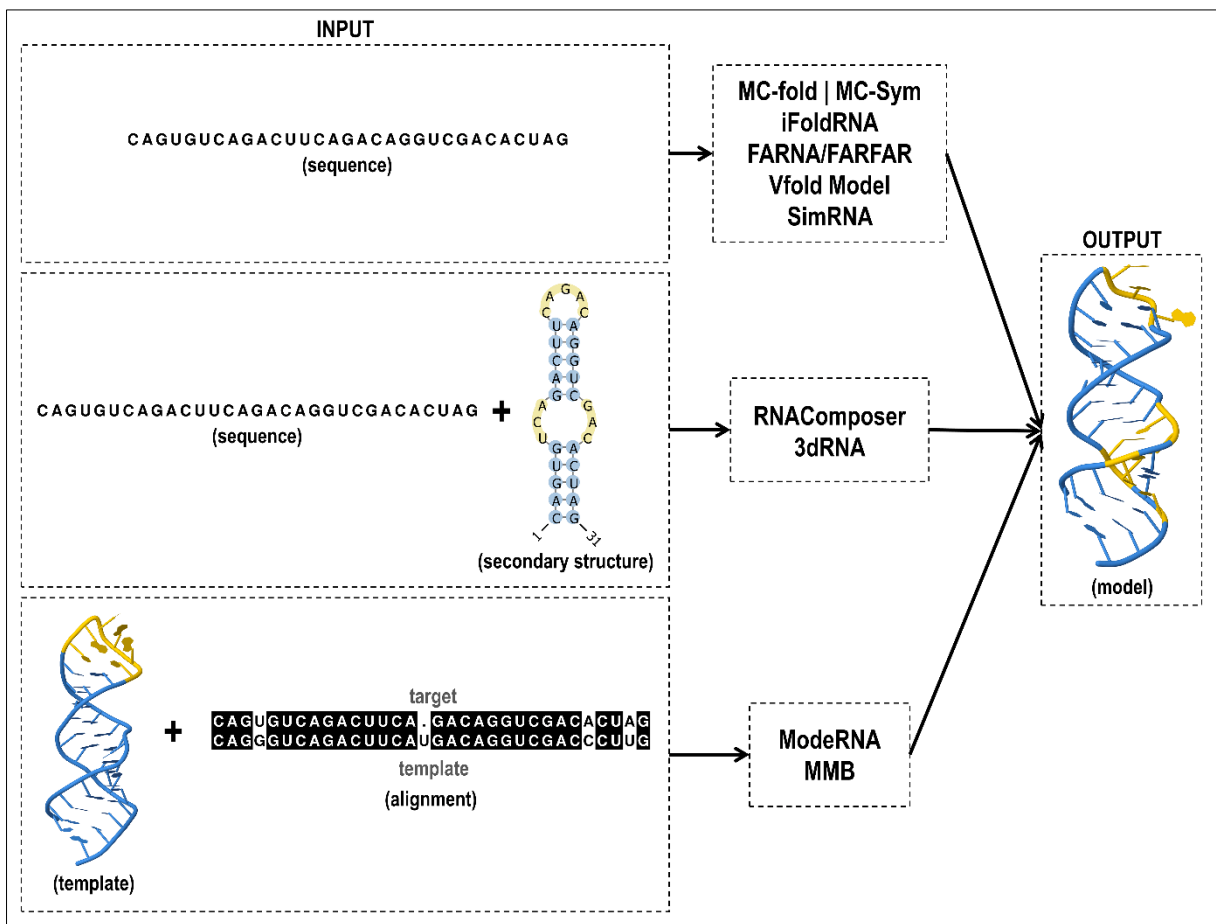
Figure 10: Input and output files of RNA tertiary structure prediction tools. **Source:** Antunes et al. 2018 (29).

Table 2: Classes of riboswitches presenting experimentally resolved 3D structures. **Source:** Antunes et al. 2018 (29).

Class	# structures	PDB ID	Rfam accession
AdoCbl-variant	4	4frg:b; 4frg:x; 4frn:a; 4frn:b	RF01689
c-di-GMP-I	17	3irw:r; 3iwn:a; 3iwn:b; 3mum:r; 3mur:r; 3mut:r; 3muv:r; 3mxh:r; 3ucu:r; 3ucz:r; 3ud3:r; 3ud4:r; 4yaz:r; 4yaz:a 4yb0:r; 4yb0:a; 4yb1:r	RF01051
c-di-GMP-II	2	3q3z:v; 3q3z:a	RF01786
Cobalamin	2	4gma:z; 4gxy:a	RF00174
FMN	6	3f2q:x; 3f2t:x; 3f2w:x; 3f2x:x; 3f2y:x; 3f30:x	RF00050
glmS	39	2gcs:b; 2gcv:b; 2h0s:b; 2h0w:b; 2h0x:b; 2h0z:b; 2ho6:b; 2ho7:b; 2nz4:p; 2nz4:q; 2nz4:r; 2nz4:s; 2z74:b; 2z75:b; 3b4a:b; 3b4b:b; 3b4c:b; 3g8s:p; 3g8s:q; 3g8s:r; 3g8s:s; 3g8t:p; 3g8t:q; 3g8t:r; 3g8t:s; 3g96:p; 3g96:q; 3g96:r; 3g96:s; 3g9c:p; 3g9c:q; 3g9c:r; 3g9c:s; 3l3c:p; 3l3c:q; 3l3c:r; 3l3c:s; 4meg:b; 4meh:b	RF00234
Glycine	19	3owi:a; 3owi:b; 3oww:a; 3oww:b; 3owz:a; 3owz:b; 3ox0:a; 3ox0:b; 3oxb:a; 3oxb:b; 3oxd:a; 3oxd:b; 3oxe:a; 3oxe:b; 3oxj:a; 3oxj:b; 3oxm:a; 3oxm:b; 3p49:a	RF00504
Lysine	16	3d0u:a; 3d0x:a; 3dig:x; 3dil:a; 3dim:a; 3dio:x; 3diq:a; 3dir:a; 3dis:a; 3dix:a; 3diy:a; 3diz:a; 3dj0:a; 3dj2:a; 4erj:a; 4erl:a	RF00168
MFR	16	3ski:a; 3ski:b; 3skl:a; 3skl:b; 3skr:a; 3skr:b; 3skt:a; 3skt:b; 3skw:a; 3skw:b; 3skz:a; 3skz:b; 3slm:a; 3slm:b; 3slq:a; 3slq:b	RF01510
preQ1-II	1	2miy:a	RF01054
PreQ1-III	1	4rzd:a	RF02680
Purine	40	1y26:x; 1y27:x; 2b57:a; 2ees:a; 2eet:a; 2eeu:a; 2eev:a; 2eew:a; 2g9c:a; 2xnw:a; 2xnz:a; 2xo0:a; 2xo1:a; 3ds7:a; 3ds7:b; 3fo4:a; 3fo6:a; 3g4m:a; 3gao:a; 3ger:a; 3ges:a; 3gog:a; 3got:a; 3la5:a; 3rkf:a; 3rkf:b; 3rkf:c; 3rkf:d; 4fe5:b; 4fej:b; 4fel:b; 4fen:b; 4feo:b; 4fep:b; 4lx5:a; 4lx6:a; 4tzx:x; 4tzy:x; 4xnr:x; 5c7u:b	RF00167
SAM	28	2gis:a; 2ydh:a; 2ygh:a; 3gx2:a; 3gx3:a; 3gx5:a; 3gx6:a; 3gx7:a; 3iqn:a; 3iqp:a; 3iqr:a; 3v7e:c; 3v7e:d; 4aob:a; 4b5r:a; 4kqy:a; 5fjc:a; 5fk1:a; 5fk2:a; 5fk3:a; 5fk4:a; 5fk5:a; 5fk6:a; 5fkd:a; 5fke:a; 5fkf:a; 5fkg:a; 5fkh:a	RF00162
SAM-I-IV-variant	2	4l81:a; 4oqu:a	RF01725
THF	10	3sd3:a; 3suh:x; 3sux:x; 3suy:x; 4lvv:a; 4lvw:a; 4lvx:a; 4lvy:a; 4lvz:a; 4lw0:a	RF01831
TPP	24	2cky:a; 2cky:b; 2gdi:x; 2gdi:y; 2hoj:a; 2hok:a; 2hol:a; 2hom:a; 2hoo:a; 2hop:a; 3d2g:a; 3d2g:b; 3d2v:a; 3d2v:b; 3d2x:a; 3d2x:b; 3k0j:e; 3k0j:f; 4nya:a; 4nya:b; 4nyb:a; 4nyc:a; 4nyd:a; 4nyg:a	RF00059
ydaO-yuaA	4	4qlm:a; 4qln:a; 4w90:c; 4w92:c	RF00379
ykoK	3	2qbz:x; 3pdr:x; 3pdr:a	RF00380

1.8 Molecular Dynamics

There is a complex network of chemical entities that can dynamically create life at the molecular level. For example, proteins and nucleic acids adopt a specific structure according to their function; ions are transported through the membranes, enzymes can trigger a cascade of chemical reactions, etc. Due to the complexity of biological systems, computational methods have increasingly become essential for the study of life sciences. The fastest, most potent and sophisticated biological systems can be exploited through computer simulations (120).

Molecular dynamics (MD) emerged as one of the first simulation methods applied to fluid dynamics by Alder and Wainwright (121) and by Rahman (122) in the late 1950s and early 1960s. Due to advances in technology and algorithm improvements, MD is a valuable tool in many areas of physics and chemistry. Since the 1970s MD has been widely used to study the structure and dynamics of macromolecules.

Experiments play a central role in science. The richness of experimental results provides a basis for understanding the chemical mechanism of life. Modern technieques, such as X-ray diffraction or NMR, allow the determination of the structure and the elucidation of the function of large molecules of biological interest. However, the experiment is only possible in conjunction with underlying models and theories. Computational simulations have infuenced the relationship between experimental and theoretical essays. The essence of a simulation is the use of the computers to model a physical system (123).

According to the IUPAC (International Union of Pure and Applied Chemistry), "molecular dynamics is a simulation procedure consisting of the computation of the motion of atoms in a molecule or individual atoms or molecules in solids, liquids and gases, according to Newton's laws of motion. Forces acting on the atoms, required to simulate their motions, are generally calculated using molecular mechanics force fields" (124). Briefly, the name of the force field is assigned to the description of a system of many particles by the overlapping of simple mathematical terms. The numerical solution of the equations of motion from initial coordinates and velocities for the motion of each atom subjected to the force field is obtained through a computational algorithm. They are all based on the Newton's equations resolution,

through a power series expansion of the position of each particle for each increment in time (125).

Each force field establishes a set of mathematical equations dedicated to reproducing aspects of molecular behavior, such as the stretching of chemical bonds, the deformation of a bonding angle or twisting of a dihedral. These equations, in turn, are parameterized to reproduce the behavior of the compounds of interest (126). The classical force fields are composed of two classes of potential functions (**Figure 11**) (127):

(i) Bonded interactions are represented by covalent bonds (stretching potential), angular potentials, proper and improper dihedrals.

(ii) Non-bonded interactions consist of Lennard-Jones's potential and Coulomb's interaction and are computed based on a list of unbound neighboring atoms within a particular cut-off radius.

For each system, it is recommended to analyze which force field is the most appropriate for its use because each one contains a set of specific empirical potentials for different kinds of molecules involved (DNA, proteins, carbohydrates, etc.) (128).

Molecular Dynamics simulation opens the possibility for a more detailed study of macromolecule systems at the microscopic scale and times of the order of nanoseconds or less. With this methodology it is possible to monitor the time evolution of the systems as well as the determination of dynamic, structural and thermodynamic properties.

1.8.1 RNA Structural Molecular Dynamics

The use of atomic models and the accomplishing of RNA simulations are still considered a challenge. The main difficulty are related to the molecular size and the time associated with structure formation. In the process of RNA folding, stem-loops may be formed in the microsecond range, while global architectures may need seconds to minutes to fold. Regarding size, even at the secondary structure level, molecules with more than a dozen nucleotides have a multitude of possible states allowing a huge room for base pairing, giving rise to different intermediate structures separated by energy barriers. Also, there are problems related to the high charge of the RNA backbone, leading to fundamental interactions with solvent and ions, and the intrinsic nature of hydrogen bonding in base pairing, which would require quantum mechanics calculations to obtain more accurate results (129).

About nucleic acid simulations, current classical force fields have been developed paying more attention to the parameterization of nucleotides. Hanke and Gohlke (130) have shown that the AMBER 99 force field, with parmbsc0 and parmχOL modifications, can represent the dynamic behavior of the apo structure of the guanine riboswitch aptamer domain.

In a recent review of the molecular simulations of RNA structure (131), it was provided with an exhaustive overview of simulations performed on three available groups of riboswitches: *(i)* purine riboswitches, including adenine and guanine-sensing riboswitches; *(ii)* preQ1 riboswitches that sense 7-aminomethyl-7-deazaguanine (preQ1); and *(iii)* SAM riboswitches sensing SAM.

$$E_{total} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$
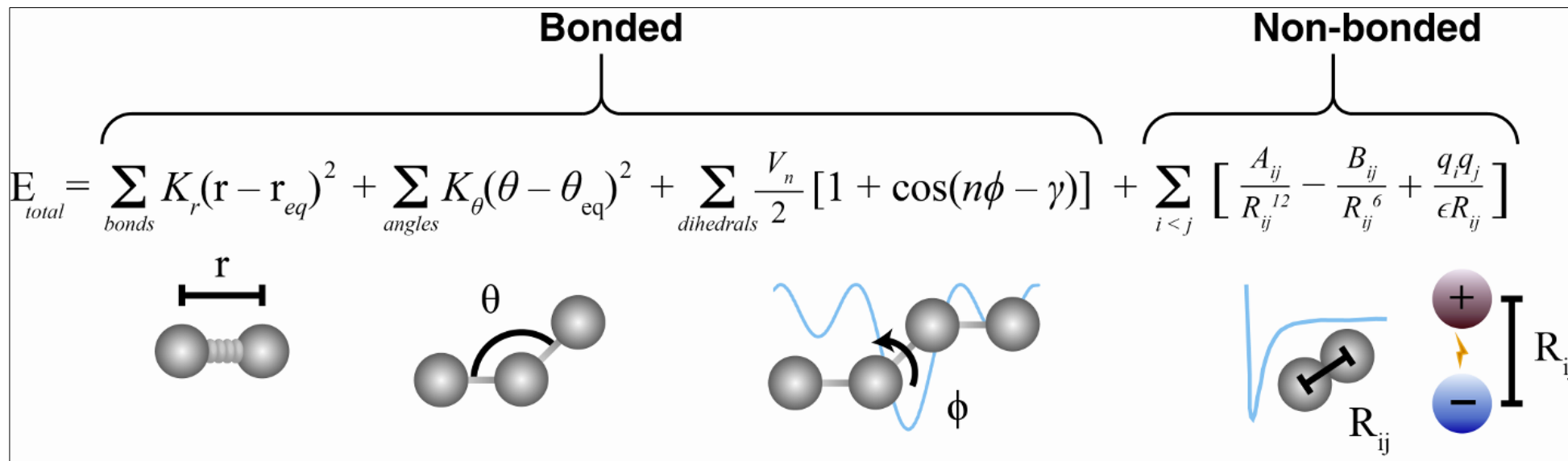
Figure 11: An example of an equation used to approximate the atomic forces that govern molecular movement. The atomic forces that govern molecular movement can be divided into those caused by interactions between atoms that are chemically bonded to one another and those caused by interactions between atoms that are not bonded. Chemical bonds and atomic angles are modeled using simple springs, and dihedral angles (that is, rotations about a bond) are modeled using a sinusoidal function that approximates the energy differences between eclipsed and staggered conformations. Non-bonded forces arise due to van der Waals interactions, modeled using the Lennard-Jones potential, and charged (electrostatic) interactions, modeled using Coulomb's law. **Source:** Durrant and McCammon, 2011 (127).

30

1.9 Correlation network analysis

The study of complex networks is an approach that covers several areas of knowledge, such as computer science, mathematics, physics, biology and sociology (132–135). Networks are represented by a graph that presents a nontrivial topographic structure, composed of a set of vertices (or nodes) connected by edges, which establish a relationship between two vertices according to the modeled problem (132,136–139). The study of networks in the form of graphs is one of the bases of discrete mathematics that begun with Euler in 1736 (140). Euler proposed a solution to the problem of the Seven Bridges of Konigsberg; the study aimed at determining whether it was possible to cross the seven bridges of the city of Königsberg without a single bridge being crossed more than once, which proved impossible(140).

However, not every graph can be considered a complex network. Specific topographical properties, which are not present in simple graphs, are necessary for such classification (134,141–143). Among the features are: clustering, resistance, degree correlation and mixing patterns.

(i) Clustering: the formation of clusters intrinsic to networks are studied through the transitivity phenomenon. This occurs when a vertex A is connected to a vertex B, and vertex B is connected to a vertex C, increasing the chances of vertex A also being connected to vertex C. In analogy to the social context of network, if A is a friend with B and B is friends with C; A and C are potential friends (134,141–143).

(ii) Resistance: it indicates network's ability to withstand node removal without loss of functionality. This property is directly related to the degree distribution of vertices, since the withdrawal of vertices can lead to loss a connection between vertex pairs and/or a significant increase in the path to be traveled between one vertex and another (134,141–143).

(iii) Degree correlation: The correlation indicates whether the edges in a network can make associations between vertices with similar degrees. This correlation is mainly used in networks with pattern variations, to investigate the probability of connecting the vertices of different types (134,141–144).

(iv) Mixing patterns: Some types of networks present a mixture of different patterns where vertices can represent different kinds of objects. In this case, the probability of having a connection between vertices will depend on the specific

vertex type. For example, in food web networks, there are vertices representing plants, herbivorous animals, and carnivorous animals. In this context, it is expected that there are edges connecting herbivores to plants and herbivores to carnivores. However, no significant connections between herbivores and herbivores or between carnivorous animals and plants are expected (134,141–143).

Another critical aspect of network theory is the study of the centrality measures. In complex networks, and especially in a biological context, some nodes will be more important than others in the propagation of communication (133,136,139,145,146). The network centrality has been extensively studied over the years, and some analysis has been developed to describe how essential or central a given node can be. The analyses can be based on the number of connections, the distance between connections, the relevance of the connection to a specific function, among other means (133,134,136,139,145,146). Among the leading measures of centrality, the following ones stand out:

(i) Degree centrality: it considers the number of connections that a given node can have. Thus, according to this measure, the more connections, the higher the importance of this node for the network (133,136,144–146).

(ii) Eigenvector centrality: the concept presented in degree centrality is expanded to the eigenvector centrality. Eigenvector centrality considers not only the number of connections that a given node has but also the connections of its neighbors (133,139,144,145).

(iii) Betweenness centrality: it describes the influence of a given node on the information propagation. Thus, if a signal is propagated along a shorter path, it is expected to experience less interference that a longer path. In some cases, as in allosteric mechanisms in macromolecules, the use of pathways that pass through cross residues that interfere with signal propagation may be interesting for the allosteric mechanism (133,145,146).

(iv) Degeneration: is based on the description of the number of times a given node appears in the paths (2, 5, 8).

2 OBJECTIVES

2.1 Main objective

To investigate the dynamics and functional behavior of *apo* and *holo* TPP riboswitches in aqueous solution.

2.2 Specific objectives

- Understand the functional behavior of bacteria and plant TPP riboswitches.
- Search for candidates for TPP riboswitch in the human genome.
- Propose 3D models and analyze the structural behavior of human TPP riboswitch candidates.

3   MATERIAL AND METHODS

3.1   Analysis of crystallographic structures

For this study, TPP riboswitch 3D-structures of *Escherichia coli* (TPPsw[ec]) and *Arabidopsis thaliana* (TPPsw[at]) obtained by X-ray crystallography were selected from the PDB (104). Both structures are bound to TPP and present high-resolution crystal, of 2.05 Å (PDB ID: 2GDI (147)) and 2.25 Å (PDB ID: 3D2G (148)), for TPPsw[ec] and (TPPsw[at]), respectively. Corresponding sequence and secondary structures information were also taken from the PDB files and analyzed with 3DNA software suite (149). TPPsw[ec] and TPPsw[at] sequences were then aligned using the SARA-Coffee mode of T-Coffee program (150), and figures of sequence alignment were rendered using ALINE (151). Graphical representation of 2D and 3D structures were generated using VARNA (152) and UCSF Chimera, respectively.

3.2   Search for TPP riboswitch in the human genome

Infernal (INFERence of RNA ALignment) version 1.1.2 (92), with default settings, was used for searching TPP riboswitch candidates in the human genome. Infernal's algorithm implements covariance models (CMs), a particular case of stochastic context-free grammar (SCFGs), to create a probabilistic model that accounts for RNA sequence and secondary structure conservation that can be used to search for a particular structural pattern in user-provided sequences (153).

First, the software utilizes a set of reliable sequence alignments, along with a shared secondary structure annotation (Stockholm format), to create the CM model specific to that target RNA family. Then, it uses a dynamic programming algorithm to find a similar sequence and structural patterns in a set of target sequences. Infernal is available at http://eddylab.org/infernal/.

The program output consists of a ranked list of the hits with the most significant matches to the query. Each hit represents a region of local similarity of the CM to a subsequence of a full target database sequence. An alignment between the matched sequence and the model sequence with its corresponding confidence value for each residue in the alignment is also shown.

3.2.1 Creation of training sequences

Infernal requires a set of training sequences for the creation of CMs that must be provided by the user. For the construction of the training sequence, a set of reliable sequences of TPP aptamer was selected where only TPP riboswitch aptamer sequences with solved 3D structures were used. These sequences were obtained from the PDB (104).

Multiple sequence alignment and common secondary structure annotation was constructed using LocARNA, with default settings. The program can be found at the Freiburg RNA tools web server http://rna.informatik.uni-freiburg.de/LocARNA/Input.jsp (154). The LocARNA multiple alignments are shown, along with the predicted structure, by RNAalifold (155). LocARNA computes pairwise alignments using dynamic programming. Multiple alignments were constructed from pairwise alignments using a progressive alignment strategy.

Using the Infernal software, the alignment of TPP riboswitch aptamer in Stockholm format and the *cmbuild* program were used to create the CM; the model was calibrated by the *cmcalibrate* program. The resulting model was used in the search for TPP riboswitch in the human genome.

3.2.2 Searching for TPP riboswitch in the human genome

To search for TPP riboswitch in the human genome, we used the methodology previously applied by our research group (156), which utilized the UniGene database as the start point.

The UniGene (157) is a project aiming at the clustering of different transcripts of every gene from a given genome. Thus, we used the transcriptome sequences from the Unigene *build 236* (Hs.seq.uniq.gz file; Apr 25 2013) to perform the search for matches of TPP riboswitches in the human genome through the application of Infernal software. The Unigene can be retrieved at ftp://ftp.ncbi.nih.gov/repository/UniGene/Homo_sapiens/.

3.3 Modeling of the 3D structures of the human TPP riboswitch candidates

3D structures of human TPP riboswitch candidates were built with ModeRNA software version 1.7.1 (158). ModeRNA is a program for comparative RNA modeling

that builds models using the atomic coordinates of a known RNA molecule (template) and the alignment between the target and template sequences. The program interprets the sequence alignment as a set of instructions and uses it to build a 3D model structure by mimicking the template structure, with the subsequent introduction of the variable parts.

Three models of putative human TPP riboswitches were built. The first two models, called CANcomp[1] and CANcomp[2], were constructed using the complete sequence. A third model named truncated CANtrunc from now on, consisted of the former sequence in which 18 nucleotides were deleted due to the low identity with the CM created by the Infernal software (described in topic 3.2.1).

All molecular interactions were represented using the amber 99sb-ildn force field (159) and the models were energy minimized using the GROMACS version 5.1.2 package (160). A short minimization procedure of 150 steps (100 steps of steepest descent plus 50 steps of conjugate gradient was performed. The description for creating the templates will be presented below.

3.3.1 Modeling of truncated candidate

We modeled the human TPP riboswitch candidate disregarding the sequence of 18 nucleotides (UAGCAGAUGAGCUCUCCA) that have a low identity with the CM created by the Infernal software (described in topic 3.2.1). Initially, we used four different 3D structures of TPP riboswitch of PDB as templates for comparative modeling. Structures with different sequences, high-resolution crystal and complexed with TPP were selected. The templates 2GDI (147) (55% identity), 2HOJ (161) (58% identity) and 3K0J (162) (55% identity) are from *Escherichia coli* while 3D2G (148) (50% identity) is from *Arabidopsis thaliana*. Templates and target sequences were then aligned with LocARNA. Four 3D models were built, one for each template. The model with the highest level of identity and the lowest RMSD value with its corresponding template was selected for further analysis.

3.3.2 Modeling of complete candidates

The complete model, containing the 18 nucleotides with low identity with the CM created by the Infernal software (described in topic 3.2.1), was modeled using two different approaches. First, the portion of 18 critical nucleotides was produced by *de novo* method, and then it was merged with the complementary region corresponding

to the truncated model, which was built using comparative modeling, with ModeRNA software.

In the first step, we performed the prediction of the secondary structure of the sequence corresponding to the 18 nucleotides using Vfold model (163). Vfold predicted a 2D structure from the nucleotide sequence, and the free energy landscape was employed to build the ensemble of 2D structures with the identification of the lowest free energy state. The web server (164) is freely accessible at http://rna.physics.missouri.edu/. From the 2D structure, RNAComposer server (165,166) was used for to predict of the 3D structure of the fragment. This server predicts RNA structure based on the secondary structure written in dot-bracket notation provided by the user. The secondary structure is divided into fragments containing overlapping canonical base pairs to build the model. The fragments are related to 3D elements found in RNA FRABASE database (167,168), which is a dictionary containing RNA 3D structure elements derived from structures deposited in the PDB. RNAComposer automatically assembles the 3D elements using overlapping canonical base pairs followed by the energy minimization in the torsion angle space and subsequently in the Cartesian atom coordinate space. For the construction of the model, the job was submitted to http://rnacomposer.cs.put.poznan.pl/.

In the second step, two models of human TPP riboswitch candidates were created using comparative modeling. The first model, called CANcomp[1], was constructed using as templates the crystal structure 2GDI along with fragment modeled by the *de novo* method. The second model (CANcomp[2]) had as templates the fragment modeled *de novo* method, and the CANtrunc model (described in topic 3.3.1).

3.4 Molecular dynamics simulations

Molecular dynamics (MD) simulations were carried out using the GROMACS version 5.1.2 package (160), and RNA interactions were represented using the amber *99sb-ildn* (159) force field. Bonded and Lennard-Jones molecular parameters for TPP have been obtained using the generalized amber force field (GAFF) (169) and AM1-BCC (170) tools while atomic partial charges were added using ANTECHAMBER (171). ACPYPE (172) program was employed to create a GROMACS compatible topology file. Electrostatic interactions were treated using the particle mesh Ewald (PME) algorithm with a cut-off of 10 Å.

MD trajectories were monitored to investigate possible differences in the molecular behavior between *apo* and *holo* TPPsw[ec] and TPPsw[at]. In the *apo* systems, TPP was removed from the X-ray crystal structure and replaced with solvent water. The molecular behavior of the CANtrunc, CANcomp[1] and CANcomp[2] models, in the *holo* forms, were also simulated. Two initially positioned magnesium ions in the crystal structure were kept in both the *apo* and *holo* systems and also contributed to charge neutrality system. These ions are essential for the ligand binding and confer stability to the riboswitch as well (173). Each system was simulated under periodic boundary conditions in a triclinic box whose dimensions were automatically defined considering a distance of 1 nm from the outermost RNA atoms in all cartesian directions. The simulation box was filled with TIP3P water molecules (174).

Simulations were performed in three phases: *(i)* Energy minimization, *(ii)* thermalization and equilibration, and *(iii)* trajectory production.

Energy minimization procedure was performed through 5000 steps and a gradient tolerance < 1.0 kJ mol^{-1} nm^{-1} of the steepest descent and conjugate-gradient algorithms. These steps were carried out with heavy atom restraints by applying a harmonic potential with force constant of 1000 kJ mol^{-1} nm^{-2} for the steepest descent algorithm. Applications of the conjugate-gradient algorithm does not allow the application of restraints.

In the second phase, starting atomic velocities were assigned to all the atoms of the system using a Maxwell-Boltzmann distribution, corresponding to an initial temperature of 20 K. Then, the systems were gradually heated up to 300 K over 500 picoseconds (ps) utilizing the Langevin thermostat. During this stage, all heavy atoms were harmonically restrained by applying a force constant of 1000 kJ mol^{-1} nm^{-2}.

Systems were subsequently equilibrated during twenty successive 100 ps long equilibration simulations where position restraints approached zero progressively. After this period, the systems were simulated with no restraints all at 300 K for different time scales (**Table 3**). All simulations were performed in the NPT ensemble. The V-rescale thermostat and Berendsen barostat were used for temperature (300 K) and pressure control (1atm), respectively.

Table 3: Simulation conditions for all systems.

System	Time (ns)	Traj. Num.	Ions	Waters
TPPsw[ec] free	1000	2	38	8737
TPPsw[ec] bound	1000	2	39	8706
TPPsw[at] free	1000	2	38	8560
TPPsw[at] bound	1000	2	39	8539
CANtrunc	1000	2	39	9662
CANcomp[1]	250	2	48	11617
CANcomp[2]	250	1	48	13888

3.5 Trajectory analysis

As in the trajectory analysis, we were interested only in the structural aspects of the systems, regardless the temporal correlation, two independent MD simulations were concatenated, and trajectory analyses were conducted for crystal structures and modeled systems.

To investigate structural changes of the TPP aptamers, root-mean-square deviation (RMSD) values were calculated separately for the whole RNA and its substructures after fitting to their respective parts taking the initial structure of the production dynamics as a reference. Hydrogen bond formation was defined using a geometric criterion with VMD software. It was considered a hit when the distance between two polar heavy atoms, with at least one hydrogen atom attached, was less than 3.5 Å. Motif Identifier for Nucleic acids Trajectory (MINT) (175) program was used to evaluate the number of Watson–Crick (WC)-edge and non-WC-edge hydrogen bonds (and their sum) per nucleotide throughout the simulations.

3.6 Principal components analysis

The study of large-scale domain motions is essential for characterizing the conformational dynamics of macromolecules. As a matter of fact, functional motions are usually described by a few numbers of degrees of freedom that can be calculated

using principal component analysis (PCA) (176,177). PCA analysis was carried out for all systems using Bio3D (178) library as implemented in R (179). Rotational and translational motions were removed before calculation of the covariance matrix by least-squares superposition to the corresponding average structures. The 3N×3N covariance matrices of C5' atomic positions (Cartesian coordinates) were then calculated for each state. The conformations explored during the MD simulations were applied using hierarchical clustering in R (*hclust*) with the *complete* linkage method based on the PC1-PC2 subspace, where PC1 and PC2 denote the projections onto the two first eigenvectors.

3.7 Correlation network analysis

The cross-correlation and network analyses were carried out using the Bio3D and the *igraph* R packages (180). Initially, the dynamic cross-correlation matrices (DCCM) were calculated separately for each simulation using as inputs the corresponding MD trajectory superimposed onto the initial structure. Then, each group of two matrices per riboswitch state was utilized to obtain a consensus matrix. A proximity/contact map filter was applied in the construction of the correlation network for residues that remained within 4.5 Å from one another for at least 75% of simulation time. Briefly, graphs were obtained considering C5' atoms as nodes and the connection between nodes *i* and *j* were weighted using the absolute values of cross-correlations ($C_{(i,j)}$) coefficients:

$$w_{(i,j)} = -log\left(\left|C_{(i,j)}\right|\right). \qquad [1]$$

We also calculated the relative importance of each node for communication using centrality measures. According to the definition of the betweenness centrality, the relevance of a given node is defined by its presence in shortest communication paths connecting nodes over the entire network (181).

$$c_B(n) = \sum_{i \neq j \neq k \in N} \frac{\sigma(i,j|n)}{\sigma(i,j)} . \qquad [2]$$

According to the above equation, the betweenness centrality of a node depends on the total number of the shortest paths between nodes *i* and *j* that pass-through *n*

($\sigma(i,j|n)$ and $\sigma(i,j)$, which is the total number of shortest paths between nodes *i* and *j* (regardless of whether they pass or not through *n*).

Another measure of centrality can be given by the eigenvector that accounts for the global relevance of each residue based on the connections with neighboring nodes. In other words, nodes with high eigenvector centrality are those connected to other central residues (182).

$$c_{Ei}(v) = \alpha \sum_{\{u,v\}\in E} c_{Ei}(u). \qquad [3]$$

The vector $c_{Ei} = (c_{Ei}(1), ..., c_{Ei}(N_v))^T$ is obtained as a solution to the eigenvalue problem $A c_{Ei} = \alpha^{-1} c_{Ei}$, where A is the adjacency matrix for the network graph G. More mathematical details can be found at Kolaczyk, 2009 (182).

The Square inner product (SIP) was used to compare the overall similarity of the centrality profiles calculated for the systems. It varies between 0 and 1 and is defined as

$$SIP = \frac{(w_A^T w_B)^2}{(w_A^T w_A)(w_B^T w_B)} , \qquad [4]$$

where w_A and w_B are *N*-length vectors containing the fluctuation value for each atom in proteins A and B, respectively (183).

The Yen's algorithm (184) was used to calculate the shortest pathways connecting two nodes in the network. Path lengths are defined as the sum of the edge weights connecting a pair of nodes in a given pathway. The first 1000 shortest paths were collected and employed to calculate the node degeneracy value, which represents the percentage of pathways from the overall ensemble in which a given node is present.

4   RESULTS

4.1   Unraveling structural behavior of TPP riboswitches in bacteria and plants species

4.1.1   RNA content analysis

Nucleotide sequence alignment between TPPsw^{ec} and TPPsw^{at} revealed they share 68% identity (**Figure 12A**), along with high conservation in their secondary structures (**Figure 12B**). According to the SimTree server (185), the secondary structure analysis indicated a normalized score of 0.68 in a scale from 0 to 1, where 1 indicates a perfect match and 0 no match at all.

A detailed inspection of the secondary structure contents showed the aptamers domains display identical J3-2 junctions and an identical number of pairings in the P1, P4, and P5 helices. On the other side, the P3 helix was the least conserved substructure, presenting four nucleotides and an additional base pair in TPPsw^{ec}. It is noteworthy that although the number of nucleotides in the P1 helix of TPPsw^{ec} is fewer than in TPPsw^{at}, the number of base pairing remains the same, being four for each case. Differences can also be found in the J2-4 junction and P2 helix, in which an additional pairing in TPPsw^{ec} is formed. Globally, the 2D riboswitch structures are highly conserved between species, although minor differences are observed, mainly concerning the content of base pairings, being TPPsw^{ec} two pairings longer than TPPsw^{at}.

Both prokaryotic and eukaryotic TPP aptamers share a common 3D structure and organization. The structural superposition between TPPsw^{ec} and TPPsw^{at} crystal structures resulted in a root-mean-square deviation (RMSD) of approximately 0.6 Å (**Figure 13C, D**). The aptamer consists of switch helix (P1) and two sensor arms (P2/P3 and P4/P5), that form a long-range tertiary rearrangement capable of stabilizing the interaction between the L5 loop and the P3 stem (**Figure 13A, B**). The conformational organization of TPP riboswitch is mediated by TPP, which binds to the TPP pyrimidine ring in the arm P2/P3 while the TPP pyrophosphate group locates in the P4/P5 arm (**Figure 13E, F**).

Figure 12: Sequence alignment **(a)** and secondary structure **(b)** between TPP riboswitch of *Escherichia coli* (PDB ID: 2GDI) and *Arabidopsis thaliana* (PDB ID: 3D2G). **(a)** Black filled positions of the alignment represent conserved residues. **(b)** Conserved nucleotides were colored in gray. Stems, loops, and junctions were identified according to the caption in the figure.
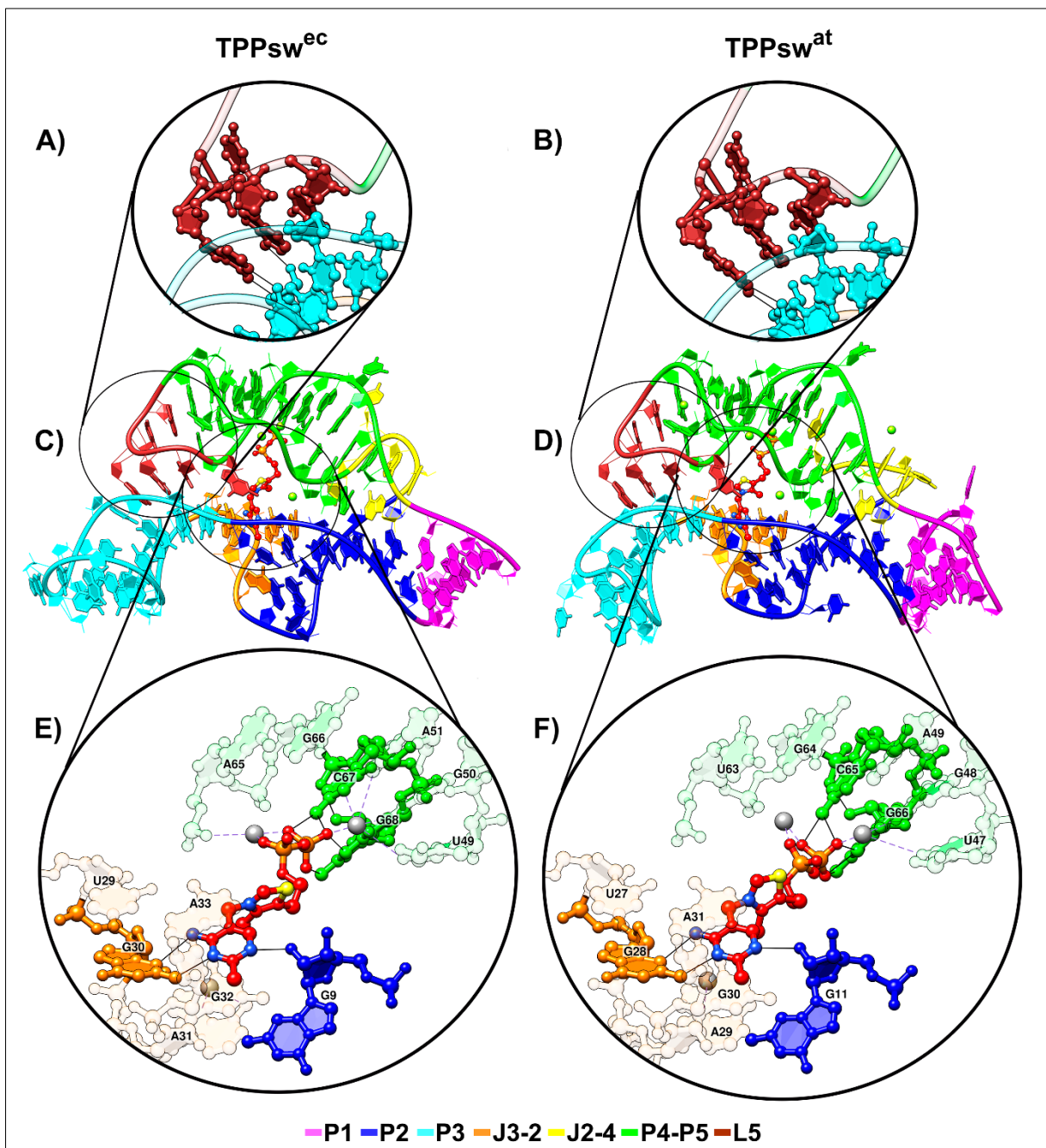
Figure 13: Tertiary structure of the TPP riboswitches of *Escherichia coli* (PDB ID: 2GDI) and *Arabidopsis thaliana* (PDB ID: 3D2G). Cartoon and sticks representation of P3-L5 substructures **(a, b)**, whole aptamer **(c, d)**, and TPP binding site **(e, f).** Stems, loops, and junctions are figure caption. Thick black lines and gray dashed lines indicate hydrogen bonds formation and interactions with the MG^{+2} ion respectively **(a, b, e, f)**. **(a, b)** Nucleotides A59(57) and A60(58) from loop L5 connect with residues G27(25) and C28(26) from the P3 minor groove. **(e, f)** The aminopyrimidine ring of TPP formed hydrogen bonds with G30(28) and the 2-OH' of G9(11). Direct contacts to nonbridging oxygens of β-phosphate of TPP were also established via N4 of C67(65) and N1 of G68(66). All other pyrophosphate-RNA contacts were mediated through two Mg2+ ions (colored in gray).

44

4.1.2 Global and local stability of the aptamer domain

The structural stability of riboswitches in aqueous solution, in both *apo* and *holo* configurations, was evaluated by comparing the average RMSD values calculated over the production MD simulations, taking the initial structures as references (**Table 4**).

During MD simulations, both *apo* and *holo* forms presented deviations around 3Å. Apparently, no noticeable differences have been observed between the RMSD values obtained for both TPPsw^{ec} and TPPsw^{at} upon complexation. Indeed, both average RMSD values per region and the overall fluctuation pattern were similar for all systems. These values are consistent with RMSD values seen in other studies reporting simulations of small RNA aptamers (186–189) which ranged between 3-4 Å approximately.

A more in-depth look at particular regions of the RNA aptamer evidenced some differences in their dynamical behavior. For example, the P1 region presented higher RMSD values than for all other substructures. Also, this region was more flexible in the *holo* forms. It worth mentioning that nucleotides composing P1 helix are located in the terminal regions of the aptamer, with broader movement during dynamics. Unlike the P1 segment, P3 helix displayed higher RMSD values in the *apo* forms. However, the comparison of fluctuations between *apo* and *holo* riboswitches showed that TPP binding resulted in the stabilization of the P3 helix, but at the same, it contributed to the P1 helix disruption (**Figure 14**). Curiously, the nucleotide U36 from P2 helix of TPPsw^{ec} displayed more significant fluctuations in the *holo* form (*holo*: 5.4 Å; *apo*: 2.3 Å). No other significant changes were observed in the P2 substructure in *apo* and *holo* TPPsw^{at}.

Table 4: Root mean square deviations (Å) of free and bound states of TPPsw^{ec} and TPPsw^{at} as a whole and for substructures.

	Free		Bound	
	TPPsw^{ec}	TPPsw^{at}	TPPsw^{ec}	TPPsw^{at}
Whole	2.83 ± 0.45	3.18 ± 0.34	2.62 ± 0.41	2.73 ± 0.37
P1	5.91 ± 1.56	4.97 ± 0.67	6.44 ± 1.71	5.22 ± 0.71
P2	1.70 ± 0.33	2.37 ± 0.40	2.81 ± 0.80	2.06 ± 0.27
P3-L3	2.99 ± 0.66	3.16 ± 0.48	2.88 ± 0.58	2.87 ± 0.50
P4-P5	1.70 ± 0.25	2.03 ± 0.17	1.65 ± 0.27	1.76 ± 0.35
J3-2	1.37 ± 0.29	1.78 ± 0.26	1.85 ± 0.42	1.94 ± 0.43
J2-4	3.42 ± 0.38	2.96 ± 0.40	3.74 ± 0.26	1.63 ± 0.61
L5	1.43 ± 0.29	1.41 ± 0.31	1.53 ± 0.41	1.52 ± 0.48
TPP	-	-	1.38 ± 0.30	1.41 ± 0.36

"-" no RMSD values
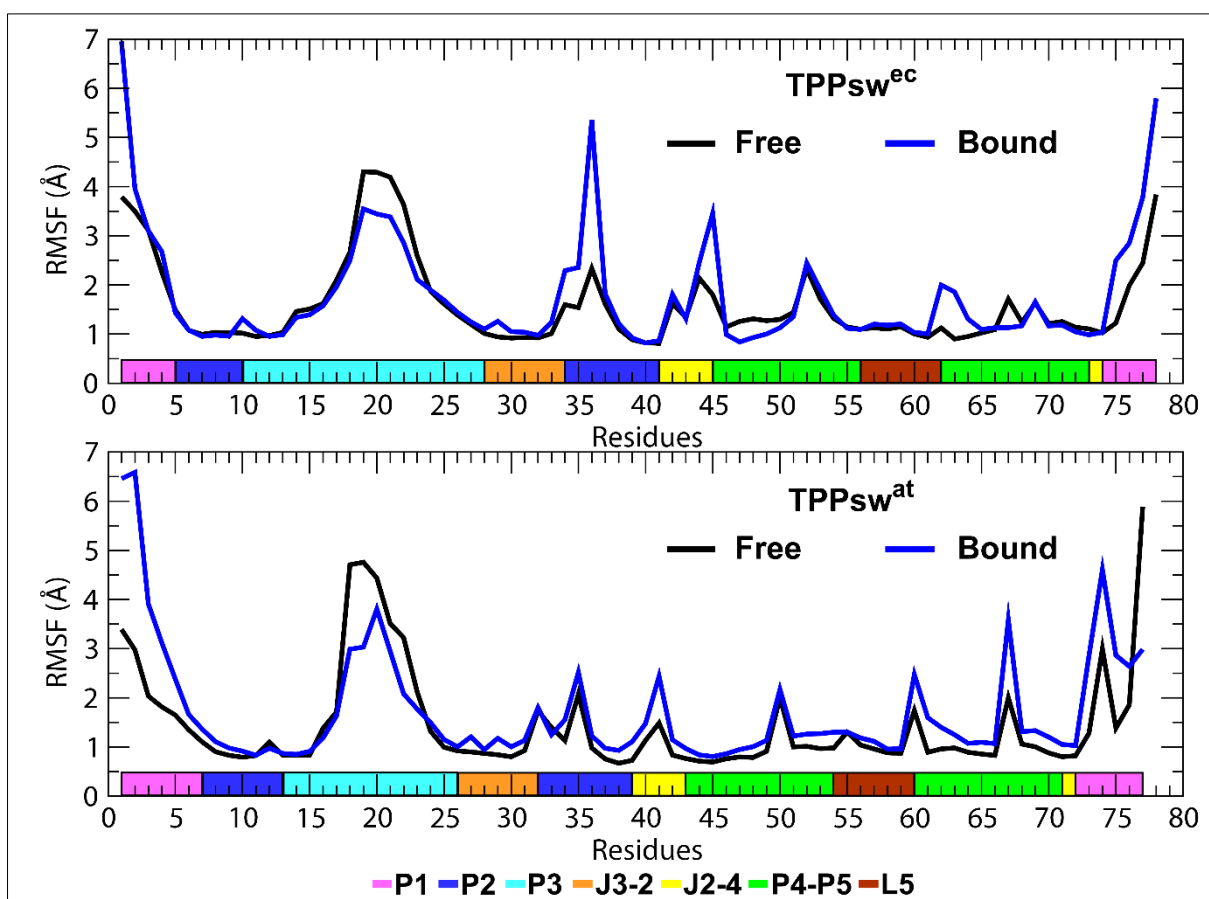
Figure 14: All heavy atoms Root Mean Square Fluctuations for the TPP free and bound states of TPPsw^{ec} and TPPsw^{at}. Secondary structure regions are depicted below the plots and colored according to the figure caption.

4.1.3 Monitoring P3-L5 interaction

Despite being remotely located from the TPP binding site, the interaction between P3-L5 substructures is essential for metabolite binding (41,162). The conformation of the P3-L5 region is kept formed via two non-Watson-Crick (WC) base pairs. Nucleotides A59(57) and A60(58) from loop L5 connect with residues G27(25) and C28(26) from the P3 minor groove (**Figure 13A, B**) and are essential for the P3-L5 interaction (147). Numbers in parenthesis indicate the nucleotide position in the TPPsw[at] systems.

We also monitored the formation of non-Watson-Crick hydrogen bonds between P3 and L5 during MD simulations (**Table 5**). Notably, the G27(25)•A59(57) in *apo* TPPsw[ec] was the least stable one that remained connected for approximately 68% of the simulation time. Upon TPP binding, the stability of this interaction increased up to 74% while no corresponding difference was observed between *apo* and *holo* TPPsw[at] systems (with an occupancy of ~ 81% in both cases).

In addition, the C28(26)•A60(58) was more stable than G27(25)•A59(57) and presented occupancies higher than 83% in all systems. In this way, the base pair G27•A59 of TPPsw[ec] appears to be influenced by the presence of the ligand, which seems to have contributed for stabilization.

Table 5: Hydrogen bonds occupancy of non-Watson-Crick pairs involved in the P3-L5 interaction of free and bound states of TPPsw[ec] and TPPsw[at].

Nucleotides		% of frames			
		Free		Bound	
P3	L5	TPPsw[ec]	TPPsw[at]	TPPsw[ec]	TPPsw[at]
G27(25)	A59(57)	68.31%	81.77%	74.41%	80.88%
C28(26)	A60(58)	92.66%	89.34%	89.39%	83.62%

4.1.4 TPP-RNA interaction

To verify the structural stability of the TPP-RNA complexes during MD simulations, we calculated the occupancy of hydrogen bonds and the average distances between atoms involved in the RNA-TPP-MG interaction (**Table 6**). We confirmed the existence of conserved interactions in all systems. The aminopyrimidine ring of TPP formed hydrogen bonds with N2 and N3 of G30(28) and the 2-OH' of G9(11). Direct contacts to non-bridging oxygens of β-phosphate of TPP were also formed via N4 of C67(65) and N1 of G68(66). All other pyrophosphate-RNA contacts were mediated through the two Mg^{2+} ions (147,162) (**Figure 13E, F**).

Analysis of the distribution of the total number of RNA-TPP hydrogen bonds revealed that most of the conformations presented interactions varying from 3 to 4 in TPPsw[ec] and between 2 and 3 in the TPPsw[at] (**Figure 15**). The number of sampled structures presenting four or more hydrogen bonds was consistently higher in the TPPsw[ec] system. The occupancy of the two hydrogen bonds formed between G30(28) and TPP was similar for both systems. Curiously, the occupancy of G9(11) for TPPsw[at] was less than a half (25.67%) than for TPPsw[ec] (55.58%).

In the *holo* systems, the MG1 ion remained closer than 1.90 Å from O3 β-phosphate of TPP, with no specific interaction with RNA. In TPPsw[at], the magnesium atom MG2 got closer to MG1 (3.35 ± 1.13 Å) and interacted more firmly with O1 oxygen atom from TPP (1.90 ± 0.06 Å), leaving the O5 oxygen atom from β-phosphate free to interact with RNA with consequent formation of a bifurcated H-bond. The O5 oxygen atom from TPP interacted with C65 e G66, with hydrogen bond occupancies of 35.11% e 25.65%, respectively.

Each TPPsw[ec] nucleotide involved in pyrophosphate recognition formed more than one hydrogen bond because they interacted with other oxygen atoms from TPP, being C67•TPP–O1 and G68•TPP–O5 examples of this, with occupancies of 41.30% and 30.48%, respectively.

Table 6: Occupancy and distance of RNA-TPP interaction.

Atoms		Occupancy (%)		Distance (Å)	
RNA	**TPP**	**TPPsw^{ec}**	**TPPsw^{at}**	**TPPsw^{ec}**	**TPPsw^{at}**
G9(11) – 2O'	TPP – N2	55.58	25.67	2.86 ± 0.19	3.03 ± 0.35
G30(28) – N2	TPP – N1	91.34	90.30	2.99 ± 0.14	3.06 ± 0.17
G30(28) – N3	TPP – N	64.48	69.97	3.08 ± 0.17	3.02 ± 0.14
C67(65) – N4	TPP – O1	41.30	0.05	3.05 ± 0.40	5.18 ± 0.97
C67(65) – N4	TPP – O3	2.68	9.69	4.02 ± 0.33	4.67 ± 0.77
C67(65) – N4	TPP – O5	1.12	35.11	4.70 ± 0.70	4.08 ± 1.07
G68(66) – N1	TPP – O1	6.38	0.00	3.71 ± 0.51	4.17 ± 0.29
G68(66) – N1	TPP – O3	0.12	0.29	3.31 ± 0.25	3.24 ± 0.31
G68(66) – N1	TPP – O5	30.48	25.65	3.64 ± 0.62	3.65 ± 0.41
MG1	MG2	-	-	6.58 ± 0.97	3.35 ± 1.13
MG1	TPP – O3	-	-	1.85 ± 0.04	1.89 ± 0.06
MG2	TPP – O1	-	-	3.38 ± 1.53	1.90 ± 0.06

"-" no occupancy values
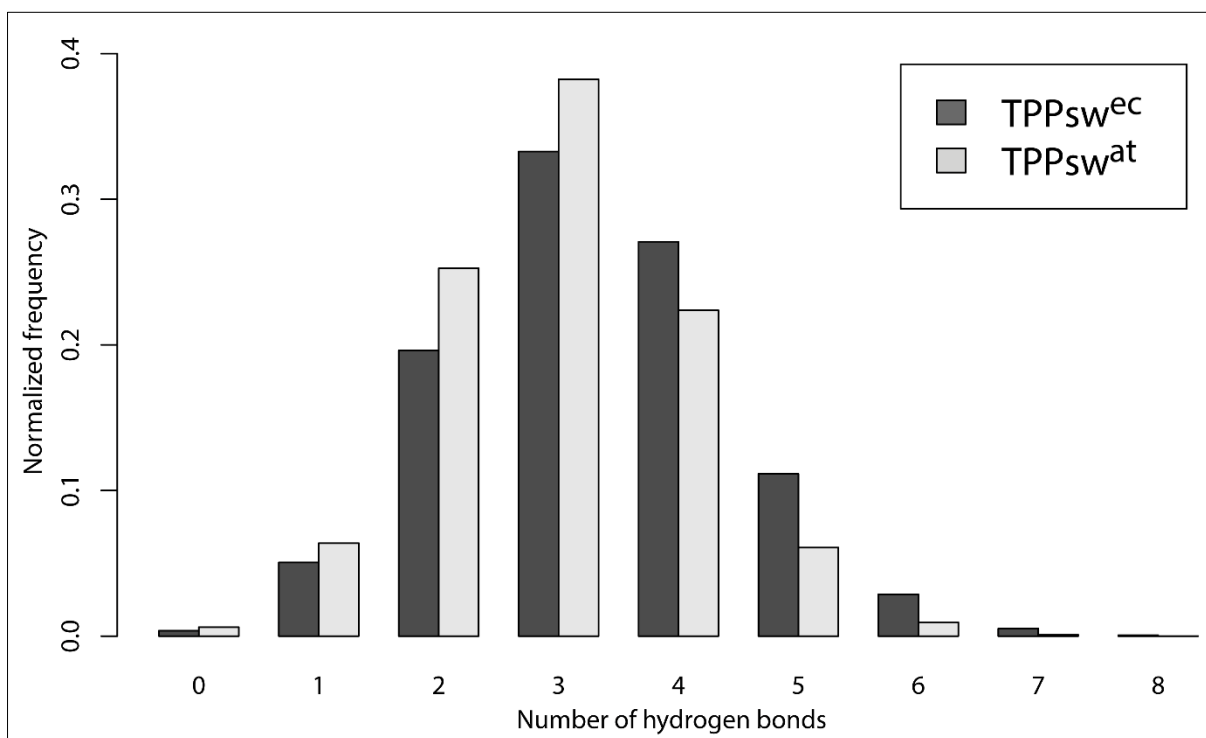
Figure 15: Distribution of the number of hydrogen bonds formed between RNA aptamer and TPP of TPPsw^{ec} and TPPsw^{at} systems. Bars were normalized by hydrogen bonds occurrence during the simulation time.

4.1.5 *Principal component analysis suggests different dominant motions in apo and holo states*

To identify statistically relevant motions of TPP riboswitches in the aqueous solution, we performed PCA using the MD trajectories. Overall, the first three components (named PC1, PC2, and PC3) captured the dominant motions, presenting the highest contributions to total fluctuations. The first three PCs accounted for 54.9% and 52.6% of the overall variance in *apo* TPPsw^{ec} and TPPsw^{at}, respectively. In the *holo* states, the contribution of the first three PCs was slightly higher: 61.7% and 58.7% for TPPsw^{ec} and TPPsw^{at}, respectively.

We compared the projections of the trajectories onto the subspace spanned by the first three principal components. The four systems showed a uniform and overlapping PC subspace (**Figure 16**). RMSFs and structural projections along the three PCs are shown in **Figure 17**.

Inspection of the atomic fluctuations along PC1 revealed substantial higher flexibility of P3 in the *apo* systems, confirming the structural stabilization of this region due to ligand binding. However, this region presented higher fluctuations along PC2 in the *holo* state in the TPPsw^{ec} system, but not in TPPsw^{at}. Therefore, stabilization of the most statistical relevant motions as a consequence of the ligand binding was undoubtedly more pronounced in TPPsw^{at}, as lower fluctuations were noticed along both principal components.

Furthermore, TPP binding resulted in increased flexibility of P1 helix. In *holo* TPPsw^{ec}, the ligand promoted an increase in the flexibility of nucleotides 34-37 of the P2 helix. This fact was not observed in TPPsw^{at}.
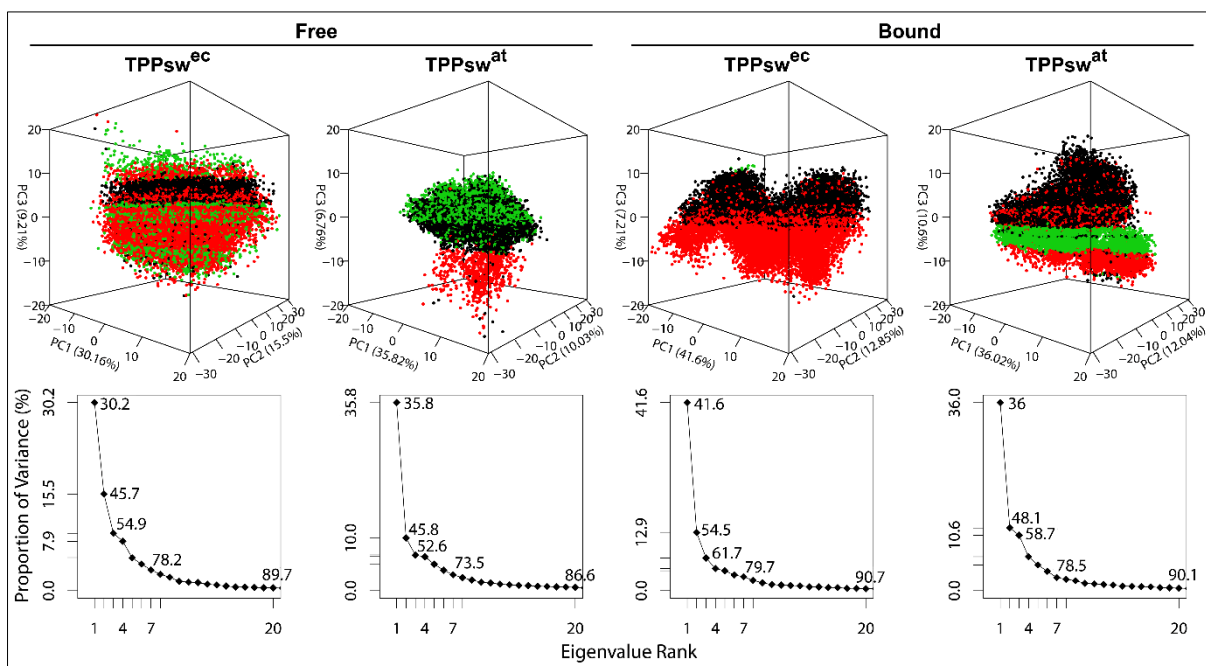
Figure 16: Principal component analysis (PCA). The first three eigenvectors were extracted from the essential dynamics and compared among themselves. The variance captured by eigenvectors is also shown. The plots indicate the extension in the conformational space, determined by the first three eigenvectors, where each dot represents one frame from the trajectory. Colors represent the three conformations clustering in the PC space.
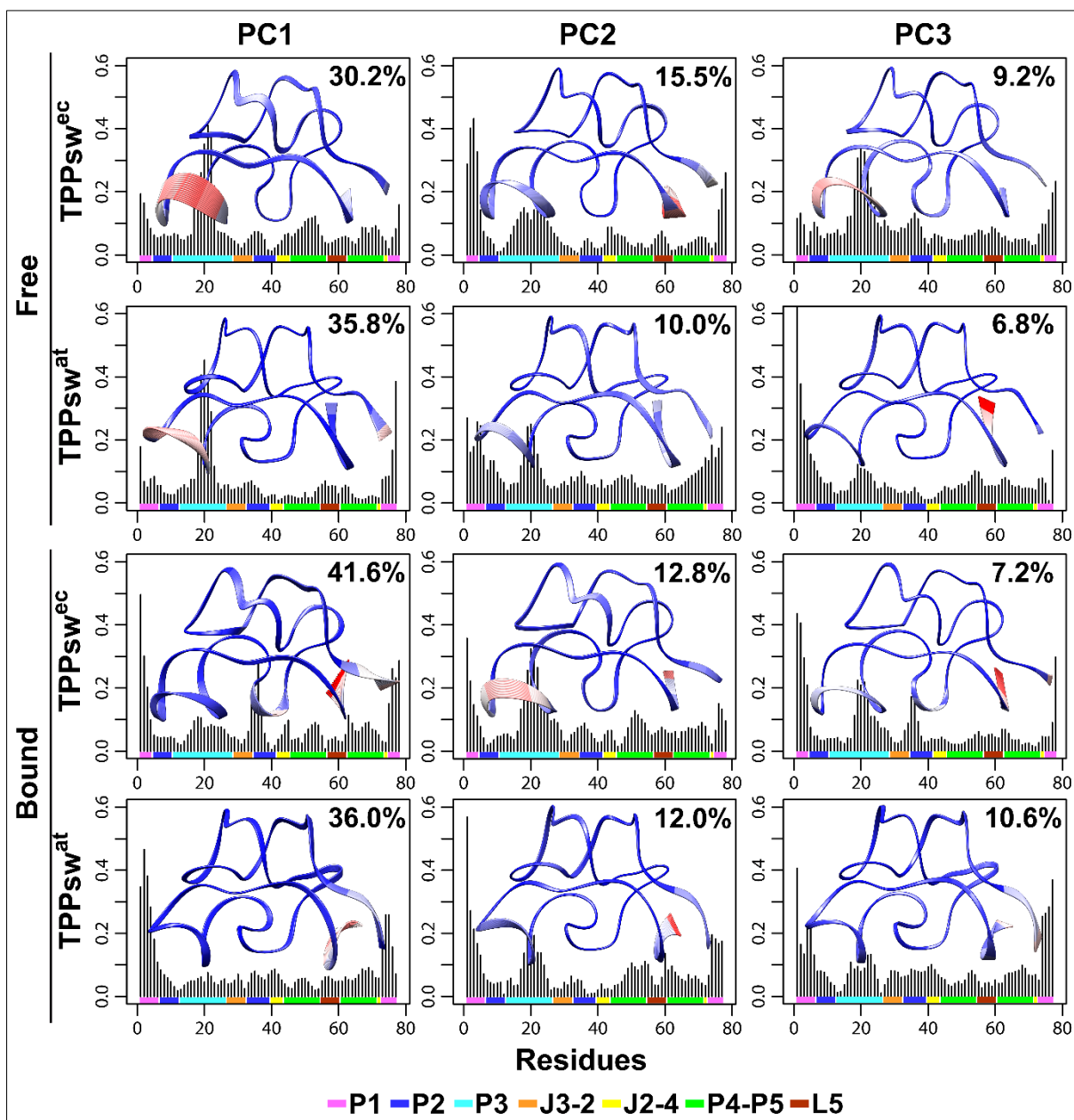
Figure 17: RMSF along the first three principal components. The fraction of variance captured each component is shown in the top right part of each graph. Interpolated structures obtained by displacements along each vector are displayed within each graph. Blue indicates overlapping regions with little or no motion. Red areas represent mobile regions. The secondary structure elements are given in the lower margin of the plots and colored according to the figure caption.

52

4.1.6 Correlation network analysis reveals distinct responses to ligand binding in TPPsw^{ec} and TPPsw^{at}

We analyzed the correlations between pairs of nucleotides to investigate how TPP binding affects the dynamic couplings in TPPsw^{ec} and TPPsw^a. We calculated dynamic cross-correlation matrices (DCCM) for each simulated system, as described in the methods section A similar cross-correlation pattern was observed for both systems in their *apo* states. However, both the extent of regions displaying anticorrelations and their magnitudes were greater in TPPsw^{ec}, mainly in the P3–P4-P5–L5 region (**Figure 18**).

It was observed that TPP binding stabilized the P3-L5 interaction in TPPsw^{ec}, as evidenced by weaker anticorrelations, which are likely to be associated with the separation of these regions in the *apo* state. This result is in line with other analysis revealing increased stability of P3 motions in the *holo* state (**Figure 17**). Interestingly, no noticeable effects were observed in the correlation pattern in this region for TPPsw^{at}. In TPPsw^{at} system, TPP binding resulted in increased anticorrelations at the P1-P2 helices, indicating a possible destabilization of interactions that utterly resulted in an increased separation between them. In contrast, the dynamic coupling pattern in this region was not altered in TPPsw^{ec}.
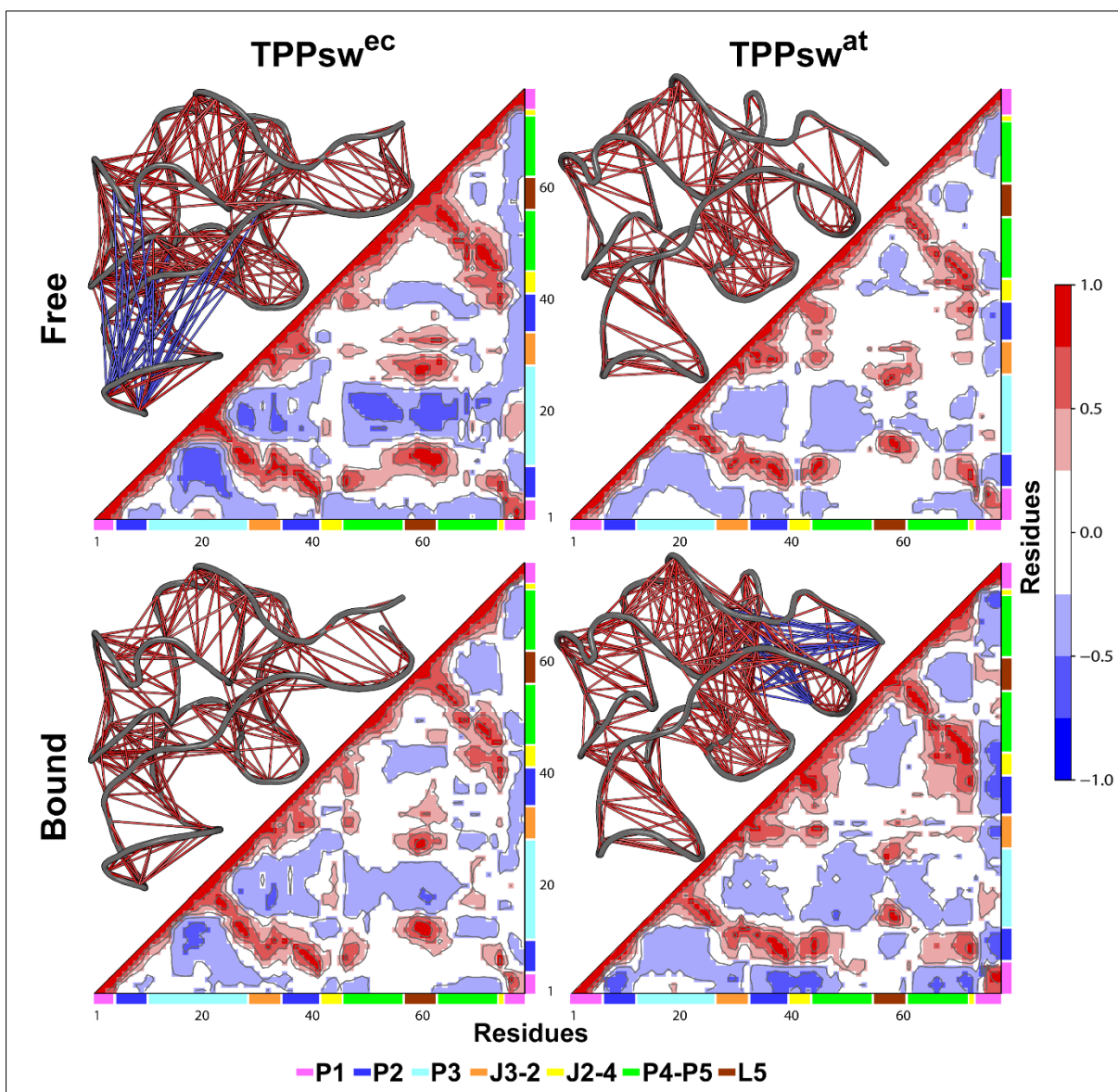
Figure 18: DCCMs of free and bound states of TPPsw[ec] and TPPsw[at]. Next to each matrix, the corresponding 3D structures with lines connecting pairs of correlated residues are shown. For clarity sake, only the pairs presenting $(|C_{ij}|) > 0.6$. are represented.

Next, we performed a correlation network analysis by constructing weighted graphs in which each residue was represented by a single node and the weight of the connection between pairs of nodes was proportional to their respective correlation coefficients previously calculated. To quantify the relative importance of each residue in the network, we computed the betweenness centrality *per* nucleotide for each simulated system (**Figure 19**). This metric is used to identify critical nodes for communication over the network. Residues presenting high betweenness values are considered "bottlenecks" of information as they are mostly found in shortest communication paths (137).

We calculated the square inner product (SIP) to compare the overall similarity of the betweenness centrality profiles calculated for the *apo* and *holo* states. According to this analysis, high SIP values are associated with weak modulation of intramolecular communication introduced by TPP binding. Indeed, we obtained a higher SIP for TPPsw^{at} (0.89) than for TPPsw^{ec} (0.65), thus reinforcing more noticeable TPP related effects in TPPsw^{ec} systems. In both *holo* systems, we noticed increased centrality values at the P4-P5 helices. Interestingly, nucleotides G27, C28, and A59 were critical for P3–L5 interaction and displayed higher betweenness values in *holo* TPPsw^{ec}, indicating that ligand binding favors efficient communication through these nucleotides (**Figure 18**). This feature was not observed in TPPsw^{at}, in which C26 and A58 centrality values were higher in the *apo* state.
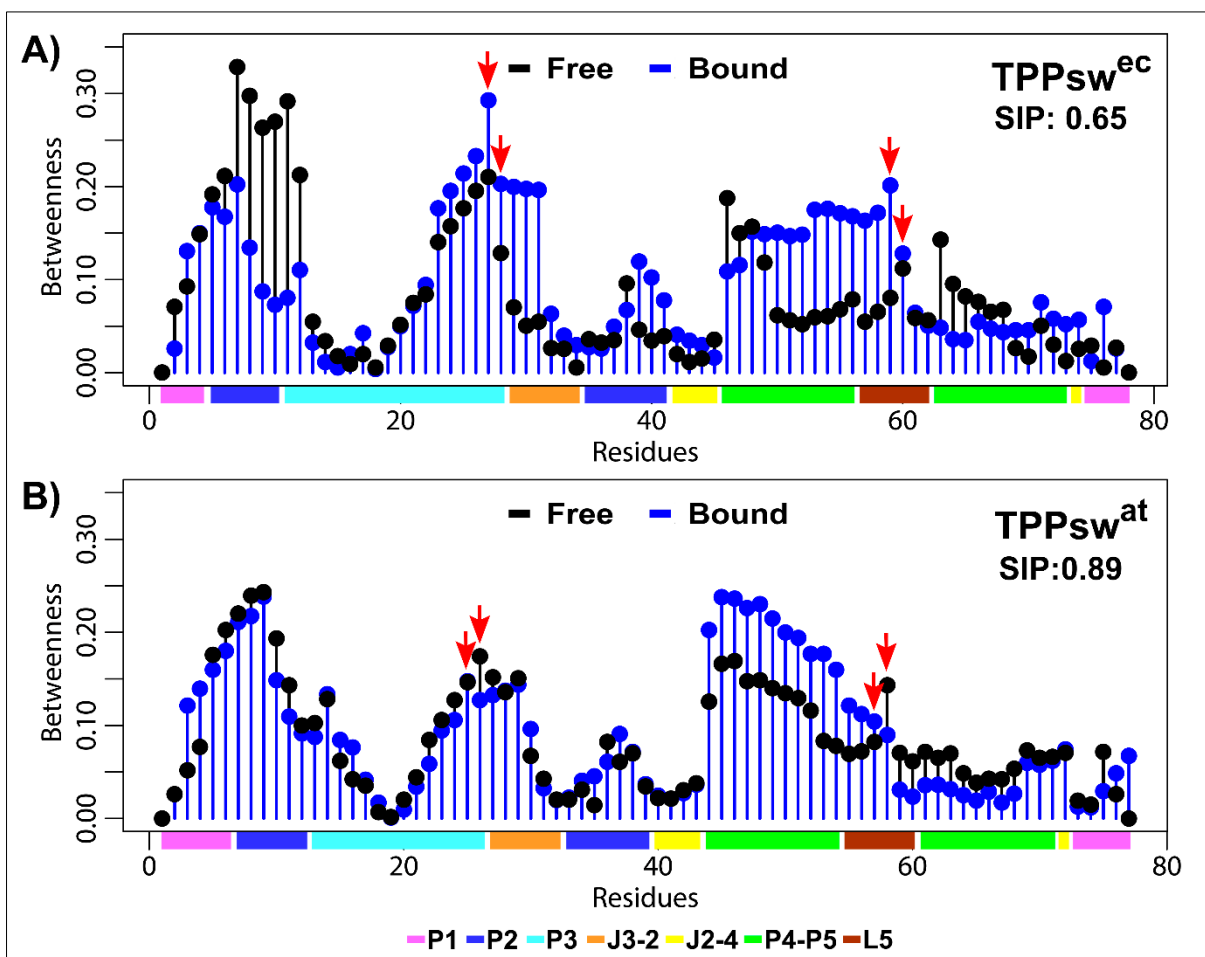
Figure 19: Betweenness centrality of the node for each residue of free and bound TPPsw[ec] **(a)** and TPPsw[at] **(b)**. Red arrows indicate residues G27(25)–A59(57) involved between P3-L5 interaction. Square Inner Product (SIP) is shown above each graph.

4.1.7 Communication pathways between P3-L5

To get a deeper understanding of the critical residues governing P3-L5 interactions, we computed the 1000 shortest paths between G27(25) [located in P3] and A59(57) [located in L5] (**Figure 20**). The normalized node degeneracy metric reveals the percentage of paths accessing each node. We observed critical residues for communication with degeneracy values > 0.35, conserved in all systems. While the majority of these residues belongs to L5 (U58(56), A60(58) and G61(59)), two of them (C12(14) and C28(26)) are located in P3 (**Figure 20A**).

The comparison of the distribution of node degeneracies obtained for TPPsw[ec] shows a narrower distribution at the P3 region in the holo state, indicating that TPP binding reduced the presence of a few residues in the shortest paths. Whereas in the apo state a large number of residues are accessed, in the holo state communication mostly involves nucleotides 59-61. In contrast, the distributions obtained for both TPPsw[at] states were very similar (**Figure 20A**).

We calculated the number of nodes per path to evaluate further and characterize the influence of TPP binding on P3-L5 interactions (**Figure 20B**). This analysis was based on the hypothesis that communication involving fewer nodes along the pathway is likely more efficient. Indeed, in *holo* TPPsw[ec], the P3-L5 communication required fewer nodes (**Figure 20B)**. Interestingly, while TPP binding did not modify the global distributions of betweenness centrality distributions obtained for both species (**Figure 20C** left boxes), opposing trends were perceived concerning the average centrality calculated for the residues participating in shortest paths (**Figure 20C** right boxes). Whereas for TPPsw[ec] TPP binding resulted in increased shortest paths centralities, it led to a slight decrease of average betweenness in TPPsw[at].

To further support this analysis, we computed the eigenvector centralities for the overall network and the shortest paths (**Figure 20C** lower boxes). Again, TPP binding resulted in higher centrality in the shortest paths only for TPPsw[ec]. The eigenvector centrality accounts for the global relevance of each residue based on the connections with neighboring nodes. In other words, nodes with high eigenvector centrality are those connected to other central residues. Therefore, the selection of specific P3 residues imposed by ligand binding in TPPsw[ec] resulted in stronger communication along pathways accessing a selection of neighboring residues with high centrality. In agreement with our previous analysis (**Figure 18-20**), a

corresponding effect was not observed in TPPsw[at], which strongly suggests the weaker influence of TPP upon the dynamic communication.
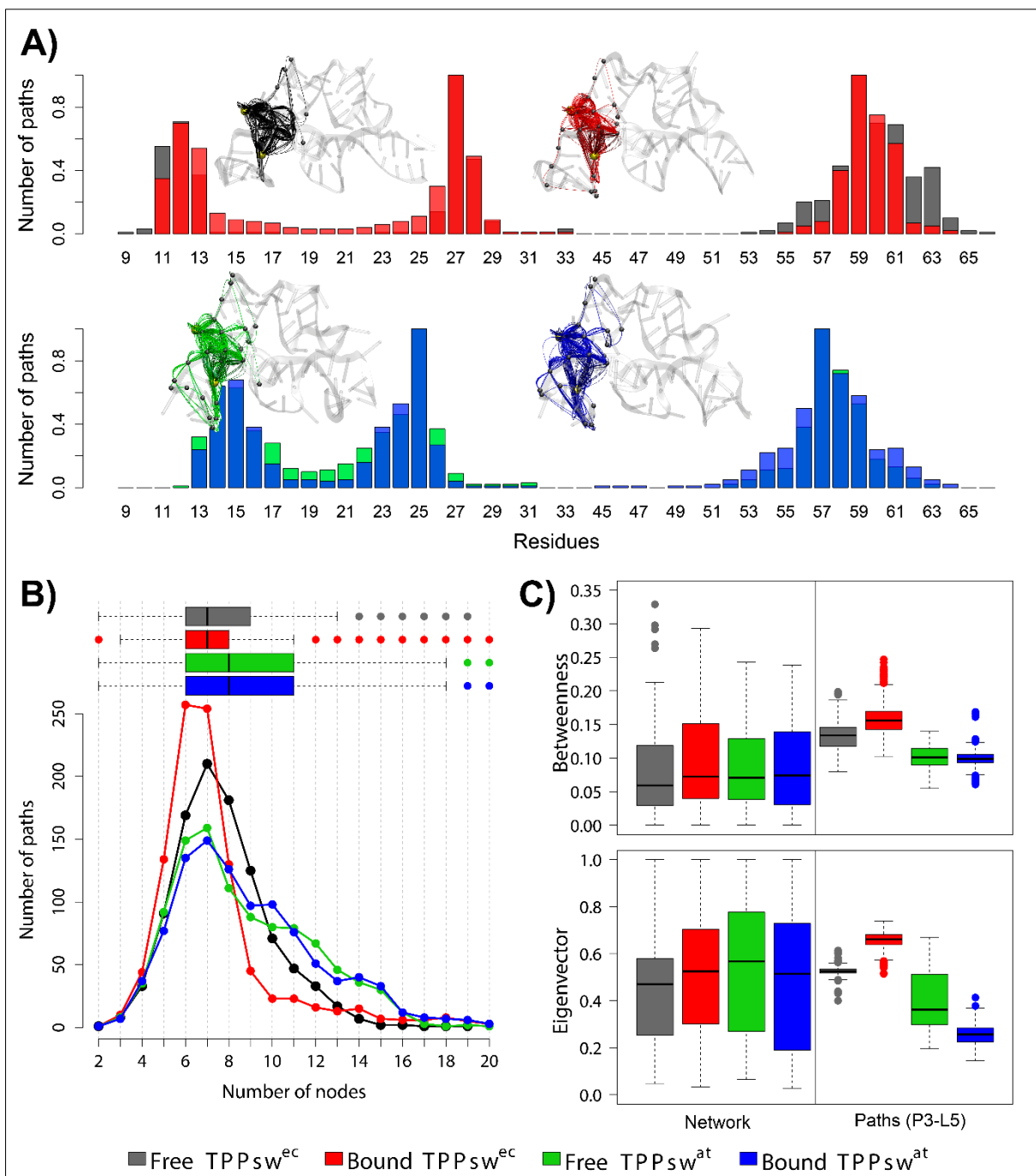
Figure 20: Shortest paths of communication connecting G27(25) and A59(57) residues of free and bound TPPsw[ec] and TPPsw[at]. **(a)** Normalized node degeneracy graph and visualization of sub-optimal paths in a correlation network. **(b)** Number of nodes per path. **(c)** Boxplot of betweenness and eigenvector centrality of the paths compared to the one corresponding to the complete network. Each system was colored according to the figure caption.

4.2 Searching, modeling and unraveling structural behavior of the human candidate TPP riboswitch

4.2.1 Searching for TPP riboswitch in the human genome

To obtain a set of reliable TPP aptamer sequences for CM construction, we used only TPP riboswitch sequences with elucidated 3D structures. The PDB contains 24 structures identified under 17 different identification codes. Out of these, only four are different in their ribonucleotide sequences (**Table 7**).

For the construction of the multiple sequence alignment and common secondary structure annotation, only the four different sequences were used (**Figure 21A**). The alignment preserved the conservation of the five stems characterizing a TPP aptamer (**Figure 21B**).

Table 7: TPP riboswitch structures deposited in the PDB.

PDB ID	Chain ID	Organism	Resolution (Å)
2CKY	A/B	*Arabidopsis thaliana*	2.90
3D2G	A/B	*Arabidopsis thaliana*	2.25
3D2V	A/B	*Arabidopsis thaliana*	2.00
3D2X	A/B	*Arabidopsis thaliana*	2.50
2GDI	X/Y	*Escherichia coli*	2.05
4NYA	A/B	*Escherichia coli*	2.65
2HOJ	A	*Escherichia coli*	2.50
2HOK	A	*Escherichia coli*	3.20
2HOL	A	*Escherichia coli*	2.90
2HOM	A	*Escherichia coli*	2.89
2HOO	A	*Escherichia coli*	3.00
2HOP	A	*Escherichia coli*	3.30
4NYB	A	*Escherichia coli*	3.10
4NYC	A	*Escherichia coli*	3.15
4NYD	A	*Escherichia coli*	2.90
4NYG	A	*Escherichia coli*	3.05
3K0J	E/F	*Escherichia coli*	3.10

*The highlighted lines identify structures of the same ribonucleotides sequence.
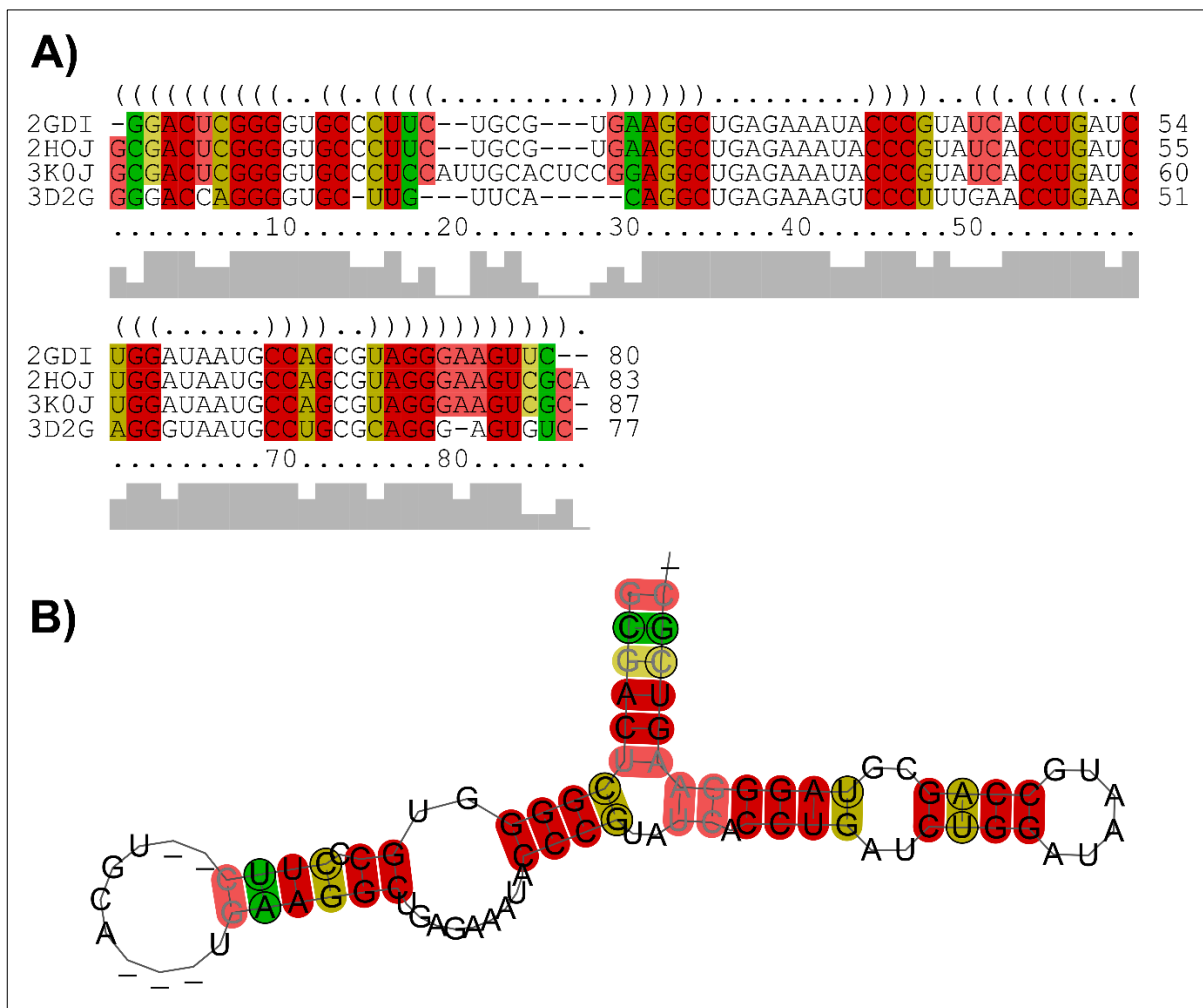
Figure 21: Alignment of TPP riboswitch sequences with elucidated 3D structure. **(a)** Alignment multiple sequence and **(b)** secondary structure of alignment consensus. Color coding of structural conservation is indicated for each base pair. Red, olive, and green correspond to the base pairs C-G, G-C, and A-U, respectively. The color saturation symbolizes the structural conservation of each base pair, decreasing in saturation as the number of unpaired nucleotides observed at that position increases.

The TPP riboswitch candidate found in the human genome is localized in the fibulin 2 gene (FBLN2). The candidate presented a score of 24.5 according to the covariance model. The threshold value of the Rfam v.13.0 score is 25.95 for TPP riboswitch classification. Despite being below the cutoff, FBLN2 was the only candidate found, regarding the training sequence used to create the CM.

The FBLN2 gene encodes a protein of the extracellular matrix, which is essential during the organogenesis (reviewed by Timpl (190)). The genomic context analysis of the FBLN2 gene reveals at least one possible event of alternative splicing in the second FBLN2 exon, which is supported by the EST DB221811 alignment (**Figure 22**). As can be depicted in Figure 22B, the EST DB221811, which was obtained from trachea, lacks the riboswitch candidate region.

Infernal software output had a particular type of alignment between FBLN2 and CM, called a *local end*. Local ends happen when a large insertion or deletion is used in the optimal alignment at a reduced penalty and allow Infernal to be tolerant to the insertion and/or deletion of RNA substructures not modeled by the CM. In this case, 6 CM positions were skipped to a *local end* insertion, and 18 residues were inserted in the FBLN2 sequence (**Figure 23**). The *local ends* enable remote homology detection.
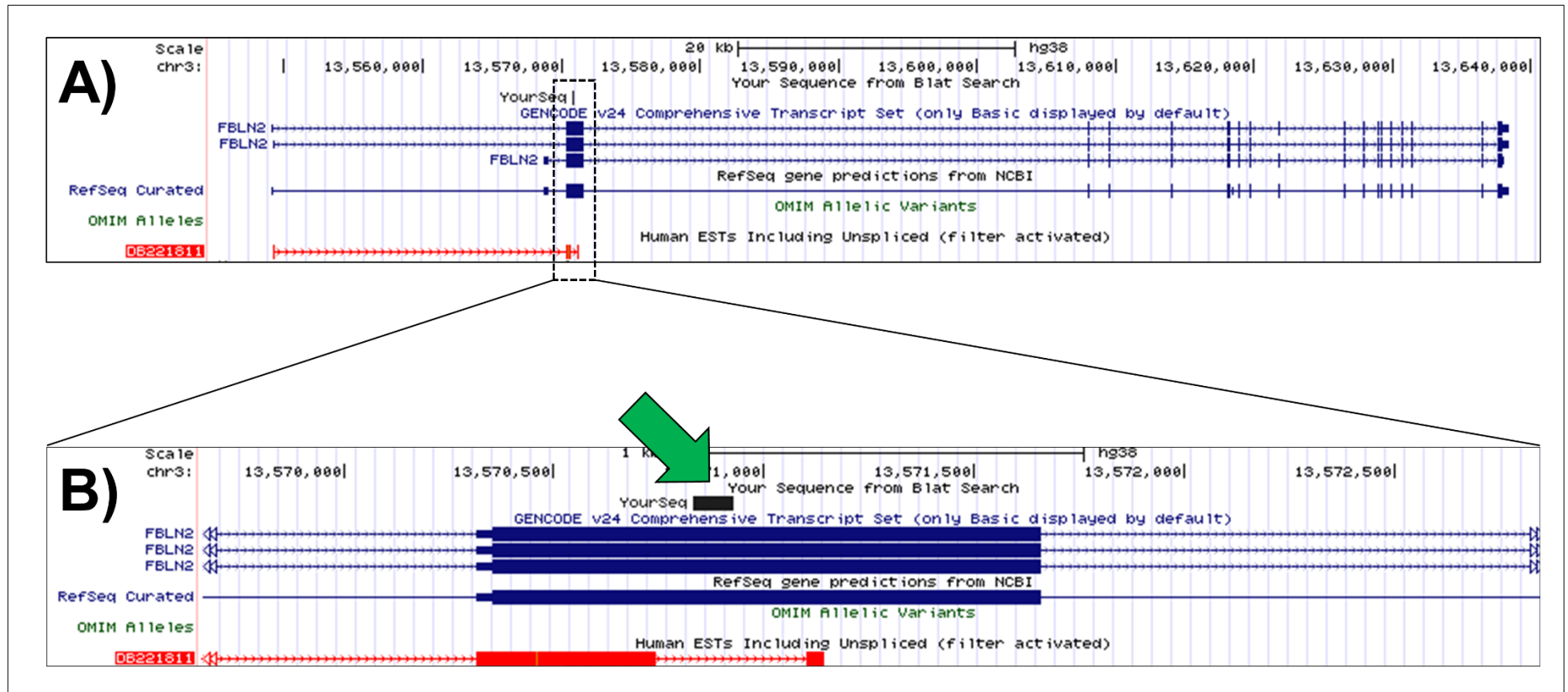
Figure 22: The FBLN2 human gene. **(a)** The overview of the FBLN2 human gene. **(b)** Exon 2 and the riboswitch predicted sequence. The green arrow highlights the predicted region and in red is shown the EST DB221811 lacking this region. **Source:** UCSC Genome Browser (hg38).
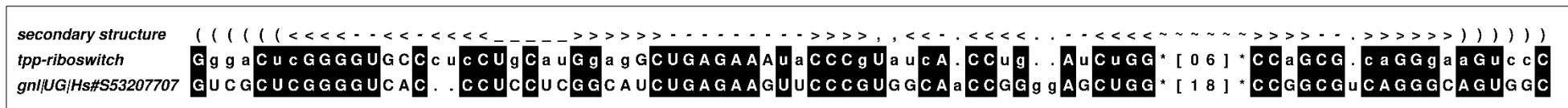
Figure 23: Alignment for hit FBLN2 of the Infernal program. The first line shows the predicted secondary structure, in WUSS format, of the target sequence. Base pairs in simple stem-loops are annotated with "<>" characters. Base pairs enclosing multifurcations (multiple stem-loops) are annotated with "()". For single-stranded residues, characters mark hairpin loops; "-" characters mark interior loops and bulges; "," characters mark single-stranded residues in multifurcation loops. Insertions relative to this consensus are annotated by a "." character. The second and third lines are the CM (*tpp-riboswitch*) and FBLN2 (*gnl|UG|Hs#S53207707*) sequences, respectively. Black filled positions of the alignment represent conserved residues. Dashes (.) in this line indicate deletions in the target sequence concerning the model. The strings *[06]* and *[18]* correspond positions where deletions and insertions occurred, respectively.

4.2.2 Modeling of the three-dimensional structure of the human TPP riboswitch candidate

To gain some insight into the structure of human TPP riboswitch candidate, we built tridimensional models based on the crystallographic structure of TPP riboswitch. TPP riboswitch 3D structures are highly conserved, even in organisms from the different kingdoms (*Escherichia coli* and *Arabidopsis thaliana*) (**Figure 24**). Bearing this in mind and in order to obtain more accurate information about the position of the conserved residues, we performed a multiple alignment using the four different TPP riboswitch sequences and the targets (**Figure 25**). For targets, the P3 and L5 regions were less conserved with the TPP riboswitch sequences.

The named truncated model (CANtrunc), was constructed by utilizing the crystal structure of the TPP riboswitch of *E. coli* (PDB ID: 2GDI), displaying 55% identity, as aligned in **Figure 25**. The CANtrunc and template present high structural conservation with an RMSD of 0.66Å. Interestingly, CANtrunc displays binding arrangement conservation containing the T-loop-like turn formed by the conserved U29-G30-A31-G32-A33 segment. Residues G11-G30 and C64-G65, which interact directly with aminopyrimidine ring of TPP and β-phosphate of TPP, respectively, were also conserved (**Figure 26**). It is noteworthy that all the residues cited were conserved in the multiple alignment between targets and TPP riboswitches.
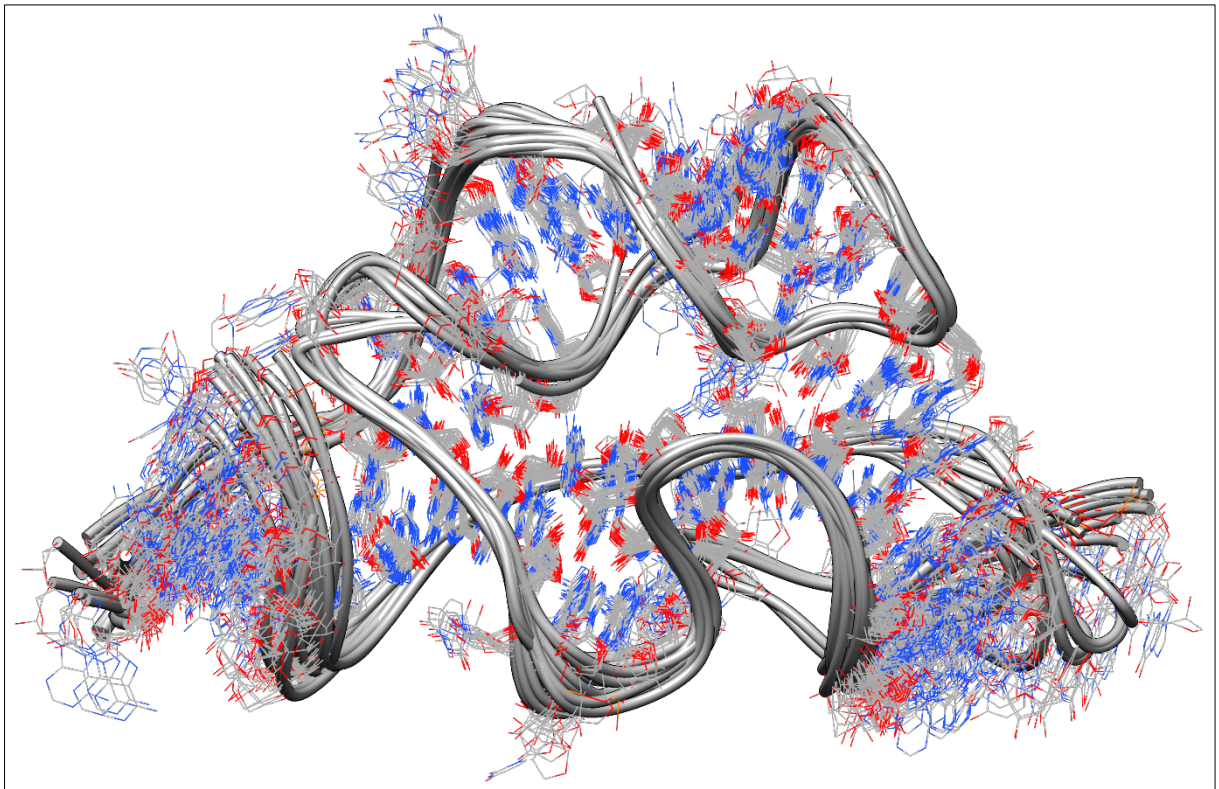
Figure 24: Structural alignment of the 24 structures of TPP riboswitch deposited in the PDB.
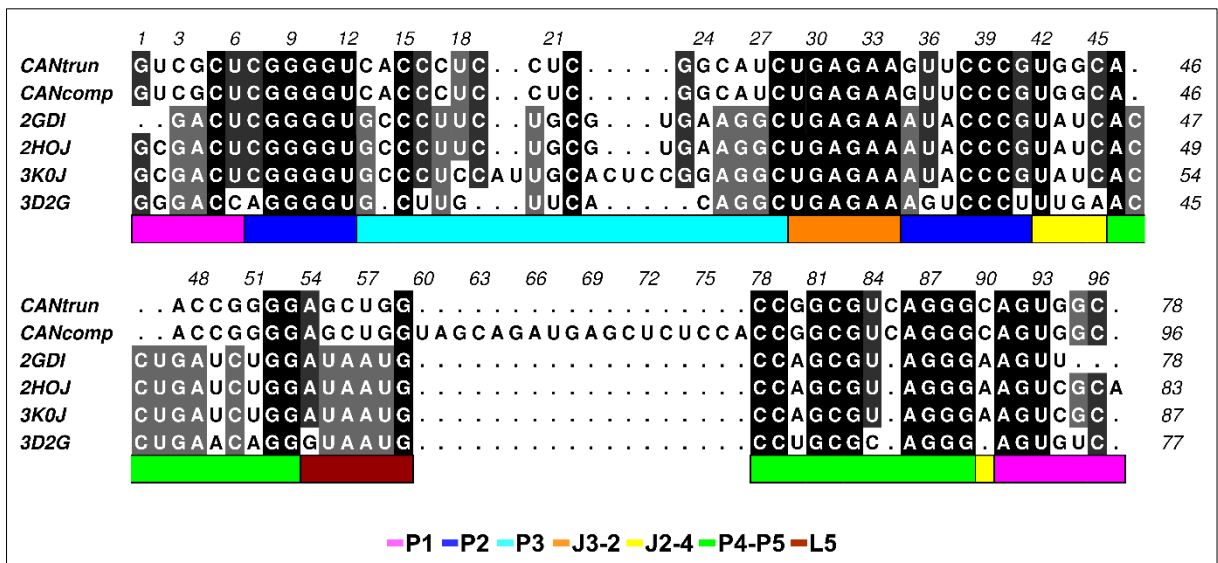
Figure 25: Sequence multiple alignment of TPP riboswitch candidates and templates. Black filled positions of the alignment represent conserved residues. Stems, loops, and junctions were identified according to the legend caption in the figure.
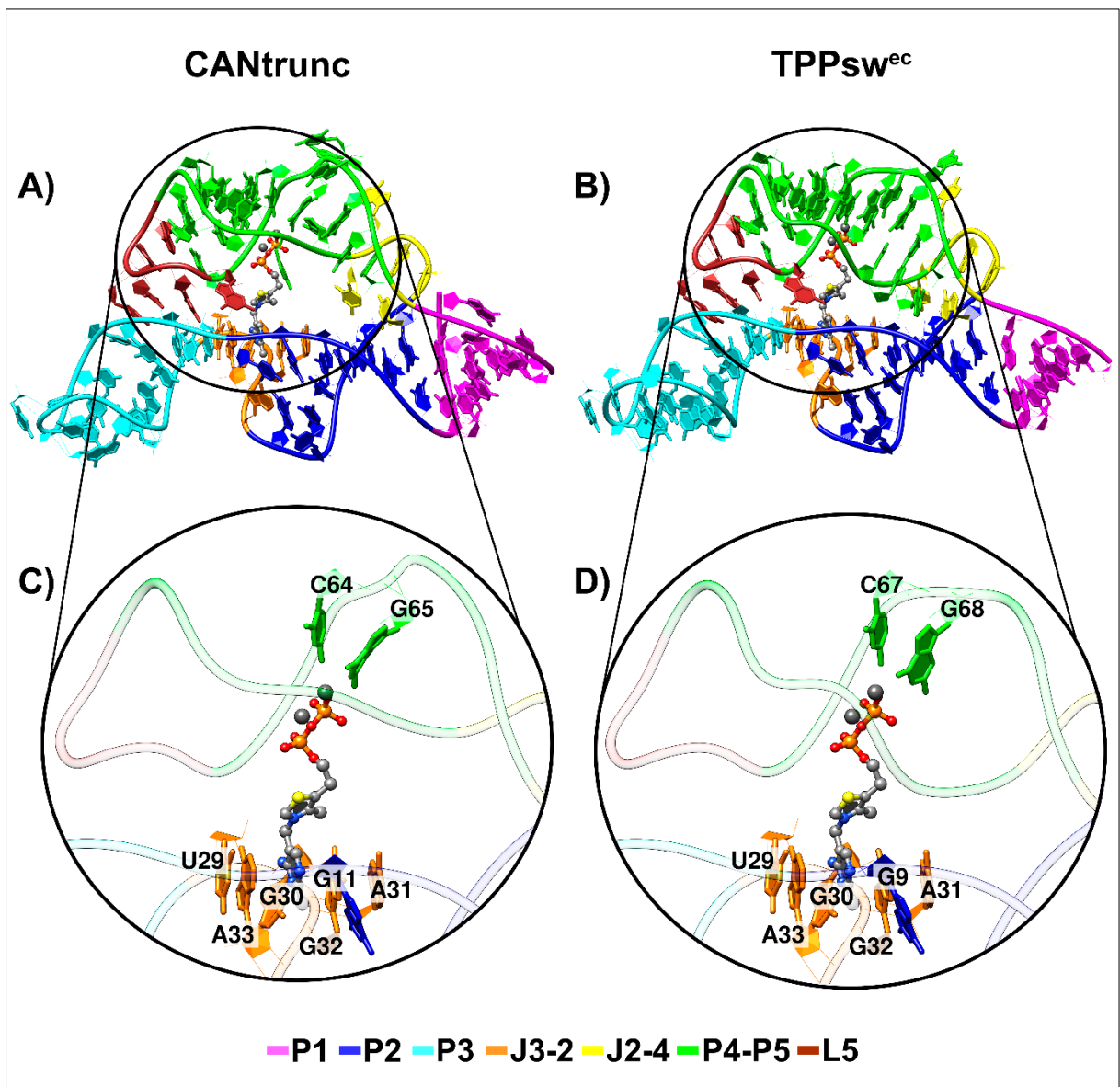
Figure 26: Three-dimensional model of truncated TPP riboswitch candidate (CANtrunc) developed by comparative modelling **(a)**, and crystal structure of the TPP riboswitch of *E. coli* (PDB: 2GDI) used as template **(b)**. TPP binding arrangement is containing T-loop-like turn formed by the conserved U29-G30-A31-G32-A33 segment. Residues G11(9)-G30 and C64(67)-G65(68) interact directly with aminopyrimidine ring of TPP and β-phosphate of TPP, respectively **(c, d)**. Stems, loops, and junctions are figure caption.

The portion of 18 critical nucleotides of the complete models (CANcomp[1] and CANcomp[2]), which was produced by *de novo* method, consists of 10 base pairings divided into two stems (G55, C56, U57, G58, G59, U60 - A77, C78, C79, G80, G81, C82 and; A64, G65, A66 - U72, C73, U74) (**Figure 27AB**).

In CANcomp[1] model, the corresponding region to the crystal structure 2GDI (residues 1 to 59, and 78 to 96) displayed 55% identity (**Figure 25**). For modeling, the crystal structure including the fragment modeled by the *de novo* method (residues 60-77) were used as templates. The RMSD of CANcomp[1] with the crystal structure and fragment modeled by the *de novo* method were 2.96 Å and 0.96 Å, respectively (**Figure 27C**).

For the construction of the CANcomp[2], we used as templates the CANtrunc model (residues 1 to 59, and 78 to 96), and the fragment modeled *de novo* method (residues 60-77). The RMSD of CANcomp[1] with the CANtrunc model and fragment modeled by the de novo method were 1.15 Å and 0.97 Å, respectively (**Figure 27C**).

The main difference between CANcomp[1] and CANcomp[2] is that in the first case the fragment was part of an expanding arm sensor containing P4/P5/fragment, while the second model presented the fragment backwards to the core of the TPP aptamer.

A)

55 57   60   63   66   69   72   75   78   81

G C U G G U A G C A G A U G A G C U C U C C A C C G G C

((((((. (. (((. . . . .)))) .))))))

B)

C)
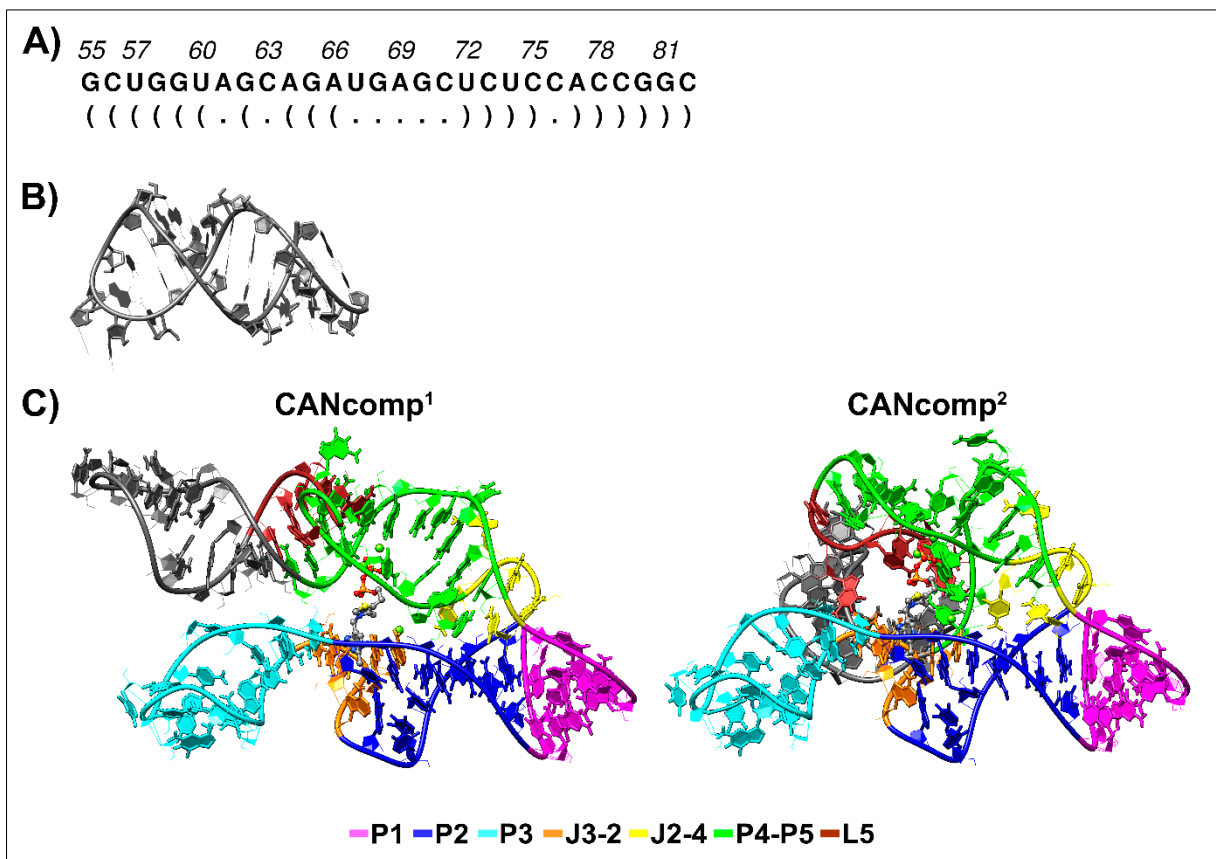
Figure 27: Three-dimensional models of complete TPP riboswitch candidate (CANcomp[1] and CANcomp[2]) developed by comparative modeling. Sequence and structure secondary in dot-bracket-notation **(a)** and tertiary structure **(b)** of the portion of 18 critical nucleotides fragment modeled by the *de novo* method. **(c)** The CANcomp[1] model had the TPP riboswitch of *E. coli* (PDB: 2GDI) along with fragment modeled by the *de novo* method were used as templates. The CANcomp[2] had as templates the CANtrunc model and the fragment modeled *de novo* method. Stems, loops, and junctions are figure caption.

4.2.3 Global and local stability of the human TPP riboswitch candidate structures

The structural stability of CANtrunc, CANcomp[1], and CANcomp[2] in aqueous solution was evaluated by comparing the average RMSD values (**Table 8**) and RMSF (**Figure 28**) calculated over the MD production simulations taking the initial structures as references.

During MD simulations, both CANcomp[1] and CANcomp[2] models presented deviations around 6 Å. CANtrunc showed a smaller variation (5 Å) than the complete models. The *de novo* fragment contributed to these higher values of RMSD, varying in average 6.92 ± 1.99 Å and 4.86 ± 1.46 Å, respectively. The P3 region presented higher RMSD values than for all other substructures it was extremely flexible in the CANcomp[1] (9.20 Å) and CANcomp[2] (9.11 Å).

Table 8: Root mean square deviations (Å) of the human TPP riboswitch candidate systems as a whole and substructures.

	CANtrunc	CANcomp[1]	CANcomp[2]
Whole	5.09 ± 0.77	5.96 ± 0.91	6.09 ± 0.90
P1	4.28 ± 0.82	4.06 ± 1.12	5.30 ± 0.90
P2	4.18 ± 0.88	3.20 ± 0.60	4.03 ± 0.53
P3-L3	5.67 ± 1.13	9.20 ± 1.95	9.11 ± 2.20
P4-P5	5.54 ± 0.85	4.64 ± 0.79	6.03 ± 0.76
J3-2	2.91 ± 0.77	4.40 ± 0.77	4.60 ± 0.79
J2-4	5.39 ± 1.15	3.11 ± 0.52	5.67 ± 0.67
L5	5.57 ± 1.93	5.83 ± 0.92	5.74 ± 079
TPP	3.44 ± 0.58	3.08 ± 0.81	4.54 ± 0.59
Fragment	-	6.92 ± 1.99	4.86 ± 1.46
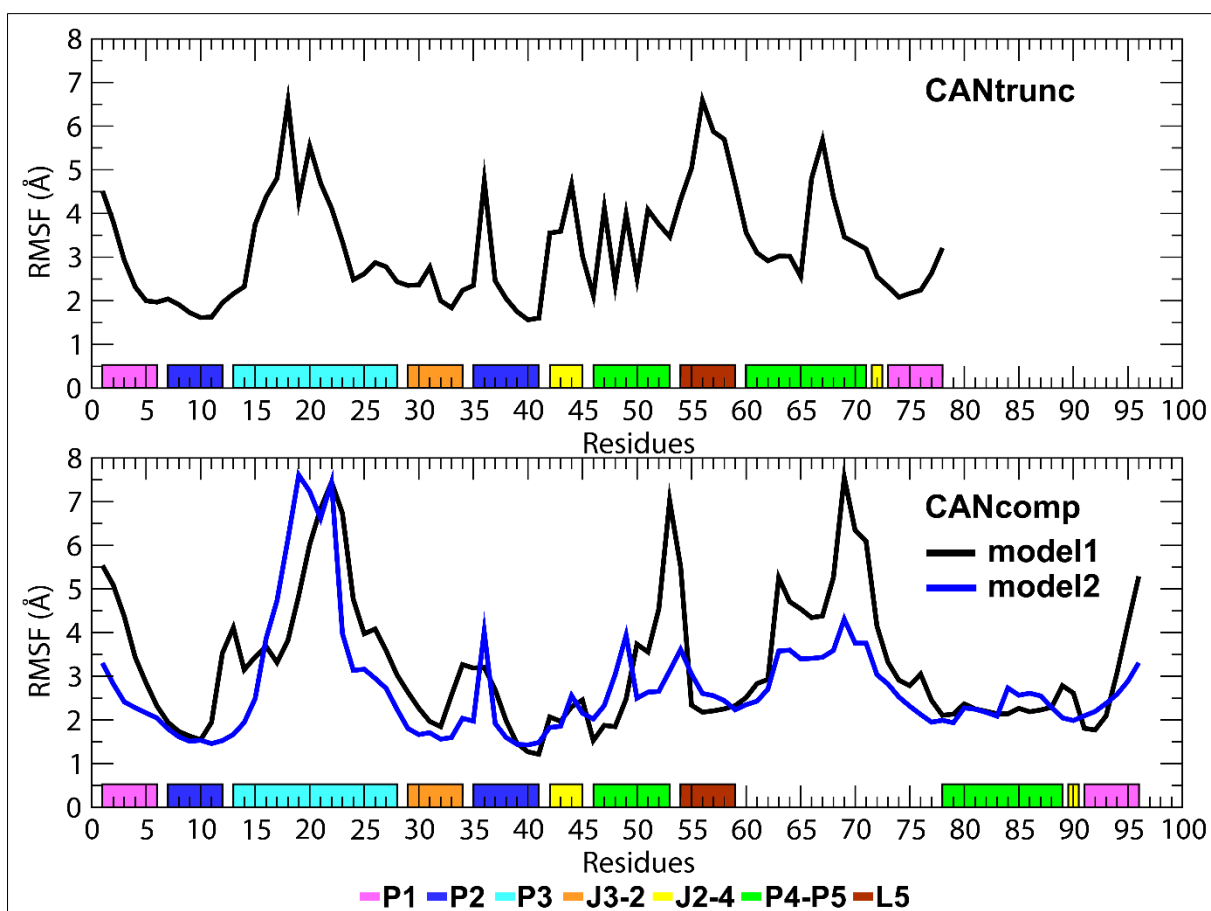
Figure 28: Heavy atoms Root Mean Square Fluctuations of the human TPP riboswitch candidate systems. Secondary structure regions are depicted below the plots and colored according to the figure caption.

4.2.4 TPP-RNA interaction of the human TPP riboswitch candidate structures

The systems presented conservation of the binding site when compared to the crystallographic structure. To verify the stability of RNA-TPP during MD simulations, we calculated the occupancy of hydrogen bonds and the average distances between atoms involved in the RNA-TPP-MG interaction (**Table 9**). Analysis of the distributions of the total number of RNA-TPP hydrogen bonds revealed that most of the conformations presented interactions varying from 3 to 4 in CANtrunc, 2 to 3 in CONcomp[2], and between 1 and 2 in the CONcomp[1].

We confirmed the existence of conserved interactions in the aminopyrimidine ring of TPP formed hydrogen bonds with G30 and the 2-OH' of G11 in CANtrunc and CONcomp[2] systems. The occupancy of the two hydrogen bonds formed between G30 e TPP was similar for CANtrunc and CONcomp[2] systems. Curiously, these frequencies were similar to the ones obtained in the simulations of crystallographic structures of TPP riboswitch (**Table 6**). The frequency of G9(11) for CANtrunc was less than (42.83%) than for CANcomp[2] (52.80%). The latter had a comparable value to the TPPsw[ec] system (55.58%).

The most significant difference of TPP-RNA interaction is related to the direct contacts to non-bridging oxygens of β-phosphate of TPP. Only C64, in the CANtrunc system, interacted directly with the TPP with an occupancy of 72.83%. In the other systems, pyrophosphate-RNA contacts were mediated through two Mg^{2+} ions.

In the CANcomp[1] system, the hydrogen bond occupancy was less than 50%, and the distance between the TPP and RNA binding site nucleotides was higher than 4 Å. Despite few direct contacts, the driving force responsible for maintaining TPP inside the riboswitch binding site was the non-bonded interactions.

Table 9: Occupancy and distance RNA-TPP interaction of human TPP riboswitch candidates.

Atoms		Occupancy (%)			Distance (Å)		
RNA	**TPP**	**CANtrunc**	**CANcomp[1]**	**CANcomp[2]**	**CANtrunc**	**CANcomp[1]**	**CANcomp[2]**
G11– 2O'	TPP – N2	42.83	6.70	52.80	3.07 ± 0.45	4.38 ± 1.07	2.89 ± 0.15
G30 – N2	TPP – N1	87.52	47.44	89.77	3.07 ± 0.17	4.27 ± 0.99	3.07 ± 0.15
G30 – N3	TPP – N	61.79	31.61	59.42	3.00 ± 0.13	3.98 ± 1.45	2.09 ± 0.11
C64 – N4*	TPP – O1	72.83	0	0	4.45 ± 2.16	7.34 ± 0.42	6.87 ± 0.58
MG1	MG2	-	-	-	5.89 ± 0.87	4.60 ± 0.14	3.48 ± 0.07

* C82 – N4 in the CANcomp systems.

4.2.5 Principal component analysis of the human TPP riboswitch candidate structures

We performed PCA using the MD trajectories to compare statistically relevant motions of candidates with crystallographic structure. The first three PCs accounted for 69.1%, 61.6% and 58.7% of the overall variance in CANtrunc, CANcomp[1] and CANcomp[2], respectively. RMSFs and structural projections along the three PCs are shown in **Figure 29**.

Unlike the *holo* systems of TPPsw[ec] and TPPsw[at] (**Figure 17**), the presence of the ligand in the candidates caused minor fluctuations in the P1 helix. TPP binding resulted in increased flexibility of P3 helix in all candidates. CANcomp[2] system presented higher fluctuations in the P3 helix than the other candidates. This structural feature was also found in the unbound TPPsw[ec] and TPPsw[at] systems.

CANtrunc system revealed substantial high flexibility of P4-P5-L5 regions. These regions presented lowest sequence identity with the crystallographic structure. For the complete models, the region that presented higher fluctuations along PC1-2 in the CANcomp[1] system corresponded to the supplementary fragment. In CANcomp[2], this fluctuation was less pronounced.
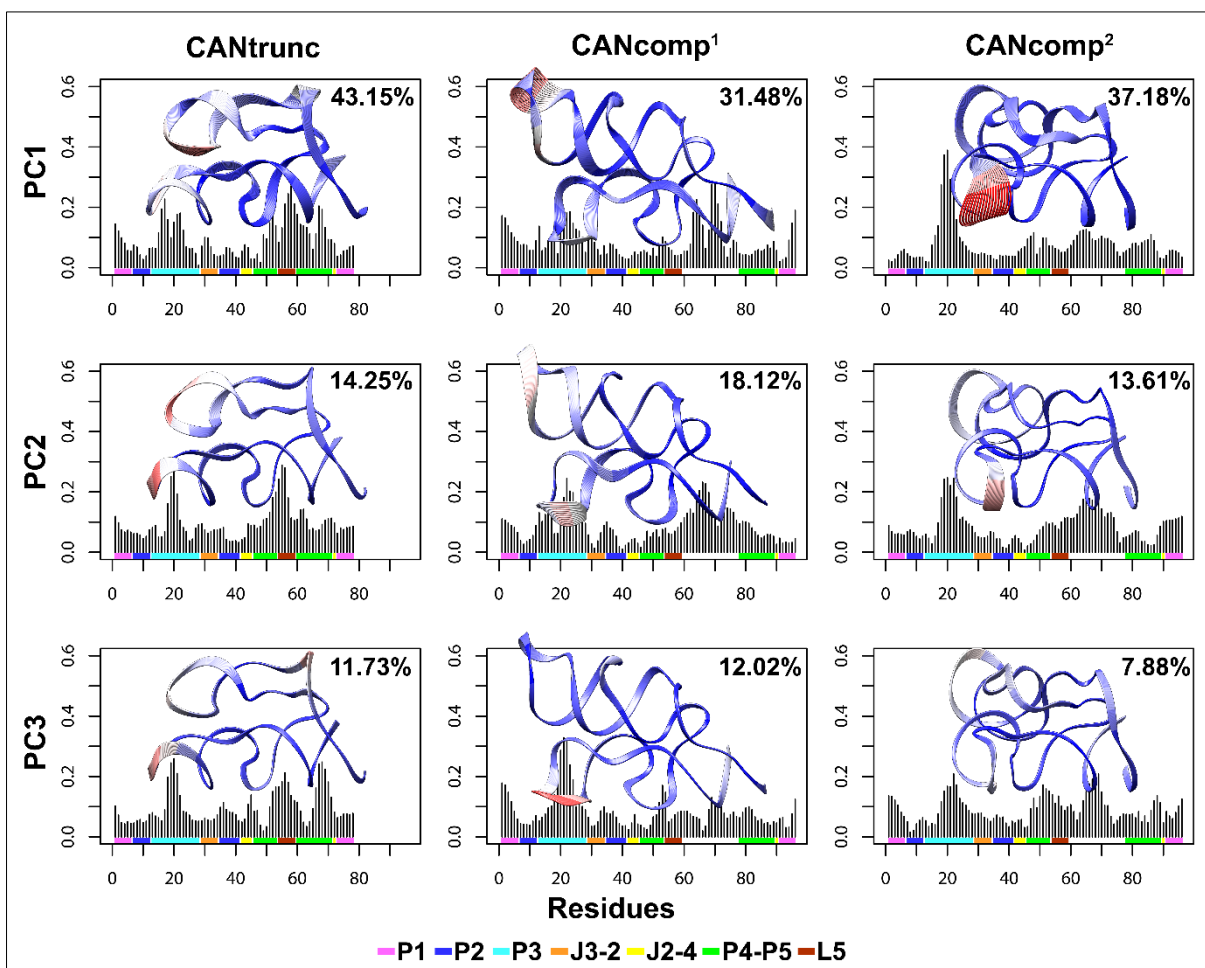
Figure 29: RMSF along the first three principal components. The fraction of variance captured each component is shown in the top right part of each graph. Interpolated structures obtained by displacements along each vector are displayed within each graph. Blue indicates overlapping regions with little or no motion. Red areas represent mobile regions. The secondary structure elements are given in the lower margin of the plots and colored according to the figure caption.

73

4.2.6 Correlation network analysis of the Principal component analysis of the human TPP riboswitch candidate structures

We analyzed the correlations between pairs of nucleotides to investigate how similar are the candidate systems with the crystallographic structure. The matrices of the candidates presented a correlation pattern similar to each other and to the unbound TPPsw[ec] (**Figure 30** and **Figure 18**).

The existence of the fragment in CANcomp[1] and CANcomp[2] contributed to increase anticorrelations. While in CANtrunc the P3–P4-P5–L5 region displayed anticorrelations, CANcomp1 and CANcomp2 anticorrelations included the fragment (P3–P4-P5–L5–fragment).

We calculated the SIP to compare the overall similarity of the betweenness centrality profiles for the candidate systems (**Figure 31**). We obtained a higher SIP among the complete candidates (0.83) than for the truncated one. CANtrunc had a SIP of 0.80 and 0.74 with CANcomp1 and CANcomp2, respectively. According to this analysis, high SIP values are associated with a similar communication among the models.

When comparing the betweenness centrality of candidates with TPPsw[ec] and TPPsw[at] systems (**Figure 19**), we noticed increased centrality values at the last 19 residues about TPPsw[ec] and TPPsw[at] systems. This suggests a model of different action in the communication through these nucleotides for candidates and TPP aptamer structures.
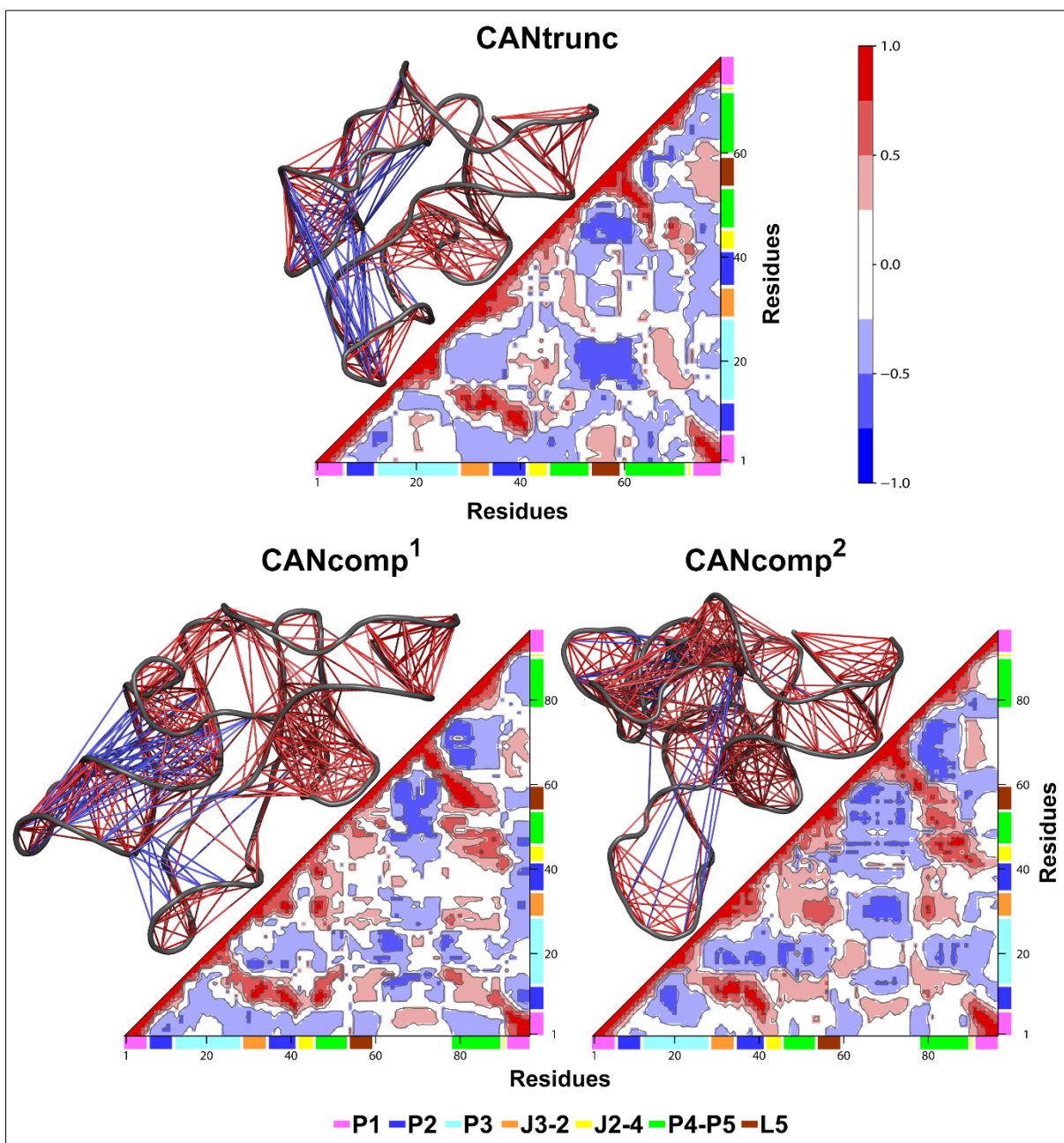
Figure 30: DCCMs of the human candidates TPP riboswitch. Next to each matrix, the corresponding 3D structures with lines connecting pairs of correlated residues are shown. For clarity sake, only the pairs presenting ($|C_{ij}|$) > 0.6. are represented.
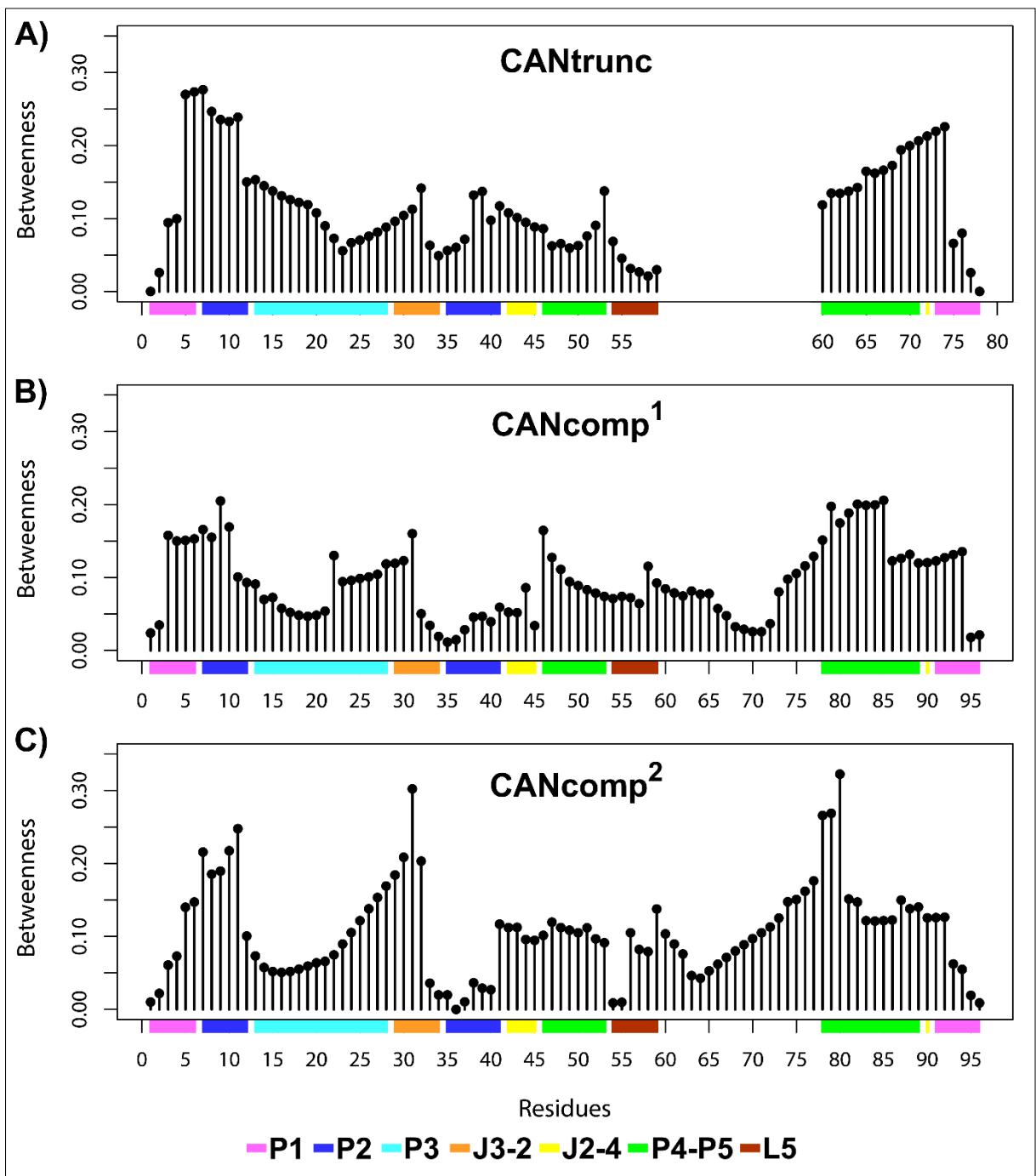
Figure 31: Betweenness centrality of the node for each residue of the human candidates TPP riboswitch.

5 DISCUSSION

5.1 Evidence from complementary methods suggests that TPP riboswitch of plants present more subtle regulation mechanisms than bacteria

Plant and bacterial riboswitches TPP aptamers share similar core structures and bind to the same ligand. However, minor structural and dynamical differences between them can be found, especially concerning the behavior of P3 helix. Particularly in plants, the length of distal P3 extension varies among TPP aptamer representatives of the same species, as observed in *Physcomitrella patens* (61). The P3 distal portion is not required for ligand binding or L5-P3 interaction (148,191), but might act as an anchor for the aptamer as already pointed out by Anthony et al. (192). Also, the authors claim that the correct folding could help in the competition with other RNA structures with different regulation mechanisms.

However, despite the P3 stem significantly variable in length in plants, the TPP aptamer is structurally stable. This stability might lead to slower arm movement than the observed in the helix arm of *E. coli* TPP riboswitch (193). Cross-correlation analysis corroborates this hypothesis because stronger negative correlations were noticed in the *apo* TPPsw[ec] involving substructures P3–P4-P5–L5 (**Figure 18**). Our findings also suggest that communication pathways between P3-L5 may be different in E. *coli* and *A. thaliana*. The communication between P3-L5 in TPPsw[ec] can be very efficient in the *holo* state, while in TPPsw[at] the corresponding effect was weakened, thus suggesting a slower response to TPP binding in plants than in bacteria (**Figure 20**).

The *A. thaliana* crystallographic structure used as starting point for our simulations contains a shortened P3 stem formed by 14 nucleotides. On the other hand, the corresponding structure in *E. coli* is composed of 18 nucleotides. The P3 helix of TPPsw[at], although smaller than the one in TPPsw[ec], showed no significant modifications in the presence of the ligand, indicating that the size of P3 can be oblivious to plants and its influence about slow folding can be negligible.

Guedich et al. wondered whether the slow TPPsw[at] folding would be related to a single nucleotide. The authors concluded that U35, located on the P2 helix, is crucial for shaping a TPP-binding competent riboswitch (194). In our analysis, the equivalent pyrimidine nucleotide in TPPsw[ec] is U36 (**Figure 12B**). The magnitude of the fluctuations at this position was 2-fold higher in the *holo* state than in the *apo* TPPsw[ec]

state (**Figure 14**). In contrast, in the TPPsw[at] system, similar fluctuations were perceived regardless of a ligand binding. Furthermore, PCA data also supported these outcomes by showing that the segment 34-37 of the *holo* TPPsw[ec] displayed the most significant motion amplitude along PC1 (**Figure 17**).

Grounded on these findings, we hypothesize that different interactions found in the microenvironment surrounding nucleotide U36 of TPPsw[ec] (and U35 in TPPsw[at]) are related to different TPP responses. In TPPsw[ec], this nucleotide is neighbored at 3' by a non-canonical A37-G9 base pair. A similar context is observed for U35 in TPPsw[at], which is delimited by a non-canonical G34-G11 base pair but on 5' instead. Nucleotides G9 and G11 of TPPsw[ec] and TPPsw[at], respectively, form hydrogen bonds with the aminopyrimidine ring of TPP. Interestingly, our simulations have shown that hydrogen bond occupancy between G9(11) and N2 of TPP was less than a half for TPPsw[at] (25.67%) than for TPPsw[ec] (55.58%); this suggests that slight differences in the environment may directly interfere the stability of TPP-aptamer interactions.

Finally, TPP riboswitches of *Arabidopsis thaliana* present subtler and slower regulation mechanisms than *Escherichia coli* (192–194). Here, we have shown through molecular dynamics simulations and networking analysis that minor structural differences in the aptamer enable enhanced intramolecular communication in the presence of TPP in TPPsw[ec], but not in TPPsw[at]. Weaker responses to changes in the TPP concentration may be related to the autotrophic mode of nutrition, which demands the endogenous synthesis of thiamine. Unlike in plants, bacteria can grow under rich conditions that allow them to satisfy their full demand for compounds like thiamine exogenously (61). In this way, we provide new insights into RNA behavior of TPP riboswitch, which may have adapted to the different metabolic demands of each group of organisms to accomplish distinct TPP binding modulation.

5.2 Evidence from complementary methods suggests the existence of a potential TPP riboswitch candidate in the human genome

In the particular case of riboswitches, a single RNA sequence is capable of adopting, at least, two stable secondary structures to regulate the expression of a given gene. These structures are conserved throughout evolution despite sequence variations (195). There are several metabolites capable of binding to different riboswitches, but to date, the only riboswitch described in eukaryotes is regulated by TPP (60).

Riboswitches have the potential to act as targets for antibiotic and chemotherapeutic drugs and can be developed for the creation of transgenic organisms. For example, the pyrithiamine, an analogue of thiamine, has been determined to be toxic to bacteria and fungi, targeting TPP riboswitches and thereby repressing thiamine biosynthesis. Thus, it becomes essential to investigate the existence of riboswitches described in the human genome to avoid unwanted interactions. Different computational tools were developed to search for novel riboswitches; this allows the identification of robust candidates before experimental validation is made.

Through bioinformatics analysis, we identified a potential candidate for TPP riboswitch in the human genome located in the FBLN2 gene. This gene belongs to the fibulin family that encodes an extracellular matrix protein. Fibulin 2 binds calcium and other various extracellular ligands. This protein may play a role in organ development, in particular, during the differentiation of heart, skeletal and neuronal structures (196).

Interestingly, the predicted FBLN2 riboswitch has overlap with an EST that lacks this region (**Figure 22B**) and should be considered when this candidate will be studied in depth. This splice variant of FBLN2 may show an alternative way to bypass the riboswitch influence in transcripts of this gene.

Riboswitches control the expression of several genes involved in the transport and biosynthesis of critical metabolites for bacteria, fungi, and plants. Our results showed no relation to the metabolism of thiamine in humans. However, the structural and dynamic analysis of FBLN2 revealed that, even in simulations of 1 μs extension, TPP remained bound to RNA.

Of the three FBLN2 models created, CANtrunc was the one that best resembled the crystallographic structure. This fact is due to the model and template have the same

number of residues (78 nt). This feature favored a better preservation of the TPP binding site. The hydrogen bond analysis (**Table 9**) evidenced that CANtrunc maintained specific RNA interactions with the aminopyrimidine ring and pyrophosphate of TPP.

The inclusion of the 18 nt long fragment may have influenced the loss or decrease of the direct interaction between the ligand and those nucleotides considered as conserved in the TPP riboswitch class. The CANcomp[1] system did not maintain the specific bonds with the ligand, which prompted us to discard this structure. Moreover, it was observed that the presence of the fragment had little or none influence in the region interacting with the aminopyrimidine ring of TPP in the CANcomp[2] model.

The fragment is located in the P4/P5 arm that interacts with TPP pyrophosphate group. The addition of 18 nt caused increased flexibility in this arm, as demonstrated in the results of RMSF and PCA (**Figure 28** and **Figure 29**), and indirect influence of the interaction of RNA-pyrophosphate of TPP.

FBLN2 can be considered as a possible candidate for TPP riboswitch but depends on experimental evaluation to be confirmed. In-line probing experiments could be conducted to get insights into the secondary structure of RNA, and surface plasmon resonance spectroscopy essays must be performed to evaluate the affinity of metabolite binding to riboswitch aptamer domain.

Combination of experimental and in silico methods have proved of great value. The hybridity of bioinformatics with In-line experiments and surface plasmon resonance spectroscopy were employed in TPP riboswitch characterization in *Alishewanella tabrizica* and *Alishewanella aestuarii* (197). We used molecular modeling to analyze structural behavior in addition to Bioinformatics to identify a TPP riboswitch candidate. This approach proved valid to generate insights into the structure and dynamics of the target and could be applied to other structural RNA motifs.

6 CONCLUSIONS

- Distinct interactions found in the microenvironment surrounding nucleotide U36 of TPPsw^{ec} (and U35 in TPPsw^{at}) are related to different responses to TPP.

- The slight differences in the environment directly interfered in the stability of TPP-aptamer interactions by altering the hydrogen bonding pattern.

- The communication between P3-L5 in TPPsw^{ec} was very efficient in the *holo* state, while in TPPsw^{at} the corresponding effect was weakened.

- The P3 helix of TPPsw^{at} showed no significant modifications in the presence of the ligand, suggesting that the size of P3 can be oblivious to plants and its influence about slow folding can be negligible.

- The networking analysis showed that minor structural differences in the aptamer enable efficient intramolecular communication in the presence of TPP in TPPsw^{ec}, but not in TPPsw^{at}.

- FBLN2 gene was identified as a potential candidate for TPP riboswitch in the human genome.

- CANtrunc model was the one that best resembled the crystallographic structure and maintained specific RNA interactions with the aminopyrimidine ring and pyrophosphate of TPP.

- The CANcomp^{1} system did not maintain the specific intermolecular bonds with the ligand, which allowed us to discard this structure as a potential candidate.

- CANcomp^{2} system maintained the particular bonds with the aminopyrimidine ring of TPP.0

- The presence of the fragment destabilizes the P4/P5 arm and indirectly influences the interaction of RNA with pyrophosphate group of TPP.

- FBLN2 can be considered as a possible candidate for TPP riboswitch. Notwithstanding, an experimental evaluation must be accomplished to be confirmed as one.

7 REFERENCES

1. Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561–3.

2. Albert B, Bray D, Hopkin K. Fundamentos da Biologia Celular. Artmed; 2011. 864 p.

3. Mattick JS. Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep. 2001 Nov;2(11):986–91.

4. Qu Z, Adelson DL. Evolutionary conservation and functional roles of ncRNA. Front Genet. 2012;3:205.

5. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res. 2011;39(suppl 1):D146–51.

6. Barrick JE, Breaker RR. The distributions, mechanisms, and structures of metabolite-binding riboswitches. Genome Biol. 2007;8(11):R239.

7. Edwards AL, Batey RT. Riboswitches: A common RNA regulatory element. Nat Educ. 2010;3(9):9.

8. Breaker RR. Riboswitches and the RNA world. Cold Spring Harb Perspect Biol. 2012;4(2):a003566.

9. VITRESCHAK AG. Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. RNA. 2003 Sep;9(9):1084–97.

10. Winkler W, Nahvi A, Breaker RR. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature [Internet]. 2002 Oct 31 [cited 2016 Sep 17];419(6910):952–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12410317

11. Winkler WC, Cohen-Chalamish S, Breaker RR. An mRNA structure that controls gene expression by binding FMN. Proc Natl Acad Sci. 2002 Dec;99(25):15908–13.

12. Winkler WC, Nahvi A, Sudarsan N, Barrick JE, Breaker RR. An mRNA structure that controls gene expression by binding S-adenosylmethionine. Nat Struct Biol. 2003 Sep;10(9):701–7.

13. Wang JX, Lee ER, Morales DR, Lim J, Breaker RR. Riboswitches that Sense S-adenosylhomocysteine and Activate Genes Involved in Coenzyme Recycling. Mol Cell. 2008 Mar;29(6):691–702.

14. Ames TD, Rodionov DA, Weinberg Z, Breaker RR. A Eubacterial Riboswitch Class That Senses the Coenzyme Tetrahydrofolate. Chem Biol. 2010

Jul;17(7):681–5.

15. Regulski EE, Moy RH, Weinberg Z, Barrick JE, Yao Z, Ruzzo WL, et al. A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism. Mol Microbiol. 2008 May;68(4):918–32.

16. Mandal M, Breaker RR. Adenine riboswitches and gene activation by disruption of a transcription terminator. Nat Struct Mol Biol. 2004 Jan;11(1):29–35.

17. Batey RT, Gilbert SD, Montange RK. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. Nature. 2004 Nov;432(7015):411–5.

18. Roth A, Winkler WC, Regulski EE, Lee BWK, Lim J, Jona I, et al. A riboswitch selective for the queuosine precursor preQ1 contains an unusually small aptamer domain. Nat Struct Mol Biol. 2007 Apr;14(4):308–17.

19. Wacker A, Buck J, Mathieu D, Richter C, Wohnert J, Schwalbe H. Structure and dynamics of the deoxyguanosine-sensing riboswitch studied by NMR-spectroscopy. Nucleic Acids Res. 2011 Aug;39(15):6802–12.

20. Sudarsan N, Lee ER, Weinberg Z, Moy RH, Kim JN, Link KH, et al. Riboswitches in Eubacteria Sense the Second Messenger Cyclic Di-GMP. Science (80-). 2008 Jul;321(5887):411–3.

21. Nelson JW, Sudarsan N, Furukawa K, Weinberg Z, Wang JX, Breaker RR. Riboswitches in eubacteria sense the second messenger c-di-AMP. Nat Chem Biol. 2013 Oct;9(12):834–9.

22. Serganov A, Huang L, Patel DJ. Structural insights into amino acid binding and gene control by a lysine riboswitch. Nature. 2008 Oct;455(7217):1263–7.

23. Mandal M. A Glycine-Dependent Riboswitch That Uses Cooperative Binding to Control Gene Expression. Science (80-). 2004 Oct;306(5694):275–9.

24. Ames TD, Breaker RR. Bacterial aptamers that selectively bind glutamine. RNA Biol. 2011 Jan;8(1):82–9.

25. Cromie MJ, Groisman EA. Promoter and Riboswitch Control of the Mg2+ Transporter MgtA from Salmonella enterica. J Bacteriol. 2010 Jan;192(2):604–7.

26. Chawla M, Credendino R, Poater A, Oliva R, Cavallo L. Structural Stability, Acidity, and Halide Selectivity of the Fluoride Riboswitch Recognition Site. J Am Chem Soc. 2015 Jan;137(1):299–306.

27. Klein DJ. Structural Basis of glmS Ribozyme Activation by Glucosamine-6-

Phosphate. Science (80-). 2006 Sep;313(5794):1752–6.

28. Peselis A, Serganov A. Themes and variations in riboswitch structure and function. Biochim Biophys Acta [Internet]. 2014 Oct [cited 2016 Sep 17];1839(10):908–18. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24583553

29. Antunes D, Jorge NAN, Caffarena ER, Passetti F. Using RNA Sequence and Structure for the Prediction of Riboswitch Aptamer: A Comprehensive Review of Available Software and Tools. Front Genet [Internet]. 2018 Jan 19 [cited 2018 Jun 9];8:231. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2017.00231/full

30. Mandal M, Breaker RR. Gene regulation by riboswitches. Nat Rev Mol Cell Biol [Internet]. 2004 Jun [cited 2016 Sep 17];5(6):451–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15173824

31. Serganov A, Nudler E. A Decade of Riboswitches. Cell [Internet]. 2013 Jan;152(1–2):17–24. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867412015462

32. Garst AD, Batey RT. A switch in time: detailing the life of a riboswitch. Biochim Biophys Acta. 2009;1789(9–10):584–91.

33. Tucker BJ, Breaker RR. Riboswitches as versatile gene control elements. Curr Opin Struct Biol [Internet]. 2005 Jun [cited 2016 Sep 17];15(3):342–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15919195

34. Hammann C, Westhof E. Searching genomes for ribozymes and riboswitches. Genome Biol [Internet]. 2007 [cited 2016 Sep 17];8(4):210. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17472738

35. Ottink OM, Rampersad SM, Tessari M, Zaman GJR, Heus HA, Wijmenga SS. Ligand-induced folding of the guanine-sensing riboswitch is controlled by a combined predetermined induced fit mechanism. RNA. 2007 Dec;13(12):2202–12.

36. Haller A, Rieder U, Aigner M, Blanchard SC, Micura R. Conformational capture of the SAM-II riboswitch. Nat Chem Biol. 2011 Jun;7(6):393–400.

37. Coppins RL, Hall KB, Groisman EA. The intricate world of riboswitches. Curr Opin Microbiol. 2007 Apr;10(2):176–81.

38. Garst AD, Edwards AL, Batey RT. Riboswitches: structures and mechanisms. Cold Spring Harb Perspect Biol [Internet]. 2011 Jun [cited 2016 Sep 17];3(6). Available from: http://www.ncbi.nlm.nih.gov/pubmed/20943759

39. Reining A, Nozinovic S, Schlepckow K, Buhr F, Fürtig B, Schwalbe H. Three-state mechanism couples ligand and temperature sensing in riboswitches. Nature. 2013 Jul;499(7458):355–9.

40. Nahvi A. Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes. Nucleic Acids Res. 2004 Jan;32(1):143–50.

41. Lang K, Rieder R, Micura R. Ligand-induced folding of the thiM TPP riboswitch investigated by a structure-based fluorescence spectroscopic approach. Nucleic Acids Res. 2007 Aug;35(16):5370–8.

42. Lai EC. RNA Sensors and Riboswitches: Self-Regulating Messages. Curr Biol. 2003 Apr;13(7):R285–91.

43. Nudler E, Mironov AS. The riboswitch control of bacterial metabolism. Trends Biochem Sci [Internet]. 2004 Jan [cited 2016 Sep 17];29(1):11–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14729327

44. Thore S, Leibundgut M, Ban N. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. Science [Internet]. 2006 May 26 [cited 2016 Sep 17];312(5777):1208–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16675665

45. Machtel P, Bąkowska-Żywicka K, Żywicki M. Emerging applications of riboswitches – from antibacterial targets to molecular tools. J Appl Genet. 2016 Mar;1–11.

46. Hollands K, Proshkin S, Sklyarova S, Epshtein V, Mironov A, Nudler E, et al. Riboswitch control of Rho-dependent transcription termination. Proc Natl Acad Sci U S A. 2012 Apr;109(14):5376–81.

47. Gilbert SD, Rambo RP, Van Tyne D, Batey RT. Structure of the SAM-II riboswitch bound to S-adenosylmethionine. Nat Struct Mol Biol. 2008 Feb;15(2):177–82.

48. Thomas AA, Le Huerou Y, De Meese J, Gunawardana I, Kaplan T, Romoff TT, et al. Synthesis, in vitro and in vivo activity of thiamine antagonist transketolase inhibitors. Bioorganic Med Chem Lett. 2008;18(6):2206–10.

49. Murray RK, Granner DK, Mayes PA, Rodwell VW. Harper's Illustrated Biochemistry [Internet]. Vol. 16, Molecular Physiology. 2003. 693 p. Available from: http://www.mhprofessional.com/product.php?cat=39&isbn=0071625917

50. Finglas PM. Dietary Reference intakes for thiamin, riboflavin, niacin, vitamin B6, folate, vitamin B12, pantothenic acid, biotin and choline. Trends Food Sci Technol [Internet]. 2000;11(8):296–7. Available from:

http://linkinghub.elsevier.com/retrieve/pii/S0924224401000103

51. Prinzo ZW. Thiamine deficiency and its prevention and control in major emergencies. Micronutr Ser [Internet]. 1999 [cited 2018 Aug 7]; Available from: http://bases.bireme.br/cgi-bin/wxislind.exe/iah/online/?IsisScript=iah/iah.xis&src=google&base=DESAST RES&lang=p&nextAction=lnk&exprSearch=12400&indexSearch=ID

52. Whitfield KC, Smith G, Chamnan C, Karakochuk CD, Sophonneary P, Kuong K, et al. High prevalence of thiamine (vitamin B1) deficiency in early childhood among a nationally representative sample of Cambodian women of childbearing age and their children. Tickell KD, editor. PLoS Negl Trop Dis [Internet]. 2017 Sep 5 [cited 2018 Aug 7];11(9):e0005814. Available from: http://dx.plos.org/10.1371/journal.pntd.0005814

53. Collie JTB, Greaves RF, Jones OAH, Lam Q, Eastwood GM, Bellomo R. Vitamin B1 in critically ill patients: needs and challenges. Clin Chem Lab Med [Internet]. 2017 Jan 26 [cited 2018 May 4];55(11):1652–68. Available from: http://www.degruyter.com/view/j/cclm.2017.55.issue-11/cclm-2017-0054/cclm-2017-0054.xml

54. Comín-Anduix B, Boren J, Martinez S, Moro C, Centelles J, Trebukhina R, et al. The effect of thiamine supplementation on tumour proliferation. A metabolic control analysis study. Eff thiamine Suppl tumour proliferation A Metab Control Anal study [Internet]. 2001;268(15):4177–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11488910

55. Winkler W, Nahvi A, Breaker RR. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature [Internet]. 2002;419(6910):952–6. Available from: http://www.nature.com/doifinder/10.1038/nature01145

56. Ontiveros-Palacios N, Smith AM, Grundy FJ, Soberon M, Henkin TM, Miranda-Ríos J. Molecular basis of gene regulation by the THI-box riboswitch. Mol Microbiol. 2008;67(4):793–803.

57. Kubodera T, Watanabe M, Yoshiuchi K, Yamashita N, Nishimura A, Nakai S, et al. Thiamine-regulated gene expression of< i> Aspergillus oryzae thiA</i> requires splicing of the intron containing a riboswitch-like domain in the 5′-UTR. FEBS Lett. 2003;555(3):516–20.

58. Li S, Breaker RR. Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. Nucleic Acids Res. 2013;41(5):3022–31.

59. Croft MT, Moulin M, Webb ME, Smith AG. Thiamine biosynthesis in algae is regulated by riboswitches. Proc Natl Acad Sci. 2007;104(52):20770–5.

60. Bocobza S, Adato A, Mandel T, Shapira M, Nudler E, Aharoni A. Riboswitch-dependent gene regulation and its evolution in the plant kingdom Riboswitch-dependent gene regulation and its evolution in the plant kingdom. 2007;2874–9.

61. Wachter A, Tunc-Ozdemir M, Grove BC, Green PJ, Shintani DK, Breaker RR. Riboswitch control of gene expression in plants by splicing and alternative 3′ end processing of mRNAs. Plant Cell Online. 2007;19(11):3437–50.

62. Mukherjee S, Retwitzer MD, Barash D, Sengupta S. Phylogenomic and comparative analysis of the distribution and regulatory patterns of TPP riboswitches in fungi. Sci Rep [Internet]. 2018 Dec 3 [cited 2018 May 13];8(1):5563. Available from: http://www.nature.com/articles/s41598-018-23900-7

63. Kubodera T, Watanabe M, Yoshiuchi K, Yamashita N, Nishimura A, Nakai S, et al. Thiamine-regulated gene expression of Aspergillus oryzae thiA requires splicing of the intron containing a riboswitch-like domain in the 5'-UTR. FEBS Lett [Internet]. 2003 Dec 18 [cited 2016 Sep 17];555(3):516–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14675766

64. Li S, Breaker RR. Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. Nucleic Acids Res [Internet]. 2013 Mar 1 [cited 2016 Sep 17];41(5):3022–31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23376932

65. Bocobza SE, Aharoni A. Switching the light on plant riboswitches. Trends Plant Sci. 2008 Oct;13(10):526–33.

66. Croft MT, Moulin M, Webb ME, Smith AG. Thiamine biosynthesis in algae is regulated by riboswitches. Proc Natl Acad Sci U S A [Internet]. 2007 Dec 26 [cited 2016 Sep 17];104(52):20770–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18093957

67. Bocobza S, Adato A, Mandel T, Shapira M, Nudler E, Aharoni A. Riboswitch-dependent gene regulation and its evolution in the plant kingdom. Genes Dev [Internet]. 2007 Nov 15 [cited 2016 Sep 17];21(22):2874–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18006684

68. Wachter A, Tunc-Ozdemir M, Grove BC, Green PJ, Shintani DK, Breaker RR. Riboswitch control of gene expression in plants by splicing and alternative 3' end processing of mRNAs. Plant Cell [Internet]. 2007 Nov [cited 2016 Sep

17];19(11):3437–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17993623

69. Bocobza SE, Aharoni A. Small molecules that interact with RNA: riboswitch-based gene control and its involvement in metabolic regulation in plants and algae. Plant J. 2014 Aug;79(4):693–703.

70. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: Updates to the RNA families database. Nucleic Acids Res. 2015;43(D1):D130–7.

71. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. J Biol Chem. 2002;277(50):48949–59.

72. Sudarsan N, Barrick JE, Breaker RR. Metabolite-binding RNA domains are present in the genes of eukaryotes. 2003;(203):644–7.

73. Wachter A. Riboswitch-mediated control of gene expression in eukaryotes. RNA Biol [Internet]. 2010;7(1):67–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20009507%5Cnhttp://www.tandfonline.com/doi/abs/10.4161/rna.7.1.10489

74. Breaker ERLKFBRR. Metabolite-sensing Riboswitches as Antibacterial Drug Targets. In: Miller AAM and PF, editor. Emerging Trends in Antibacterial Discovery: Answering the Call to Arms. Caister Ac. 2011. p. 107–30.

75. Blount KF, Breaker RR. Riboswitches as antibacterial drug targets. Nat Biotechnol [Internet]. 2006 Dec [cited 2018 May 14];24(12):1558–64. Available from: http://www.nature.com/doifinder/10.1038/nbt1268

76. Lünse CE, Scott FJ, Suckling CJ, Mayer G. Novel TPP-riboswitch activators bypass metabolic enzyme dependency. Front Chem [Internet]. 2014 Jul 28 [cited 2018 May 14];2:53. Available from: http://journal.frontiersin.org/article/10.3389/fchem.2014.00053/abstract

77. Serganov A, Polonskaia A, Phan AT, Breaker RR, Patel DJ. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. Nature. 2006;441(7097):1167–71.

78. Ray PS, Jia J, Yao P, Majumder M, Hatzoglou M, Fox PL. A stress-responsive RNA switch regulates VEGFA expression. Nature. 2009;457(7231):915–9.

79. Carmeliet P. Angiogenesis in life, disease and medicine. Vol. 438, Nature. 2005. p. 932–6.

80. Shibuya M. Vascular endothelial growth factor-dependent and -independent

regulation of angiogenesis. BMB Rep. 2008;41(4):278–86.

81.   Folkman J. Angiogenesis in cancer, vascular, rheumatoid and other disease. Vol. 1, Nature Medicine. 1995. p. 27–30.

82.   Ferrara N. Role of vascular endothelial growth factor in the regulation of angiogenesis. Vol. 56, Kidney International. 1999. p. 794–814.

83.   Ferrara N. VEGF and the quest for tumour angiogenesis factors. Vol. 2, Nature Reviews Cancer. 2002. p. 795–803.

84.   Arcondéguy T, Lacazette E, Millevoi S, Prats H, Touriol C. VEGF-A mRNA processing, stability and translation: A paradigm for intricate regulation of gene expression at the post-transcriptional level. Nucleic Acids Res. 2013;41(17):7997–8010.

85.   Jafarifar F, Yao P, Eswarappa SM, Fox PL. Repression of VEGFA by CA-rich element-binding microRNAs is modulated by hnRNP L. EMBO J. 2011;30(7):1324–34.

86.   Yao P, Potdar AA, Ray PS, Eswarappa SM, Flagg AC, Willard B, et al. The HILDA complex coordinates a conditional switch in the 3′-untranslated region of the VEGFA mRNA. PLoS Biol. 2013;11(8):e1001635.

87.   Chang T-H, Huang H-D, Wu L-C, Yeh C-T, Liu B-J, Horng J-T. Computational identification of riboswitches based on RNA conserved functional sequences and conformations. RNA [Internet]. 2009 Jul [cited 2016 Sep 17];15(7):1426–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19460868

88.   Bengert P, Dandekar T. Riboswitch finder--a tool for identification of riboswitch RNAs. Nucleic Acids Res [Internet]. 2004 Jul 1 [cited 2016 Sep 17];32(Web Server issue):W154-9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15215370

89.   Abreu-Goodger C, Merino E. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. Nucleic Acids Res [Internet]. 2005 Jul 1 [cited 2016 Sep 17];33(Web Server issue):W690-2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15980564

90.   Havill JT, Bhatiya C, Johnson SM, Sheets JD, Thompson JS. A new approach for detecting riboswitches in DNA sequences. Bioinformatics [Internet]. 2014 Nov 1 [cited 2016 Sep 17];30(21):3012–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25015992

91.   Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids

Res [Internet]. 2013 Jul [cited 2016 Sep 17];41(12):e121. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23598997

92. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics [Internet]. 2013 Nov 15 [cited 2016 Sep 17];29(22):2933–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24008419

93. Singh P, Bandyopadhyay P, Bhattacharya S, Krishnamachari A, Sengupta S. Riboswitch detection using profile hidden Markov models. BMC Bioinformatics [Internet]. 2009;10(1):325. Available from: http://www.biomedcentral.com/1471-2105/10/325

94. Nawrocki EP, Eddy SR. Computational identification of functional RNA homologs in metagenomic data. RNA Biol [Internet]. 2013;10(7):1170–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23722291%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3849165

95. Clote P. Computational prediction of riboswitches [Internet]. 1st ed. Vol. 553, Methods in Enzymology. Elsevier Inc.; 2015. 287-312 p. Available from: http://dx.doi.org/10.1016/bs.mie.2014.10.063

96. Gupta A, Swati D. Exploring Riboswitches in Archaeal Metagenomes. J RNAi Gene Silenc J RNAi Gene Silenc. 2016;12(12):536–43.

97. Kang I, Kim S, Islam MR, Cho J-C. The first complete genome sequences of the acI lineage, the most abundant freshwater Actinobacteria, obtained by whole-genome-amplification of dilution-to-extinction cultures. Sci Rep [Internet]. 2017;7(June 2016):42252. Available from: http://www.nature.com/articles/srep42252

98. Leyn SA, Suvorova IA, Kholina TD, Sherstneva SS, Novichkov PS, Gelfand MS, et al. Comparative Genomics of Transcriptional Regulation of Methionine Metabolism in Proteobacteria. Kuipers OP, editor. PLoS One [Internet]. 2014 Nov 20 [cited 2017 Apr 10];9(11):e113714. Available from: http://dx.plos.org/10.1371/journal.pone.0113714

99. Nawrocki EP. Annotating Functional RNAs in Genomes Using Infernal. In 2014 [cited 2017 Apr 10]. p. 163–97. Available from: http://link.springer.com/10.1007/978-1-62703-709-9_9

100. Remmele CW, Xian Y, Albrecht M, Faulstich M, Fraunholz M, Heinrichs E, et al. Transcriptional landscape and essential genes of Neisseria gonorrhoeae. Nucleic Acids Res. 2014;42(16):10579–95.

101. Pang E, Cao H, Zhang B, Lin K. Crop Genome Annotation: A Case Study for the

Brassica rapa Genome. In Springer Berlin Heidelberg; 2015 [cited 2017 Apr 10]. p. 53–64. Available from: http://link.springer.com/10.1007/978-3-662-47901-8_5

102. Chojnowski G, Walen T, Bujnicki JM. RNA Bricks--a database of RNA 3D motifs and their interactions. Nucleic Acids Res [Internet]. 2014 Jan [cited 2016 Sep 17];42(Database issue):D123-31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24220091

103. Baker D, Sali a. Protein structure prediction and structural genomics. Science (80-). 2001;294(5540):93–6.

104. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res [Internet]. 2000 Jan 1 [cited 2017 Aug 16];28(1):235–42. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.235

105. Magnus M, Matelska D, Lach G, Chojnowski G, Boniecki MJ, Purta E, et al. Computational modeling of RNA 3D structures, with the aid of experimental restraints. RNA Biol [Internet]. 2014 [cited 2016 Sep 17];11(5):522–36. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24785264

106. Bujnicki JM. Protein-structure prediction by recombination of fragments. ChemBioChem. 2006;7(1):19–27.

107. Barash D, Gabdank I. Energy minimization methods applied to riboswitches: a perspective and challenges. RNA Biol. 2010;7(1):90–7.

108. Sloma MF, Mathews DH. Improving RNA secondary structure prediction with structure mapping data [Internet]. 1st ed. Vol. 553, Methods in Enzymology. Elsevier Inc.; 2015. 91-114 p. Available from: http://dx.doi.org/10.1016/bs.mie.2014.10.053

109. Hamada M. RNA Secondary Structure Prediction from Multi-Aligned Sequences. In 2015 [cited 2017 Feb 15]. p. 17–38. Available from: http://link.springer.com/10.1007/978-1-4939-2291-8_2

110. Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics. 2004;5:140.

111. Huang W, Kim J, Jha S, Aboul-ela F. The Impact of a Ligand Binding on Strand Migration in the SAM-I Riboswitch. PLoS Comput Biol. 2013;9(5).

112. Purzycka KJ, Popenda M, Szachniuk M, Antczak M, Lukasiak P, Blazewicz J, et al. Automated 3D RNA structure prediction using the RNAComposer method for riboswitches1 [Internet]. 1st ed. Vol. 553, Methods in Enzymology. Elsevier Inc.; 2015. 3-34 p. Available from: http://dx.doi.org/10.1016/bs.mie.2014.10.050

113. Cruz JA, Blanchet M, Boniecki M, Cruz A. RNA-Puzzles : A CASP-like evaluation of RNA three-dimensional structure prediction RNA-Puzzles : A CASP-like evaluation of RNA three-dimensional structure prediction. 2012;610–25.

114. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) - round x. Proteins Struct Funct Bioinforma. 2014;82(SUPPL.2):1–6.

115. Miao Z, Adamiak RW, Blanchet M-F, Boniecki M, Bujnicki JM, Chen S-J, et al. RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. RNA [Internet]. 2015;21(6):1066–84. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4436661&tool=pmce ntrez&rendertype=abstract

116. Miao Z, Adamiak RW, Antczak M, Batey RT, Becka AJ, Biesiada M, et al. RNA-Puzzles Round III : 3D RNA structure prediction of five riboswitches and one ribozyme.

117. Miao Z, Adamiak RW, Antczak M, Batey RT, Becka AJ, Biesiada M, et al. RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. RNA [Internet]. 2017 May [cited 2017 Dec 10];23(5):655–72. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28138060

118. Gong S, Wang Y, Wang Z, Zhang W. Computational Methods for Modeling Aptamers and Designing Riboswitches. Int J Mol Sci [Internet]. 2017;18(11):2442. Available from: http://www.mdpi.com/1422-0067/18/11/2442

119. Jager S, Schiller B, Babel P, Blumenroth M, Strufe T, Hamacher K. StreAM-Tg: algorithms for analyzing coarse grained RNA dynamics based on Markov models of connectivity-graphs. Algorithms Mol Biol [Internet]. 2017 [cited 2017 Dec 10];12:15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28572834

120. Cai W, Li J, Yip S. Molecular Dynamics. In: Comprehensive Nuclear Materials [Internet]. Elsevier; 2012 [cited 2018 May 5]. p. 249–65. Available from: http://linkinghub.elsevier.com/retrieve/pii/B9780080560335001282

121. Alder BJ, Wainwright TE. Studies in molecular dynamics. I. General method. J Chem Phys. 1959;31(2):459–66.

122. Rahman A. Correlations in the Motions of Atoms in Liquid Argon. Phys Rev. 1964;136(2A):A405--A411.

123. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Vol. 9, Nature Structural Biology. 2002. p. 646–52.

124. molecular dynamics in drug design. In: IUPAC Compendium of Chemical Terminology [Internet]. Research Triagle Park, NC: IUPAC; [cited 2018 May 2]. Available from: http://goldbook.iupac.org/MT06969.html

125. Allen MP. Computer simulation of a biaxial liquid crystal. Liq Cryst. 1990;8(4):499–511.

126. VERLI H. Bioinformática da Biologia à Flexibilidade Molecular. Vol. 53, Universidade Federal do Rio Grande do Sul. 2014. 1689-1699 p.

127. Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biol [Internet]. 2011 Oct 28 [cited 2018 May 2];9(1):71. Available from: http://bmcbiol.biomedcentral.com/articles/10.1186/1741-7007-9-71

128. Leach AR. Molecular modelling: principles and applications [Internet]. Culture Ethnicity and Conflict. 2001. 744 p. Available from: http://books.google.de/books?id=kB7jsbV-uhkC

129. Cragnolini T, Derreumaux P, Pasquali S. Ab initio RNA folding. J Phys Condens Matter. 2015;27(23):233102.

130. Hanke CA, Gohlke H. Force field dependence of riboswitch dynamics [Internet]. 1st ed. Vol. 553, Methods in Enzymology. Elsevier Inc.; 2015. 163-191 p. Available from: http://dx.doi.org/10.1016/bs.mie.2014.10.056

131. Šponer J, Bussi G, Krepl M, Banáš P, Bottaro S, Cunha RA, et al. RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. Chem Rev [Internet]. 2018 Apr 25 [cited 2018 May 15];118(8):4177–338. Available from: http://pubs.acs.org/doi/10.1021/acs.chemrev.7b00427

132. Papo D, Buldú JM, Boccaletti S, Bullmore ET. Complex network theory and the brain. Philos Trans R Soc. 2014;369(1653):20130520–20130520.

133. Motta R, De Andrade Lopes A, De Oliveira MCF. Centrality measures from complex networks in active learning. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2009;5808 LNAI:184–96.

134. Kantarcı B, Labatut V. Classification of Complex Networks Based on Topological Properties. 2013 Third Int Conf Cloud Green Comput. 2013;297--304.

135. Guillaume J-L, Latapy M. A Realistic Model for Complex Networks. 2003;12.

136. Ronqui RF. Medidas de centralidade em redes complexas: correlações, efetividade e caracterização de sistemas. 2014;

137. Sethi A, Eargle J, Black AA, Luthey-Schulten Z. Dynamical networks in tRNA:protein complexes. Proc Natl Acad Sci [Internet]. 2009;106(16):6620–5. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0810961106

138. Metz J, Calvo R, Seno ERM, Romero R a. F, Liang Z. Redes Complexas: Conceitos e Aplicações. Relatórios Técnicos do ICMC. 2007;1–45.

139. Ercsey-Ravasz M, Lichtenwalter RN, Chawla N V., Toroczkai Z. Range-limited centrality measures in complex networks. Phys Rev E - Stat Nonlinear, Soft Matter Phys. 2012;85(6).

140. Euler L. Solutio Problematis ad geometriam situs pertinentis. Vol. 8, Commentarii Academiae Scientiarum Imperialis Petropolitanae. 1736. 128-140 p.

141. Vairachilai S, Devi MKK, Raja M. Analysis of Statistical and Structural Properties of Complex networks with Random Networks. 2017;146(1):137–46.

142. Ghoshal G. Structural and dynamical properties of complex networks. 2009;

143. Mcphail M. The Statistical Properties of Complex Networks. 2014;

144. Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci. 2002;99(12):7821–6.

145. Nasiruzzaman ABM, Pota HR, Mahmud MA. Application of centrality measures of complex network framework in power grid. IECON 2011 - 37th Annu Conf IEEE Ind Electron Soc. 2011;4660–5.

146. Perra N, Fortunato S. Spectral centrality measures in complex networks. Phys Rev E - Stat Nonlinear, Soft Matter Phys. 2008;78(3):1–11.

147. Serganov A, Polonskaia A, Phan AT, Breaker RR, Patel DJ. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. Nature [Internet]. 2006;441(7097):1167–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16728979%5Cnhttp://www.nature.com/nature/journal/v441/n7097/pdf/nature04740.pdf

148. Thore S, Frick C, Ban N. Structural basis of thiamine pyrophosphate analogues binding to the eukaryotic riboswitch. J Am Chem Soc. 2008;130(26):8116–7.

149. Lu X-J, Olson WK. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc [Internet]. 2008 Jul 3 [cited 2017 Aug 17];3(7):1213–27. Available from: http://www.nature.com/doifinder/10.1038/nprot.2008.104

150. Di Tommaso P, Bussotti G, Kemena C, Capriotti E, Chatzou M, Prieto P, et al. SARA-Coffee web server, a tool for the computation of RNA sequence and structure multiple alignments. Nucleic Acids Res [Internet]. 2014 Jul [cited 2017 Aug 17];42(Web Server issue):W356-60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24972831

151. Bond CS, Schüttelkopf AW. ALINE: A WYSIWYG protein-sequence alignment editor for publication-quality alignments. Acta Crystallogr Sect D Biol Crystallogr. 2009;65(5):510–2.

152. Blin G, Denise A, Dulucq S, Herrbach C, Touzet H. VARNA: Interactive drawing and editing of the RNA secondary structure. IEEE/ACM Trans Comput Biol Bioinforma. 2010;7(2):309–22.

153. Eddy SR, Eddy S, Erdmann V, Barciszewska M, Symanski M, Hochberg A, et al. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics. 2002;3(1):18.

154. Smith C, Heyne S, Richter AS, Will S, Backofen R. Freiburg RNA Tools: A web server integrating IntaRNA, ExpaRNA and LocARNA. Nucleic Acids Res. 2010;38(SUPPL. 2):373–7.

155. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics [Internet]. 2008;9:474. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2621365&tool=pmcentrez&rendertype=abstract

156. Jorge NAN, Machado-Lima A, Passetti F. Desenvolvimento de uma metodologia de bioinformática para a busca de riboswitches no genoma humano. Universidade Federal do Estado do Rio de Janeiro – UNIRIO; 2010.

157. Pontius JU, Wagner L, Schuler GD. UniGene: a unified view of the transcriptome. NCBI Handb [Internet]. 2003;1:1–12. Available from: http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/handbook/ch21d1.pdf

158. Rother M, Rother K, Puton T, Bujnicki JM. ModeRNA: a tool for comparative modeling of RNA 3D structure. Nucleic Acids Res [Internet]. 2011 May [cited 2016 Sep 17];39(10):4007–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21300639

159. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins Struct Funct Bioinforma. 2010;78(8):1950–8.

160. Abraham MJ, Murtola T, Schulz R, P??ll S, Smith JC, Hess B, et al. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1–2:19–25.

161. Edwards TE, Ferré-D'Amaré AR. Crystal structures of the thi-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule

recognition. Structure. 2006;14(9):1459–68.

162. Kulshina N, Edwards TE, Ferré-D'Amaré AR. Thermodynamic analysis of ligand binding and ligand binding-induced tertiary structure formation by the thiamine pyrophosphate riboswitch. RNA. 2010;16(1):186–96.

163. Cao S, Chen SJ. Physics-based de novo prediction of RNA 3D structures. J Phys Chem B. 2011;115(14):4216–26.

164. Xu X, Zhao P, Chen SJ. Vfold: A web server for RNA structure and folding thermodynamics prediction. PLoS One. 2014;9(9).

165. Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, et al. Automated 3D structure composition for large RNAs. Nucleic Acids Res. 2012;40(14):1–12.

166. Biesiada M, Purzycka KJ, Szachniuk M, Blazewicz J, Adamiak RW. Automated RNA 3D Structure Prediction with RNAComposer. In 2016. p. 199–215. Available from: http://link.springer.com/10.1007/978-1-4939-6433-8_13

167. Popenda M, Błażewicz M, Szachniuk M, Adamiak RW. RNA FRABASE version 1.0: An engine with a database to search for the three-dimensional fragments within RNA structures. Nucleic Acids Res. 2008;36(SUPPL. 1):386–91.

168. Popenda M, Szachniuk M, Blazewicz M, Wasik S, Burke EK, Blazewicz J, et al. RNA FRABASE 2 . 0 : an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. 2010;

169. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comput Chem [Internet]. 2004 Jul 15 [cited 2017 Aug 17];25(9):1157–74. Available from: http://doi.wiley.com/10.1002/jcc.20035

170. Jakalian A, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J Comput Chem [Internet]. 2002 Oct 18 [cited 2017 Aug 17];23(16):1623–41. Available from: http://doi.wiley.com/10.1002/jcc.10128

171. Wang J, Wang W, Kollman PA, Case DA. Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model [Internet]. 2006 Oct [cited 2017 Aug 17];25(2):247–60. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1093326305001737

172. Sousa da Silva AW, Vranken WF. ACPYPE - AnteChamber PYthon Parser interfacE. BMC Res Notes [Internet]. 2012 Jul 23 [cited 2017 Aug 17];5:367. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22824207

173. Leipply D, Draper DE. Effects of Mg $^{2+}$ on the Free Energy Landscape for Folding

a Purine Riboswitch RNA. Biochemistry [Internet]. 2011 Apr 12 [cited 2017 Aug 17];50(14):2790–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21361309

174. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys [Internet]. 1983 Jul 15 [cited 2017 Aug 17];79(2):926–35. Available from: http://aip.scitation.org/doi/10.1063/1.445869

175. G??rska A, Jasi??ski M, Trylska J. MINT: Software to identify motifs and short-range interactions in trajectories of nucleic acids. Nucleic Acids Res. 2015;43(17):1–10.

176. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemom Intell Lab Syst. 1987;2(1–3):37–52.

177. Abdi H, Williams LJ. Principal component analysis. Vol. 2, Wiley Interdisciplinary Reviews: Computational Statistics. 2010. p. 433–59.

178. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics [Internet]. 2006 Nov 1 [cited 2017 Aug 17];22(21):2695–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16940322

179. R Development Core Team. R: A Language and Environment for Statistical Computing. R Found Stat Comput Vienna, Austria [Internet]. 2016;0:{ISBN} 3-900051-07-0. Available from: http://www.r-project.org/

180. Csárdi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Syst. 2006;1695:1–9.

181. Kolaczyk ED, Csárdi G. Statistical Analysis of Network Data with R. 2014. 207 p.

182. Kolaczyk ED. Statistical Analysis of Network Data [Internet]. Vol. 26, Learning. 2009. 251-264 p. Available from: http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-98134-5?cm_mmc=AD-_-Enews-_-ECS12245_V1-_-978-0-387-98134-5%5Cnhttp://link.springer.com/10.1007/978-0-387-88146-1

183. Skjærven L, Yao X-Q, Scarabelli G, Grant BJ. Integrating protein structural dynamics and evolutionary analysis with Bio3D. BMC Bioinformatics [Internet]. 2014;15(1):399. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-014-0399-6

184. Yen JY. Finding the *K* Shortest Loopless Paths in a Network. Manage Sci [Internet]. 1971 Jul 1 [cited 2017 Aug 17];17(11):712–6. Available from: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.17.11.712

185. Eden E, Wallach I, Yakhini Z. SimTree: A Tool for Computing Similarity Between RNA Secondary Structures [Internet]. [cited 2017 Dec 7]. Available from: http://bioinfo.cs.technion.ac.il/SimTree/

186. Gong Z, Zhao Y, Chen C, Duan Y, Xiao Y. Insights into ligand binding to preQ1 riboswitch aptamer from molecular dynamics simulations. PLoS One. 2014;9(3):1–12.

187. Aytenfisu AH, Liberman JA, Wedekind JE, Mathews DH. Molecular mechanism for preQ1 -II riboswitch function revealed by molecular dynamics. Rna. 2015;21:1898–907.

188. Wang W, Jiang C, Zhang J, Ye W, Luo R, Chen H-F. Dynamics Correlation Network for Allosteric Switching of PreQ1 Riboswitch. Sci Rep [Internet]. 2016;6(April):31005. Available from: http://www.nature.com/articles/srep31005

189. Hanke CA, Gohlke H. Ligand-mediated and tertiary interactions cooperatively stabilize the P1 region in the guanine-sensing riboswitch. PLoS One. 2017;12(6):1–29.

190. Timpl R, Sasaki T, Kostka G, Chu ML. Fibulins: A versatile family of extracellular matrix proteins. Nat Rev Mol Cell Biol. 2003;4(6):479–89.

191. Thore S, Leibundgut M, Ban N. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. Science (80-). 2006;312(5777):1208–11.

192. Anthony PC, Perez CF, Garcia-Garcia C, Block SM. Folding energy landscape of the thiamine pyrophosphate riboswitch aptamer. Proc Natl Acad Sci [Internet]. 2012;109(5):1485–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22219369%5Cnhttp://www.pnas.org/content/109/5/1485.full.pdf%5Cnhttp://www.pnas.org/cgi/doi/10.1073/pnas.1115045109

193. Duesterberg VK, Fischer-Hwang IT, Perez CF, Hogan DW, Block SM. Observation of long-range tertiary interactions during ligand binding by the TPP riboswitch aptamer. Elife. 2015;4(DECEMBER2015):1–17.

194. Guedich S, Puffer-Enders B, Baltzinger M, Hoffmann G, Da Veiga C, Jossinet F, et al. Quantitative and predictive model of kinetic regulation by E. coli TPP riboswitches. RNA Biol. 2016;13(4):373–90.

195. Ritz J, Martin JS, Laederach A, Turner D, Warden C. Evolutionary Evidence for Alternative Structure in RNA Sequence Co-variation. Chen S-J, editor. PLoS Comput Biol [Internet]. 2013 Jul 25 [cited 2017 Aug 16];9(7):e1003152. Available from: http://dx.plos.org/10.1371/journal.pcbi.1003152

196. Zhang RZ, Pan TC, Zhang ZY, Mattei MG, Timpl R, Chu ML. Fibulin-2 (FBLN2): human cDNA sequence, mRNA expression, and mapping of the gene on human and mouse chromosomes. Genomics [Internet]. 1994;22(2):425–30. Available from: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.ncbi.nlm.nih.gov/htbin-post/Omim/getmim%3Ffield=medline_uid&search=7806230

197. Mehdizadeh Aghdam E, Sinn M, Tarhriz V, Barzegar A, Hartig JS, Hejazi MS. TPP riboswitch characterization in Alishewanella tabrizica and Alishewanella aestuarii and comparison with other TPP riboswitches. Microbiol Res [Internet]. 2017 Jan 1 [cited 2018 May 23];195:71–80. Available from: https://www.sciencedirect.com/science/article/pii/S0944501316303500?via%3Dihub

198. Staple DW, Butcher SE. Pseudoknots: RNA Structures with Diverse Functions. PLoS Biol. 2005 Jun;3(6):e213.

APPENDIX A – PUBLISHED PAPER

Using RNA Sequence and Structure for the Prediction of Riboswitch Aptamer: A Comprehensive Review of Available Software and Tools

Deborah Antunes[1], Natasha A. N. Jorge[2,3], Ernesto R. Caffarena[1] and Fabio Passetti[2,3]*

[1] Scientific Computing Program (PROCC), Computational Biophysics and Molecular Modeling Group, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil, [2] Laboratory of Functional Genomics and Bioinformatics, Oswaldo Cruz Institute, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil, [3] Laboratory of Gene Expression Regulation, Carlos Chagas Institute, Fundação Oswaldo Cruz, Curitiba, Brazil

RNA molecules are essential players in many fundamental biological processes. Prokaryotes and eukaryotes have distinct RNA classes with specific structural features and functional roles. Computational prediction of protein structures is a research field in which high confidence three-dimensional protein models can be proposed based on the sequence alignment between target and templates. However, to date, only a few approaches have been developed for the computational prediction of RNA structures. Similar to proteins, RNA structures may be altered due to the interaction with various ligands, including proteins, other RNAs, and metabolites. A riboswitch is a molecular mechanism, found in the three kingdoms of life, in which the RNA structure is modified by the binding of a metabolite. It can regulate multiple gene expression mechanisms, such as transcription, translation initiation, and mRNA splicing and processing. Due to their nature, these entities also act on the regulation of gene expression and detection of small metabolites and have the potential to helping in the discovery of new classes of antimicrobial agents. In this review, we describe software and web servers currently available for riboswitch aptamer identification and secondary and tertiary structure prediction, including applications.

Keywords: riboswitch, RNA motif, riboswitch aptamer prediction, RNA secondary structure, RNA tertiary structure

INTRODUCTION

Fifty years ago, the central dogma of molecular biology proposed a preferential flow of information, stating that DNA is transcribed into RNA, which in turn is translated into proteins with structural or catalytic functions (Crick, 1970; Albert et al., 2011). Since then, new findings have indicated that this theory was incomplete. For instance, in 2007, the ENCODE Project Consortium showed that, although most of the DNA is transcribed, only a fraction of the transcriptome is translated into proteins. RNA portions that do not encode proteins were then termed non-coding RNAs (ncRNA) (Crick, 1970; Mattick, 2001; Albert et al., 2011). Those ncRNAs belonging to the same class share precise sequence and structural characteristics, which have been conserved throughout several evolutionary processes. The degree of sequence conservation is smaller than that observed for protein-coding genes, but is crucial to explain the functional heterogeneity of the ncRNAs (Amaral et al., 2011; Qu and Adelson, 2012). One of the most significant examples of conserved functional RNAs are the riboswitches (Barrick and Breaker, 2007).

100