Function Prediction of Transcription Start Site Associated RNAs (TSSaRNAs) in *Halobacterium salinarum* NRC-1

Yagoub Ali Ibrahim Adam

Inter-units Postgraduate Program in Bioinformatics
1- Faculty of Philosophy, Sciences and Languages at Ribeirão Preto
Department of Computation and Mathematics (DCM-FFCLRP)
2- Institute of Mathematics and Statistics (IME)

University of São Paulo

Ribeirão Preto, Brasil
February, 2019

Yagoub Ali Ibrahim Adam

Function Prediction of Transcription Start Site Associated RNAs (TSSaRNAs) in *Halobacterium salinarum* NRC-1

A thesis submitted to the Inter-units Postgraduate Program in Bioinformatics at the University of São Paulo for the degree of Doctorate in Sciences

Concentration Area: Bioinformatics

Inter-units Postgraduate Program in Bioinformatics
1- Faculty of Philosophy, Sciences and Languages at Ribeirão Preto - Department of Computation and Mathematics (DCM-FFCLRP)
2- Institute of Mathematics and Statistics (IME)
University of São Paulo

Advisor
Prof. Dr. Ricardo Zorzetto Nicoliello Vêncio
Department of Computing and Mathematics – FFCLRP

Ribeirão Preto, Brasil
February, 2019

Yagoub Ali Ibrahim Adam

Predição de função para TSSaRNAs (transcritos associados a sitios de início de transcrição) em *Halobacterium salinarum* NRC-1

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade de São Paulo para a obtenção do título do Doutor em Ciências

Área de Concentração: Bioinformática

Programa Interunidades de Pós-Graduação em Bioinformática
1- Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - Departamento de Computação e Matemática (DCM-FFCLRP)
2- Instituto de Matemática e Estatística (IME)
Universidade de São Paulo

Orientador
Prof. Dr. Ricardo Zorzetto Nicoliello Vêncio
Departamento de Computação e Matemática – FFCLRP

Ribeirão Preto, Brasil
Fevereiro, 2019

"This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001"

"O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001"

FICHA CATALOGRÁFICA

Yagoub Ali Ibrahim Adam

Function Prediction of Transcription Start Site Associated RNAs (TSSaRNAs) in *Halobacterium salinarum* NRC-1 / Yagoub Ali Ibrahim Adam; Supervisor: Ricardo Zorzetto Nicoliello Vêncio. Ribeirão Preto - SP, 2019.

173 p.

Predição de função para TSSaRNAs (transcritos associados a sitios de início de transcrição) em *Halobacterium salinarum* NRC-1 / Yagoub Ali Ibrahim Adam; Orientador: Ricardo Zorzetto Nicoliello Vêncio. Ribeirão Preto - SP, 2019.

173 p.

Tese (Doutorado - Programa Interunidades de Pós-Graduação em Bioinformática)

1- Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - Departamento de Computação e Matemáticade

2- Instituto de Matemática e Estatística (IME)

Universidade de São Paulo, 2019.

Programa Interunidades de Pós-Graduação em Bioinformática

Área de Concentração: Bioinformática

1- TSSaRNA 2- Predição de ncRNAs 3- *Halobacterium salinarum* NRC-1 4- RNA não codificante 5- Anotação funcional 6- Anotação estrutural 7- Rfam 8- Funções de ncRNAs 9- Estruturas RNA 10- Interações de RNA 11- Regulação baseada em RNA 12- Estruturas de ncRNAs de ordem superior 13- Interações de TSSaRNAs-LSm 14- Docagem de RNA.

Name: Yagoub Ali Ibrahim Adam
Title: Function Prediction of Transcription Start Site Associated RNAs (TSSaRNAs) in *Halobacterium salinarum* NRC-1

A thesis presented for the degree of Doctorate in Sciences
1- Faculty of Philosophy, Sciences and Languages at Ribeirão Preto;
Department of Computation and Mathematics (DCM-FFCLRP)
2- Institute of Mathematics and Statistics (IME)
University of São Paulo / Universidade de São Paulo

Approved in: 7th February, 2019.

Examination committee

Prof. Dr. Ricardo Z. N. Vêncio _____ **Institution:** FFCLRP-USP
Judgment: Approved _____ **Signature:** _____

Prof. Dr. Wilson Araújo da Silva Junior _____ **Institution:** FMRP-USP
Judgment: Approved _____ **Signature:** _____

Prof. Dr. Helder Takashi Imoto Nakaya _____ **Institution:** FCF-USP
Judgment: Approved _____ **Signature:** _____

Prof. Dr. Alexandre Rossi Paschoal _____ **Institution:** UTFPR
Judgment: Approved _____ **Signature:** _____

Prof(a). Dr(a). Silvana Giuliatti _____ **Institution:** FMRP-USP
Judgment: Approved _____ **Signature:** _____

1-FFCLRP-USP

Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - Universidade de São Paulo

(Faculty of Philosophy, Sciences and Languages at Ribeirão Preto - University of São Paulo)

2-FCF-USP

Faculdade de Ciências Farmacêuticas - Universidade de São Paulo

(Faculty of Pharmaceutical Sciences - University of São Paulo)

3-FMRP-USP

Faculdade de Medicina de Ribeirão Preto - Universidade de São Paulo

(Faculty of Medicine at Ribeirão Preto - University of São Paulo)

4-UTFPR

Universidade Tecnológica Federal do Paraná
(Federal University of Technology in Paraná)

Dedication

I dedicate this thesis project to:

My beloved father; although he is no longer of this world but his memories still my life.

My family, the symbol of love and unlimited giving.

All friends who always encourage and support.

All colleagues.

To everyone who I met and touch my heart.

To those are/were part of my daily life.

Acknowledgements

In the Name of ALLAH, the Most Merciful, the Most Compassionate all praise be to ALLAH, the Lord of the worlds; and prayers and peace be upon Mohammed His servant and messenger. First and foremost, I'm grateful to ALLAH for giving the strength to conduct this work as I could never have achieved the doctorate without His help and bless. I would like to express my appreciation and my sincere thanks to my supervisor Prof. Dr. Ricardo Z. N. Vêncio who is kindly has contributed by his time and efforts as well as offering the constructive criticism for this work to be done; really I am forever indebted to my supervisor for his support, encouragements and patience all the time during the journey of obtaining the doctorate degree.

With a great pleasure I would like to acknowledge prof.(a) Dr.(a) Tie Koide for her helps, assistance and her contribution from the beginning of this project by providing many feedback.

My sincere thanks extended to all the academic staff and the administrative officers (both past and present) at the Bioinformatics Postgraduate Inter-units Program in particular the staff at The Institute of Mathematics and Statistics of the University of São Paulo (IME) and the staff at The Department of Computation and Mathematics (DCM) - The Faculty of Philosophy, Sciences and Languages at Ribeirão Preto (FFCLRP).

My unlimited thanks also extended to the secretariat of the Post-Graduate Program in Bioinformatics -the former staff (Ms. Patricia Cristina Martorelli and Ms. Cristiane de Fátima Braulino), the current secretary (Ms. Márcia Ferreira), and all the assistance

staff- for their assistances during this doctoral period.
Also, I would to thank the secretary at The Department of Computation and Mathematics (DCM) and the secretary of the Postgraduate Committee at The Faculty of Philosophy, Sciences and Languages at Ribeirão Preto (FFCLRP) for their helps.
I am extremely grateful for both the past and the present colleagues at the Laboratory for Biological Information Processing – LabPIB- for their helps, conversations and encouragement.
I am extremely grateful for both past and present colleagues at the Microbial Systems Biology Lab -LaBiSisMi- for their helps, conversations and encouragement.
I would like to express my unlimited acknowledge to the **Coordination of Improvement of Higher Education Personnel (CAPES)**, Ministry of Education, Brazil; for funding this project.
I would like to thank all Brazilian friend as well the international students for their conversation and changing the ideas with special thanks for those are/were part of my daily life.
I thank everyone who helped me during this doctoral period with criticism, conversations and suggestions. Also, my special thanks extended for those who helped me to grow up as bioinformatician-bioiformaticist. Last but not least, my deepest thanks go to everyone who took part in the journey of making this thesis real. Their names are too numerous to mention, but many of them inspired me to continue learning and sharing with others.

So, many thanks for you all.
Yagoub Ali Ibrahim Adam

Contents

List of Figures

List of Tables

Abstract

Yagoub A. I. Adam. **Function Prediction of Transcription Start Site Associated RNAs (TSSaRNAs) in *Halobacterium salinarum* NRC-1**. PhD (Thesis)- Bioinformatics Postgraduate Interunits Program (Department of Computation and Mathematics (DCM) - Faculty of Philosophy, Sciences and Languages at Ribeirão Preto (FFCLRP) & The Institute of Mathematics and Statistics of the University of São Paulo (IME-USP)) - University of São Paulo, Ribeirão Preto, 2019. The Transcription Start Site Associated non-coding RNAs (TSSaRNAs)

have been predicted across the three domain of life. However, still, there are no reliable annotation efforts to identify their biological functions and their underline molecular machinery. Therefore, this project addresses the question of what are the potential functions of TSSaRNAs regarding their roles in addressing the cellular functions. To answer this question, we aimed to accurately identify TSSaRNAs in the model organism *Halobacterium salinarum* NRC-1 (an Archean microorganism) that incubated at the standard growth condition. Consequently, we aimed to investigate TSSaRNAs structural stability in the term of the thermodynamic energies. Moreover, we attempted to functionally annotate TSSaRNAs based on Rfam functional classification of non-coding RNAs.

Based on the statistical approach we developed an algorithm to predict TSSaRNA using next-generation RNA sequencing data (RNA-Seq). To perform structural annotation of TSSaRNAs, we investigated the structural stability of TSSaRNAs by modeling the secondary structures by minimizing the thermodynamic free energy. We simulated TSSaRNAs tertiary structures based on the secondary structures constrain using the Rosetta-Common RNA tool. The structures of the minimum free energy supposed to be biophysically stable structures. To investigate the higher order structures of TSSaRNAs, we studied the hybridization between TSSaRNAs and their cognate genes as part of RNA based regulation system. Also, based on our hypothesis that TSSaRNAs may bind to protein to trigger their function, we have investigated the interaction between TSSaRNAs and Lsm protein which known as a chaperone protein that mediates RNA function and involved in RNA processing. Our pipeline to perform the functional annotation of TSSaRNAs aimed to classify TSSaRNAs into their corresponding Rfam families based on two steps: either through querying TSSaRNAs sequences against the co-variance models of Rfam families or by querying the Rfam sequences against the co-variance models of the consensus secondary structures in TSSaRNAs.

The results showed that the prediction algorithm has succeeded to identify a total of 224 TSSaRNAs that expressed in the same strand of the mRNAs and 58 TSSaRNAs that expressed as antisense of the mRNAs. The identified TSSaRNAs molecules showed a median length of 25 nucleotides. Regarding the structural annotation of TSSaRNAs, the results showed that most of TSSaRNAs possessed thermodynamically stable secondary structures and their tertiary structures were capable of forming more complex structures through binding with other biomolecules. About the formation of higher-order structures, we have observed that most of TSSaRNAs (92.2%) were capable of hybridizing into their cognate genes also 55 TSSaRNAs indicated putative interactions with Lsm protein. Furthermore, the computation docking experiments demonstrated the TSSaRNAs-Lsm complexes associated with favorable binding energy of

a median of -542900 kcal mole^{-1}. Regarding the functional annotation of TSSaRNAs, the results showed that the majority of TSSaRNAs (42.05%) considered as potential cis-acting regulators such as cis-regulatory element and sRNAs, but still, there are potential trans-acting regulators to regulate distant molecules such as CRISPR and antisense RNA. Moreover, the results indicated that TSSaRNAs could trigger more complex function as a catalytic function such as Riboswitch or to play a role in the defense against a virus such as CRISPR.
As a conclusion; based on the results of this study we could state that TSSaRNAs have several potential functions opening the experimental validation perspective.

Keywords:
TSSaRNA, ncRNAs prediction, *Halobacterium salinarum* NRC-1, Non-coding RNA, Functional annotation, Structural annotation, Rfam, ncRNAs functions, RNA structures, RNA interactions, RNA based regulation, Higher-order ncRNAs structures, TSSaRNAs-LSm interactions, RNA docking.

Resumo

Yagoub A. I. Adam. **Predição de função para TSSaRNAs (transcritos associados a sitios de início de transcrição) em *Halobacterium salinarum* NRC-1**. Doutorado (tese) – Programa Interunidades de Pós-graduação em Bioinformática (Departamento de Computação e Matemática (DCM) – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) e Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP), Universidade de São Paulo, Ribeirão Preto, 2019.

Os RNA não codificantes associados ao sítio de início da transcrição - em inglês, transcription start site associated non-coding RNAs (TSSaRNA) - foram observados nos três domínios da vida. No entanto, sem esforço confiável de anotação para identificar suas funções biológicas e seus mecanismos moleculares. Portanto, esse projeto levanta a questão de quais são as funções em potencial dos TSSaRNAs a respeito de seus papeis nas funções celulares. Para responder esta questão, nós objetivamos em identificar de forma eficaz os TSSaRNAs no organismo modelo *Halobacterium salinarum* NRC-1 (um microrganismo do domínio Arqueia) encubado em uma condição de crescimento padrão. Consequentemente, nós investigamos a estabilidade estrutural dos TSSaRNAs em relação a energias termodinâmicas. Ainda, fizemos a anotação funcional dos TSSaRNAs baseado na classificação funcional Rfam dos RNAs não-codificantes.

Baseada em uma abordagem estatística nós desenvolvemos um algoritmo para predizer TSSaRNA usando dados de sequenciamento de RNA de nova geração (RNA-Seq). Para investigar a estabilidade estrutural dos TSSaRNAs nós modelamos as estruturas secundárias minimizando a energia livre termodinâmica para alcançar a estrutura mais estável biofisicamente. Nós simulamos estruturas terciárias de TSSaRNAs baseado nas restrições das estruturas secundárias usando a ferramenta Rosetta-Common RNA. As estruturas de energia livre mínima seriam supostamente estruturas estáveis biofisicamente. Para investigar as estruturas de ordem superior (quaternária) dos TSSaRNAs, nós estudamos a hibridização entre os TSSaRNAs e seus genes cognatos como parte de um possível sistema de regulação baseado em RNA. Ainda, baseada na hipótese que os TSSaRNAs podem ligar à proteína para habilitar sua função, nós investigamos a interação entre TSSaRNAs e proteína Lsm que é conhecida por ser uma proteína chaperone que media função do RNA e está envolvida no processamento do RNA. Nosso pipeline para executar a anotação funcional dos TSSaRNAs objetivou classificar as TSSaRNAs em suas correspondentes classes Rfam baseado em dois passos: por meio de consulta das sequências TSSaRNA em relação a modelos de covariância de famílias Rfam ou por consulta de sequências Rfam em relação a modelos de covariância das estruturas de secundárias de consenso das estruturas secundárias nos TSSaRNAs.

Os resultados mostraram que o algoritmo de detecção teve sucesso em identificar um total de 224 TSSaRNAs que expressaram na mesma direção dos mRNAs e 58 TSSaRNAs que expressaram

no sentido oposto (antisenso) dos mRNAs. As moléculas TSSaRNAs identificadas mostraram um comprimento mediano de 25 nucleotídeos. A respeito da anotação estrutural dos TSSaRNAs, os resultados mostraram que a maioria dos TSSaRNAs possuíam estruturas secundárias estáveis termodinamicamente e suas estruturas terciárias foram capazes de formar estruturas mais complexas por meio de vínculos com outras biomoléculas. Quanto à formação de estruturas de maior de estruturas de alta ordem nos observamos que a maioria dos TSSaRNAs (92.2%) são capazes, pelo menos em princípio, de hibridizar em seus genes cognatos e, também, 55 TSSaRNAs evidenciaram interagir com a proteína Lsm. Além disso, os experimentos computacionais de docking demonstratam os complexos TSSaRNAs-Lsm associados com energia de ligação favorável com uma média de - 542900 kcal mole^{-1}. Quanto à anotação funcional dos TSSaRNAs, os resultados mostraram que a maioria dos TSSaRNAs (42.05%) podem ser consideradas potenciais reguladores atuando em cis tais como elemento cis-regulamentar e sRNAs, mas ainda há pontenciais reguladores atuando em trans para regular moléculas em *loci* distantes, tais como CRISPR e RNA antisense. Além disso, os resultados mostraram que TSSaRNAs podem potencialmente ativar funções mais complexas como uma função catalítica, tal como Riboswitch ou executar um papel de defesa contra vírus, tal como CRISPR.

Como conclusão; baseado nos resultados desse estudo, nós podemos afirmar que TSSaRNAs possuem várias funções em potencial abrindo a perspecitiva de validação experimental.

Palavras-chave:

TSSaRNA, Predição de ncRNAs, *Halobacterium salinarum* NRC-1, RNA não codificante, Anotação funcional, Anotação estrutural, Rfam, Funções de ncRNAs, Estruturas RNA, Interações de RNA, Regulação baseada em RNA, Estruturas de ncRNAs de ordem superior, Interações de TSSaRNAs-LSm, Docagem de RNA.

CHAPTER 1
Introduction

1.1 Introduction

In the last decade, studying the functions of RNA molecules and their mode of action is being a central research subject for both experimental and computational biology. The classical central dogma of molecular biology has considered RNA molecules as mediators that facilitate transferring information between DNA and protein. Recently using high-throughput techniques in studying the dynamic of transcriptome has revealed a plethora of non-classical RNAs. The non-classical RNAs are regularly transcribed by transcription machinery of the living organism as stable RNA molecules. Hence, these non-classical RNA molecules are not associated with any known protein they are named as non-coding RNAs (ncRNAs). Many studies have provided evidence that ncRNAs are capable of involving in many molecular functions beyond the translation machinery. The majority of ncRNAs are still considered as mysterious molecules and functionally unknown. However, many studies showed strong evidence that ncRNAs could play critical roles in triggering many functions. Therefore, ncRNAs have been considered as versatile biological molecules that could play roles in many cellular processes. Nowadays, many types of research are focusing on the field of RNA biology to understand the functions of the existing ncRNAs classes as well as discovering new classes and explore the aspects of ncRNAs potential roles in cellular processes.

Recently a new class of ncRNAs associated with the signal of transcription start sites (TSS) has been discovered in many species. As this new class of ncRNA is enriched near to the transcription start sites of coding genes; thus, it has named as transcription start sites associated RNAs (TSSaRNAs). TSSaRNAs are considered as ubiquitous molecules in all kingdoms of life yet their function remains not well understood. In this

study we are going to use the bioinformatics approach (computational approach) in an attempt to characterize TSSaRNAs and understand their potential biological functions in *Halobacterium salinarum* NRC-1.

1.1.1 Structure of the thesis

As it is essential to explain how the thesis has structured and the purpose of the various chapters and their connection to each other here I am providing the outlines and the structure of the rest of this thesis.

This chapter is an introductory chapter which includes the review of the literature besides the purpose and significance of this study. This chapter has been organized into the following sections: section 1.2 gives an overview of the three domains of life and their classification characteristics, section 1.3 introduces ncRNAs and their biological functions, section 1.4 focuses on the computational methods to annotate ncRNAs as this section is a milestone to understand most of the posterior methodology chapter. Therefore, it has organized to subsections to give solid information as follows:

- Subsection 1.4.1 focuses on the annotation of ncRNAs based on their structures, and it gives an overview of RNA structures including primary, secondary, tertiary and higher order structures then it goes on to describe and discuss various computational methods to predict these structures. Moreover, the subsection 1.4.1.3.2 covers the energy terms that stabilize RNA tertiary structures as well as the subsection 1.4.1.4 which covers RNA interactions systems and their molecular roles in the cellular processes.
- Subsection 1.4.2 gives an overview of the annotation of ncRNAs based on their sequences including the computational methods that used to study sequences alignment and the basic local alignment search tool (BLAST).
- Subsection 1.4.3 presents Rfam classification and annotations of ncRNAs.

The section 1.5 is going to highlight TSSaRNAs and their existence in all domains of life, this section covers TSSaRNAs biogenesis as well as their potential function. The section

1.6 highlights the next-generation RNA sequencing as one of major source of raw data for bioinformaticians; this section starts by covering the basic experimental design then goes throughout covering the basic computational method to analyze the raw data such as checking the quality control of raw data, mapping raw data to reference genome as well as describing the RNA-seq metric that used by bioinformaticians to answer the biological question. Section 1.7 presents the scientific problem and the rationale to conduct this research. This chapter will end up by section 1.8 which explains the objectives of this study.

Chapter 2 Describes and discusses the computational methodology of this study.

Chapter 3 demonstrates the results.

Chapter 4 discusses the findings of this study

Chapter 5 provides the conclusion and the limitation in this study.

This thesis is end-up by the section of appendices which demonstrates and provides some miscellaneous information.

1.2 Domains of life

Since 1977 the existence of the three domains of life as a model of evolution is certain. The classification of these domains is based on the studies of differences in the ribosomal ribonucleic acid (rRNA), and on the cellular membrane and its reactivity and sensitivity to antibiotics [1]. The reasons for using the rRNA molecule as the phylogenetic indicator because of its unique function throughout all the domains. Also, rRNA has a conserved primary structure (sequences), secondary, and tertiary structure. These structures are rarely changed over time. The three domains of life are referred as Bacteria, Archaea, and Eukarya (Fig. 1.1).

The Eukarya represents a domain of various organisms. The major cellular structure feature of this domain is that Eukaryotic cells contain a cellular nucleus and some organelles surrounded by the cellular plasma membrane. The Eukarya consists of wide broad of organisms vary from unicellular organisms, fungi, plants, to animals.

The Bacteria domain represents a wide range of unicellular organisms. Bacterial cells are lack of any cellular components surrounded by the plasma membrane. The major characteristic feature of Bacteria is that their cell wall chemically consists of peptidoglycan. The third domain of life is the Archaea which represents a domain of unicellular microorganisms. The primary molecular characteristic of Archaea and Bacteria is that both domains are lack of nucleus organelle or any other organelles surrounded cellular membrane in their cytosol [1–3]. However, Archaea could be distinguished from Bacteria by analysis the chemical composition of their cell wall where Archaeal cell wall does not contain peptidoglycan. Moreover, Archaeans are not sensitive to common bacterial antibiotics. Many Archaeans are considered as extremophile organisms because they are adapted to survive in very extreme environmental conditions. These extreme conditions are very hot conditions and/or very salty environments. Moreover, Archaeans are capable and adapted to survive in various habitats including soils, oceans, and even on/in other organisms.

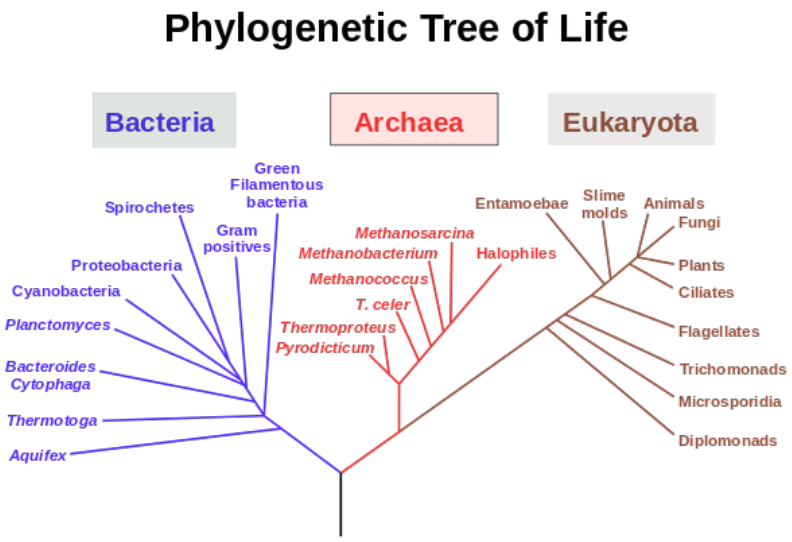


Figure 1.1: A phylogenetic tree of three domains of life

A figure shows the separation of the three domains of life: Bacteria (left), Archaea (center), and Eukaryotes (right). (source: `https://bio.libretexts.org/TextMaps/Microbiology/Book%3A_Microbiology_(Kaiser)`)

1.2.1 *Halobacterium salinarum* NRC-1

Halobacterium is a genus that belongs to the Archaean domain which consists of several species. Most of these species are tolerant of high salt concentration. Also, the species of this genus are capable of surviving in various extreme growth environments [4, 5]. Therefore, species belonging to the genus Halobacterium provide a research gate to study the mechanisms that address the molecular response, and an organism's adaptation to survive in extreme environments. The particular species *Halobacterium salinarum* NRC-1 has been used intensively as a model organism in the field of system biology. *Halobacterium salinarum* NRC-1 requires very simple growth requirements to be cultured in microbiological laboratories. *Halobacterium salinarum* NRC-1 requires only a high concentration of sodium chloride and other salts as an extreme environment; hence, *Halobacterium salinarum* NRC-1 neither requires a high incubation temperature nor an anaerobic environment. Taken advantage of the simple requirement for growth control besides the advent of sequencing technologies the whole genome of *Halobacterium salinarum* NRC-1 was sequenced in 2000 [6]. The sequenced genome revealed that the genome consists of one chromosome with 2,014,239 base pairs (bp) and two small plasmids which are: pNRC100 (191,346 bp) and pNRC200 (365,425 bp). The whole genome consists of 2,571,010 (bp) in total. This genome is characterized by a high G+C content (68%) and about 2,360 predicted genes [6]. Nowadays *Halobacterium salinarum* NRC-1 is considered as a promising model organism in system biology. Therefore, it has been studied intensively using well developed experimental techniques such as next-generation sequencing technologies, micro-arrays chips for high-throughput gene expression analysis, and the gene mutation methods [7–9]. Consequently, a huge data collection (which is referred as big data by data scientist) has been generated associated with various biological questions in different conditions. This big data associated with *Halobacterium salinarum* NRC-1 is still waiting for further analysis using bioinformatics efforts since no single study can exhaust all aspects of their data. These bioinformatics efforts will enhance our understanding of the molecular process of cellular activities at all levels.

1.3 Non-coding RNAs (ncRNAs)

The classical paradigm of central dogma in molecular biology has defined the RNA molecule (RiboNucleic Acid) as a type of nucleic acid that transcribed from a Deoxyribonucleic Acid template (DNA). RNA is proposed to function as a mediator to convey the biological information from DNA to protein. From a chemical point of view, RNA is considered as a polymer of nucleotides with three major components which are: a phosphate group, ribose sugar molecule of 5 carbon atoms, and a nitrogen base. The nitrogen base contains either purine or pyrimidine ring. There are four types of nucleotides based on the chemical properties of the nitrogen base. The four types of nucleotides are named as adenine (A), cytosine (C), guanine (G), and uracil (U). The nitrogen bases of RNA are the same as in DNA except that DNA contains thymine (T) instead of uracil (U).
Until recently in many molecular biology textbooks, RNA molecules have been classified into three major groups such as: Messenger RNA(mRNA) to serve as a blueprint (template) in protein synthesis process, Ribosomal RNA (rRNA) as a component of ribosomes that are essential for protein synthesis process, and the transfer RNA (tRNA) as an adapter for the protein translation process. Besides the three traditional types of RNA, later the next-generation sequencing technologies revealed a plethora of RNA molecules that are transcribed in all organism of the three domains of life but neither have known protein associated to it nor serve as ribosomal RNA or transfer RNA. Therefore, these types of RNA transcripts have named as non-coding RNAs (ncRNAs). ncRNAs have grouped into two major groups as infrastructure ncRNAs and regulatory ncRNAs. However, for reliable classification, many scientists prefer to group ncRNAs depending on their size into two major groups such as small ncRNAs and long ncRNAs. However, ncRNAs could be further grouped into more reliable functional groups. In general ncRNAs could be classified at least into the following groups: a) microRNAs (miRNAs) which are single-stranded RNA where the mature microRNAs are sized about (20-24 nucleotides), b) PiwiRNAs (piRNAs) which are sized about 24-31 nucleotides and are capable of forming a complex with Piwi proteins, c) small interfering RNAs (siRNAs)

which are sized about 20-24 nucleotides and they are capable of mediating mRNA silencing after the transcription process (post-transcription silencing), d) long no-coding RNAs (lncRNAs) which are characterized by their size of more than 200 nucleotides.
Upon discovering ncRNAs, versatile functions of ncRNAs have been revealed. For instance, ncRNAs could play a role as an enzyme (Ribozymes). Also, many studies have provided evidence for ncRNAs to play a key role in diverse biological functions. Moreover, many studies have proved that ncRNAs are capable of regulating many biological processes such as histone modification and transcription process. Furthermore, in the case of multicellular organism ncRNAs could play a role in the process of cellular differentiation. Nowadays, many researchers showed that ncRNAs are associated with various diseases including cancers. Due to the diversity of ncRNAs potential functions, system biologist considers ncRNAs as a "wild card" molecules that play critical roles in various cellular processes and trigger many biological functions [10, 11].

1.4 The computational aspects of ncRNAs annotation

The number of predicted ncRNAs is increasing day by day but still validation the function of the majority of ncRNAs remains a big challenge in the field of computational biology and system biology. So far, the bio-molecular functions of many ncRNAs molecules are still poorly characterized or even unknown. Due to the importance of the potential roles of the ncRNAs and the diversity of their predicted functions, the prediction of ncRNAs function has been a hot topic and an evolving research area in bioinformatics and system biology and their sub-fields.
The computational aspects to functionally annotate ncRNAs are growing into two major approaches. These approaches are the sequence-based functional annotation of ncRNAs and the structural-based functional annotation of ncRNAs. The sequence-based functional annotation is using the homology-dependent methods to infer the functions of ncRNAs. This approach is based on the hypothesis that given evidence of sequences similarity could assume an identical or similar role.

The second approach of ncRNAs annotation is the structural-based functional annotation; this approach is based on the concept that the molecular function of biomolecules is mainly driven by their higher orders folded structures. This approach compares the conversed motifs in ncRNAs stable structures to infer their potential functions. The structural-based functional annotation of ncRNAs could be succeeded even when the discrepancy in their sequences has existed.

As the set of the residues in RNA molecules contains a few numbers of nucleotides. Therefore, to get a reliable annotation for ncRNAs scientist should rely on both sequence-based and structural-based methods simultaneously. To cover the theoretical parts of the functional annotation of TSSaRNAs, this section is going to give an overview of ncRNAs annotation methods, and the structural features of ncRNAs. Also, this section is going to provide an overview of functionally annotate ncRNAs based on the well-known Rfam functional classification of ncRNAs.

1.4.1 Structural based ncRNAs annotation

To exert their molecular action, the linear transcript of the ncRNAs molecule is often folded around itself to form complementary base-pairing as secondary structure. The resulted RNA secondary structure is more likely to fold into more complex tertiary and higher order structures by adding more inter/intramolecular interaction [12].

To thoroughly understand the ncRNAs functionality a lot of efforts are paid to study and predict their structures and their interactions with other molecules hence, the computer experiments to model and simulate ncRNAs structures have considered as a hot research topic in recent years [12, 13]. Computer experiments offer the capability of studying the structures and their dynamics in a tiny time scale. Also, the computer experiments could accelerate the wet-lab experiments, and help in performing the correct experiment by reducing the technical errors when estimating the parameters. Furthermore, computer experiments reduce bench works costs in many cases.

As the structures of the ncRNAs could be decomposed into primary, secondary, tertiary structures as well as higher order structures [14]. Therefore, the following subsections

are going to give some details about the structural properties besides the computational methods to predict ncRNAs structures.

1.4.1.1 RNA primary structures

RNA primary structure is a linear single-stranded sequence of ribonucleotides. Unlike proteins, RNA primary structures could perform essential molecular functions. For instance, mRNAs are translated to proteins when they are in the primary structure. On the other hand, there is evidence that the secondary structures in coding regions of mRNAs regulate translation process and even it could inhibit the process [15].

1.4.1.2 RNA secondary structures

RNA secondary structures are formed as a result of the interaction between nucleotides. This interaction arises mostly as Watson-Crick hydrogen bond due to pairing between pyrimidine and purine bases. In some cases, non-canonical base pairing may be occurred such as G-U base pairing. Also, other types of non-canonical base pairing exist such as pairing with modified nucleotides [16]. RNA secondary structure is often considered as a stable molecule. The secondary structure of RNA shows RNA molecules as double-stranded molecules because parts of RNA single-stranded molecule are complementarily bound to other parts [17].

The RNA secondary structures could be annotated into structural elements which are also referred as structural motifs (Fig. 1.2). The most common RNA secondary structure element is called a hairpin loop. The hairpin motif is considered as the functional motif for many RNA secondary structures [18]. The hairpin consists of a stem-loop structure where the stem structure is made up by the complementary pairing of the nucleotides in a single RNA strand to look like a double-stranded RNA piece. On the other hand, the loop structure is a bubble like structure formed due to the presence of unpaired nucleotides in the stem structure. Different types of loops structures are distinguishable such as internal loop as unpaired nucleotides within a stem, external loop as loop appear between the ends of RNA molecules or multi-branch loop as a connection

of multiple stems structures [17]. One more important RNA secondary structure motif is called pseudo-knots (PK) motif. The pseudo-knot motif is formed due to the chance of the presence of complementary interaction between nucleotides in a hairpin, internal or multi-branch loop with other unpaired regions by Watson-Crick base pairing and/or non-canonical base pairing [17].

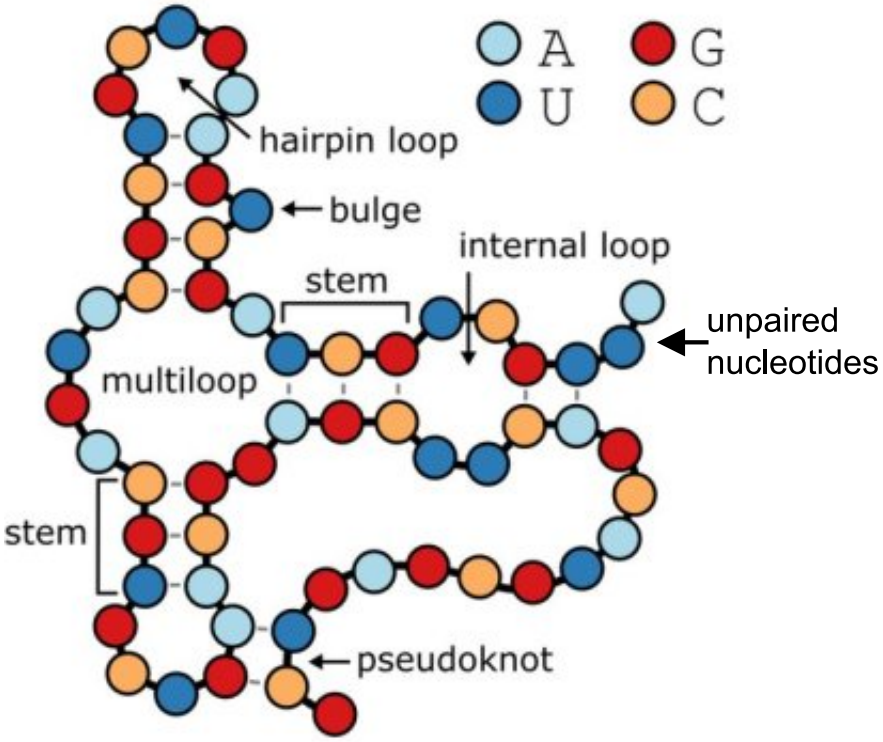


Figure 1.2: RNA secondary structure motifs

Adapted from: Oregon State University `http://bprna.cgrb.oregonstate.edu/`

1.4.1.2.1 Prediction of RNA secondary structures

Various tools to predict RNA secondary structures exist. These tools differ from each other mostly in their underline prediction algorithm. Next few subsections are going to highlight the most popular methods to predict RNA secondary structures such as the methods that based on comparative sequence analysis, the methods that based on minimizing thermodynamic energy, and the methods that based on the based on statistical sampling.

Prediction of RNA secondary structures based on sequences similarity

This method is based on the assumption that the existence of higher similarity in RNA sequences will result in similar functional structures. Therefore, RNA secondary structures could be predicted by performing comparative sequence analysis using the well-developed sequences alignment methods. The comparative sequence analysis search for a conserved region in target RNA molecules then compare the conserved region with RNA for which structure is already known. Prediction of RNA secondary structure based on the comparative sequence analysis (homologous sequences) will yield a reliable, functional structure when the structural motifs are conserved among the RNA molecules or when conserved functional residues exist [17]. The most challenging computational aspect of using a method that is based on the comparative sequence analysis is to overcome the slowness of the primary step in the prediction algorithm. Because, the primary step of aligning sequences will be time-consuming for a large number of sequences [13,17]. Many tools have been developed to predict RNA secondary structure based on the comparative sequence analysis, and some of them are included in the table (1.1).

Prediction of RNA secondary structure based on RNA thermodynamics

From the thermodynamics point of view, RNA secondary structure could be predicted without comparing to known structures or without knowing any conserved structural regions. The methods that rely on thermodynamics suppose that the most stable RNA secondary structure of any given molecule is which possess the minimum Gibbs free en-

ergy in the system [19]. These methods calculate and minimize the total free energy of RNA structure of all secondary structure elements that composed by a particular RNA molecule. Many tools to minimize the free energy of RNA molecule using dynamics programming are available [13, 17, 20, 21]. Examples of these tools are included in the table (1.1). The accuracy of the prediction of RNA secondary structures based on the thermodynamics methods rely mainly on the experimentally measured parameters for base pairing and other secondary structure elements. Any change in the estimated parameters will change the stability of predicted structures. Changes in the thermodynamics parameters could even yield a different structural conformation. Nowadays many software use the parameter estimated by Michael Zuker [22, 23] as correct parameters. However, the process of parameters estimations is still evolving.

Prediction of suboptimal RNA secondary structures

As it is possible of the existence of multiple biologically active structures for a single RNA molecule. Therefore, prediction of all structures which have similar minimum free energy is essential to obtain all active conformations of the individual RNA molecule. Different structures with small change from the lowest free energy for a single RNA molecule called suboptimal structures. Zuker addressed efforts to predict the suboptimal structures using dynamic programming methods [20]. Also, Williams and Tinoco contributed to the efforts of predicting the suboptimal structures using dynamic programming methods [21]. Mfold program is a popular program to generate optimal and suboptimal structures of an RNA molecule. More software that predicts the suboptimal structures exist such as RNAstructure, ViennaRNA Package, and many other tools. Examples of these tools are included in the table (1.1).

Prediction of RNA secondary structure based on statistical sampling

Based on the statistical mechanics' rules, we could consider the prediction of RNA secondary structure as a complex system. In this complex system, all possible structures of any particular RNA molecule would not be obtained. Therefore, RNA secondary structures problem could be addressed by sampling structure according to the concepts

of Boltzmann distribution in statistical mechanics. The statistical sampling approach allows sampling of any secondary structure element and estimates the probability of its existence according to a calculated partition function [24]. The probability of sampling a secondary structure in a particular RNA sequence at the equilibrium state could be calculated based on the given formula (Equation 1.1). This formula has been adapted from the literature [24].

$$\frac{e^{(-E(I)/RT)}}{U} \tag{1.1}$$

E(I) is the free energy of the given structure **I**,

R is the ideal gas constant,

T is the absolute temperature in kelvin unit,

U is the partition function for all secondary structures of a given RNA molecule. This partition function could be calculated using the formula in given equation (Equation 1.2).

$$U = \sum (e^{-E(I)/RT}) \tag{1.2}$$

Prediction of Pseudoknots

The structural topologies such as pseudoknots structural motifs are involved in many RNA functions. Therefore, it is interesting to be predicted as secondary structure elements. Many software showed a limitation in predicting the pseudoknots especially that use classical dynamic programming methods [17], but there are some tools are succeed to overcome the computational challenges in predicting pseudoknots such as Ipknot, KineFold, PknotsRG, pKiss, and many other tools. Examples of these tools are included in the table (1.1).

Table 1.1: Examples of open source computational tools for RNA secondary strucures prediction

Purpose	Prediction method	Prediction Tools
Prediction of RNA secondary structures	sequence comparative	Multilign, CentroidAlign CONSAN, Foldalign, KnetFold, Murlet,MXSCARNA, R-COFFEE, TurboFold, RNAforester, RNAFold.
	Statistical sampling	Sfold, UNAfold.
Prediction of RNA secondary structures with suboptimal structures	minimizing thermodynamic energy	ViennaRNA package, mfold, RNAstructure.
	Statistical sampling	ViennaRNA package, RNAstructure
Prediction of RNA secondary structures with pseudoknots	minimizing thermodynamic energy	Ipknot, KineFold, Pknots, PknotsRG, pKiss.

1.4.1.3 RNA tertiary structures

RNA secondary structures are more likely to fold into more complex RNA tertiary structures. RNA tertiary structures are made up due to the interactions between RNAs secondary structure elements to result in a folded three-dimensional shape. The tertiary structures of ncRNAs could be considered as the active form that triggers the biological function [14]. The following subsections are going to give some details about the computational methods to predict ncRNAs tertiary structures as well as demonstrating the energy that stabilize RNA tertiary structures.

1.4.1.3.1 Prediction of RNA tertiary structures

RNAs tertiary structures (3D) are essential for their bio-molecular functions. Therefore, prediction of RNA tertiary structures from the primary or the secondary structures is a milestone step in predicting and analysis ncRNAs molecular functions. RNA tertiary structures prediction using wet-lab experiments is increasing rapidly, for instance, querying the worldwide repository of experimentally predicted tertiary structures of biomolecules -Protein Data Bank (RCSB PDB)[1]- in February 2019 using the keyword RNA has hit total 11052 tertiary structures that contain RNA molecules. The RCSB repository also shows the number of predicted structures that contain only RNA tertiary structures is growing fast (Fig. 1.3). A large number of RNA tertiary structures have been predicted experimentally yet many computational efforts are needed to predict RNA tertiary structures *in silico*. Nowadays, the most common computational methods to predict RNA tertiary structures are categorized into two main groups: a) predicting RNA tertiary structures based on the homology modeling; b) predicting RNA tertiary structures based on the molecular simulation.

Based on the concept of the homology molding, RNA tertiary structures could be predicted successfully when a tertiary structure template exists. The algorithms in any homology modeling approach generally will consist of three steps: the first step is to find an identical or similar sequences between model and template, then align the similar sequences with adjusting the geometrical configuration of model depending on the template coordinates, and the final the hardest step is to count the insertion and deletion in sequences to get the closest structure [27].

The approach to predict RNA tertiary structures based on the simulation methods are referred as physical methods. These physical methods could be divided into two major classical classes which include: a) the Monte Carlo sampling approach where RNA tertiary structures modeled using the Monte Carlo sampling process. An example of the algorithm that uses a Monte Carlo sampling method to predict RNA tertiary structures is called fragment assembly of RNA (FARNA). In FARNA algorithm the process starts

[1] www.rcsb.org

by sampling and assembling short fragments of RNA about 1-3 nucleotides substructure from existing databases, then minimizes the structural energy based on knowledge-based energy function. The knowledge-based energy function could be extended to a method called Fragment Assembly of RNA with Full Atom Refinement (FARFAR) [28] to model the full atom potential energy. FARFAR knowledge-based energy function allows getting more realistic hydrogen bonds. b) The other class of the physical methods is the molecular dynamic approach where RNA tertiary structural dynamics and folding trajectory could be predicted successfully by using the general molecular dynamic methods with an improved RNA force field. Many molecular dynamics simulation packages could be used to simulate RNA tertiary structures such as chramm, gromacs, and Amber. Refer to the table (1.2) which includes the most common tools that used to predict RNA structure.

Table 1.2: Examples of open source computational tools for RNA tertiary strucures prediction

Purpose	Prediction method	Prediction Tools
Prediction of tertiary structures	sequence based methods	RNA-MoIP, MC-Pipeline
	coarse-grained structural models	iFoldRNA
	Molecular Dynamics	SimTK: NAST
	A Probabilistic Model of RNA Conformational Space	BARNACLE
	Monte Carlo Sampling	TreeFolder, Rosetta software.

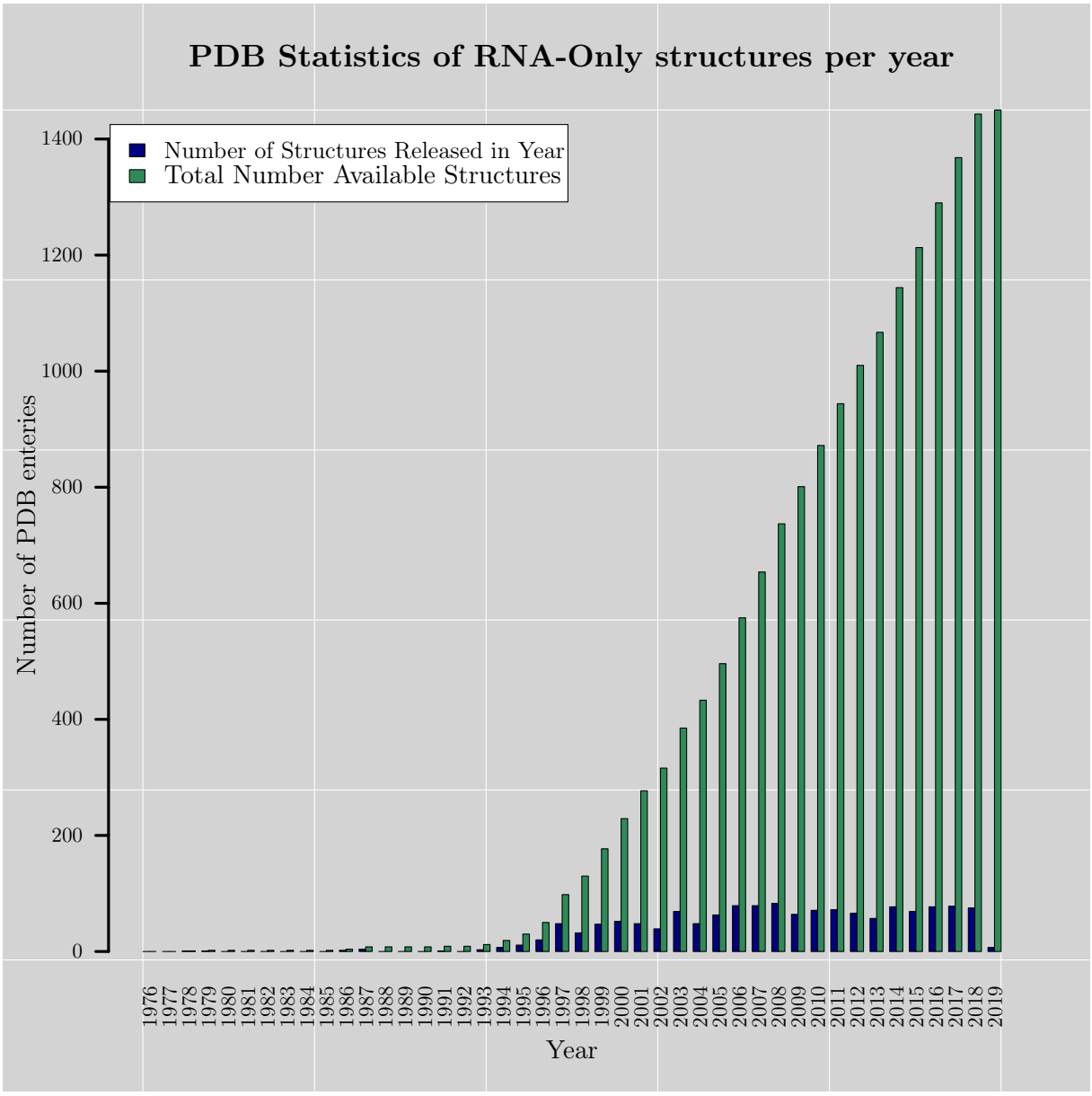


Figure 1.3: Graph displays the statistics for the growth of RNA-only tertiary structures per year (1976- February 2019) at Protein data bank repository.
To regenerate this graph visit the protein data bank (RCSB PDB) website at
`https://www.rcsb.org/stats/growth/rna`

1.4.1.3.2 Energy stabilizing RNA tertiary structures

The stability of predicted RNA tertiary structures could be interpreted from the values of the estimated energy of RNA tertiary structures. The energy terms of RNA tertiary structures could be classified into low-resolution energy and high-resolution energy [29]. The low-resolution energy calculated from RNA tertiary structures considering only the coordinates of residues without considering the full atomic details of each residue [30]. The low-resolution energy terms give only the contours of the RNA structure, but also it could give some few atomic details such as hydrogen at low resolution and van der Waals interactions [30].

Regarding the high-resolution energy terms, these terms give more atomic details by considering the interactions of all atoms in particular structure [31]. It is important to notice that different energy terms could be obtained when modeling RNA tertiary structure using different packages. Also, the values associated with each energy term could be varied as each package has its scoring function. In the case of Rosetta software, the knowledge-based scoring function of modeling RNA tertiary structures is based on the methods that used to model tertiary structures of protein [31]. The low energy terms that reported by Rosetta package are including various terms such as a radius of gyration, base–backbone interaction, backbone-backbone interactions, base to base interactions, stacking interaction, Van der Waals interactions, and other terms. The Rosetta full-atom energy terms reported by Rosetta include many physical forces and chemical potentials such as van der Waals forces, full atom hydrogen bonding, solvation and desolvation energy for both polar and apolar (non-polar) groups, the energy that arose from backbone torsion angles and many other terms [29]. As we used Rosetta package to model TSSaRNA tertiary structures, thus we have listed all RNA energy terms that could be reported by the Rosetta software in the given tables (table 1.3, and table 1.4). More details of Rosetta energy terms and their contribution on the RNA tertiary stability can be found at the following link `http://www.rosettacommons.org/docs/latest/`. It is important to notice that Rosetta energy function (force field)

is a combination of both physics-based force fields and knowledge-based force fields. Therefore, the value of energy terms that generated by Rosetta scoring function is not expressed on universal energy unit such as Kilocalorie per mole or KT energy unit (the product of the Boltzmann constant and the temperature). Instead, Rosetta software reports the energy terms on a unit called Rosetta Energy Unit (REU) [30].

It is crucial to state that, to interpret RNA structural stability the researchers often pay attention to the energy terms that are well-known by their contribution in triggering the biological functions, but at the same time they should emphasize on the overall estimated energy rather than rely on only one or few energy terms.

Instances of the important terms of the low-resolution energy terms that reported by Rosetta package for a particular RNA tertiary structure are including: a) the term of radius of gyration which is considered as the root-mean-square distance of RNA structural components from its center of the mass, we could think about radius of gyration as "How secondary structures are compactly packed in to the tertiary structure for each RNA", thus low radius of gyration indicate evidence of more compact structures. b) the term of Van der Waals interactions which is used to check the existence of any clash in residues of the predicted RNA tertiary structure. The value of Van der Waals interactions allow us to know if there any overlap between residues, thus the lower value of Van der Waals interactions means good tertiary structure model and no many clashes between residues. c) the term of energy that arose from 2' hydroxyl group, the importance of this term arises as some studies revealed that the additional hydroxyl (OH) group which present in the second carbon atom of ribosome sugar plays significance biological and structural functions [32]. Therefore, it is essential to estimate the energy associated it as isolate energy term.

Instances of the important terms of high-resolution energy terms that reported by Rosetta package for a particular RNA tertiary structure are including: a) Lennard-Jones interactions, this term is decomposed into three parts: i) Attractive between atoms in different residues, ii) Repulsive between atoms in different residues, and iii) Repulsive between atoms in the same residue; b) the term of the solvation energy, this term measures the

amount of the energy that needed to dissolve a given RNA tertiary structure in a solvent. The energy associated to the solvation is useful for interpretation the type of chemical reaction and determine whether the interaction is an endothermic reaction which needs external energy to initiate the reaction or it is an exothermic reaction (spontaneous reaction) that releases energy into ambient. The endothermic reaction type of reaction is referred as a simultaneous reaction; c) the term that associated with the hydrogen bonds. The importance of this term arose due to the significance of the hydrogen bond in the biological system and its reaction.

Table 1.3: Low resolution energy terms of Rosetta package. These terms are obtained from the RosettaCommons documentation.

Energy Term/Score	Interpretation
rna_rg	Radius of gyration for RNA
rna_vdw	Low resolution clash check for RNA
rna_base_backbone	Bases to 2'-OH, phosphates, etc.
rna_backbone_backbone	2'-OH to 2'-OH, phosphates, etc.
rna_repulsive	Mainly phosphate-phosphate repulsion
rna_base_pair_pairwise	Base-base interactions[1]
rna_base_pair	Base-base interactions[1]
rna_base_axis	Force base normals to be parallel
rna_base_stagger	Force base pairs to be in same plane
rna_base_stack	Stacking interactions
rna_base_stack_axis	Stacking interactions should involve parallel bases.
atom_pair_constraint	Harmonic constraints between atoms involved in Watson-Crick base pairs specified by the user in the params file
rms	all-heavy-atom RMSD to the native structure

superscript **1** indicates both Watson-Crick and non-Watson-Crick interaction.

Table 1.4: High resolution energy terms of Rosetta package. These terms are obtained from the RosettaCommons documentation

Energy Term/Score	Interpretation
atom_pair_constraint	Harmonic constraints between atoms involved in Watson-Crick base pairs specified by the user in the params file
rms	all-heavy-atom RMSD to the native structure
fa_atr	Lennard-jones attractive between atoms in different residues
fa_rep	Lennard-jones repulsive between atoms in different residues
fa_intra_rep	Lennard-jones repulsive between atoms in the same residue
lk_nonpolar	Lazaridis-karplus solvation energy, over nonpolar atoms
hack_elec_rna_phos_phos	Simple electrostatic repulsion term between phosphates
hbond_sr_bb_sc	Backbone-sidechain hbonds close in primary sequence
hbond_lr_bb_sc	Backbone-sidechain hbonds distant in primary sequence
hbond_sc	Sidechain-sidechain hydrogen bond energy
ch_bond	Carbon hydrogen bonds
geom_sol	Geometric Solvation energy for polar atoms
rna_torsion	RNA torsional potential.

1.4.1.4 Higher order RNA structures (RNA interactions)

RNA tertiary structures most probably go through more complex higher order structures by interacting and binding with other biomolecules (Protein, DNA, RNA) to result in a quaternary RNA structure as a higher level of complex RNA structure. The higher-order structures of RNA are essential components in cells and play critical roles in many fundamental biological processes [14]. For instance, in the translation process, ribosomal RNAs bind to ribosomes to build the translation machinery. Moreover, many researches have revealed that RNA-biomolecules interactions play essential roles in triggering and regulating many critical cellular processes [33]. These interactions occur as part of various regulation machinery to regulate many cellular processes starting by regulating transcription process throughout controlling many complex regulatory types of machinery such as playing as a major regulator of gene expressions as well as involving in the complex cellular communications' mechanism. Nowadays studying RNA interaction with DNA, RNA, or protein is being a hot research topic for computational biologist and bioinformaticians. Many researchers are investigating the fundamental mechanism of RNA-biomolecules interactions alongside prediction of RNA binding sites and the conditions for interactions. Studying RNA-biomolecules interactions is being a topic of research interest to experimental biologist by generating new experimental data or by validating the findings revealed from the computational experiments. The next few paragraphs provide an overview of the interactions of RNA molecules with various biomolecules and their roles in the cellular process.

RNA-DNA interactions often occur as a dynamical hybridization (refereed as RNA-DNA triplex structures). The triplex structure is a helix consists of three nucleic acid strands as non-canonical nucleic acid structures. Furthermore, many bioinformaticians are referring to the RNA-DNA interactions as R-loop structures [34]. RNA-DNA interactions play various roles in controlling the transcription process by the action of chromatin remodeling or by controlling DNA stability [34]. Besides controlling the transcription process of a given organism, RNA-DNA interactions could play as trans-acting regula-

tors to regulate the transcription process of foreign genome such as CRISPR-mediated DNA cleavage [35].

Regarding RNA-RNA interactions it is well-known as an essential regulator in gene expression machinery. RNA-RNA regulatory systems regulate the gene expression via a cis-acting mechanism to control a close gene such as regulating the closest mRNA or via trans acting mechanism to regulate genes far away from the genomic location of transcribed RNA. It is also well-known that RNA-RNA hybridizations engage in controlling the gene expression beyond the transcription process by participating in post-transcription regulation, translation regulation or by influencing mRNA degradation and stability. Various RNA-RNA interactions systems exist naturally in all cell types for instances the interaction of small ncRNAs with their targets which including the binding of snRNAs and snoRNAs to their targets and the hybridization of microRNAs and sRNAs with their target mRNAs [36]. Besides the natural cellular system of RNA-RNA interactions, there is an artificial system of RNA-RNA interactions via siRNA-mRNA hybridization to control the expression level of targeted mRNA [36].

Regarding the interactions between RNA and Protein, it is clear that RNA-protein interactions are essential in all vital biological processes starting from transcription and translations, throughout regulating all cellular functions [37]. In all cellular systems, there are a plethora of proteins that are capable of binding to RNA molecules to maintain the critical cellular activities, hence these are proteins called RNA binding proteins (RBP). RNA-Protein interactions trigger many vital biological processes through binding to specific binding sites on single-stranded RNA molecules, binding to motifs on the secondary structures, binding motifs on the folded RNA tertiary structure, or even bind non-specifically to RNA molecules to trigger its function. Many examples of RNA-binding protein are well characterized for instance ribosome protein which binds to ribosomal RNA. A particular example of RNA-binding protein which plays a significant role in RNA processing is called LSm protein. LSm proteins belong to the family of Sm-like proteins which are existed across all domains of life as eukaryotic LSm (up to 16 LSm proteins), eukaryotic Sm (up to 7 Sm proteins), eukaryotic SMN/Gemin proteins,

archaeal LSm proteins (up 3 LSm proteins), and the unique bacterial Hfq protein. The major characteristic of LSm protein is its tendency to form a complex ring structure of LSm subunits. LSm protein is implicated in various RNA processing including messenger RNA splicing, messenger RNA decay, stabilizing RNA, or play as a chaperone molecule to trigger RNA-RNA or RNA-protein interactions [38]. In the Archaea domain, in particular, the species that belong to the genus Halobacterium including *Halobacterium salinarum* NRC-1 encode a single LSm protein (UniProt ID: B0R5R2)[2]. This single LSm protein is phylogenetically related to the LSm1 subfamily. Our research group in the laboratory for biological information processing -LabPIB-[3][4] is working on LSm-RNA interactions prediction and experimental detection regarding different classes of RNAs.

1.4.1.4.1 Prediction of higher order RNA structures

The computational task to predict RNA higher order structures is still a challenging task due to its computing time and the accuracy of its result; however, there are many promising tools to overcome these challenges. The tools that predict RNA higher order structures often use the algorithms that are applied to solve somewhat similar tasks. For instance, to predict the nucleic acids interactions (RNA-DNA or RNA-RNA) bioinformaticians often use the algorithms which were initially being used to predict RNA secondary structures mainly that depend on the dynamic programming methods or depend on the molecular simulation methods [25].

In the case of RNA-Protein interactions, it has well-established bench-work protocols, but it is a much more challenging computational task due to the difficulties that arise when implementing an algorithm that used to solve similar problems such protein-protein interaction. The current computational approach to predict RNA-Protein interactions often consists of two steps: the first step is to find where the small molecule (usually the RNA) as ligand geometrically fit to bind the larger molecule (usually the Protein) as a receptor. Finding the pocket where the ligand fit to the receptor is a crucial step and

[2] http://www.uniprot.org/uniprot/B0R5R2
[3] http://dcm.ffclrp.usp.br/pesquisa$_l ab.php?codlab = 2codcurso = 2$
[4] http://labpib.fmrp.usp.br/

it depends on both the geometrical shape and the charge on the surface. Most of the existing tools for docking experiments use the Fourier transformations method to find the interaction pocket [26]. After finding the best fit geometrical interaction between RNA and protein, the second step is to calculate the interaction (binding) energy to estimate the thermodynamic stability of the RNA-Protein complex. The calculation step depends on the geometrical shape, charges, and solvent. Most of the tools that calculate the binding energy in RNA-Protein complex are simplifying the calculation of binding (dissociation) free energy by considering only the contribution of both solvation energy and the electrostatic statistic energy and neglecting many other energy terms. The neglected energy terms are very sophisticated to be calculated such as the molecular mechanics' energy, the entropic changes in the system and many others [5]. Apbs (Adaptive Poisson-Boltzmann Solver) tool[6] is an example of a computational tool that calculates the binding energy in RNA-Protein complex.

1.4.2 Sequence based ncRNAs annotation

It is common to use the assumption that the sequences arose from the same ancestor perform a similar function. Hence, the potential function of many ncRNAs could be inferred by providing evidence of the existence of functionally annotated molecules in databases that sharing the same ancestor with the targeted ncRNA. The terms homologous (homologous sequences) is used to refer to the existence of a common ancestor between sequences. Sequence homology is inferred when there is a high similarity between two or more sequences to suppose both similarity and descending from a common ancestor [39]. There are two terms associated and imply the concept of homology which are orthologs and paralogs. Orthologs occur when the homologous sequences are in different species and arose from an ancestral gene. In the case of orthologs, it is not necessary that these orthologs sequences are responsible for the same function, but there is a high probability that they perform a related function. The other term is called paralogs

[5] http://apbs-pdb2pqr.readthedocs.io/en/latest/examples/binding-energies.html
[6] http://www.poissonboltzmann.org/

which indicates homologous sequences within the same species which is generally known as sequence/gene duplication to perform somewhat similar function [39].
It is important to the point that, if two sequences are sharing somewhat identical or similar nucleotides that do not indicate these sequences directly are descended from the common ancestor, but a very high similarity among along sequences is providing strong evidence of homology. The degree of similarity is estimated by calculating the alignment score [39]. Most of the alignment tools use a very sophisticated and rigorous statistical method to calculate the alignment score. However, a straightforward naive intuitive way to estimate an alignment score is by using an additive formula. This additive formula is represented by the following equation (Equation 1.3).

$$AlignmentScore = \#Identity + \#mismatch - \#gapPenality \tag{1.3}$$

Where: Identity refers to the identical nucleotides in both sequences. Mismatch occurred when a nucleotide had changed to one with similar properties. The mismatch often occurs in protein residues. Gap occurs when there is an insertion or deletion in one or more residues.

1.4.2.1 Sequence alignment

The task of computing sequence similarity to infer the functional homology is known as a sequence alignment. The basic idea behind sequence alignment is to recognize somewhat a significant similarity between a target sequence with an already functionally annotated sequence or by comparing the target ncRNAs to the databases of functionally known molecules. When performing sequence alignment of only two sequences is technically known as pairwise alignment. Extending the pairwise alignment to include aligning many sequences -sequences in databases- will be referred as multiple sequence alignments [39]. The task of sequence alignment could be solved computationally by using the general dynamic programming algorithm. The implementation of dynamic programming divides the sequence alignment into two broad major classes such as a) global alignment where the similarity between sequences is compelled to be extended over the full length of the

sequences. Global sequence alignment usually leads to many gaps. Moreover, global sequence alignment is less sensitive for highly divergent sequences. b) the second class of sequences alignment is known as local alignment. In this class, the result of the dynamic programming focuses only on regions of higher similarity between the sequences. Therefore, local alignment focuses mainly on a subset of sequences rather than the entire sequences. Focusing only on a subset of sequences result in a reasonable reduction in the computing time. Generally, bioinformaticians start performing sequence alignment by local alignment if the result showed promising similarity then they could extend it into the global alignment. When no hits in local alignment, then it will be meaningless to perform global alignment on the given sequences. The following subsection provides an overview of the blast algorithm which was used intensively to annotate the ncRNAs by searching sequences similarity.

1.4.2.2 Basic local alignment search tool (BLAST)

Based on the assumption that for any homologous sequences there will be a short ungapped region of high similarity between them thus the basic local alignment search tool has been introduced by Altschul et al. in 1990 [40]. Blast is a heuristic method for local alignment designed for database searches. Nowadays, researchers could perform blast easily either by using the Blast NCBI web-server or its standalone tool to run it on the local machine. Blast algorithm is capable of computing pairwise alignment of input sequences against all known sequences in a database then detect the best scoring hits which passed the statistical threshold parameter to considered as a significant homolog. The statistical threshold in the blast is called E-value (Expect value). The value of E-value is corresponding to the number of random hits that might be obtained when searching a database of a certain size. Conceptually E-value is different from the usual P-value, but numerically it is similar for low values. Generally, the two sequences are considered homologous if the corresponding E-value is very small. Since 1997 a new version of the blast has been developed and called blast2. Blast2 (the new version of the blast) is faster than the original blast algorithm and allows better alignment and extension for the local

alignment. Refer to the table (1.5) for the common blast algorithm.

Table 1.5: The common algorithms of blast

Algorithm	Sequence query	Target Database
blastn*	Nucleotide	Nucleotide
blastx	Translated nucleotide in all six frames	Protein
tblastx	Translated nucleotide in all six frames	Translated nucleotide in all six frames
blastp	Protein	Protein

*Mostly used when working with ncRNAs

1.4.3 Rfam based ncRNAs annotation

A huge number of ncRNAs have been discovered and still day by day scientists discover more and more new ncRNAs molecules in various organism associated with different cellular conditions. As a result of the emerging needs of organizing the large information arose from ncRNAs, Rfam team has constructed a comprehensive database called Rfam [41]. Rfam is a database that contains a collection of ncRNAs families. The ncRNAs families are represented in three forms which are: multiple sequence alignments, consensus secondary structures, and covariance models. Rfam team classifies ncRNAs depending on the information retrieved from European Nucleotide Archive (ENA)[7]. ENA provides a reliable ncRNA functional annotation based on the sequencing information, and/or from the experimental evidence in the literature. Rfam database is hosted at `http://rfam.xfam.org`. The recent release of the Rfam database is version 13.0 (released in September 2017) which includes 2686 annotated ncRNAs families where 115 of them are new families.

1.4.3.1 Rfam major classes

Rfam team has classified the ncRNAs sequences into functional families (Fig. 1.4). The classification is based on the existence of conserved sequences in multiple alignments of

[7]https://www.ebi.ac.uk/ena

ncRNAs and/or on the existence of consensus secondary structure motifs. The resolution of Rfam classification is not totally strict about avoiding false positive and false negative sequences due to the biologically overlapping functions of some ncRNAs. However, Rfam families are still considered as good enough to annotate ncRNAs sequences based on the similarity between the queried sequences and the seed sequences that represent each family. Rfam has grouped the underline ncRNAs families into three major classes which are: cis-regulation elements, genes, and intron.

The major class of cis-regulatory elements consists of a non-coding DNA motif that lay on the upstream regions of the targeted gene. This class often regulate the transcription or the translation processes of downstream genes. Rfam team has organized the cis-regulatory elements into at least following motifs: (a) internal ribosome entry site (IRES) which regulates the ribosomal entry thus it regulates the translation process; (b) Ribosomal frameshift element which regulates the translation process of targeted genes; (c) leader element which regulates the expression process by binding tRNA; (d) Riboswitch which regulates the gene expression of targeted genes after selectively bind specific metabolites; (e) thermoregulator associated with heat shock response.

The major class of ncRNAs genes consists of ncRNAs molecules that are transcribed from various genome coordinates. This class is not restrictedly associated with the upstream region of targeted genes. Moreover, some ncRNAs genes are capable of triggering their own function regardless of the possibility of regulating of targeted genes. Rfam has classified the major class of ncRNA genes into the following sub-classes: CRISPR, antisense, miRNA, rRNA, ribozyme, sRNA, snoRNA, splicing, and tRNA.

The major class of intron consists of ncRNAs molecules that are part of mRNA but do not code for protein. The ncRNAs molecules from the intron region could be capable of regulating the expression of targeted genes. Therefore, Rfam has considered these ncRNAs molecules from the intron region as a major class of ncRNA. In addition to regulating mRNA, the intron region could code for specific ncRNA class such as snoRNA or miRNA.

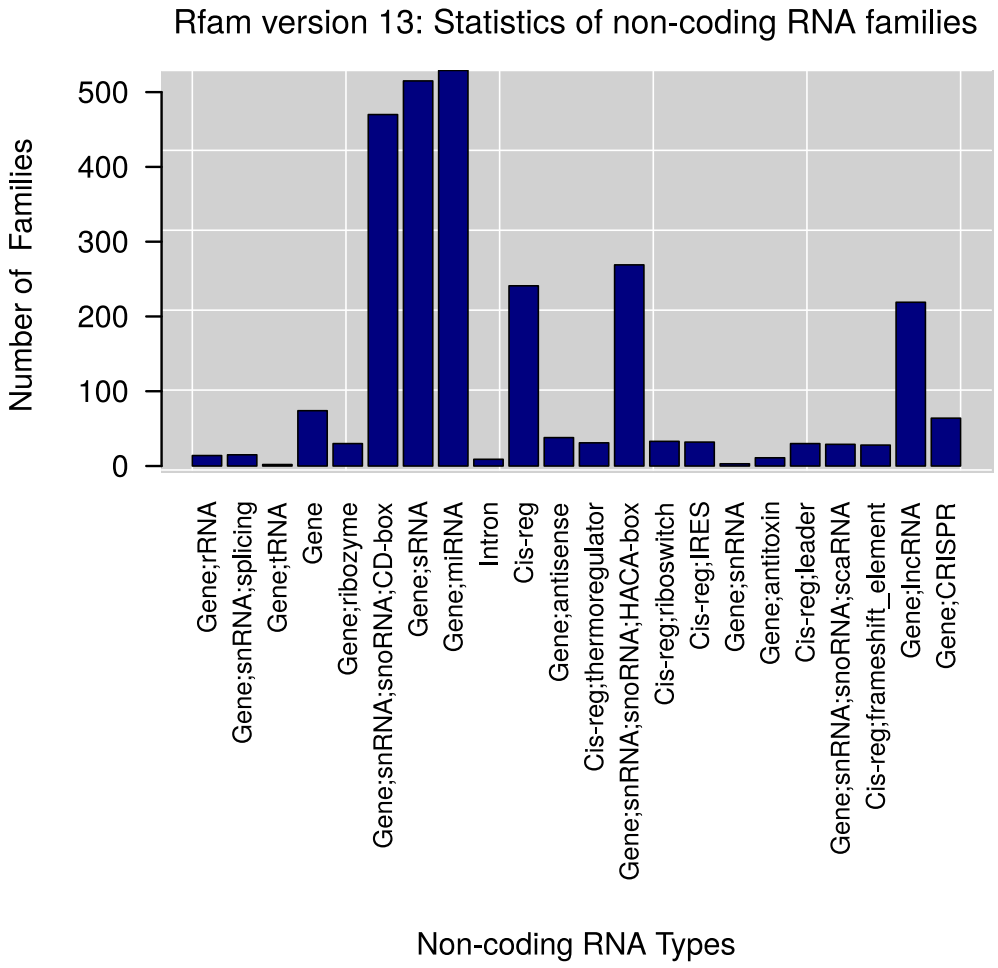


Figure 1.4: Representation of ncRNAs classes in Rfam v13.0

1.4.3.2 Covariance models (CMs)

The covariance model (CM) has introduced in 1994 by two separate research articles. One article is published by Sakakibara and his colleagues [42] while Eddy and his colleague Durbin publish the other [43]. CM is a powerful model that suited to represent the sequences and its structures in RNA multiple sequences alignments. CM has the possibility to represent the structural information even for sequences of a single RNA molecule. From

the statistics point view, CM is considered as a probabilistic model based on the concept of stochastic context-free grammar, which in turn is a generalization of the famous hidden Markov model. CM is capable of giving information about the conserved sequences in any RNA multiple alignment profiles and capture their consensus secondary structures. CM represents multiple alignment profiles as a binary tree where each node represents a secondary structure motif (but Pseudoknots cannot be represented). Without going to the sophisticated details of the CM, we could precisely characterize the grammars in CM in three items: a) a set of abstract nonterminal variables, these variables explain the secondary structures motifs; b) a set of terminal variables, these variables explain the actual sequences that appear in RNA sequences, i.e., A, C, G, and U; c) a set of production probabilities which are known as emission probabilities as well.
Rfam team model each Rfam ncRNA family into its corresponding covariance model from selected sequences that represent the particular family. Providing a CM and an RNA sequence we could calculate the probability that the target sequence belongs to this model. Hence, we could successfully annotate the ncRNAs sequences to their corresponding Rfam families.

1.4.4 ncRNAs annotations: challenges and future directions

It is expected that ncRNAs annotation will overcome the current challenges on both structural based and sequences based annotations approaches and constructing a tool which takes into consideration the stable biophysical structures and its corresponding biological function based on the sequence-based motifs. Also, by introducing the field of the RNA chemical biology, it is expected investigators will pay attention to the ncRNAs residual modifications and their structural and biological roles. Regarding the Rfam annotation approach, it is expected that Rfam team will improve the algorithm of CM to consider the pseudoknots motifs and the structural dynamics.

1.5 Transcription Start Site Associated RNAs

Recent studies that investigate the dynamics of the transcriptome in living organisms using advanced high-throughput technologies has revealed the existence of widespread small ncRNAs expressed near to transcription start sites of the majority of annotated coding genes. These molecules are considered as a new class of ncRNAs, subsequently, are named as transcription start site associated RNAs (TSSaRNAs) [44–46]. The ncRNAs that belong to this new class are considered as ubiquitous molecules among all the domains of life. The major two features to characterize TSSaRNAs are: a) their length that ranged of 16 to 200 nucleotides (the lengths of long ncRNAs are greater than 200), b) the second feature which is the most fundamental feature to classify TSSaRNAs as a distinct class of ncRNAs is that TSSaRNAs are strictly enriched near to transcription start site of coding genes. In the scientific literature, TSSaRNA is considered as a class of ncRNAs that belongs to a group of ncRNAs molecules called transcription boundary-associated RNA (TBARs) [47, 48]. TBARs consist of several distinct classes of ncRNAs. Several research groups have given these classes of TBARs various names. Therefore, it is possible to find a distinct class of TBARs with more than one name. We have provided a list of common names that are used to describe TBARs in the scientific literature in the table (table 1.6). In the case of TSSaRNAs, since transcription start sites are usually related to promoter region thus TSSaRNAs could be overlapped or could be referred synonymously with various terms such as transcription initiation RNAs (tiRNAs) [49], RNAs associated with transcription start sites [49, 50], promoter-associated RNA [51], and tiny RNAs associated with transcription initiation and splice sites [52]. By convention, TSSaRNAs could be defined as a class of small RNAs associated with transcription start sites with the possibility to arise in both orientations, i.e., sense and antisense orientations (Fig. 1.5). The next three paragraphs give an overview of TSSaRNAs detection in the three domains of life as well as highlight the major characteristic of TSSaRNAs in each domain.

Table 1.6: A table provides a list of the common names that used to describe the transcription boundary-associated RNAs (TBARs).

It is important to note that any distinct ncRNA molecule of TBARs could have given various arbitrarily names. A comprehensive review about the classification of TBARs has been provided by Yu and his colleagues ([47]).

TBARs major group	Names of ncRNAs classes
TBARs that are associated to the upstream boundary	Promoter-associated RNAs; Promoter upstream transcripts; Upstream antisense RNAs; Stable unannotated transcripts; Cryptic unstable transcripts; Upstream non-coding transcripts; Transcription start site-associated RNAs; Transcription initiation RNA.
TBARs that are associated to the downstream boundary	Transcription termination site-associated RNAs; Terminus-associated long RNAs; Terminus-associated small RNAs.

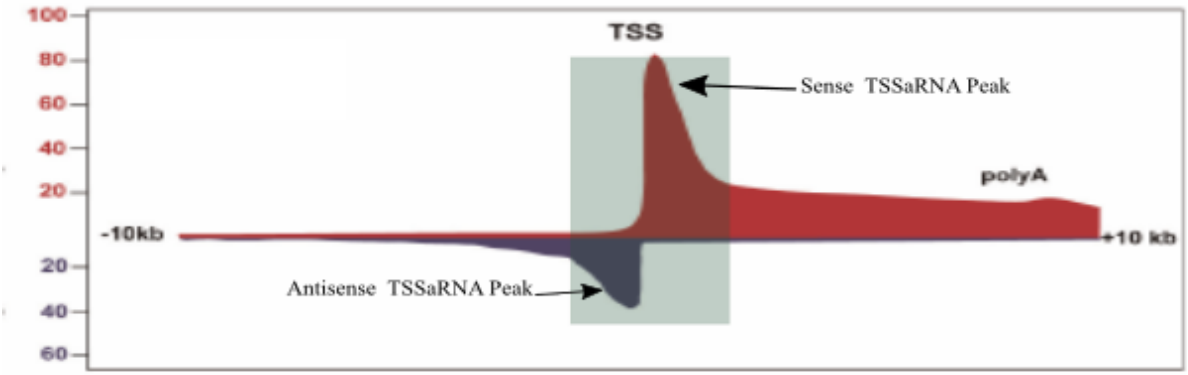


Figure 1.5: Illustration of TSSaRNA peaks in both sense and anti-sense orientation. This figure shows the possibility of the existence of both sense and antisense TSSaRNAs associated with single TSS. This figure adapted from [53]. The y-axis shows the expression enrichment. The peaks indicate the existence of TSSaRNAs. The x-axis shows the relative distance to the TSS.

1.5.1 TSSaRNAs in Eukaryote domain

In 2007 Guenther and his colleagues observed that the transcription machinery in human cells could start the process without producing mature mRNAs. They considered this phenomenon is resulted due to transcriptional/post-transcriptional regulatory mechanisms [54]. One year later (2008) Seila and his colleagues reported the existence of short RNAs associated with transcription start sites. Seila and his team considered their finding as a novel class of ncRNAs in mouse cells and human embryonic stem cell line [55]. This new class is referred as transcription start sites associated RNAs (TSSaRNAs). By using deep sequencing, Seila and his group observed that TSSaRNAs were associated with the majority of annotated mouse coding genes. Furthermore, they were able to detect this new class in various mouse cell types besides human embryonic stem cell line (hues6). They provided evidence of the existence of both sense and antisense TSSaRNAs. Moreover, they observed that TSSaRNAs are mostly associated with moderate and highly expressed genes; however, TSSaRNAs could be observed in low expressed genes as well. About the statistics of TSSaRNAs length, they observed the average length size of TSSaRNAs was 20 nucleotides while the most frequent length size was 17 nucleotides long. However, by performing cloning to enrich TSSaRNAs, they were able to detect only TSSaRNAs with length size in the range of 20-90 nucleotides long. Therefore, they suggested that TSSaRNAs with a range of 20–90 nucleotides are the most dominant and the most stable within the cell. As we have mentioned before TSSaRNAs in literature could be referred as tiny RNA associated with transcription start, we would say that in 2008 Taft and his colleagues reported tiny RNAs associated with transcription start in human, chicken, and Drosophila within a window of 60 nucleotides upstream and 120 nucleotides downstream transcription start sites. They characterized the length size of these transcripts with a modal length size of 18 nucleotides [49]. Recently in 2017, Choi and his colleagues detected TSSaRNAs in the human cell with length size up to 200 nucleotides [56].

Considering the detection of TSSaRNAs in plant cells, Wang and his colleagues reported

in 2011 the existence of TSSaRNAs in *Arabidopsis thaliana* [57]. By using multiple datasets, they were able to detect sense and antisense TSSaRNAs within a window of 300 nucleotides around TSSs. They parametrized the window to start from 100 nucleotides upstream of TSSs up to 200 nucleotides downstream of TSSs. Wang and his colleagues have not commented clearly about the maximum length size of sense TSSaRNAs, but they characterized the length size of antisense TSSaRNAs as in the range of (21–24) nucleotides.

1.5.2 TSSaRNAs in Bacteria Domain

After TSSaRNAs have been discovered in eukaryotes, later in 2012 Yus and her colleagues confirmed the existence of TSSaRNAs in two phyla of domain Bacteria. By performing sequencing without fragmentation as well as using tiling array technology, authors were able to detect TSSaRNAs in *Mycoplasma pneumoniae* and *Escherichia coli* [45]. They observed that by using deep sequencing with fragmentation, they could obtain clear peaks of TSSaRNAs even in low expressed genes. The size of the detected TSSaRNAs ranged of 16-65 nucleotides with a mean of 45 nucleotides. Yus and her colleagues succeeded to report the existence of only sense TSSaRNAs in *Mycoplasma pneumoniae* and *Escherichia coli*. Therefore, it is important to point out that Yus and her team were not able to provide any evidence of the existence of antisense TSSaRNAs in domain Bacteria.

1.5.3 TSSaRNAs in Archaeal Domain

After TSSaRNAs have been discovered in both Eukaryote and Bacteria domains, it has been interested to investigate the existence of TSSaRNAs in Archaeal domain. Hence, Zaramela and colleagues from our research group investigated the existence of TSSaRNAs in several Archaeal organisms including: *Halobacterium salinarum*, *Pyrococcus furiosus*, *Methanococcus maripaludis*, and *Sulfolobus solfataricus* [46]. Their result showed evidence for the existence of TSSaRNAs in Archaeal domains. Moreover, their result stated the length size distribution of detected TSSaRNAs in particular in the species

Halobacterium salinarum was with the median length size of 27 nucleotides while the overall length size ranged of 16-146 nucleotides.

1.5.4 TSSaRNAs: biogenesis and potential function

The TSSaRNAs biogenesis and its underline mechanism are still unclear. However, it has been suggested that the mechanism of RNA polymerase pausing during the transcription process could play an important role in TSSaRNAs biogenesis in animal cells, Bacteria and Archaea [46]. The hypothesis of RNA polymerase pausing as a key mechanism to produce TSSaRNAs is based on the observations that RNA polymerase pausing signals have precisely coexisted over sense and antisense TSSaRNAs peaks [58]. The pausing of RNA polymerase occurs with high frequencies during the transcription process of highly transcribed genes and mostly associated with active promoters [59]. It has been observed that RNA polymerase is paused after transcribing from 25 up to 60 nucleotides [60]. Upon the pause process occurs, two scenarios are expected. The first scenario is that the elongation process will resume after the pausing step to transcribe the whole gene [61]. On the other scenario, the transcription machinery will release the nascent RNA transcript as a mature non-coding RNA molecule. In theory, the step of RNA polymerase pausing could appear at any point during the transcription of genes [62]. However, the highest degree of pausing has as been observed to be near to the transcription start site (Fig. 1.6). The regulatory roles of RNA polymerase pausing and its molecular functions is still uncertain, but several studies have suggested the pausing process occurs as response to environmental condition [59, 61, 63] (Fig. 1.7). Also, the pausing process occurs as check-point stage for gene expression process via termination the elongation process [59, 64] (Fig. 1.8), or via regulating the chromatin remodeling [65]. The mechanism underline RNA polymerase pausing is not clearly understood yet, but many factors supposed to contribute in establishing RNA polymerase pausing. These factors vary from sequence-based motifs to regulators proteins. Studying pausing of RNA polymerase II in Drosophila embryo revealed many factors involved in the pausing mechanism. These factors are including GAGA, TATA, initiator, and downstream promoter element as

sequence-based motifs besides sensitivity-inducing factor (DSIF) and negative elongation factor (NELF) as protein regulators [66] (Fig. 1.8). At abstraction level, RNA pausing mechanism could be considered as the mechanism among different organisms as combinations of RNA polymerase, sequence-based motifs, and proteins regulators. It is also expected that different type of motifs and regulators could exist depending on the organism. For instance, in Human, the GGUG sequence-based motif has observed as a consensus sequence associated with promoter-associated non-coding RNAs [67].

One more phenomenon which has been suggested to play a role in TSSaRNAs biogenesis is called RNA polymerase backtracking. The term backtracking refers to a process where RNA polymerase moves backward instead of moving forward during transcribing nascent mRNA [68]. RNA polymerase backtracking occurs at certain points due to specific signals or due to weak RNA-DNA alignment [69]. The roles of RNA polymerase backtracking still not fully covered yet. However, it has been suggested that RNA polymerase backtracking occurs as regulatory steps in elongation step or a response to internal and/or external signals [70]. RNA Polymerase Backtracking could arrest the transcription process which pauses the transcription that results in an immature RNA transcript which could be released from the transcription machinery [71]. The mechanism of RNA polymerase backtracking is still considered a source of TSSaRNAs biogenesis, however, in 2011 Valen and his colleagues investigated the biogenesis of small RNAs in the human genome and they suggested that TSSaRNAs molecules as mostly are derived from nascent mRNAs by stalled RNAPII against nucleolysis [72].

Considering TSSaRNAs biogenesis in plant cells, Wang and his colleagues observed both sense, and antisense TSSaRNAs are produced in wild type cells, but TSSaRNAs are totally abolished in a triple mutation of DICER-LIKE2, DICER-LIKE23, and DICER-LIKE24 in siRNA biogenesis pathways. Therefore, their results have given strong evidence that siRNA pathways are responsible for producing TSSaRNAs in plant pathways [57].

Scientists have given suggestion and evidence about the hypothesis behind TSSaRNAs biogenesis. However, we still consider that TSSaRNAs as versatile molecules of small

RNAs which consist of tiny RNAs, miRNAs sized RNA molecules and medium-size RNAs. Therefore, we consider various biogenesis mechanisms that could be involved in TSSaRNAs biogenesis. These mechanisms could be slightly or totally different among various organisms.

The functions of RNA transcripts that yield from pausing of RNA polymerase and non-coding RNA that transcribed near to transcription start sites are still not totally characterized. Some studies showed the abundance of TSSaRNAs in the cell is very low [58]. However, it has been suggested TSSaRNAs could play roles in regulating the transcription process and could be involved in epigenetic regulation [46]. In this project, we are speculating that TSSaRNAs could be related to activation or suppression of cognate gene expression or even considering TSSaRNAs transcripts could exert their own bimolecular functions.

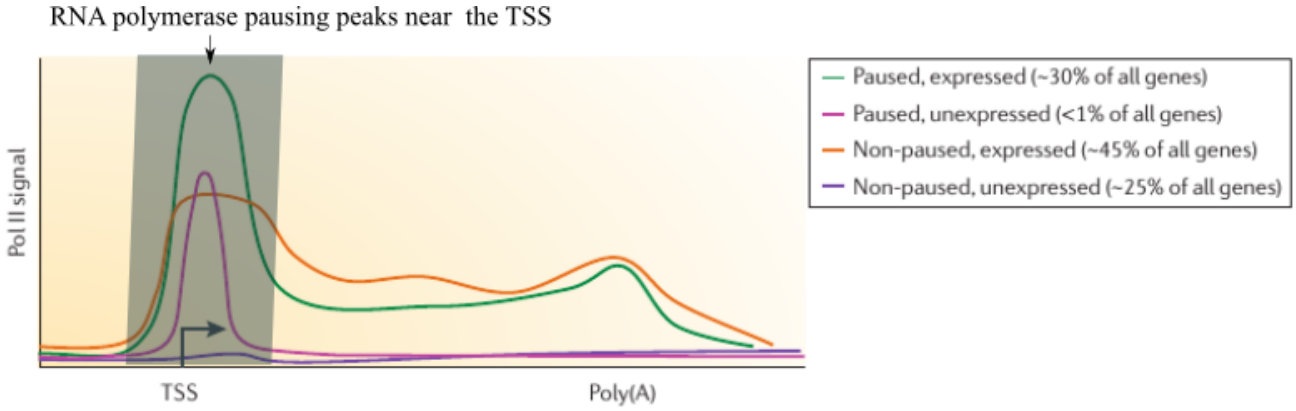


Figure 1.6: RNA Polymerase Pausing in Genes with Different Activity

This figure shows the occurrence of RNA polymerase pausing in genes with different activities. The y-axis shows Pol II ChIP-seq signal. The size of peaks reflects the occurrence of RNA Polymerase Pausing. The x-axis shows the relative distance to the TSS. The shaded area demonstrates the RNA polymerase pausing process occurs most probably around the TSS signal. This figure adapted from [59].

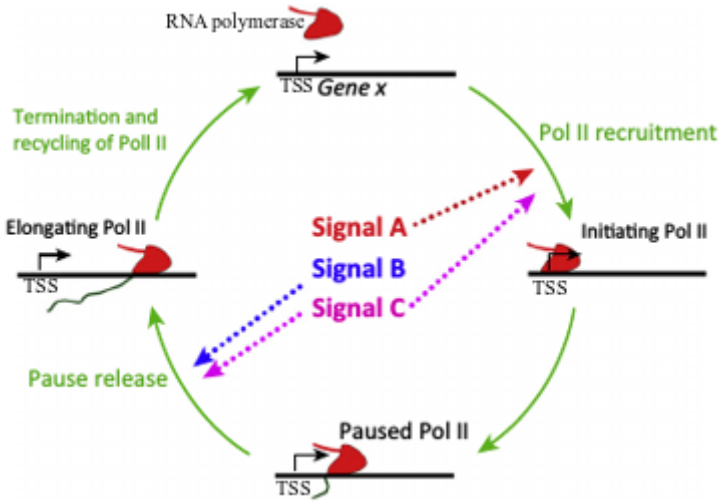


Figure 1.7: RNA Polymerase Pausing as Response to Environmental Signals

This figure shows the occurrence of RNA polymerase (Pol II) pausing as a response to environmental signals. RNA Pol II is depicted in red. Signal (A) regulates transcriptional initiation. Signal (B) regulates pause release. Signal (C) regulates both transcriptional initiation and pause release. The nascent RNA transcript is shown in black. This figure adapted from [61].

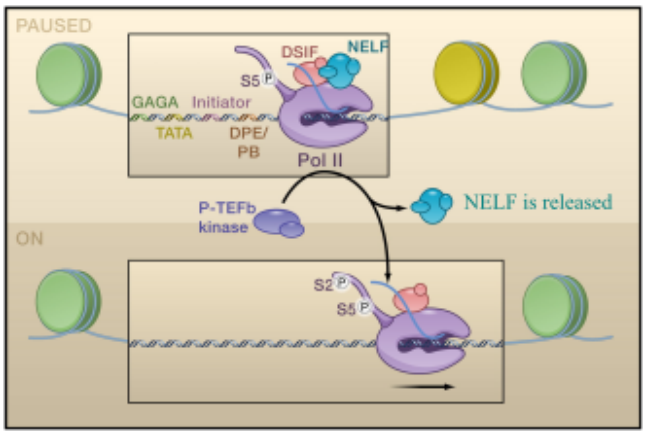


Figure 1.8: RNA Polymerase Pausing Mechanism

This figure shows RNA polymerase pausing mechanism as a combination of RNA Polymerase, sequence-based motifs (GAGA, TATA), and protein factors (the negative transcription elongation (NELF) & the sensitivity-inducing factor (DSIF)). The Pausing machinery consists of DSIF/NELF complex. The positive transcription elongation factor (P-TEFb) disassociates DSIF/NELF complex; consequently, the transcription elongation process will continue. This figure adapted from [66].

1.6 RNA Sequencing

Before performing any computational experiment, understanding the raw data is critical to all downstream analysis steps to obtain reliable findings. Understanding the data is not only guarantee of using the best analytical model, but it further helps researchers to think critically to avoid bias and any technical errors. It is also highly recommended to understand the experimental design which helps to figure out the necessity of quality control checking and helps to understand the limitations in the data. As our raw data in this project is generated by single-stranded RNA sequencing (strand specific) technology. Thus this section gives a high-level overview of RNA sequencing experiments and the general computational steps to process the raw data.

1.6.1 RNA sequencing overview

RNA sequencing is considered one of the biggest revolution in molecular biology researches. RNA sequencing is known as next-generation sequencing (NGS) as the RNA sequencing technology has been arisen to the market after the Sanger DNA sequencing. The main purpose of using next-generation sequencing is to study the dynamic of transcriptome over time. There are many advantages of using NGS technologies over the probes based technologies such as microarrays. The most fundamental advantages of using NGS to study the transcriptome instead of array technologies consist of two points: the first point is that NGS does not require any prior knowledge about the genome of the organism under investigation [73]; second, NGS is much cheaper than array-based technologies as since the end of the year 2007 the cost of sequencing is decreased dramatically (Fig. 1.9). The reduction in the NGS cost allowed many laboratories around the world to perform sequencing experiment [74]. Nowadays many scientists consider NGS as the best tool to generate reliable transcriptomic raw data because NGS help to conduct experiments with high resolution and greater sensitivity. Moreover, NGS technologies have the possibility to study a large number of samples per time known as high-throughput experiments [75].

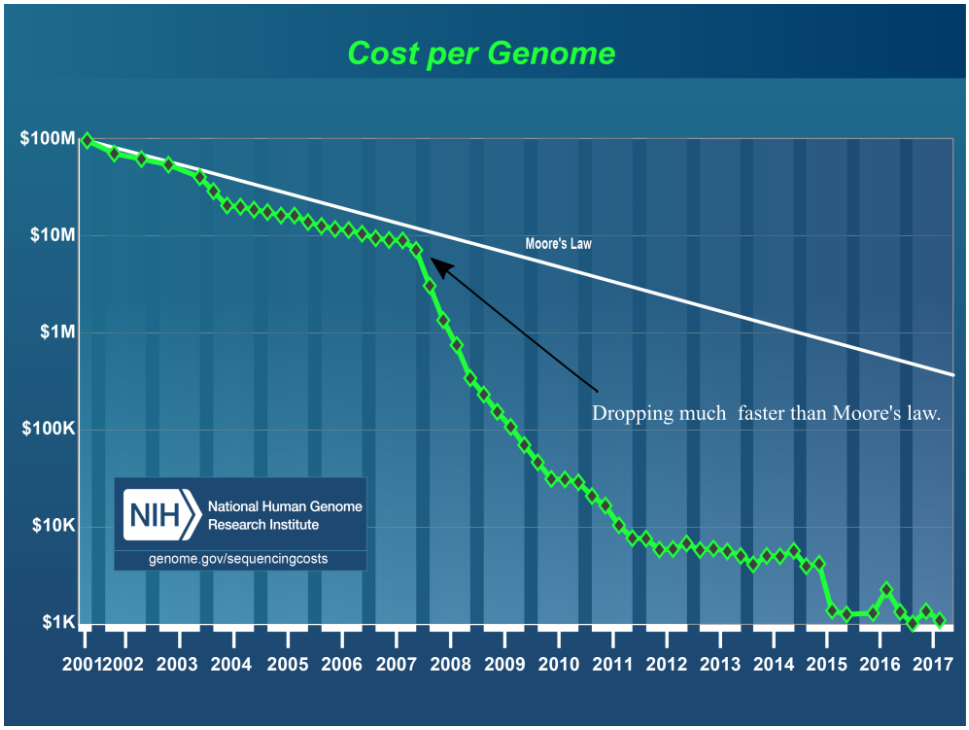


Figure 1.9: Figure shows the cost of RNA-Seq experiment per genome

Notice that since the end of the year 2007 the cost of RNA-seq per genome has been reduced dramatically. This figure has been adapted from the National Human Genome Research Institute `https://www.genome.gov/sequencingcostsdata/`

1.6.2 RNA-Seq platforms

NGS marketing is evolving rapidly, and day by day the companies are providing new technologies besides improving the capacity and productivity of the existing platform. NGS could be classified into the first generation platform, the second generation platform, and the third generation platform [76]. Furthermore, scientists now are referring to the fourth generation of NGS, but the fourth generation could be considered as a just modification of the third generation rather than new generation.

The classical Sanger sequencing technology and capillary sequencing technology are considered the first generation of sequencing. Sanger sequencing is based on sequencing

DNA molecules through PCR chain termination method. Sanger technique opened the window for the sequencing, but it is considered too expensive to be budget by traditional laboratories. Also, it was time-consuming and less sensitivity compared to the second and the third sequencing technologies [75]. The second generation of NGS Technologies includes platforms such as Illumina, 454 Roche, and SOLiD. These platforms sequence short of cDNA "RNA reverse transcribed into DNA" instead of direct RNA sequencing. The second next-generation technologies offer advantages over Sanger sequencing such as reduction in the sequencing cost and providing the possibility to perform high-throughput experiments. The limitations of the second next-generation technologies consist of long-running time to get the results besides the limitation in the read length size [76]. The second next-generation technologies sequence very short reads comparing to Sanger technology. The third generation platforms include Helicos, PacBio, and Ion Torrent. These technologies offer the possibility to sequence direct RNA molecules as well as cDNA with high resolution. Furthermore, these platforms are capable of sequence even in a single molecule resolution. The third-generation platforms perform the sequencing in a very short time where the experiment could be finished just in a few hours instead of spending days [76].

1.6.3 RNA sequencing work-flow

The full technical details of RNA sequencing and their experimental design are beyond the scope of this thesis, but we will give an overview of the work-flow of custom NGS experiments. The information offered here will be fundamental as much as needed to be known by bioinformaticians and data analyst. For those who are interested in more details, we recommend them to refer to the online technical manuals of the popular platforms in the market.

Once, the experimental biologist came up with the scientific question and the purpose of the performing RNA sequencing; the first step it will be figuring out about the experimental design as a collaborative task with computational biologist/Bioinformaticians. The stage of the experimental design results in answering some questions such as the

number of technical and biological replicates which required for the experiments, which sequencing technique is more appropriate for obtaining a trusted result, and which sequencing kit will be suitable to be used in the lab according to the lab budget and lab environment.

Once the scientist came up by the most appropriate experimental design, the next step is to perform the bench work of NGS experiments which start by isolation RNA and purification of interested populations of RNA (total RNA, mRNA, small RNA ...etc.). We are using the term, but many sequencers are actually sequencing cDNA molecules. Therefore, the isolated RNAs would be reverse transcribed do DNA then use the generated DNA strands as templates to synthesize double strand DNA. The last stage in many bench work before automating sequencing is called library preparation in this stage technician could add barcode or adapters into each sample (multiplexing different samples in each batch of the sequencer). The prepared libraries enter in the step of automation sequencing which aims to parallel amplify the preprepared library massively. This stage depends on the type of sequencing platform, but many labs use Illumina kits.

Depending on the scientific question and the chosen platform, experimental biologist and bioinformaticians could choose either strand-specific libraries or non-strand-specific protocol. Nowadays many laboratories prefer to use strand-specific protocol because this protocol is designed to provide more information about the polarity of transcripts as well as from which strand RNA-Seq reads came from [77]. There is a limitation of non-strand-specific protocols regarding the lack of information about the orientation of the libraries. However, still, it could be used to study the antisense transcripts and antisense regulation by using a more rigorous and sophisticated bioinformatics protocol.

1.6.3.1 Sequencing options

The two most common options for custom next-generation sequencing experiments are either single-end RNA sequencing or paired-end RNA sequencing. In the single-end, RNA sequencing protocol the sequencing platforms sequence only from one end to another end of the read libraries. The single-end RNA sequencing protocol is the first choice for

sequencing experiments as its cost is comparatively low. Also, this protocol generates excellent sequences for transcriptomics purpose but it not good enough to study the transcription level in repetitive elements.

The other option of sequencing is called paired-end sequencing, in this option, the sequence starts at one end of the reads then sequence the other end result in sequencing both ends with a gap (called insertion). The paired-end sequencing is more expensive as well as it is time-consuming compared to the single-end sequencing, but it provides high quality reads to perform the mapping process (alignment of reads to reference genome). Besides providing high-quality alignment process, the paired-end sequencing protocol is also more effective in identification and resolving complex genomic phenomena such detection of the massive genomic mutations including deletions, insertions and inversions and distinguishing the repetitive sequence element in the genome. The advantages of the paired-end protocol let it become the best option to assemble genome which is known as *de novo* genome sequencing. Depending on the size of insertion fragment paired-end sequencing could be referred as mate-pair sequencing when the libraries across more considerable distances with a big insertion size. The mate-pair sequencing is considered as the most suitable for various applications including *de novo* genome sequencing, Genome finishing, and structural variant detection.

We could add one more sequencing options as later next-generation sequencing has been developed to study the RNA/DNA-protein interaction *in vivo* using an option called ChIP-seq (Chromatin ImmunoPrecipitation sequencing) technology. The typical ChIP-Seq experiment starts by cross-linking the RNA/DNA-protein binding complexes followed by next-generation sequencing analysis. The power of this sequencing option that it is able to identify the protein-binding site with high resolution down to the nucleotide level [78].

1.6.4 Computational aspect of RNA-seq data analysis

It is obvious that the next-generation sequencing technologies rapidly become popular and widely used in studying transcriptome in different conditions including measuring the

level genes expression but still the downstream RNA-seq data analysis and interpreting the results is facing set of challenges. The challenges in RNA-seq data analysis arisen due to the limitation in computational infrastructure for RNA-seq raw data storage and processing, transcriptome complexity and lacking robust statistical and mathematical model to handle RNA-seq raw. To overcome these serious challenges in RNA-seq analysis expert bioinformaticians are needed to smoothly and technically pass the problem and get biologically meaningful results. The typical RNA-seq analysis work-flow consists of several steps, and the data can be investigated qualitatively and/or quantitatively. In summary, we could distinguish three universal steps to analyze RNA-seq raw data which include: the quality control assessment step, alignment reads to a genome, and mining the biologically relevant information for downstream analysis.

1.6.4.1 Quality control assessment

Any typical RNA-seq experiments consist of multiple steps starting from the extraction of RNA and reverse transcription to overcome the low stability of RNA molecules, amplification of fragments, adding adapters then sequencing. As same as all other experimental techniques in molecular biology, RNA-seq sequencing technologies, and its experimental steps are prone to certain experimental biases, technical biases, errors, and artifacts which affect all downstream analysis [79, 80]. Therefore, before analysis, any RNA-Seq data comprehensive quality control (abbreviated as QC) assessments against potential technical and biological errors is critical and important to obtain scientifically reliable results as well as it would help in the good interpretation of the findings.
The most common errors prone in any RNA-Seq raw data include: a) the existence of low quality and low confidence sequences; b) the presence of duplication in RNA-Seq reads as well as unknown bases; c) contamination of sequences with adapters and barcode per sample and even some reads from other organisms; d) the existence of intrinsic biases such as: per sequence G-C contents, nucleotide composition bias, Per base unknown bases (N) content, overrepresented sequences, and Per base sequence quality (Fig. 1.10). As quality control is a critical step in RNA sequencing (RNA-Seq); thus, many tools for

checking and performing RNA-Seq quality control assessment exist and available as open source tools. Examples of some tools for RNA-seq data processing including the quality control assessment are shown in the Digital Appendix (3).

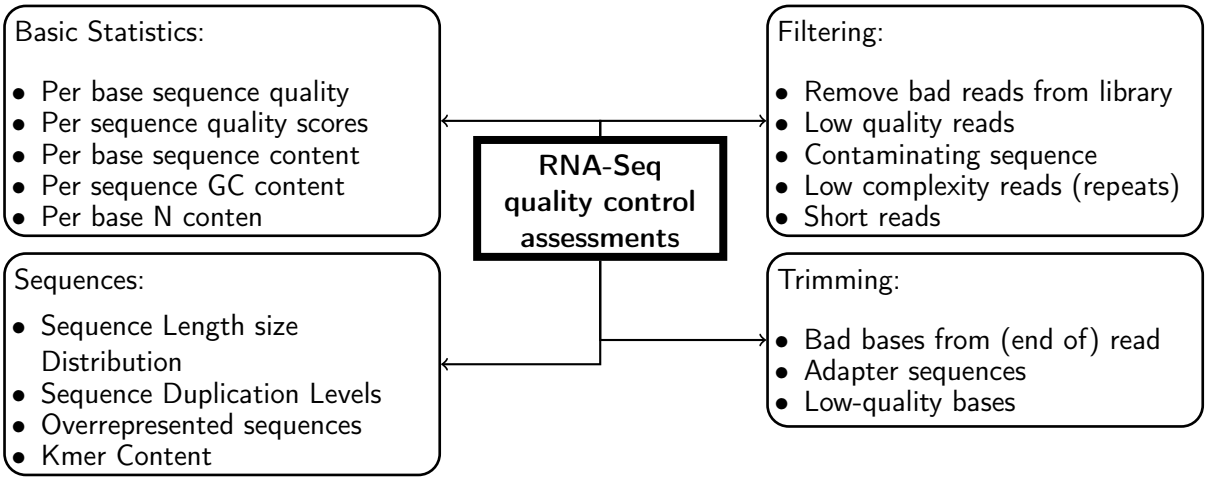


Figure 1.10: A diagram shows RNA-Seq quality control assessments

1.6.4.2 Alignment reads to genome

Aligning RNA-seq reads to a reference genome and extract the potential aligned location is referred as the mapping process. Many programs are available to align RNA-seq reads to a given reference genome and depending on their underline algorithm they are varied on their algorithms speed, performance, and accuracy [81]. To perform the mapping process, the genome could be obtained either from public repositories or could be assembled from available RNA-Seq reads. The major challenges for mapping process are: the size of reference genome, overcoming the polymorphisms in reads which refers to the mapped reads that are coming from repetitive element in given genome (low complexity region/microsatellite/repeat), mapping to assembled genome that miss known transcribed regions, mapping reads that come from exon-intron boundaries, and whether mapping search for the exact match or just looking for the sequence similarity by allowing gap and mismatch up to given cutoff score during mapping process [82].

1.6.5 Mining the biologically relevant information and downstream analysis

Generally, bioinformaticians and system biologists aim to gain two types of relevant biological information from any RNA-seq processes. This information could be categorized as genomic annotation and transcriptomics quantization. By performing genomics annotation researchers aim to identify various genomics signals such as identifying the transcriptional start sites (TSS), identifying splicing sites boundaries (exon/intron), annotation of poly-A sites, etc. On the other hand, transcriptomics quantization usually used to compare the dynamic of transcriptome between two conditions which is referred as measuring differential gene expression (DGE), but it could also be used to detect alternative genomics signals such as detection of alternate transcriptional start sites and/or alternative splicing.

1.6.6 RNA-seq metrics

Once bioinformaticians got the mapped RNA-seq reads, the next step is to convert the mapped reads into the appropriate metrics for downstream statistical analyses. The most common metric parameters for mapped reads are:

1.6.6.1 Metrics on mapped reads

After performing the mapping step, bioinformaticians usually are interested in measuring the quality of the mapping process. This quality assessment will infer the degree of suitability of RNA-Seq reads to be used into the downstream analysis. The most common measurements at this stage are (a) total read number. (b) the number of unique reads. (c) the number of duplicate reads. (d) duplication percentage or duplication rate. (e) percentage of the status of mapped reads (uniquely mapped read and unmapped reads and polymorphism of reads). (f) in case of paired-end options bioinformaticians are interested as well to know the statistics of paired reads orientation to figure out the status of the paired reads to determine if the mapping results in concordant pairs, singleton read and discordant pairs. (g) in case the experiment is not designed to sequence total RNA then it will be useful to investigate the degree of contamination with undesired classes.

(h) the percentage of GC content in the mapped reads.

1.6.6.2 Coverage

The term coverage reflects the number of times the target genome is being covered by reads such as 1X (only one time), 2X,...,100X, and so on. The value of coverage is important to be calculated even prior performing the experiments as some biological questions need a high coverage to get the right answers, for instance, the higher number of coverage is required when the purpose of the study is to estimate the single nucleotide polymorphism(SNP). The coverage could be estimated using the Lander/Waterman formula. The Lander/Waterman formula is represented by the following equation (Equation 1.4).

$$Coverage = \frac{LN}{G} \tag{1.4}$$

Where: **G** is the haploid genome length (genome size), **L** is the read length size, and **N** is the number of reads. As the read length is irrelevant when speaking about RNA-seq experiments with unequal read size such experiments designed to sequence total cellular RNA therefore, in this case, bioinformaticians could refer to the read length as the average read length or the median length. In some cases the read length could be calculated as the full sum of each reads length multiplied by its frequency.

1.6.6.3 Depth (per base coverage)

Per base cover or the depth of coverage is reflecting how many times a given genomic position –single base- has been covered by independent RNA-seq reads. This metric could be associated directly to the level of the gene expression at a given location, and it could be visualized as the histogram of coverage (such as Fig. 1.5 for example or similar found all over the literature). Many bioinformaticians use per feature coverage instead of per base coverage to figure out the expression level of a given feature, i.e., exon or gene. In this case, some normalization should be applied to get rid of potential errors (this type of error will be briefly covered in the section of reads quantification).

1.6.6.4 Breadth of coverage

The traditional concept of the coverage is assuming that the RNA-seq reads are uniformly distributed among the genome, but that is not true for the transcriptome as some areas are highly transcribed at given condition while other regions are low transcribed or even no transcription process. The term breadth of coverage calculates the percentage of reference genome that is covered under given depth; for instance to calculate the percentage of the genome that coverage at the 1X depth we could use the following formula (Equation 1.5):

$$\frac{Size\ of\ Assembled\ Genome}{Size\ of\ Reference\ Genome} * 100 \tag{1.5}$$

Note that we could calculate the size of the assembled genome at different depth 2X,10X,... etc.

1.6.6.5 Metrics for differential expression Analysis

Using RNA-seq data to describe the differential gene expression is out of the scope of this thesis. However, it is important to note that the computational analysis of RNA-seq data could be ended up by comparing the expression level of genes between multiple experiments. The assumption of the differential gene expression is based on the hypothesis that the level of expression of a particular gene is correlated to the amount of the RNA fragments transcribed from that gene. Therefore, by performing a statistical test on quantified reads per a genomic feature based on a certain probabilistic model such as poison distribution, negative binomial, normal distribution and so on, bioinformaticians could estimate the genes that are expressed in particular conditions or certain cell types. Quantification of reads per a feature often results in extreme values per certain features within and/or between conditions due to some systematic bias or technical errors. Therefore, normalization of reads count is a crucial step [83]. The common normalization metrics which could be used to perform differential genes analysis are reads per kilobase of transcript per million mapped reads (RPKM), transcripts per million (TPM), house-keeping gene normalization, and quantile normalization [83].

1.7 Rationale

1.7.1 Description of the problem and Rationale

The existence of TSSaRNAs upstream of many genes in various organisms all over the three domains of life is now a fact. However, the functions of this class of non-coding RNA are still not well characterized regarding their roles on the complex biological machinery as well as on their effect on downstream genes targets. Moreover, a study done by our research team (The Biological Information Processing Lab – LabPIB - & The Microbial Systems Biology Lab –LaBiSisMi-) has provided strong evidence of the existence of TSSaRNAs in several Archaeans transcriptomes including *Halobacterium salinarum* NRC-1. Therefore, this research project starts with the hypothesis that the existence of TSSaRNAs among the three domains of life is associated with a universal molecular function. The present study has proposed to enhance further our understanding of the potential roles of TSSaRNAs in the cellular system based on both sequence and structural information via an *in silico* approach.

Consequently, in this study, we are mainly interested in establishing an optimized prediction algorithm to increase the prediction accuracy, as well as attempting to overcome the challenges of structural prediction of these TSSaRNAs followed by addressing the questions of how their structures participate in their molecular functions. Also, as part of this study, we are willing to perform functional annotation of predicted TSSaRNAs based on Rfam functional classification of the ncRNAs.

In summary, we could say that this project has been proposed with the motivation of accurately predict TSSaRNAs in *Halobacterium salinarum* NRC-1 then mitigate the lack of knowledge in their function by classifying them according to Rfam functional classification as well as study their structural characteristics.

1.8 Research Objectives

1.8.1 Aims of the study

1. Our first objective is to design a reliable algorithm to predict TSSaRNAs from any given RNA-seq data sets.
2. Using molecular modeling techniques we aimed to predict reliable secondary and tertiary structures of TSSaRNAs.
3. We investigated the multiple roles of TSSaRNAs molecules as a part of the RNA-based regulation system in cells.
4. We predicted TSSaRNAs that function as protein binding motifs in particular binding to the Lsm regulatory protein.
5. We aimed to predict the function of TSSaRNAs in *Halobacterium salinarum* NRC-1 based on Rfam annotation of ncRNAs.

1.8.2 Study design, and Research procedures and strategy

This study is a computational approach to functionally annotate TSSaRNAs in *Halobacterium salinarum* NRC-1. The computational method is often referred as an *in silico* experiment. In the scientific literature, there are two definitions associated with term *in silico* in the field of bioinformatics [84, 85]. According to the first definition, *in silico* experiments are a set of computational operations used to investigate and understand biological phenomena based on the analysis of the previously generated data. According to the second definition, the *in silico* experiments are computer simulations to model and perform the biological experiments on a computing machine "virtual-world experiments". There is some criticism of conducting a research work using only computational approach. Critics believe that the finding of the virtual experiments would not be significant unless it has been confirmed by wet lab experiments "real-world experiment via benchwork". However, in many aspects, the *in silico* experimentation is indispensable in the field of bioinformatics. Thus, the *in silico* experiments could be performed in parallel with the benchwork experiments. The *in silico* experiments provide the opportunity to study the

biological phenomena in small-time scale. In many cases, the computational experiments could be done first to reduce the benchwork costs.

The research strategy in this study involves multiple computational steps. The first step is to obtain a reliable and trusted raw data. The raw data could be retrieved from public bioinformatics databases. The second step is to perform upstream data analysis. Upstream raw data analysis consists of the following tasks: a) checking the quality control of the raw data, b) cleaning the raw data by removing the low quality reads and adapter sequences, c) mapping the reads into the reference genome. We use the mapped reads to predict TSSaRNAs. To functionally annotate TSSaRNAs, we consider both structural-based and sequence-based TSSaRNAs annotations. The structural-based annotation considers the secondary structures, tertiary structures, and higher orders structures of TSSaRNAs. The sequence-based annotation considers the functional classification of TSSaRNAs based Rfam families. To classify TSSaRNAs into their corresponding Rfam families, we consider both sequence information of each TSSaRNAs molecule and consensus secondary structures of the closely related TSSaRNAs. The overall research strategy of this project has been demonstrated in the figure (Fig. 1.11).

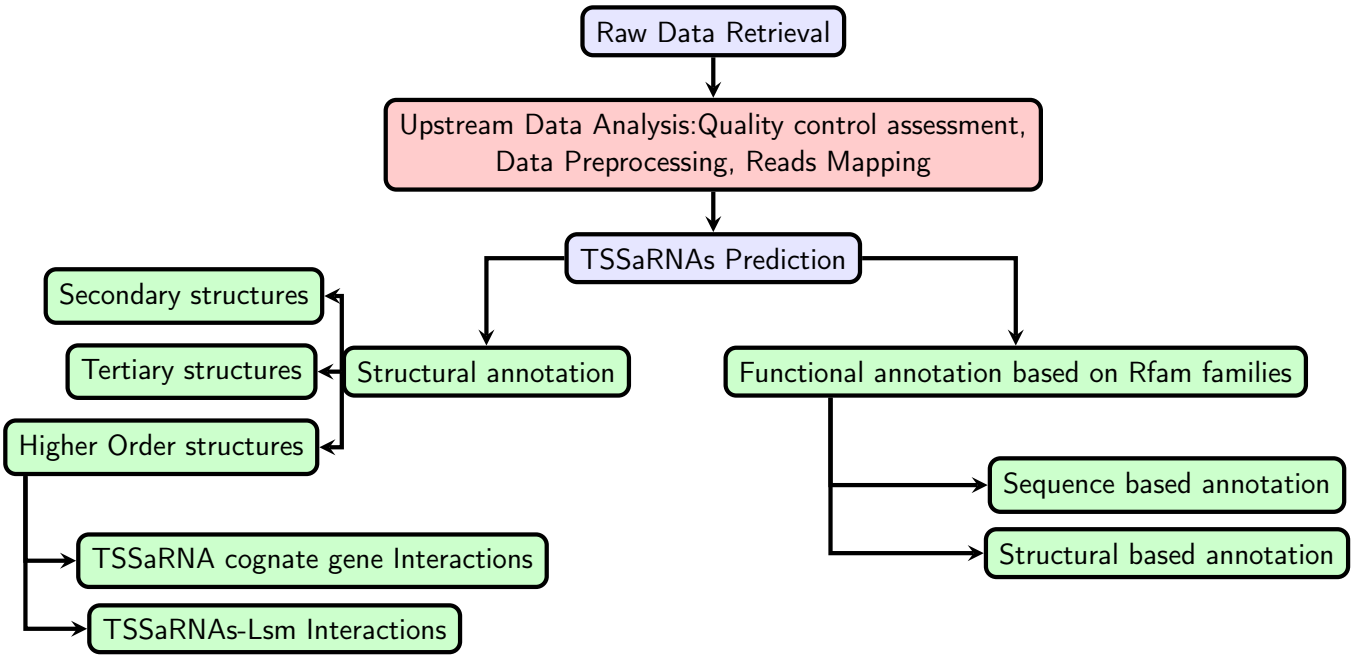


Figure 1.11: A diagram shows the research strategy in this project.

The research strategy in this *in silico* experiment consists of the following tasks: i) retrieving trusted and reliable data sets, (ii) performing upstream data analysis by checking the quality control of the raw data, cleaning the data, and mapping the reads into the reference genome (iii) predicting TSSaRNAs (iv) Annotating TSSaRNAs based on their structural information (v) Annotating TSSaRNAs based on Rfam functional classification. The detailed description of the steps for each task has been provided in the next chapter (The chapter of material and methods).

CHAPTER 2
Material and Methods

2.1 TSSaRNAs Prediction

To predict TSSaRNAs in *Halobacterium salinarum* NRC-1, we designed a prediction algorithm based on statistical distribution of mapped RNA reads around the transcription start site of annotated genes. The annotated transcription start site signals obtained from the reference sequence (RefSeq) of UCSC Genome Browser database [86–88]. The reasons behind using RefSeq annotation instead of other available annotation resources because it has known that RefSeq is a popular, reliable database that is freely available on the internet. Moreover, RefSeq provides curated non-redundant information regarding the genomic coordinates for genes linked to their products in many organisms. One more reason for using RefSeq is that RefSeq team integrates information provided by several scientific research groups across the globe to keep annotation information updated according to computations and experimental evidence.

TSSaRNAs prediction task is a multi-step task, the steps to predict TSSaRNAs start by retrieving the appropriate and the suitable raw data followed by the step of the quality control assessment and data cleaning, and end-up by TSSaRNAs prediction algorithm (refer to the Fig. 2.1). The following subsections will cover the steps of TSSaRNAs prediction as follow: retrieving the datasets, quality control assessment, and mapping reads to the references genome, then TSSaRNAs prediction algorithm.

2.1.1 Datasets retrieval

As we mentioned in chapter 1 section (1.6), the high-throughput sequencing experiments generate powerful information about the dynamics of genes expression and the patterns of the expressions throughout the genome of a given organism. In this study, we considered RNA-seq raw data as the most appropriate data sets to achieve the aims of this study in predicting TSSaRNAs. We used a special type of RNA-Seq experiments called strand-specific RNA-Seq. The reason behind considering the strand-specific experiments in this study is that the underline experimental protocol in strand-specific experiments is designed to provide more information about the polarity of transcripts as well as from which strand RNA-Seq reads are originated.

We retrieved three strand-specific RNA-seq datasets form Sequence Read Archive (SRA) repository[1] [89, 90]. The querying process has done using the unique accession number for each experiment. The unique accession numbers are SRX844124, SRX433542 & SRX441605. The scientist generated the retrieved data-sets in LaBiSisMi[2] (an acronym for Microbial Systems Biology Lab, in Portuguese) that hosted in Faculty of Medicine at Ribeirão Preto campus of the University of São Paulo. All data-sets have been generated using TrueSeq protocol from Illumina[3] to sequence total and small RNA molecules extracted from *Halobacterium salinarum* NRC-1. The organism had incubated in the standard growth condition by providing incubation temperature of 37°C and agitation of 125 revolutions per minute. Each experiment was designed to have a biological replicate such as two biological replicates in SRX844124 and SRX433542 while three biological replicates in SRX441605. The total length size of RNA reads that obtained from sequencer had varied between 20–230 nucleotides (Table 3.1).

[1] http://www.ncbi.nlm.nih.gov/sra/
[2] http://labisismi.fmrp.usp.br/index.php/en/
[3] https://www.illumina.com/

2.1.2 Raw data quality control assessment

As we pointed and emphasized in the chapter 1 section (1.6), the quality control assessment for RNA-seq raw data is a crucial step and indispensable for any downstream data analysis. Therefore, the first step after obtaining the raw data was to evaluate the quality control assessment. In this study, we used FastQC Tool version 0.11.4 [91] to check the quality control of the obtained RNA-Seq data sets. FastQC has maintained by the Babraham Institute[4]. Bioinformaticians commonly use FastQC in quality control assessments as it is a system independent cross-platform open source tool.

There are many advantages of using FastQC for checking the quality over other open source tools. These advantages have summarized in the followings: FastQC is considerably fast and memory efficient tool, besides its ability to accept common RNA-Seq file format such as BAM, SAM or FastQ as input files to assess the raw reads quality control including the sequence quality per base, the nucleotide composition per sequence, sequence duplication levels, adapters, Kmer content, and GC bias. Furthermore, FastQC generates results in HTML format with graphs and tables of reads statistics for processing manually.

In the case of the existence of adapters sequences we processed the raw data by removing the contaminated sequences using FASTX-Toolkit[5].

2.1.3 Mapping reads to the references genome

In TSSaRNAs prediction pipeline we used Bowtie2 program [92, 93] to perform the mapping process of RNA-seq reads into the reference genome. Researchers commonly use Bowtie2 as it is ultra fast mapping tool. The underline alignment algorithm in Bowtie2 is based on the Burrow's Wheeler Transform method (BWT) which optimizes the overall program's memory usage. As we needed the reference genome for the mapping process, therefore, we obtained the reference genome as a complete genome sequence of *Halobacterium salinarum*(taxonomy ID: 64091) in fasta format from The National Cen-

[4]http://www.bioinformatics.babraham.ac.uk/projects/FastQC/
[5]http://hannonlab.cshl.edu/fastx_toolkit/

ter for Biotechnology Information (NCBI)[6].

The first step on the mapping process was the indexing of the reference genome; then we proceeded RNA-seq reads alignment step. The output of the mapping process for each experimental sample is a text alignment file format known as a Sequence/Alignment Map (SAM). Each SAM file consists of an informative header section and an alignment section with at least 11 fields delimited by tabs. The fields in SAM files are referring to the reference sequence to which reads were mapped, the start position for each mapped read in the reference genome, and a Phred scaled quality score of the mapped reads besides many other details [94]. We converted all SAM files into BAM format (binary SAM) using Samtools package version 0.1.19 [94]. The BAM files overcome memory storage limit by reducing the size of the SAM files.

To prepare the mapped reads for the next step as an input file for TSSaRNAs prediction algorithm, we converted SAM/BAM files into bed format using Bedtool package version 2.25.0 [95, 96].

2.1.4 TSSaRNAs prediction

TSSaRNAs prediction algorithm has based on the parametric statistics of mapped sequenced reads. It has known that any typical RNA-seq experiment generates a huge number of reads that are capable of aligning to the reference genome. Processing a huge number of mapped reads will be a time-consuming task for any traditional computing system. To reduce the time which will be consumed by the prediction steps we limited the prediction statistics on only a predefined window around the transcription start site (TSS) signal of each gene. Limiting the statistics on the predefined windows also helps in overcoming the challenges in the data storage of such a huge number of mapped reads. The TSSaRNAs prediction algorithm consists of the following steps: (a) first step is to define a window around the transcription start site for each gene. The size of the downstream side in this window is assigned to be 50 nucleotides as a default parameter or as a quarter of the gene's length in case of short genes. Here, the term short gene is

[6] http://www.ncbi.nlm.nih.gov/

referring to any gene that its length is less than 200 nucleotide. Regarding the size of the upstream side in the window, it has assigned to be 50 nucleotides as a default value or equal to the distance to the nearest upstream gene to prevent overlapping between genes. (b) the second step is to cluster all mapped reads that are starting within the predefined window for each gene to get a vector. The vector's length is equal to the number of reads clusters, and the values associated with the vector's indexes are the counts of mapped reads that starting at a given position. (c) after obtaining the vector, the next step is the cluster/peak calling. We considered the cluster of reads that associated with the highest count of reads as a cluster that might contain a potential TSSaRNA. It is possible to exist more than one cluster associated with the highest number of reads. It the latter case we would consider the cluster that is the nearest to the transcription start site. (d) as a final step, we defined the potential TSSaRNA as the molecule with the highest frequency within each targeted cluster and at the same time passing the cutoff criteria by possessing reads to count more than the statistical average of all molecules within the cluster.

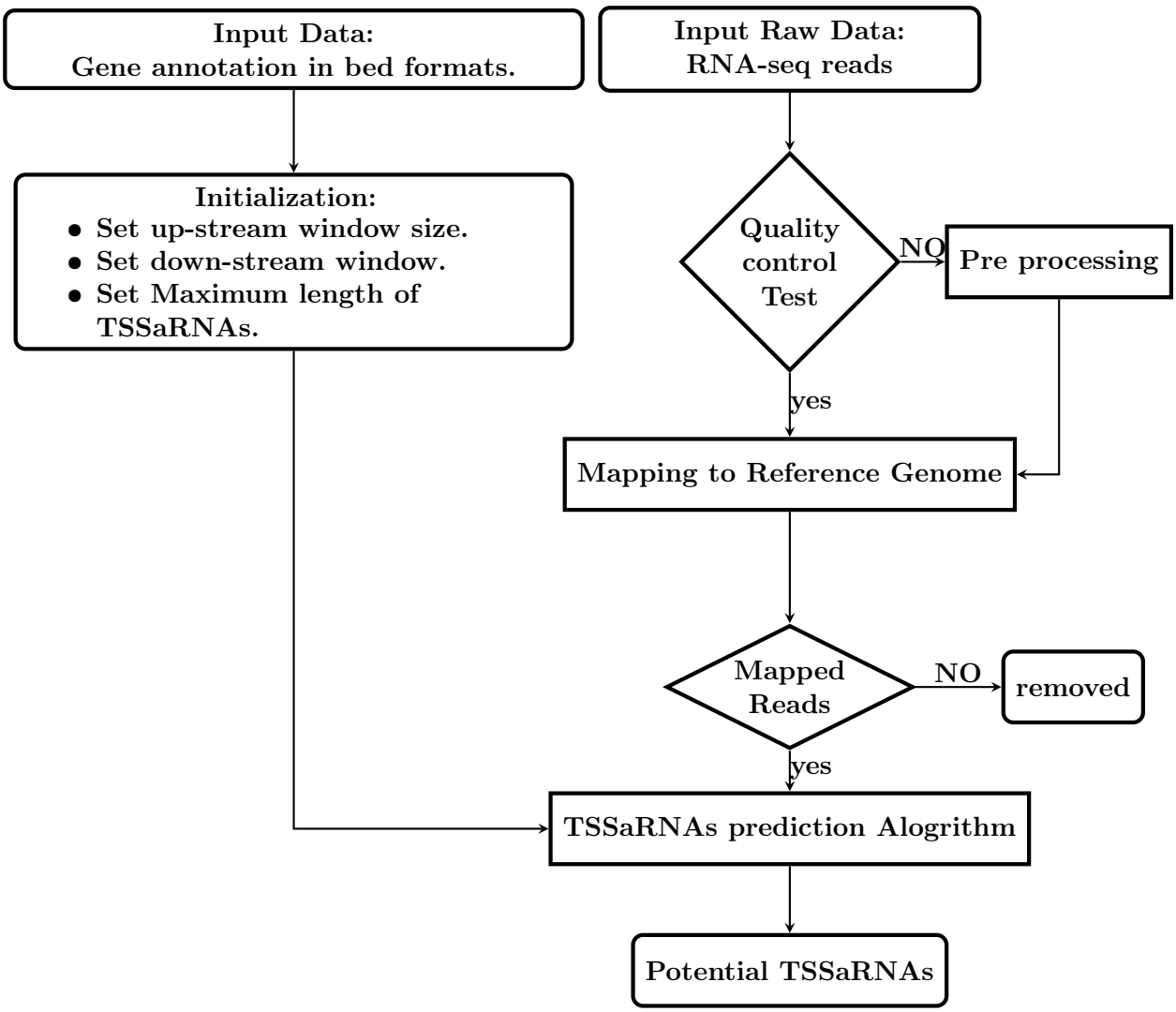


Figure 2.1: Diagram shows TSSaRNAs Prediction steps

2.2 TSSaRNAs structures prediction

The prediction of RNA structures in most scientific literature is referred to *de novo* predictions of RNA structures, but in some cases, it is referred as well to *ab initio* (homology-based) structure predictions. The term *de novo* prediction refers to all steps and methods of prediction higher order structures of RNA from simple sequences of nucleotides while the term *ab initio* prediction refers to the process of prediction unknown structures from a template with known structure by considering the homology between two structures.

In this section, we are going to cover the methodology that we used to predict TSSaRNA secondary and tertiary structures as well as higher order structures such as TSSaRNA-mRNA hybridization and TSSaRNAs-Lsm interactions.

2.2.1 Selection of TSSaRNAs sequences

The complete sequences for all predicted TSSaRNAs have been extracted from the complete genome of *Halobacterium salinarum* NRC-1. The genome obtained from NCBI repository[7] as DNA sequences. We transcribed *in silico* DNA sequences into RNA sequences to get TSSaRNAs molecules to use in the next step of prediction of TSSaRNAs secondary and tertiary structures.

2.2.2 TSSaRNAs Secondary Structure Prediction

TSSaRNA secondary structures prediction performed using Mfold off-line package (v.3.6). Mfold predicts RNA secondary structures by searching the state associated with the minimal free energy using the well known and most popular algorithm that established by Zuker(1989).

Mfold Prediction Parameters

Halobacterium salinarum NRC-1 lives in extreme environments characterized by high concentration of salt. Therefore, we changed the ionic concentration to be same as the

[7] https://www.ncbi.nlm.nih.gov/

intracellular environment. The values of intracellular concentration of magnesium and sodium ions have obtained from the literature [97,98] such as: Na^+ molar concentration to be as 1.4 M and Mg^{++} molar concentration to be as 0.12 M. Furthermore, to obtain just the optimal folding structure we changed the parameter of the maximum number of folding to be computed to be as 1. For full Mfold parameters and their values see Mfold manual.

2.2.3 Digitalizing TSSaRNAs secondary structures

To annotate the predicted TSSaRNAs based on their secondary structure elements (motifs) (Fig.1.2), we converted the structural representation of TSSaRNAs into 17 numerical variables. The secondary structures numerical variables are consisting of: a) TSSaRNAs total length; b) the count of secondary structure elements per TSSaRNA. The space of the secondary structure elements consist of the following items: single strand nucleotides in external loop, closing helices in external loop, helices elements, hairpins, multiloop, bulges, interior loop, and stacks in TSSaRNAs secondary structure; c) the values of free energies in each secondary structure elements as well as the initial and the final free energy.

2.2.4 Modeling TSSaRNAs tertiary Structures

To model TSSaRNAs tertiary structures *de novo*, we used Rosetta tool[8] [28,99] version (3.6). This version released on 8 of April 8, 2016, and it is available freely for academic purposes. Rosetta is considered as a flexible and a multi-purpose scientific tool. Rosetta tool consists of software for modeling, structure prediction, design, and remodeling proteins and nucleic acids. Rosetta software predicts RNA tertiary structures *de novo* based on the Monte Carlo sampling approach by assembling short fragments from the database. Rosetta database contains RNA crystal structures. The algorithm used by Rosetta is called the Fragment Assembly of RNA (FARNA) algorithm.
In this study, we aimed to predict only the optimal structure by minimizing energies

[8]https://www.rosettacommons.org

in RNA tertiary structures. The sampling process is repeated up to 1.000.000 cycles. We used such significant cycles numbers because the prediction accuracy depends on the number of the sampling process as Rosetta prediction based on a Monte Carlo prediction algorithm. Refer to Rosetta parameters table for more details.

2.2.4.1 Digitalizing TSSaRNAs tertiary structures

To annotate TSSaRNAs tertiary as well as to interpret their structures stability, we converted the predicted TSSaRNAs tertiary structural representation from graph representation into a digital matrix containing low and high-resolution energies for each predicted TSSaRNA tertiary structure.

2.2.5 Predicting TSSaRNAs higher-order structures

Prediction of the secondary and tertiary structures of ncRNAs is an important step to understand their structural stability. However, when investigating their biological function, it is crucial to consider the higher order structures to understand the actives sites and the potential mode of action of ncRNAs. The term higher order RNA structure is used to refer to any structure beyond the tertiary structure and mainly yield by interacting with other biomolecules or due chemical modifications. In this study, we aimed to investigate the higher order structures of TSSaRNAs by investigating TSSAsRNA-mRNA interactions and TSSaRNAs-Protein interactions.

2.2.5.1 TSSaRNA-mRNA interactions

Many ncRNAs trigger their biological function by interacting with other RNA for example miRNA in eukaryotes and sRNA in bacteria. RNA-RNA interaction is an important process in the biological system, yet it remains a challenging point for predictions due to the limitation in the current computational algorithm or due to the computational costs when predicting RNA-RNA interaction from a large dataset.

By speculating TSSaRNAs could play a role as a regulatory element to regulate the cognate genes as a target mRNA; here in this study, we investigated the potential interaction

between TSSaRNAs and their cognate genes. We used IntaRNA [25] tool version 2 to predict the interactions. IntaRNA is a general purpose tool that maintained by Prof. Backofen's bioinformatics group at Freiburg University[9]. IntaRNA is available as web-server as well as a standalone tool to be run on local machines. IntaRNA is considered as an ultra-fast tool to predict mRNA target sites for given ncRNAs by minimizing the thermodynamics energy profiles of the hybridization. To compute the target site for interactions, IntaRNA uses heuristic computational methods by computing and determining the probability for the unpaired region on ncRNA then hybridize the predicted unpaired region with the mRNA. By using the heuristic computational methods, IntaRNA could determine all possible interactions between two molecules except the interactions which yield by pseudo-knots. The thermodynamic interacting energy had calculated from two sources of contribution. The first is the hybridization free energy of the interacting subsequences which computed as folding energy. The second source of the interactions energy is the free energies that needed to unfold the interaction stacks sites in both ncRNA and mRNA molecules.

2.2.5.2 TSSaRNA-Lsm interactions

We mentioned in chapter 1, RNA-Protein interactions trigger essential roles in many biological functions. In this study, we have hypothesized the existence of TSSaRNAs-Proteins interaction to trigger the potential function of TSSaRNAs. We have speculated that TSSaRNAs posse regulatory roles; therefore, we investigated the possibility of interaction of TSSaRNAs with Lsm protein. Lsm protein is known as a protein involved in RNA processing and mediates RNA function. The investigation process consisted of two steps: in the first step, we had investigated the potential TSSaRNAs-Lsm binding based on wet-lab experimental data obtained from the experimental RIP-seq data of Lsm-RNA interaction which is generated previously by the scientist in the LaBiSisMi[10]. It is important to point out that the results of these experiments could demonstrate

[9] http://rna.informatik.uni-freiburg.de/IntaRNA/Input.jsp
[10] http://labisismi.fmrp.usp.br/index.php/en/

the evidence of binding between Lsm protein and RNAs molecules that expressed at the genomic coordinates of TSSaRNAs. However, we could not emphasize that Lsm protein binds only with TSSaRNAs. Although, we could speculate that Lsm has the capability to binds to all RNA molecules that expressed within the coordinates of the binding signal including mRNAs and TSSaRNAs.

Once we obtained TSSaRNAs that showed experimentally evidence to interact with Lsm protein we investigated the potential binding site of the complex as well as the geometrical topologies of interaction by performing computational docking experiments. Also, we studied the binding energy of the complex to estimate the thermodynamic stability of TSSaRNA-Lsm complex.

We obtained the processed data (log2-fold-change) enrichment at any genomic position) from Dr. Livia Zaramehla Ph.D. thesis work in LaBiSisMi laboratory at FMRP-USP [100]. This Ph.D. work was not published in peer-reviewed journals yet but rather only as Ph.D. thesis so is, in some sense, original data from the group. The wet-lab experiment has done to investigate genome-wide RNA-Lsm complexes using RIP-seq technology (RNA-immunoprecipitation (RIP) followed by high-throughput sequencing (seq)). The organism *Halobacterium salinarum* NRC-1 was cultured in an enriched nutrient medium and incubated in the standard growth condition by providing incubation temperature of 37°C and agitation of 125 revolutions per minute. To perform RIP-seq experiment our colleagues at LaBiSisMi collaboratively with Dr. Elisabeth Wurtmann from Institute for Systems Biology(ISB)[11] had generated a modified experimental protocol based on the current well-established protocols of ChIP-chip assays (i.e Chromatin immunoprecipitation (ChIP) followed by microarray hybridization (chip)), RIP-Chip and CLIP (i.e crosslinking immunoprecipitation) (refer to the reference [100] for more details). To obtain the target immunoprecipitated RNA, the sample was incubated at 37°C for two hours with 15μL of Proteinase K. As the last stage, to verify the presence of Lsm protein the scientist at LaBiSisMi performed SDS-PAGE and Western Blot experiments (for the detailed steps refer to the reference [100]).

[11] https://baliga.systemsbiology.net/hs2013/?q=education/interviews/wurtmann

We performed the docking experiments using ZDOCK docking tool version 3.0.2 [101]. ZDOCK tool is considered as one of the best open source tool for molecular docking based on the Fourier transform methods. ZDOCK is generated to predict protein-protein interactions, but it could be used to predict RNA-Protein interactions or even RNA-RNA interactions after the parametrization of RNA molecules. We overcome the lack of RNA parameters such as lack of the partial charge of atoms and the Van der Waals radii by using the parameters generated by Junichi Iwakiri and his colleagues [26]. We used RNA parameters that generated based on the CHARMM22 molecular dynamic force fields as ZDOCK original parameters for protein is based on the CHARMM22 force fields. Regarding the Lsm protein as there is no crystal structure of *Halobacterium salinarum* NRC-1 yet, therefore, we used the crystal structure of Lsm protein from the thermoacidophilic archaeon *Sulfolobus acidocaldarius*. The protein blast between *Halobacterium salinarum* NRC-1 Lsm protein (UniProtKB: B0R5R5 which is encoded by the gene VNG1496G) and *Sulfolobus acidocaldarius* Lsm protein (UniProtKB: A0A0U3FHU0) showed 33% (23/68) amino acid sequence identity and 54% amino acid sequence similarity associated with E-value equal to $2 * 10^{-9}$. We downloaded the Lsm protein from the protein data bank by querying using the PDB identifier(ID) 5MKL. The Lsm crystal structure (5MKL) was determined via Diffraction-based technique (resolution of 2.086Å; to see other parameters refer to the Digital Appendix (4)) [102]. Before using Lsm protein for docking, we cleaned the 5MKL PDB file by removing the solvent molecules and as 5MKL has four biological assemblies for the Lsm protein, therefore, as a preprocessing step we extracted only the first biological assembly from the 5MKL PDB file to use it as the receptor molecule for the docking experiment.

To calculate TSSaRNAs-Lsm binding energies we used APBS (Adaptive Poisson-Boltzmann Solver) version 1.4.1 [103]. For each molecule, we had set a grid dimension (275x275x275 nm) to ensure the accuracy in calculating the binding energy. Moreover, as RNA molecules are highly charged molecules; therefore, we used the non-linearized Poisson-Boltzmann method to calculate the potential of molecules.

2.3 TSSaRNAs functional annotation

Our pipeline to associate TSSaRNAs into their Rfam families consists of two main steps. The first step was querying TSSaRNAs as sequences against the covariance models of Rfam families if TSSaRNA hit a Rfam family then we have succeeded to annotate TSSaRNA into corresponding Rfam family, the second step is an alternative step where we could not succeed to annotate TSSaRNA through the step one. The reason for performing the alternative step is to do not lose Rfam families in the conserved structural element of TSSaRNAs. The alternative annotation step started by grouping TSSaRNAs based on hierarchical clustering method into closely related sequences with consensus secondary structure then we queried Rfam sequences against the covariance model of each TSSaRNAs group. For more details about the pipeline refer to the flowchart(Fig. 2.2) which explains all steps.

2.3.1 TSSaRNAs annotation based on single molecule information

We used infernal package ("INFERence of RNA Alignment")[12] [104] version 1.1.2 to annotate TSSaRNAs by detecting all known Rfam families which had represented in their sequences. In this step, the input files consisted of TSSaRNAs sequences in fasta format and the covariance models from Rfam database version 13.0. As we aimed to get the maximum numbers of Rfam families that represented in each TSSaRNA molecule, therefore, we used the default parameters in the infernal package without any changes.
It is import to note that we were more restrictive on annotating CRISPR families. The CRISPR families have considered as the systems of the adaptive immunity against foreign genetic elements in both Archaea & Bacteria. Although, as far as we know the CRISPR families have not discovered on *Halobacterium salinarum* NRC-1 yet. Therefore, in case of some TSSaRNAs had classified into the corresponding CRISPR families we considered a further annotation step by performing local sequence alignment between CRISPR seed sequences of a given CRISPR family and TSSaRNAs sequences. To perform the local

[12]http://eddylab.org/infernal/

sequence alignment experiment and visualize the alignments result we used locARNA on-line tool[13] [105, 106] with its default parameters.

2.3.2 TSSaRNAs annotation based on consensus secondary structures

As Rfam database represents ncRNA families based on their consensus secondary structures, therefore, it will be considered to query Rfam database using consensus secondary structures of close related TSSaRNAs in case a single sequence of TSSaRNA did not hit any Rfam family. This step helps to recover any missing TSSaRNAs annotation when relying only on the querying TSSaRNAs sequences against the CMs models of Rfam families. To query Rfam families against TSSaRNAs consensus secondary structures first we clustered TSSaRNAs based on their structural similarities then we constructed the CM models of selected then finally we queried the sequences of Rfam families against TSSaRNAs CM models. Below are the details for each step.

2.3.2.1 Clustering TSSaRNAs sequences

We used the RNAclust tool [107, 108] version 1.3 to get the cluster-tree of TSSaRNAs sequences that did not succeed to be annotated into their corresponding Rfam families when we queried their sequences against Rfam CMs models. RNAclust groups RNA molecules into clusters based on their secondary structures motifs by computing all possible pairwise alignments using LocARNA tool [105, 106]. After calculating the pairwise alignment for all TSSaRNAs, RNAclust calculates the probability matrix of secondary structure distribution using the RNAfold program from the ViennaRNA package [109]. Based on the calculated probability matrix, RNAclust can identify the relevant nodes in the cluster-tree which shared a consensus secondary structure. Besides, using RNAclust tool we used RNAsoup program (Spot grOUPs in RNA cluster-tree) [110] to detect the maximum numbers of cluster-tree nodes based on evaluating the squared error of the minimum free energy for every single TSSaRNA within given cluster node. This step allows calculating the minimum free energy of the consensus secondary structure in a

[13] http://rna.informatik.uni-freiburg.de/LocARNA/Input.jsp

given node.

2.3.2.2 Construction of the covariance models of TSSaRNAs cluster-tree nodes

To construct the CMs models for the unannotated TSSsRNAs datasets, first, we pre-processed the selected TSSaRNAs cluster-tree nodes obtained from RNAsoup program by converting the multiple RNA sequence alignment format into Stockholm alignment format using t_coffee [111] multiple sequence alignment package version 11.00.8cbe486[14]. Then, we used infernal package version 1.1.2 to build the covariance models of selected TSSaRNAs nodes which fulfilled the selection criteria such as that a given node has consensus secondary structures and its minimum free energy is thermodynamically stable (negative value).

2.3.2.3 Querying TSSaRNAs CM against Rfam families

After getting the co-variance models for the selected nodes, we calibrated the E-values of all models using infernal package version 1.1.2 before querying them against the Rfam sequences. This step is critical as it would adjust the statistical parameters (E-values) for each model [104]. As we aimed to reduce the false positives results which associate a large number of Rfam families with singling TSSaRNA CM model we performed the query with E-value < 0.001 as the query parameter.

2.4 Functional annotation of cognate genes

This step aimed to determine the gene ontology categories that enriched in the list of the cognate genes of TSSaRNAs. To get the gene ontology categories of the cognate genes, we performed gene classification for the list of the cognate genes based on PANTHER on-line classification system [112]. The PANTHER on-line classification system calculates the overall proportion of the genes from the given list that associated with significant terms based on biological process, cellular components or the molecular functions.

[14] http://tcoffee.crg.cat/

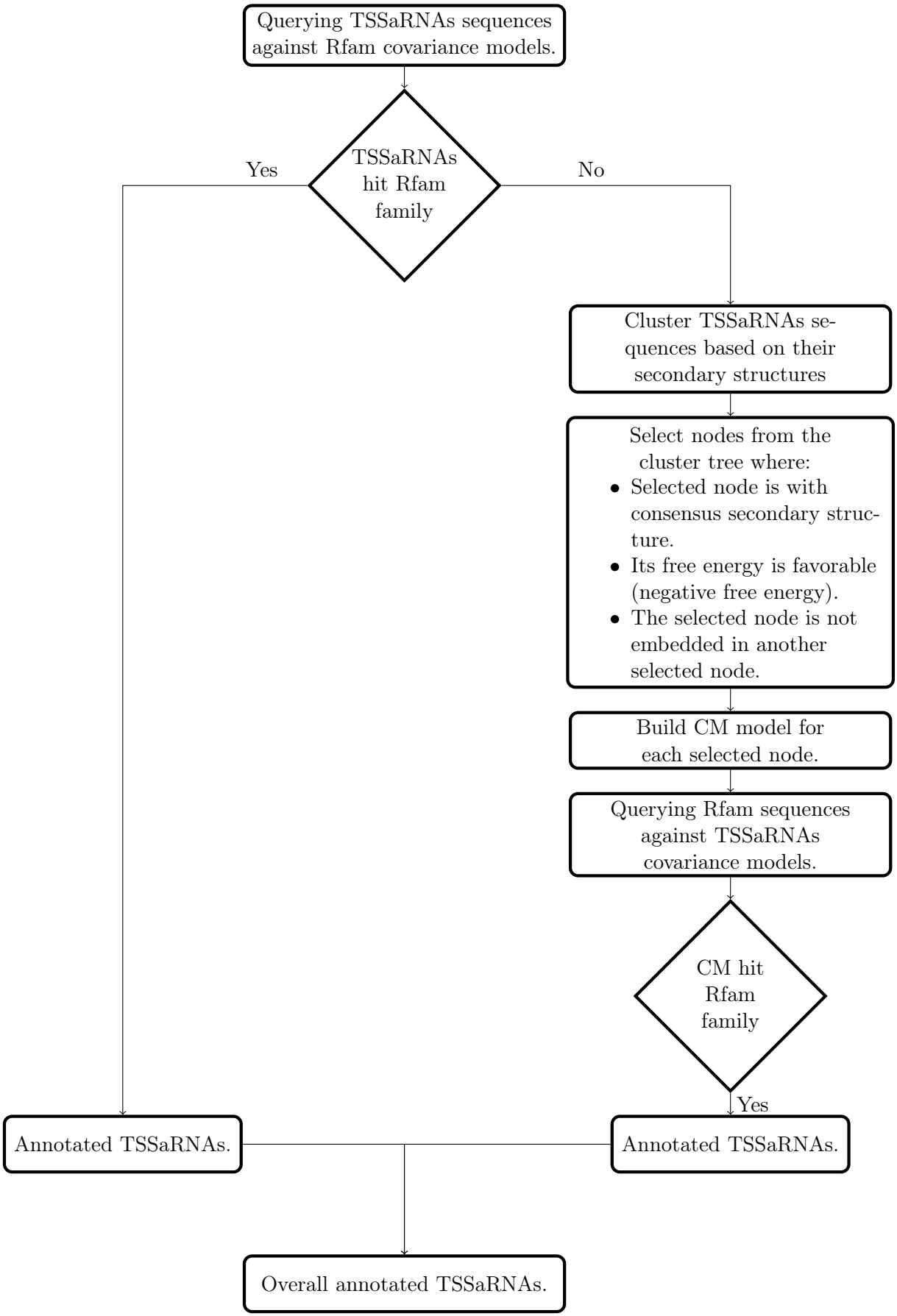


Figure 2.2: A diagram shows the steps to annotate TSSaRNAs based on Rfam functional classification of ncRNAs.

CHAPTER 3

Results

3.1 Raw data quality control assessment

Using the FastQC tool, we have assessed the quality control of the three raw data sets. The result obtained from the quality control assessment showed that: in respect to the contamination with adapters, the data sets (SRX844124) and (SRX441605) were free from any contamination of adapters sequences. The data set (SRX433542) showed contamination with Illumina small RNA adapters (Fig. 3.1) therefore, the adapters removal process has proceeded on the data set (SRX433542) before involving it into TSSaRNA prediction algorithm.

By assessing the Phred score per position, results showed that all data sets were with good Phred scores. The good Phred scores were associated mainly with the 5' end of the reads (Fig. 3.2). Phred score is measuring the quality of identification of the sequenced bases (base call accuracy). Phred score is used to estimate the probability error of the base calling. The probability error indicates the probability of a particular base is incorrect. The mathematical formula of the Phred score could be written as follows:

$$Q = -10 \log10(P)$$

Where Q is the Phred score, P is the probability error.

The FastQC tool uses the Phred score of 28 that corresponds to the probable error of 0.001584 as a cutoff value to pass the quality control. Besides good Phred scores, all

the data sets showed good content of the four types of nucleotides G, U, A, and C per position across all bases (Fig. 3.3).

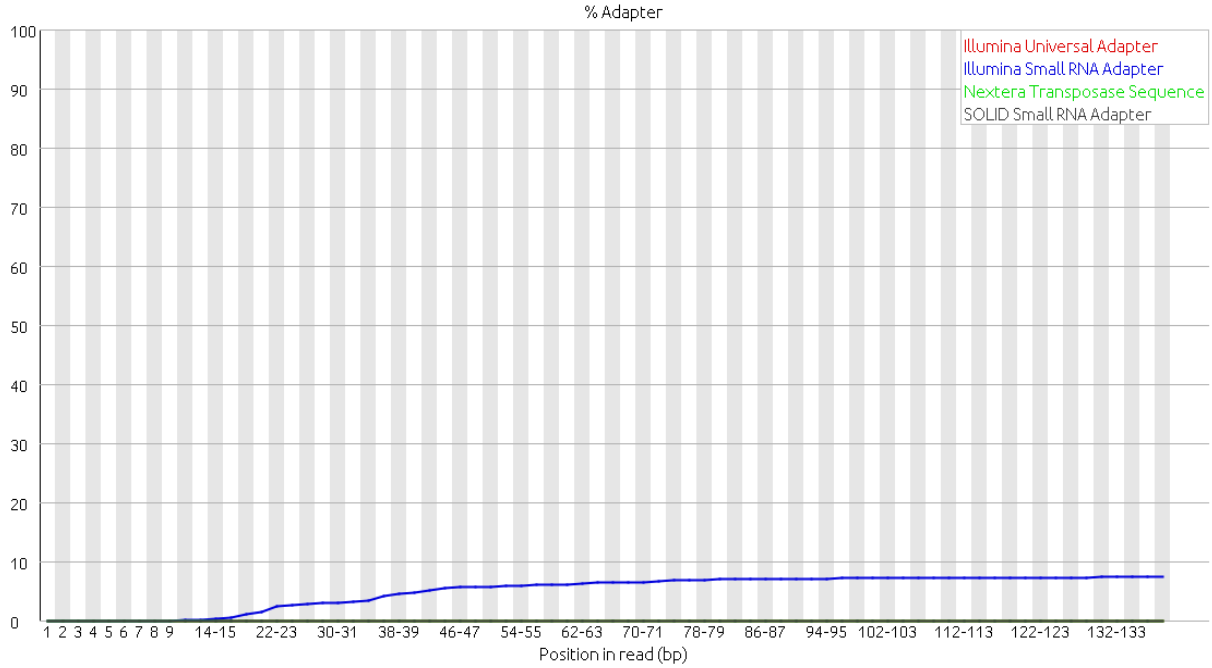
Regarding the distribution of reads length within each experiment, we have paid more attention to this term, and we have considered it as the bottleneck to fulfill the criteria of passing the quality control requirements as having reads with different length is critical to predicting TSSaRNAs. The quality checking of reads length showed that: the data sets (SRX844124) and cleaned (SRX433542) both were with proper enough distribution of reads length within the experiments(Fig. 3.4) while the data set (SRX441605) contained only short reads sized of 35 nucleotides length for majority of the reads therefore, we have not considered this dataset for TSSaRNAs prediction.

In summary, the first data set (SRX844124) is good enough to be used directly into TSSaRNAs prediction algorithm. The second data set (SRX433542) contaminated with adapters reads which needed to remove before using this dataset into TSSaRNAs prediction algorithm. The third data set (SRX441605) has excluded due to the limitation on its read length distribution. The basic statistics of reads in the three data sets which had used in this study has shown on the table (3.1).

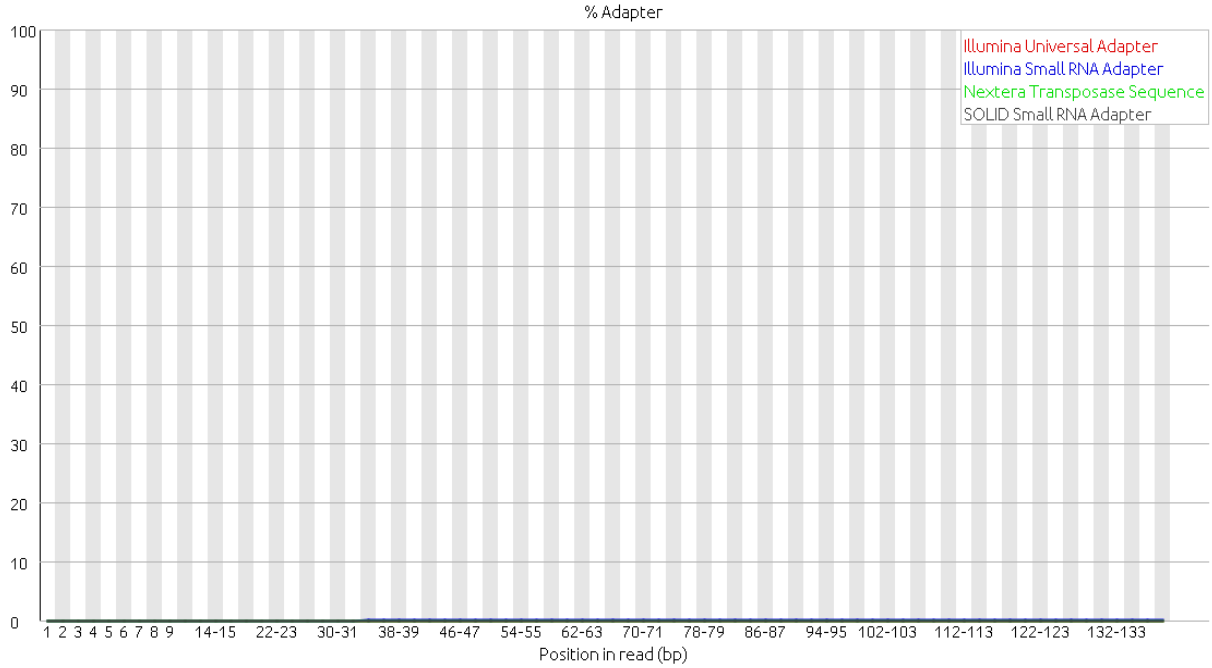
Table 3.1: Basic statistics of reads in the three data sets used in this study.

Experiment ID	Experiment I SRX844124		Experiment II SRX433542		Experiment III SRX441605		
Sample ID	SRR1760797	SRR1760798	SRR1187083	SRR1187084	SRR1743299	SRR1743300	SRR1743301
Total Reads	293943	4958743	2996953 1940715*	3238933 2065829*	1179648	13757181	22522951
Sequences flagged as poor quality	0	0	0	0	0	0	0
Length	5 - 151	5 - 151	151 (20 - 151)*	151 (20 - 151)*	35	35	35
%GC	62%	62%	40% 61%*	40% 61%*	46%	55%	54%

(*) indicates for value obtained after trim adapters.

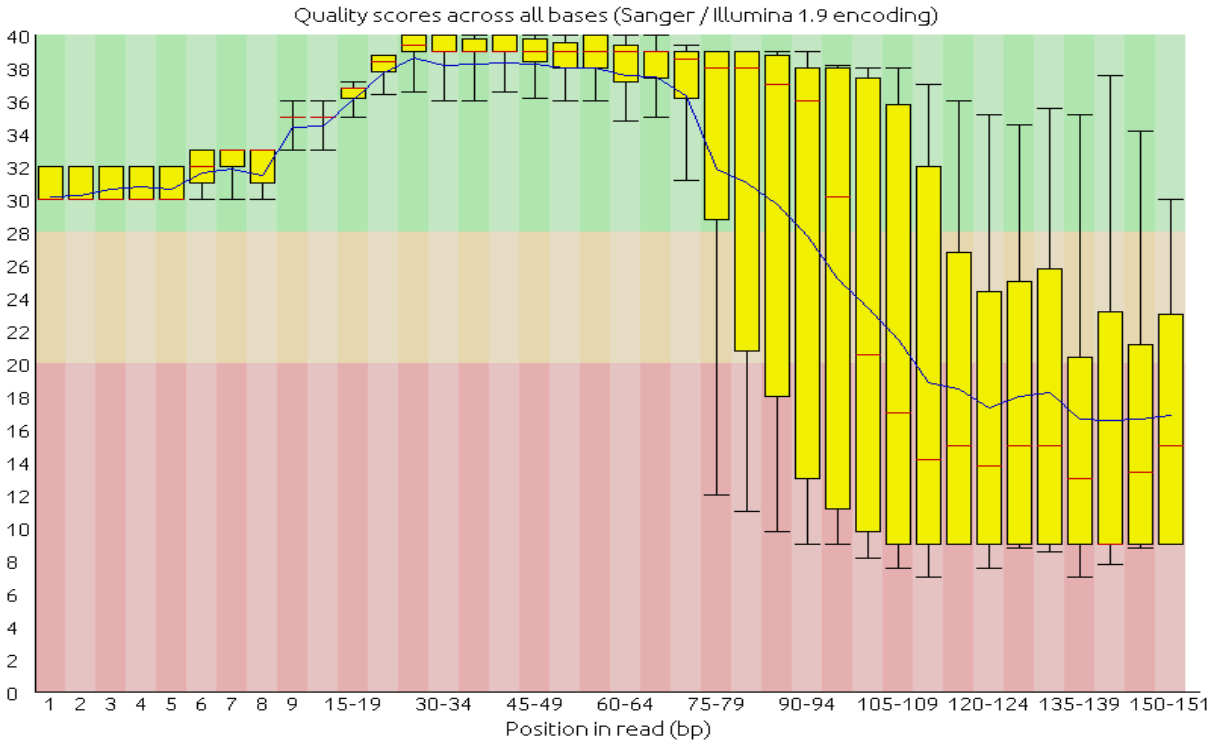


(a) Experiment II: SRR1187083 before adapters removal.

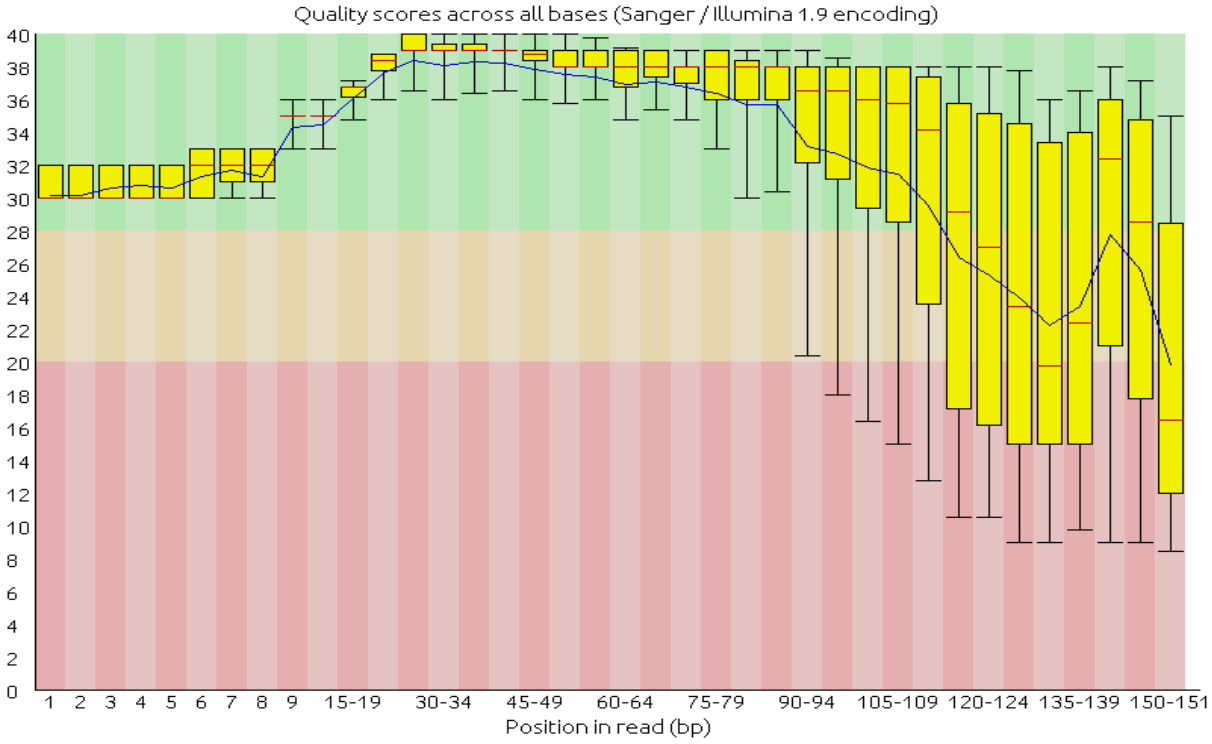


(b) Experiment II: SRR1187083 after adapters removal.

Figure 3.1: Sequence Quality Control: Adapters Content.
This figure shows an example of the adapter contents assessment for the first samples of the second experiment. The blue line in the sub-figure (a) shows the contamination with "Illumina Small RNA Adapter".

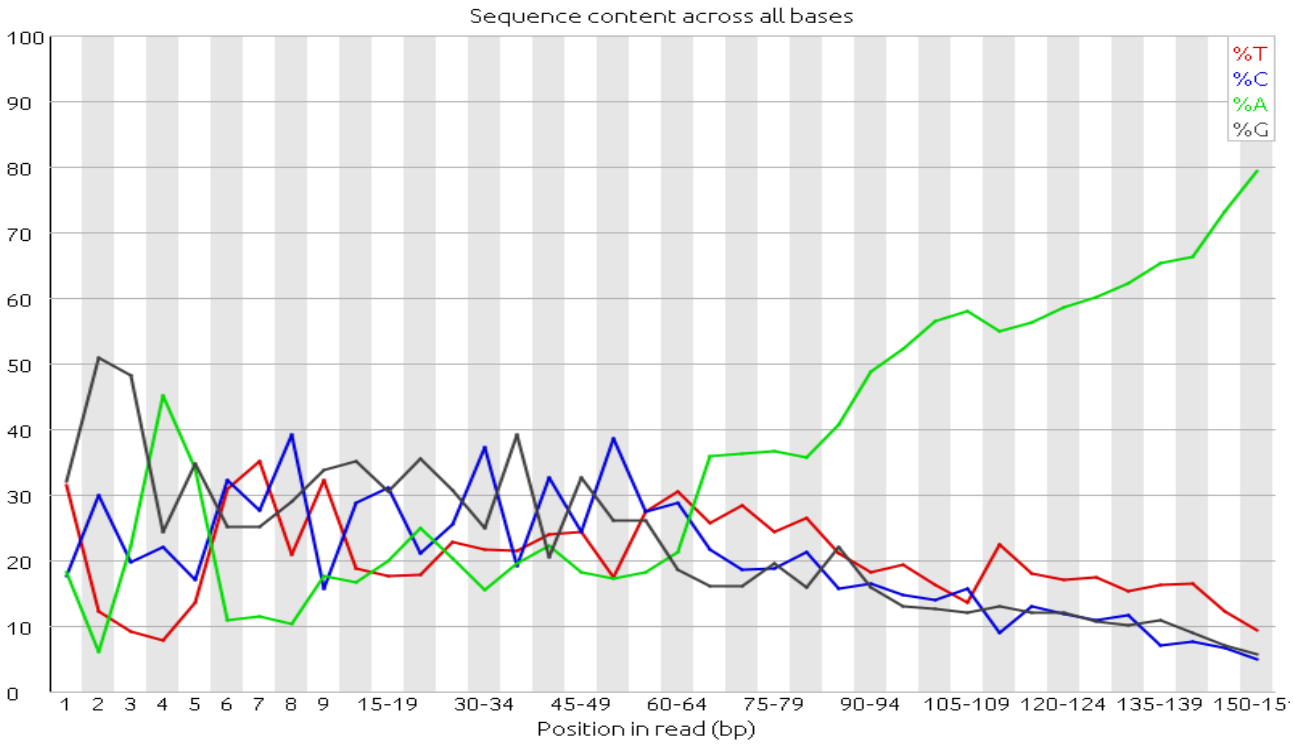


(a) Experiment II: SRR1187083 before adapters removal.

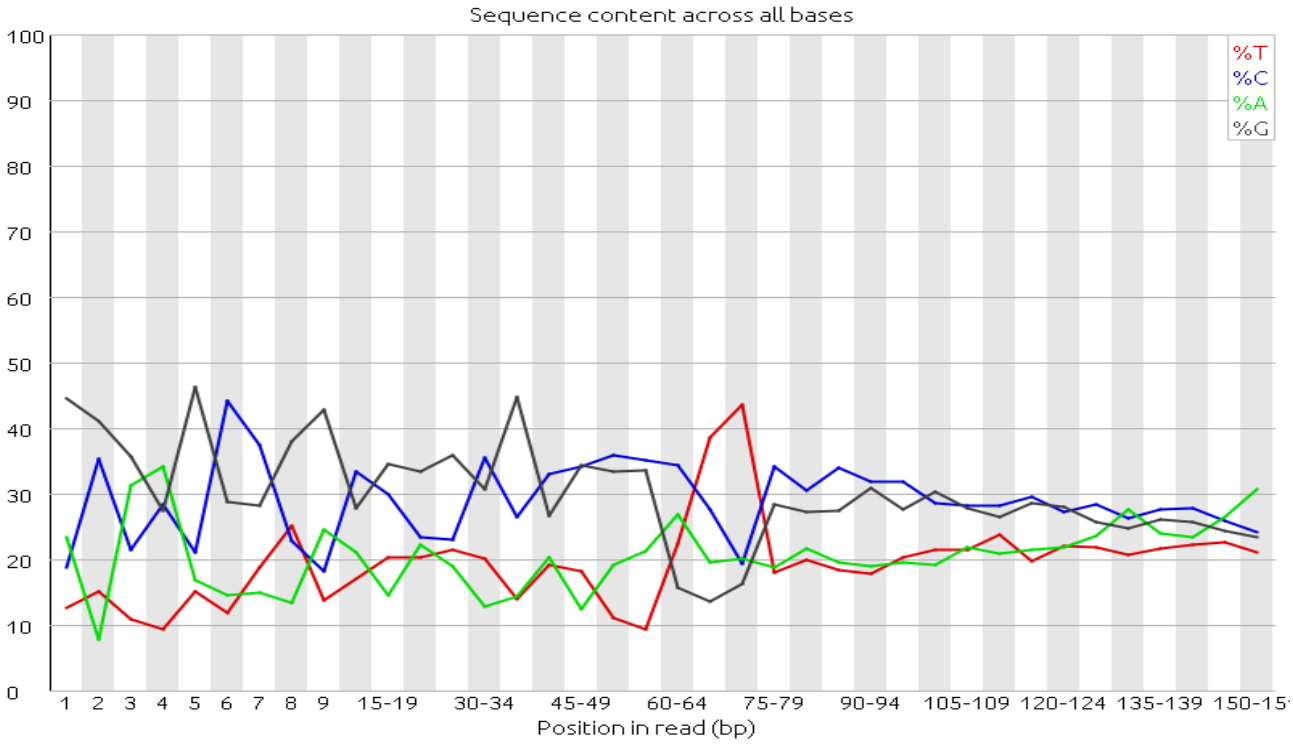


(b) Experiment II: SRR1187083 after adapters removal.

Figure 3.2: Sequence Quality Control: Quality scores across all Bases.
The figure shows an example of the quality control assessment for per-base quality scores across all reads before and after the adapters removal process in the first samples of the second experiment. The y-axis shows the quality scores(perfect quality score (green), reasonable quality score (orange), poor quality (red)). The score statistics for each position are depicted as follows: the median is shown by a central red line, the inter-quartile range (25-75%) is represented in a yellow box, the 10% point (lower whisker), the 90% point (upper whisker), the mean quality is depicted in blue line.

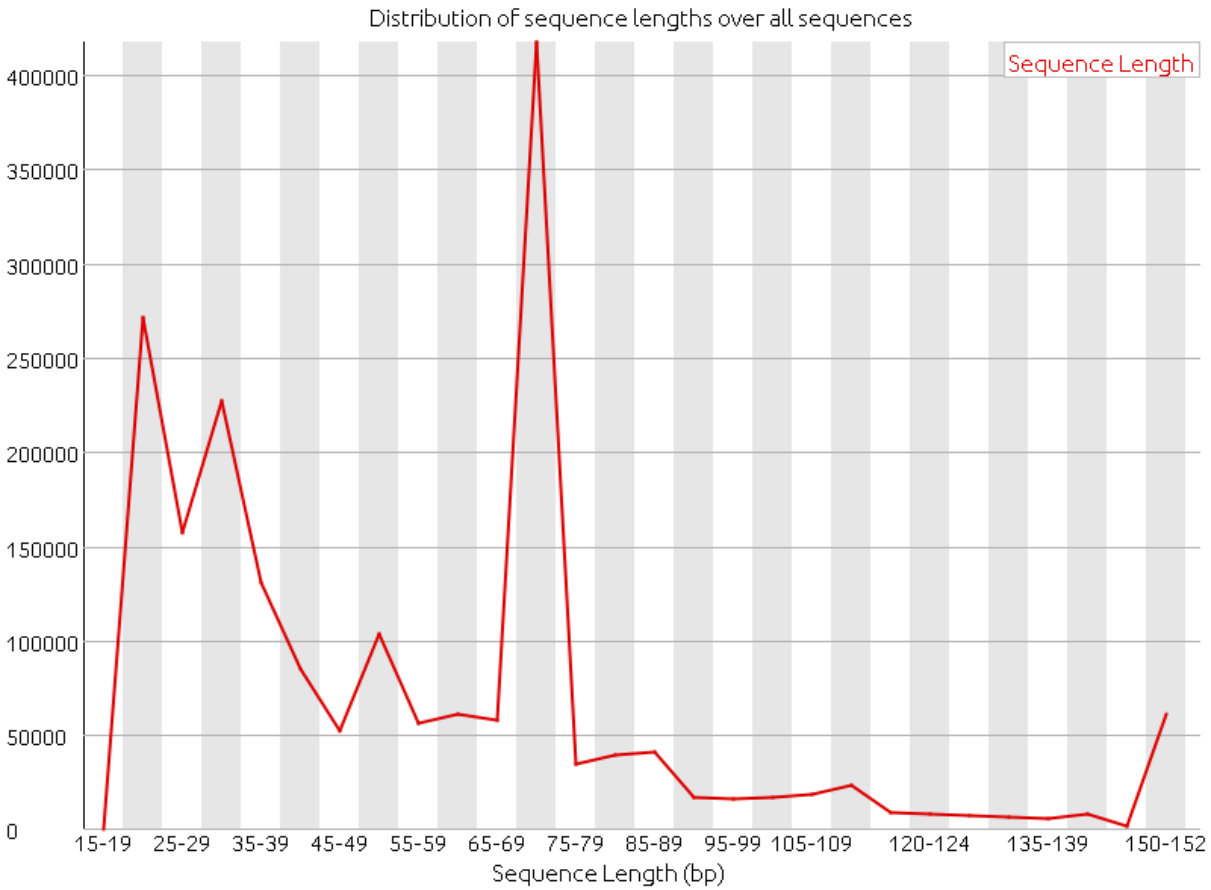


(a) Experiment II: SRR1187083 before adapters removal.



(b) Experiment II: SRR1187083 after adapters removal.

Figure 3.3: Sequence Quality Control: per base sequence content.
The figure shows an example of the quality control assessment for per base sequence contents before and after adapters removal process in the first sample of second experiment. The sub-figure (a) shows big differences between A and T towards the end of reads.



(a) Experiment II: SRR1187083 after adapters removal.

Figure 3.4: Sequence Quality Control: Distribution of Sequence length.
The figure shows an example of the quality control assessment of the distribution of the sequence length size of the first samples in the second Experiment.

3.2 TSSaRNA prediction

3.2.1 Mapping reads to the reference genome

The results obtained from the mapping process revealed that all data sets contain only unpaired reads and a high percentage of these reads have uniquely mapped for one genomic position. The overall alignment score showed quite enough high score (more than 90%) for both data sets (SRX844124) and cleaned (SRX433542) (Table 3.2).
In summary, the data sets (SRX844124) and (SRX433542) showed good enough alignment score to be used into TSSaRNAs prediction algorithm. The third data set (SRX441605) was processed in this stage although we already decided not to use it into our prediction algorithm. The result of the mapping process of the third data set (SRX441605) showed a high percentage of reads that have not had mapped into the reference genome of *Halobacterium salinarum* NRC-1. This result gave us evidence about the contamination of reads in the third experiment with RNA from other organism supporting its exclusion.

3.2.1.1 The initialization step

The result of the initialization stage (Fig. 2.1) showed that more than 80% of the 2622 RefSeq annotated genes possessed upstream window size equal to 50 nucleotides. On the other hand, around 90% of the RefSeq 2622 annotated genes maintained downstream window size identical to 50 nucleotides.

Table 3.2: Results of mapping RNA-seq reads to the reference genome.

Experiment ID	Experiment I SRX844124		Experiment II SRX433542		Experiment III SRX441605		
Sample ID	SRR1760797	SRR1760798	SRR1187083	SRR1187084	SRR1743299	SRR1743300	SRR1743301
Total Reads	293943	4958743	1940715	2065829	1179648	13757181	22522951
Unpaired Reads	293943 100.00%	4958743 100.00	1940715 100.00%	2065829 100.00%	1179648 100.00%	13757181 100.00%	22522951 100.00%
Aligned 0 times	132620 (4.51%)	159513 (3.22%)	188707 (9.72%)	103901 (5.03%)	647733 (54.91%)	1619394 (11.77%)	7033747 (31.23%)
Aligned 1 time	2305184 (78.43%)	3775847 (76.15%)	1500409 (77.31%)	1669219 (80.80%)	365914 (31.02%)	12045007 (87.55%)	15246189 (67.69%)
Aligned >1 times	501539 (17.06%)	1023383 (20.64%)	251599 (12.96%)	292709 (14.17%)	166001 (14.07%)	92780 (0.67%)	243015 (1.08%)
Overall alignment rate	95.49%	96.78%	90.28%	94.97%	45.09%	88.23%	68.77%

3.2.2 Predicted TSSaRNAs

To summarize the results of the predicted TSSaRNAs, we have considered the statistical properties of TSSaRNAs with respect to the total number of predicted TSSaRNAs per sample, the deviation of TSSaRNAs starting point from the transcription start sites, TSSaRNAs length sizes, and the number of predicted TSSaRNA per gene.

3.2.2.1 The first data set (SRX844124)

TSSaRNAs prediction algorithm has identified total 1063 sense TSSaRNAs candidates from the two samples of the first experiment. The concordance of TSSaRNAs between the two samples were 493 TSSaRNAs candidates associated with 493 cognate genes (18.8% of the considered annotated genome of 2622 genes). The disagreement TSSaRNAs were 168 TSSaRNAs which were predicted uniquely in the sample SRR1760797 and 403 TSSaRNAs which were identified uniquely in the sample SRR1760798. Considering the antisense TSSaRNAs, TSSaRNAs prediction algorithm has identified a total of 274 antisense TSSaRNAs candidates from the two samples of the first experiment. The concordance in the antisense TSSaRNAs between the two samples were 117 TSSaRNAs. Contrarily, the disagreement of antisense TSSaRNAs were 35 TSSaRNAs which predicted uniquely in the sample SRR1760797 and 112 TSSaRNAs which were identified uniquely in the sample SRR1760798.

The results of the statistical properties in the two samples of the first experiment regarding the deviation of TSSaRNAs starting point from the transcription start sites, the results showed that around 40% of the sense TSSaRNAs initiated exactly from the same point of the transcript start site signal of their cognate genes (Fig. 3.5 sub-figure a,b). On the other hand, the results showed that the antisense TSSaRNAs initiated uniformly within the predefined window (Fig. 3.6 sub-figure a,b).

Concerning length size distribution of predicted sense TSSaRNAs, the results showed a median size of 25 nucleotides. Furthermore, we have observed that around 90% of the sense TSSaRNAs sized less than or equal to 100 nucleotides (Fig.3.7 sub-figure a,b).

Considering the sizes of the antisense TSSaRNAs, the results showed a median size of 20 nucleotides. Similar to the sense TSSaRNAs, the results showed around 90% of the antisense TSSaRNAs sized less than or equal to 100 nucleotides (Fig. 3.8 sub-figure a,b). The maximum number of sense TSSaRNAs per cognate gene was 3 TSSaRNAs. In details, only one cognate gene possessed 3 TSSaRNAs in each experiment while 16 cognate genes possessed 2 TSSaRNAs in the first sample (SRR1760797) this number is reduced to be seven cognate genes that possessed2 TSSaRNAs in the second sample (SRR1760798). The percentage of the predicted TSSaRNA that their cognate gene associated with only one TSSaRNA was 97.3% and 99% of TSSaRNAs in the first and second sample respectively.

Regarding the maximum number of antisense TSSaRNAs per cognate, the result showed that in the first sample (SRR1760797) all the cognate genes associated with only one antisense TSSaRNA. On the other hand, in the second sample (SRR1760798) only three cognate genes were associated with two antisense TSSaRNAs while all the rest of cognate genes associated with only one antisense TSSaRNA.

3.2.2.2 The second data set (SRX433542)

The TSSaRNAs prediction algorithm resulted in a total of 680 sense TSSaRNAs candidates from the two samples of the second experiment. The concordance of the sense TSSaRNAs between the two samples was 336 sense TSSaRNAs candidates associated with 336 cognate genes (12.8% of the considered annotated genome of 2622 genes). The disagreement in predicted sense TSSaRNAs were 139 that identified only from the sample (SRR1187083) and 205 sense TSSaRNAs which were identified uniquely from the sample (SRR1187084). Considering the antisense TSSaRNAs, TSSaRNAs prediction algorithm resulted in a total of 123 antisense TSSaRNAs candidates from the two samples of the second experiment. The concordance of the antisense TSSaRNAs between the two samples were 67 antisense TSSaRNAs candidates. The disagreement in the predicted antisense TSSaRNAs were 26 molecules that identified only from the sample (SRR1187083) and 30 antisense TSSaRNAs that identified uniquely in the sample (SRR1187084).

The results of the statistical properties in the two samples of the second experiment regarding the deviation of TSSaRNAs starting point from the transcription start sites, showed that about 40% of sense TSSaRNAs in both samples had TSSaRNAs that initiated exactly from the same point of the transcript start site signal of their cognate genes (Fig. 3.5 sub-figure c,d). On the other hand, the antisense TSSaRNAs initiated uniformly within the predefined window (Fig. 3.6 sub-figure c,d).

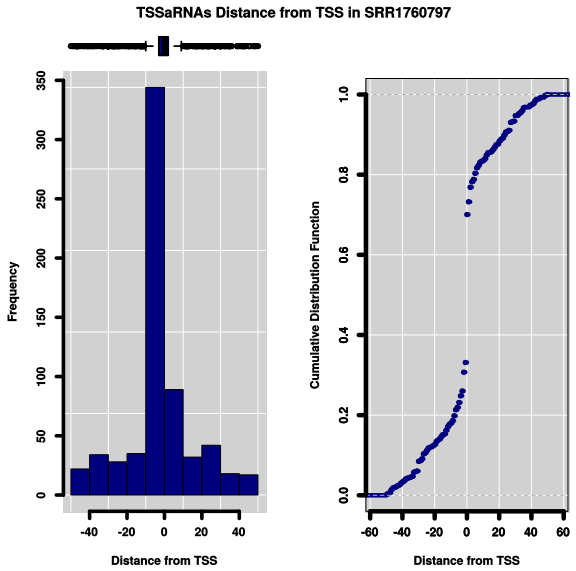
Considering the distribution of sense TSSaRNAs length size, the results showed a median size of 30 nucleotides. Moreover, we observed that around 90% of sense TSSaRNAs sized less than or equal to 80 nucleotides (Fig.3.7 sub-figure c,d). The distribution of antisense TSSaRNAs sizes showed a median length size of 20 nucleotides. Also, as same as in the sense TSSaRNAs we observed that around 90% of antisense TSSaRNAs sized less than or equal to 100 nucleotides (Fig. 3.8 sub-figure c,d).

The maximum number of predicted sense TSSaRNAs per cognate gene was 2 in both samples. In details, 12 cognate genes in the first sample (SRR1187083) possessed 2 TSSaRNAs while nine cognate genes in the second sample (SRR1187084) possessed 2 TSSaRNAs. The percentage of the predicted sense TSSaRNA that their cognate gene associated with only one TSSaRNA was 97.5% and 98.3% of TSSaRNAs in first and second sample respectively.

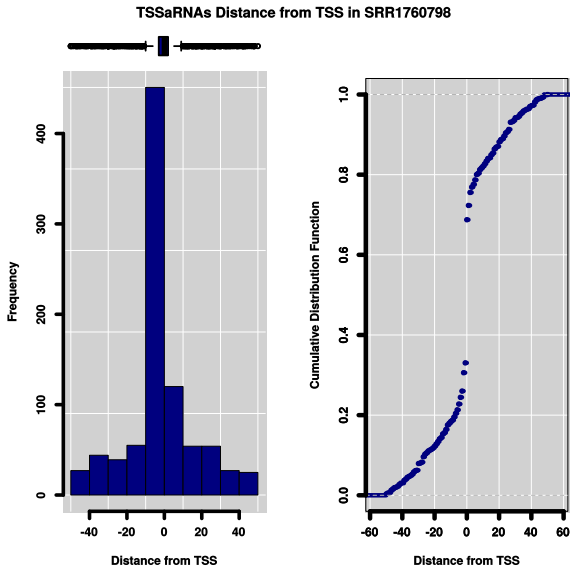
Regarding the maximum number of predicted antisense TSSaRNAs per cognate gene, the result showed that the first sample of the second experiment (SRR1187083) all the cognate genes associated with only one antisense TSSaRNA expect one cognate gene. On the other hand, all the cognate genes in the second sample (SRR1187084) had associated with only one TSSaRNA.

3.2.2.3 The third data set (SRX441605)

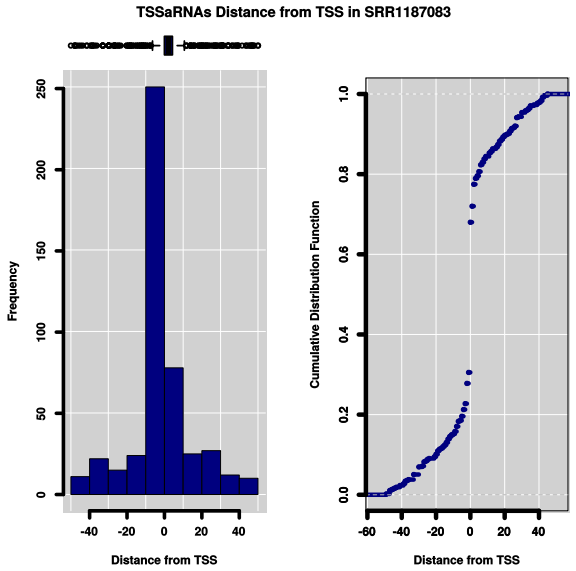
We have excluded the third data set from the pipeline due to the poor quality control and due to the evidence of the contamination with other organism's RNAs.



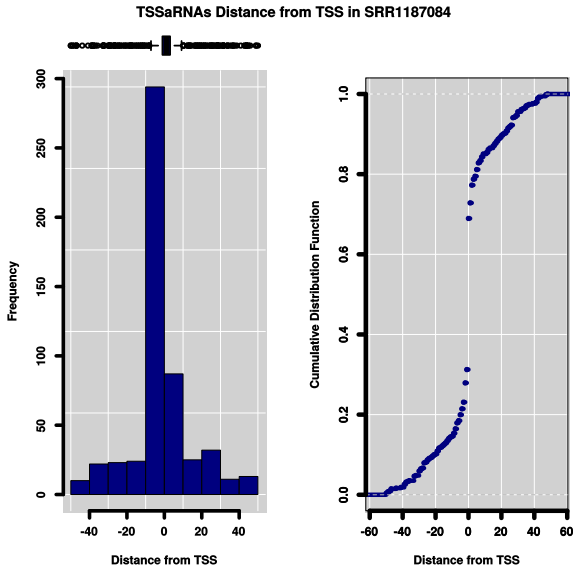
(a) Experiment I: Sample(SRR1760797).



(b) Experiment I: Sample(SRR1760798).

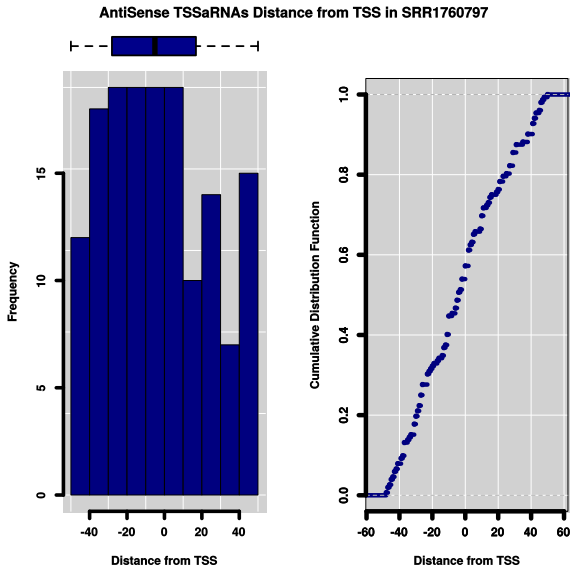


(c) Experiment II: Sample Id (SRR1187083).

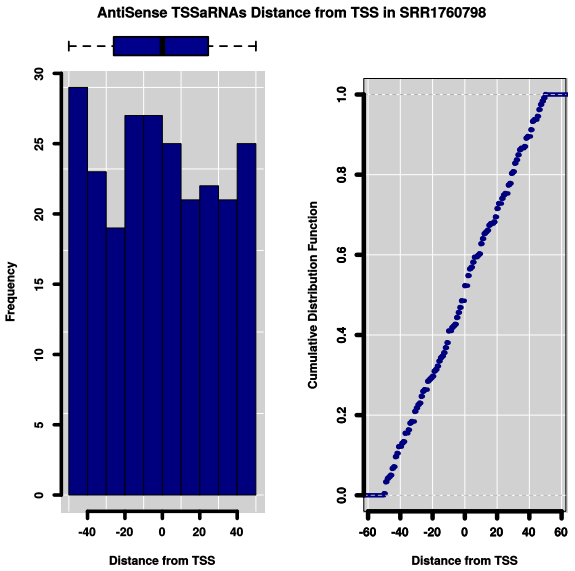


(d) Experiment II: Sample Id (SRR1187084).

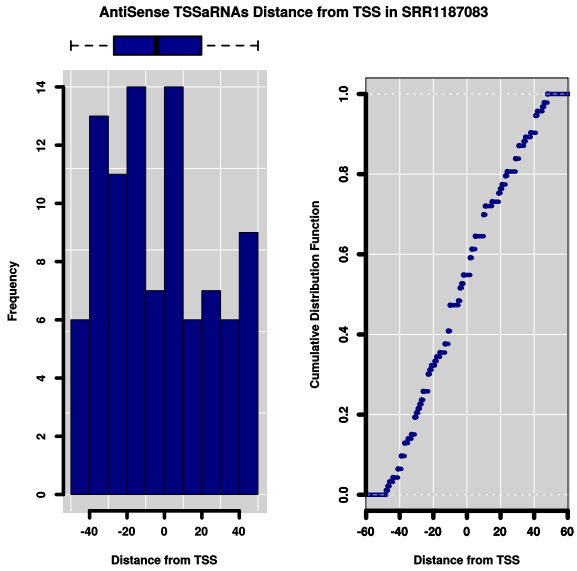
Figure 3.5: The distribution of distances of predicted sense TSSaRNAs from the transcription start sites.



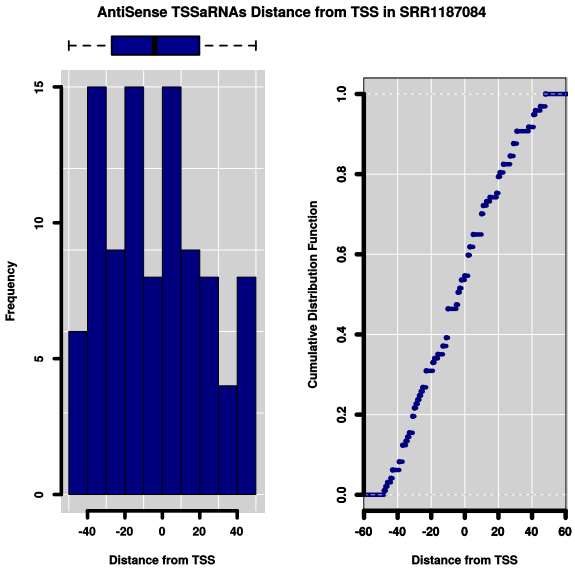
(a) Experiment I: Sample(SRR1760797).



(b) Experiment I: Sample(SRR1760798).

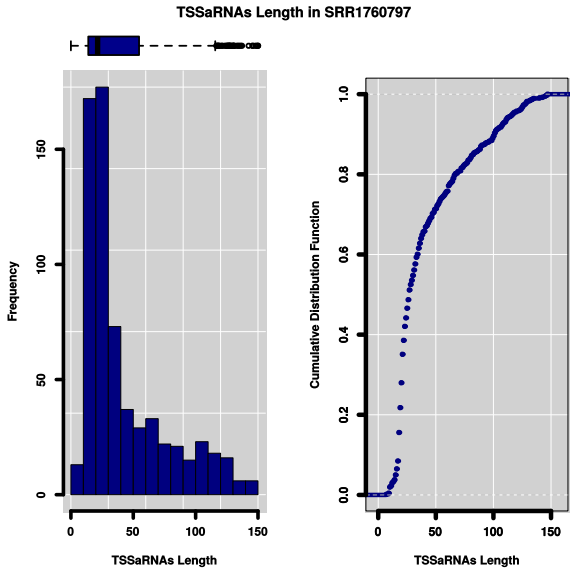


(c) Experiment II: Sample Id (SRR1187083).

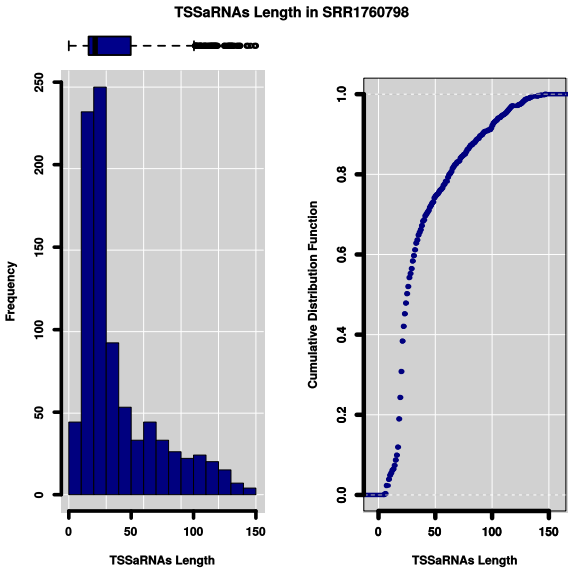


(d) Experiment II: Sample Id (SRR1187084).

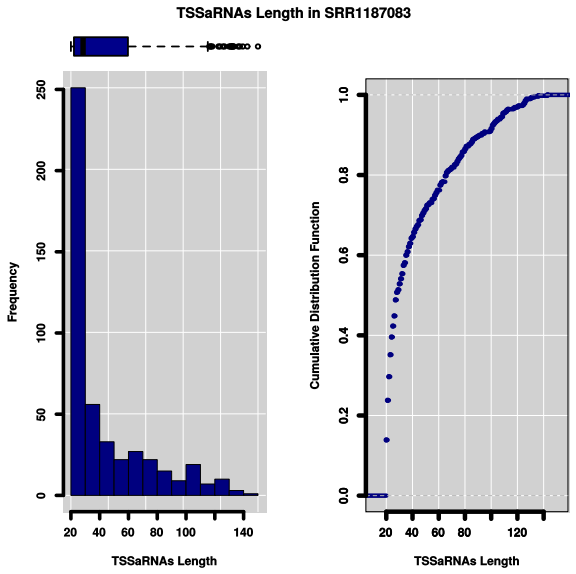
Figure 3.6: The distribution of distances of predicted antisense TSSaRNAs from the transcription start sites. This figure demonstrates the antisense TSSaRNAs were initiated uniformly within the predefined window.



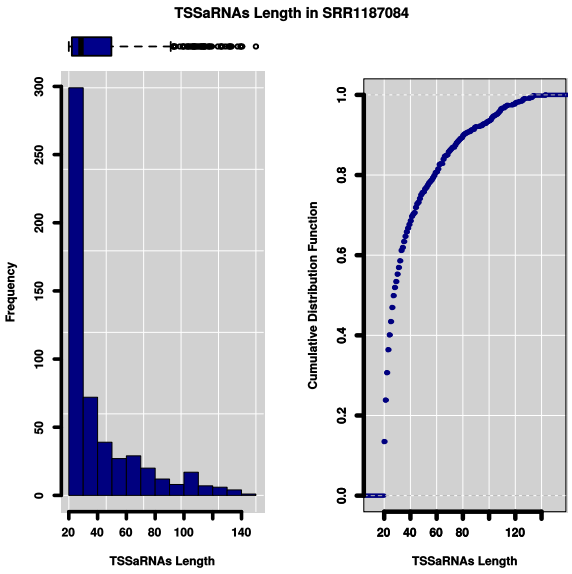
(a) Experiment I: Sample(SRR1760797).



(b) Experiment I: Sample(SRR1760798).

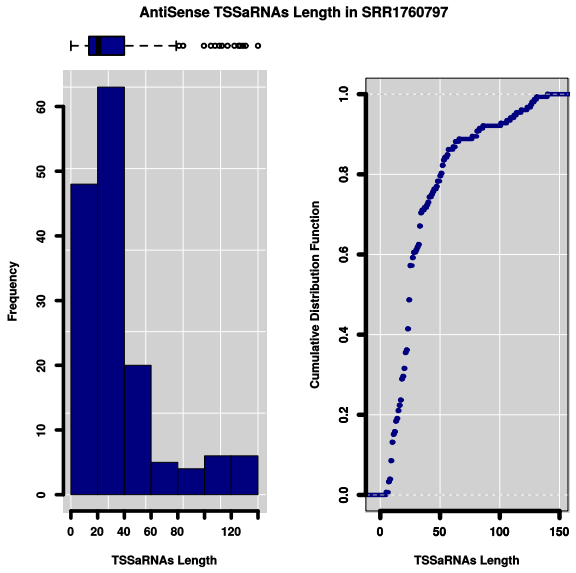


(c) Experiment II: Sample Id (SRR1187083).

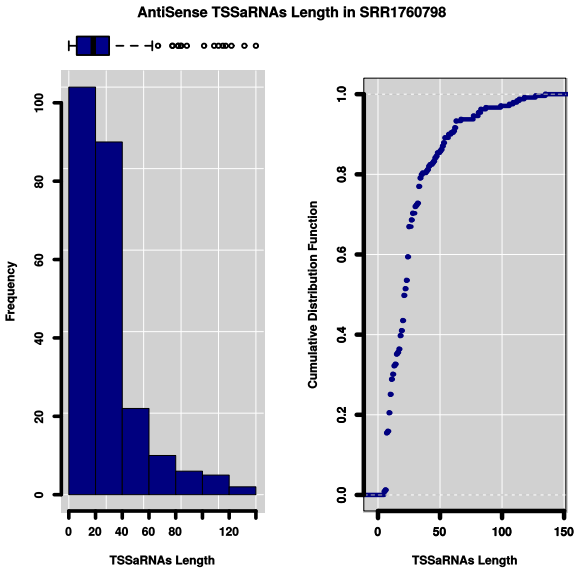


(d) Experiment II: Sample Id (SRR1187084).

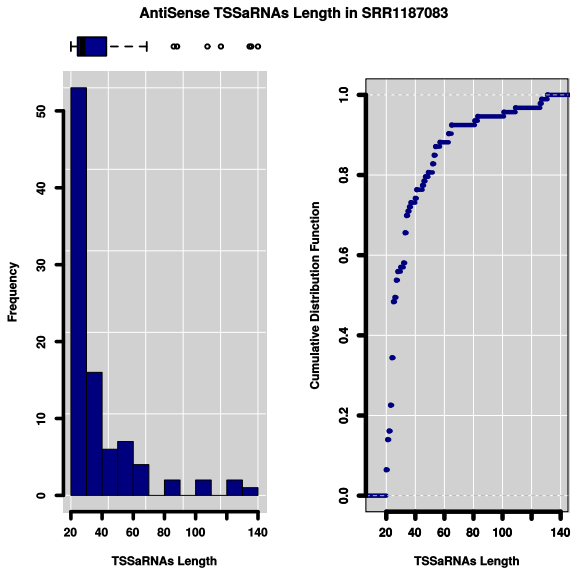
Figure 3.7: The distribution of sense TSSaRNAs' length size.



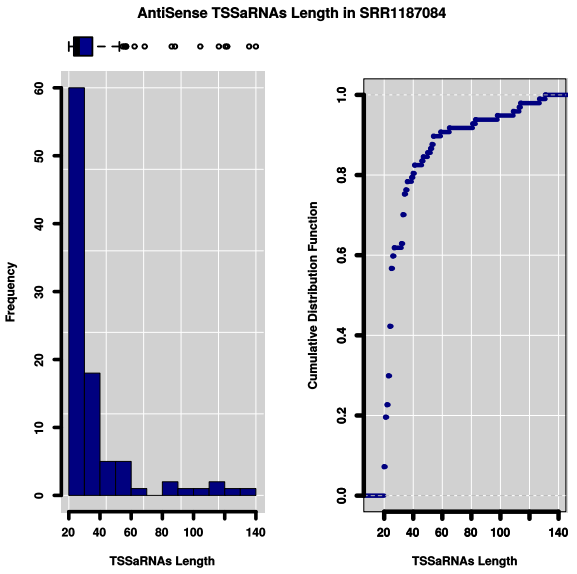
(a) Experiment I: Sample(SRR1760797).



(b) Experiment I: Sample(SRR1760798).



(c) Experiment II: Sample Id (SRR1187083).



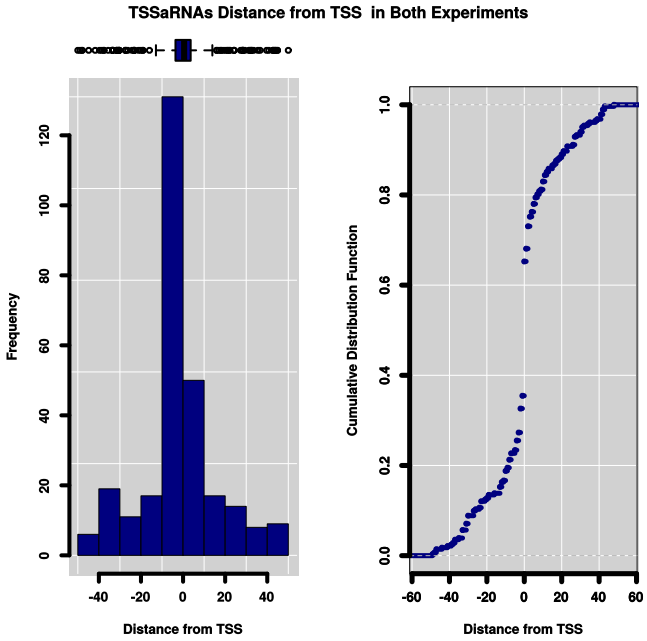
(d) Experiment II: Sample Id (SRR1187084).

Figure 3.8: The distribution of antisense TSSaRNAs' length size.

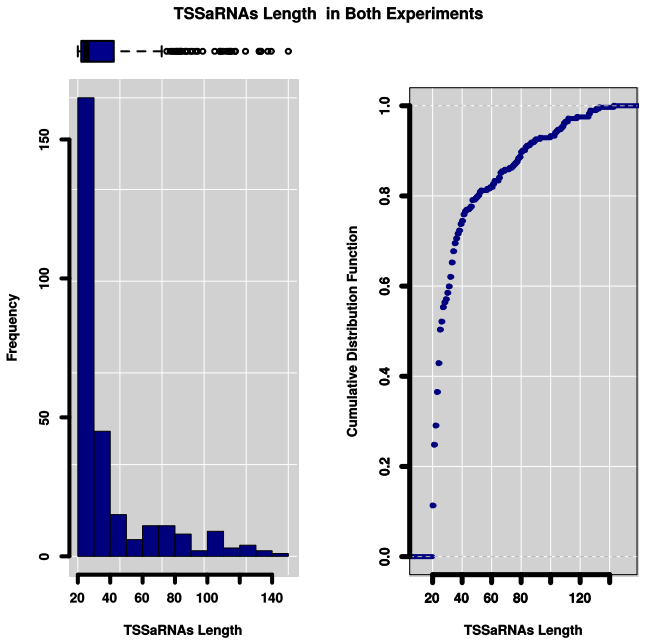
3.2.3 The final TSSaRNAs candidates (concordance TSSaRNAs)

A total of 224 sense TSSaRNAs and 58 antisense TSSaRNAs candidates have annotated as final TSSaRNAs from both experiments as TSSaRNA candidates. Final TSSaRNAs are the molecules that overlapped between all the samples in both experiments. The length of the final TSSaRNAs candidate showed a median size of 25 nucleotides (Fig.3.9 -b). In respect of the distance from the transcription start sites of cognate genes, the results showed that about 40% of total predicted TSSaRNAs initiated exactly from the same point of the transcript start site signal of their cognate genes (Fig.3.9 -a). Moreover, the results showed that there was no much coexistence of sense and antisense TSSaRNAs hence only eight cognate genes had associated with both sense and antisense TSSaRNAs. Regarding the number of TSSaRNAs per cognate gene, the results showed that all cognate genes had associated with only one TSSaRNAs in a particular orientation except the mentioned eight cognate genes that associated with both sense and antisense TSSaRNAs. Therefore, 10.45% of the 2622 annotated genes have TSSaRNA associated to them, 8.24% in the same orientation overlapping the gene, 1.91% antisense to them, and 0.31% associated to both sense and antisense TSSaRNA. For more detailed results of predicted TSSaRNAs refer to the Digital Appendix 1: Predicted TSSaRNAs.

Although, we used a very restrictive definition of TSSaRNA regarding to the count of TSSaRNAs within its cluster(#TSSaRNA> average of reads within the cluster); when comparing the obtained result with the 652 TSSaRNAs which published by Zaramel et al. [46] there were 196 (87.5%) of the sense TSSaRNAs in concordance. Touching on to the comparison of antisense TSSaRNAs, there was no need to mention the overlapping between the antisense TSSaRNAs with the previous data. With reference to our best knowledge, the antisense TSSaRNAs had not predicted in Archaea by any previous studies. The results showed that there was no bias regarding the orientation of the cognate genes as there were 139 cognate genes on the (+) strand while 143 cognate genes were on the (-) strand.



(a) Final TSSaRNAs candidates from the both experiments.



(b) TSSAaRNAs candidates Length size distribution.

Figure 3.9: TSSaRNAs concordance
(a) The distribution of distances from TSS (b) The distribution of TSSaRNAs' length sizes.

3.3 TSSaRNAs structures

3.3.1 Predicting TSSaRNAs secondary structures

Using the Mfold (v3.6) tool, we succeeded to predict the optimal secondary structures for all 224 sense TSSaRNAs and 58 antisense TSSaRNAs (Fig. 3.12). To understand the secondary structure features encoded by TSSaRNAs we represented each secondary structures as a vector of numerical values. The values of this vectors include a) the energy values such as the overall free energy, energy in helices, energy in hairpins, etc.; b) the count of occurrence of each secondary elements per TSSaRNA. The detailed results were as follows.

3.3.1.1 TSSaRNAs Thermodynamic Profiling

We have investigated the stability of TSSaRNAs secondary structures in the term of thermodynamic free energy using thermodynamic parameters of Mfold program. The calculated free energy expressed in kcal mole^{-1}. The terms of energy include the initial and the final free energy besides the energy that arose from each secondary structure element in TSSaRNAs. The next two paragraphs explain the obtained results based on the terms of thermodynamic energy.

Initial and Final Free Energy

The energy values that obtained for the initial free energy of TSSaRNAs secondary structures showed that 89.72% of TSSaRNA had associated with negative values ranged of (-60.6,-0.1) kcal mole^{-1}. The results showed that there were two TSSaRNAs which had associated with the initial free energy value of zero. Besides 27 TSSaRNAs associated with a positive initial free energy. The overall values of the initial free energy ranged of (-60.06,2.1) kcal mole^{-1} (Fig. 3.10 sub-figure a).

Considering the values of the final free energy, the results showed no variation with respect to the values of the initial free energy. The final free energy calculated by modeling the secondary structures in the given conditions. On the other hand, the values of the initial

free energy were calculated using the standard default parameters.

Thermodynamic stability of TSSaRNAs secondary structures elements

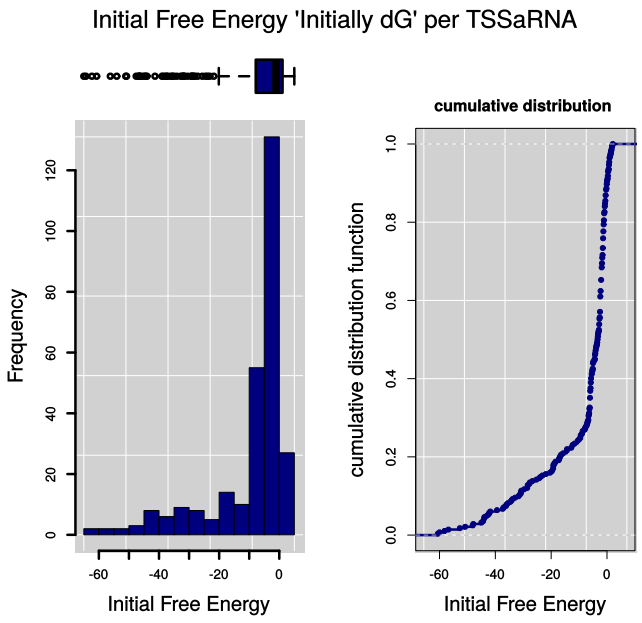
The energy associated with the stability of structural elements in the optimal secondary structures of TSSaRNAs has been extracted as free energy value expressed in kcal mole^{-1}. The results revealed that the values of the free energy that contribute to stabilizing the helices, stacks, and external loop motifs were mostly favorable free energy. For instance, in the case of helices and stacks element their free energy ranged of (-84.4,-1.4) kcal mole^{-1} (Fig. 3.10). Referring to the free energy that contributes in stabilizing external loop, the result showed that around 78.72%(222) of TSSaRNAs their external loops had associated with favorable free energy ranged of (-4.5,-0.1) kcal mole^{-1}. Differently, 9.22%(26) of TSSaRNAs their external loops associated with a free energy value of 0 kcal mole^{-1}. The rest of TSSaRNAs (12.06%(34)) their external loops had associated with positive free energy ranged of (0.1,0.6) kcal mole^{-1} (Fig. 3.11).

Regarding the free energy associated with the elements: hairpin, multi-loop, and bulge; the result showed that the values of their free energy were mostly positive (unfavorable free energy). For instance, in the case of the hairpin elements their free energy ranged of (1.7,17.6) kcal mole^{-1}. In the case of the multi-loop motifs, the results showed that most of them had associated with free energy ranged of (0,3.4) kcal mole^{-1} such as 92.20%(260) of TSSaRNAs their multi-loop motifs had associated with free energy value of 0 kcal mole^{-1}, and 7.09%(20) of TSSaRNAs their multi-loop motifs had associated with positive free energy in range of (0.3,3.4) kcal mole^{-1}. Moreover, the results showed that only two TSSaRNAs that their multi-loop motifs had associated with negative energy of -0.4 & -0.6 kcal mole^{-1}.

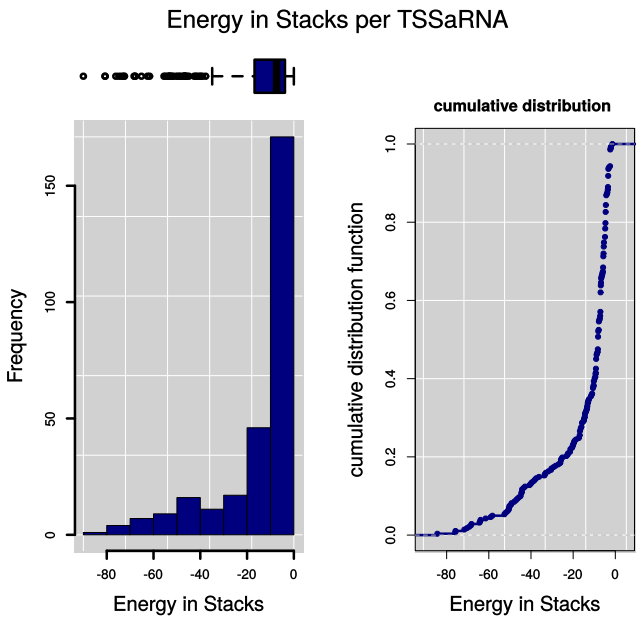
Considering the values of the free energy that associated to the bulge elements, the result showed that around 70.57% (199) of TSSaRNAs their bulge motifs had associated with a free energy value of 0 kcal mole^{-1}. The rest of TSSaRNAs (29.43%(83)) their bulge elements had associated with positive energy ranged of (0.4,9.8) kcal mole^{-1} (Fig. 3.11).

In summary, the result regarding TSSaRNAs thermodynamic energy profiling is as fol-

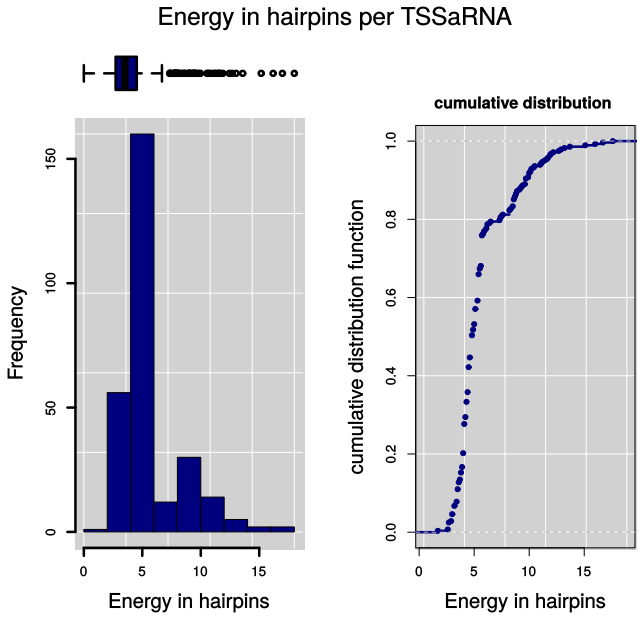
lows. The values of the free energy that had associated with the secondary structures of TSSaRNAs revealed that 89.72% of TSSaRNAs had associated with negative values of free energy. The remaining 10.28% of TSSaRNAs could be grouped into two groups: the first group consists of two TSSaRNAs (0.7%) that had associated with free energy value equal to zero, and the second group consists of 22 TSSaRNAs (9.57%) that had associated with positive values of free energy. The negative values of the free energy suggest that the secondary structures of TSSaRNAs are thermodynamically favorable structures, i.e., stable structures. The positive values of the free energy indicate that some TSSaRNAs exist naturally as linear molecules although, these TSSaRNAs could fold into stable secondary structures through a non-spontaneous process by utilizing external energy from cells. The free energy of zero implies that TSSaRNA secondary structures are at the equilibrium state, i.e., these TSSaRNA could exist naturally in the two forms (folded or linear). Regarding the energy associated with hairpins, the results showed positive values of the free energy because hairpins consist of uncomplimentary residues that contribute to a positive energy value. Also, the result revealed that the values of the free energy which had associated with multiloop and bulge elements were equal to zero because most of TSSaRNAs are lack of these elements. To see all the energies that had associated with TSSaRNAs secondary structures refer to the Digital Appendix 1: TSSaRNAs Mfold table.



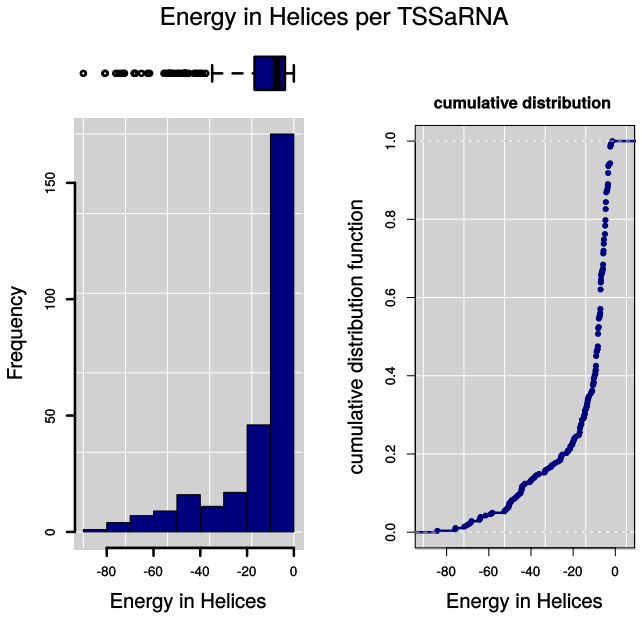
(a) Initial Free Energy.



(b) Free Energy in Stacks motifs.

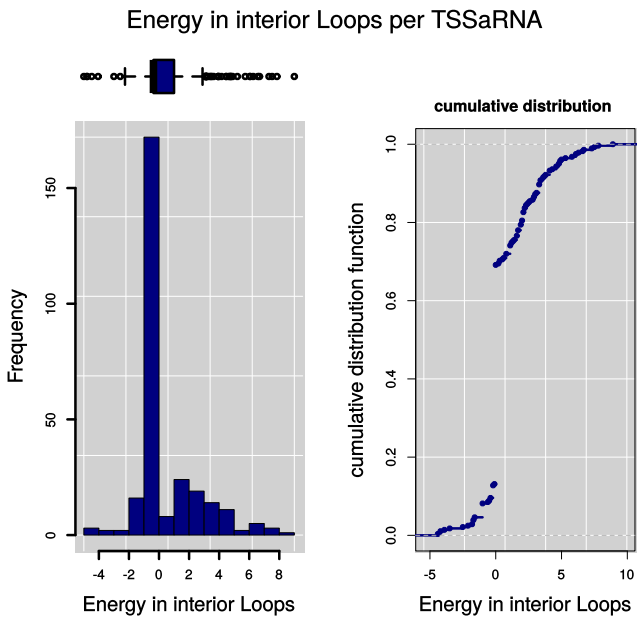


(c) Free Energy in Hairpins motifs.

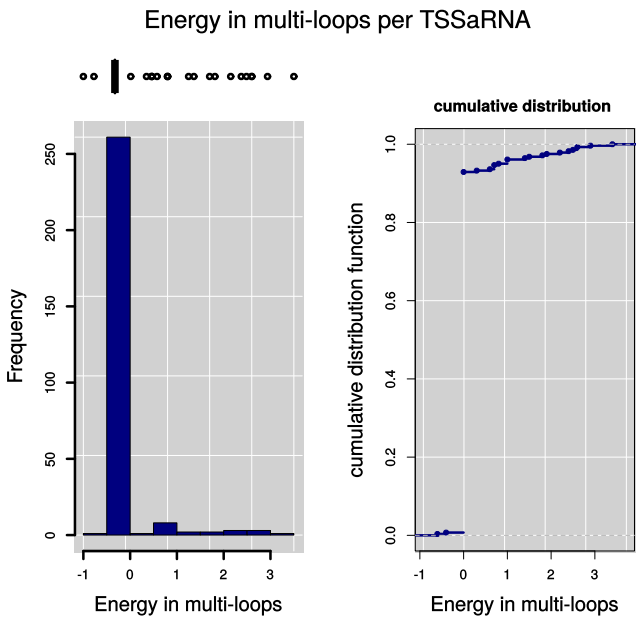


(d) Free Energy in Helices loops motifs.

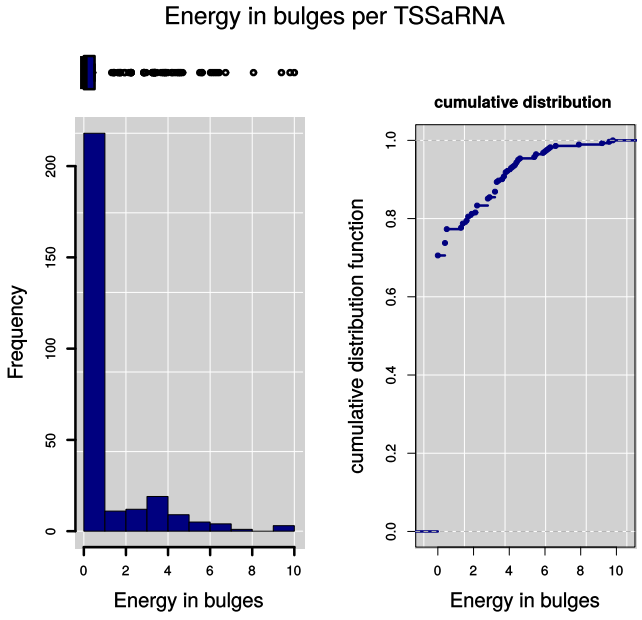
Figure 3.10: Free Energy in TSSaRNAs secondary structures I.



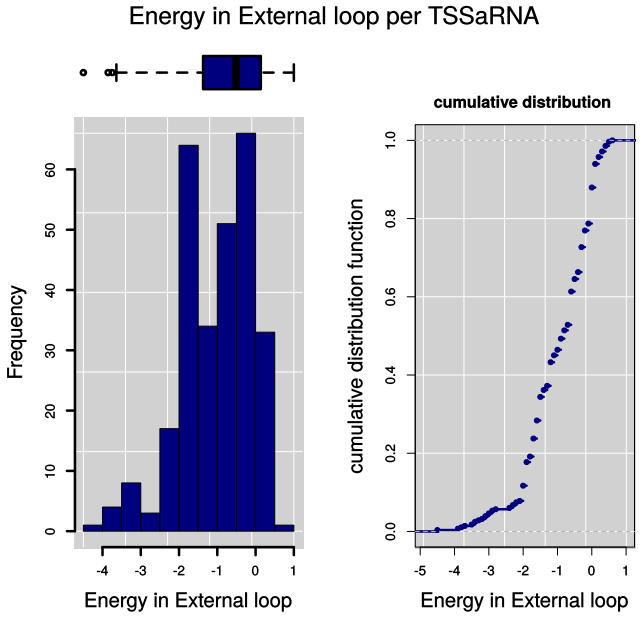
(a) Free Energy in Interior loops Motifs.



(b) Free Engeries in MultiLoops Motifs.



(c) Free Energy in Bulgs motifs.



(d) Free Energy in External loops motifs.

Figure 3.11: Free Energy in TSSaRNAs secondary structures II.

3.3.1.2 TSSaRNAs secondary structures topologies

To investigate the structural topology in TSSaRNAs that could be correlated to their potential functions, we have calculated the frequencies of the occurrence of each RNA secondary structure elements per TSSaRNA molecule. The secondary structures elements counted are: the single strand nucleotides in external loop, closing helices in external loop, internal loops, hairpin, and bulge loops. The overall results revealed that all TSSaRNAs possessed at least one of these elements: closing helix in external loop, helix element, hairpin, and stack element. The result showed that 95.39% (269) of TSSaRNAs possessed at least one single strand nucleotides in the external loop. The results showed the percentage of TSSaRNAs that lack any multi-branch loop, bulges, or interior loop elements in their optimal secondary structure were 92.20% (260), 70.57% (199), and 56.03% (158) respectively. The elements could be grouped into groups such as the elements of high occurrence frequency and the elements of low occurrence frequency. Each group of elements is discussed separately in the following.

a) Elements associated with high frequency of occurrences across all TSSaRNAs

The elements that occurred with high frequency were including single-strand nucleotides in external loop, closing helices in external loop, internal loops, hairpin, and stack. The frequencies of these elements were as follows: the occurrences of the single strand nucleotides in external loop ranged of (0-26) where only 13 of TSSaRNAs were free of any single strand nucleotide in their external loop, and 83.33%(235) of TSSaRNAs possessed 1-13 free single strand nucleotide in their external loop (Fig. 3.14 c). The result revealed that the existence of a negative relationship between the number of free nucleotides in the external loop and the number of TSSaRNAs associated with it.

In respect to the closing helices in the external loop, the results showed their occurrences ranged of 1-4. In details, 82.98%(234) of TSSaRNAs possessed only one closing helices in their external loop, 14.18%(40) of TSSaRNAs possessed two closing helices in external loop, 2.48%(7) of TSSaRNAs possessed three closing helices in external loop, and only one TSSaRNA was associated with four closing helices in the external loop (Fig. 3.14 d).

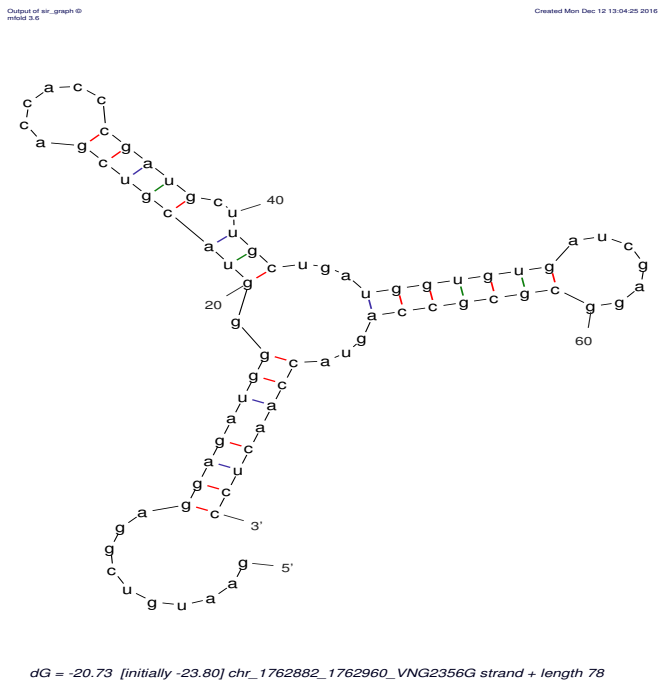
In respect to the distribution of the helices per TSSaRNA, the result showed the distribution of the helices ranged of 1-13. In details, 41.49%(117) of TSSaRNAs possessed one helix loop, 27.30%(77) of TSSaRNAs possessed two helices structure, 9.93%(28) of TSSaRNAs possessed three helices, 6.03%(17) of TSSaRNAs possessed four helices, 4.61%(13) of TSSaRNAs possessed five helices, 2.84 %(8) of TSSaRNAs possessed either (six, seven, or eight) helices, two TSSaRNAs possessed 7 helices, and only one TSSaRNA associated with either (nine, ten, twelve, or thirteen) helices elements (Fig. 3.13 c).
The occurrences of the hairpin element per TSSaRNAs was in the range of 1-4. In details, 77.30%(218) of TSSaRNAs possessed one hairpin structure, 17.38%(49) of TSSaRNAs possessed two hairpins, 4.61%(13) of TSSaRNAs possessed three hairpins elements, and only two TSSaRNAs possessed four hairpins secondary structure elements (Fig. 3.13 a).
The distribution of the stack elements among TSSaRNAs was varying between (1-33). In details, 81.21%(229) of TSSaRNAs possessed stacks in range of (1-10), 11.34%(32) of TSSaRNAs possessed stacks in range of (11-20), and 7.45%(21) of TSSaRNAs were associated with more than 21 stack elements (Fig. 3.13 d).

b) Elements with the low frequency of occurrences among TSSaRNAs

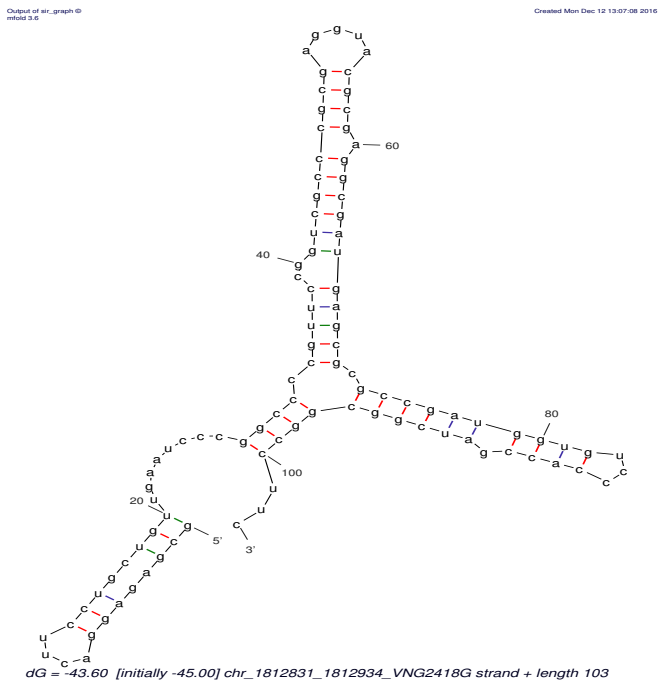
The result showed that some elements rarely occurred among the secondary structures of TSSaRNAs. These elements are the multi-branch loop, bulge, and interior loop. For instance, in case of the multi-branch loop, there was only 7.45 %(21) of TSSaRNAs that possessed one multi-branch loop besides one TSSaRNA that possessed two multi-branch loops. The rest of TSSaRNAs (260) was lack of any multi-branch loop in their secondary structures.
In respect to the occurrence of the bulges among TSSaRNAs, the result showed the frequency of the bulges was varies between 0-4. In details, 17.73%(50) of TSSaRNAs possessed one bulge, 7.45 % (21) of TSSaRNAs possessed two bulges, 2.84%(8) of TSSaRNAs possessed three bulges, and only 1.42%(4) of TSSaRNAs possessed four bulges elements (Fig. 3.13 a). The rest of TSSaRNAs (199) were lack of bulge element in their secondary structures. Regarding the interior loop, the results showed that the count of the interior

loop per TSSaRNA ranged of 0-8. In details, 56.03%(158) of TSSaRNAs were without any interior loop, 29.08%(82) of TSSaRNAs possessed one interior loop, 7.09%(20) of TSSaRNAs possessed two interior loops, 2.84%(8) of TSSaRNAs possessed three interior loops, 2.48%(7) of TSSaRNAs possessed four interior loops, 1.77%(5) of TSSaRNAs maintained five interior loops, and only one TSSaRNA possessed either six or eight interior loops, however, no any TSSaRNA associated with a seven interior loops (Fig. 3.14 a).

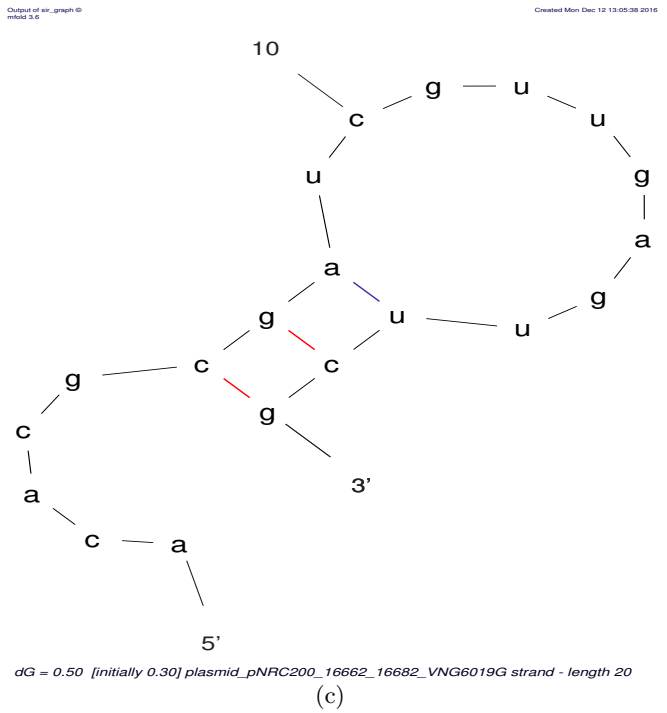
In summary, We could consider TSSaRNAs as versatile molecules that their secondary structures are composed of a combination of various elements (hairpins, bulges, stack, etc.). These elements had associated with different frequencies. The frequencies of all elements in TSSaRNAs secondary structures are provided in the Digital Appendix 1: TSSaRNAs mfold table.



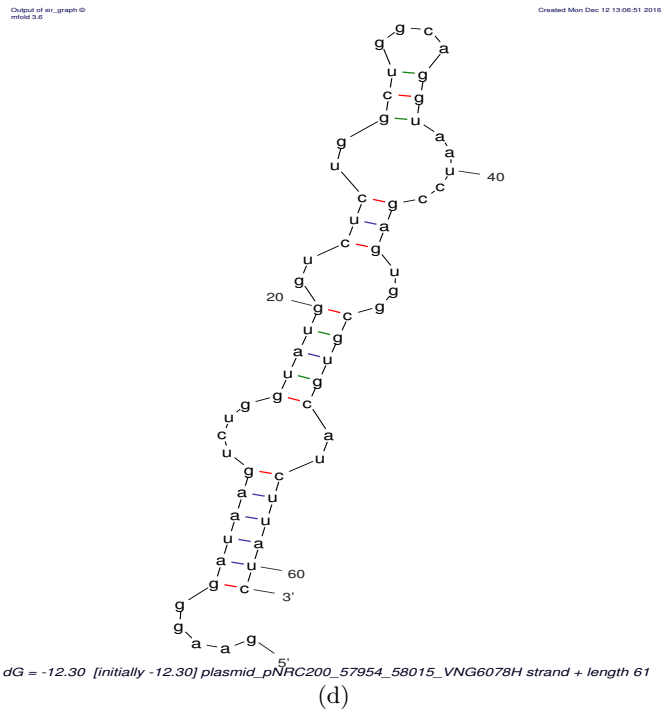
(a)



(b)

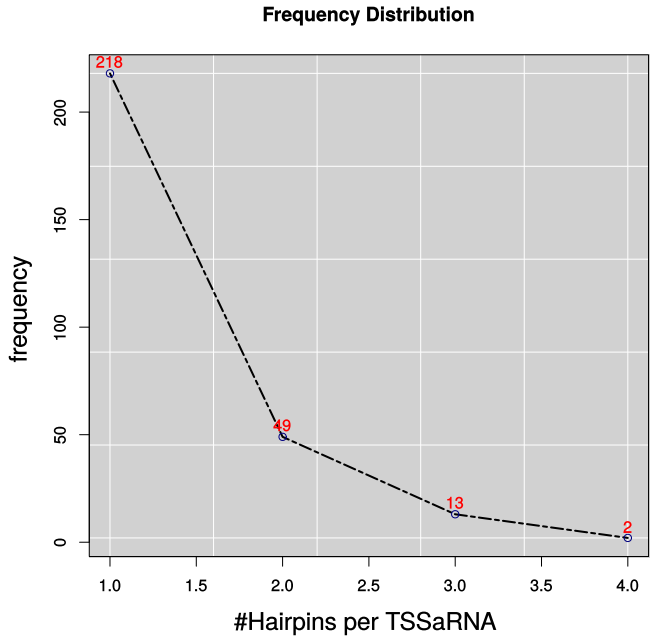


(c)

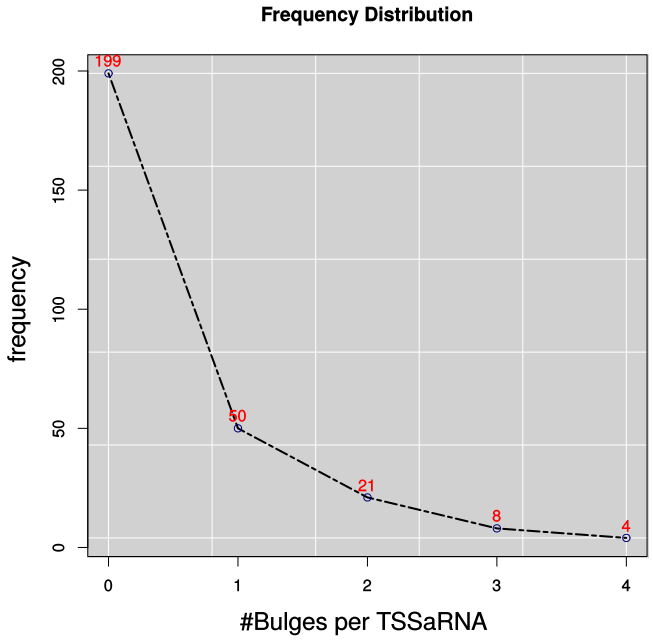


(d)

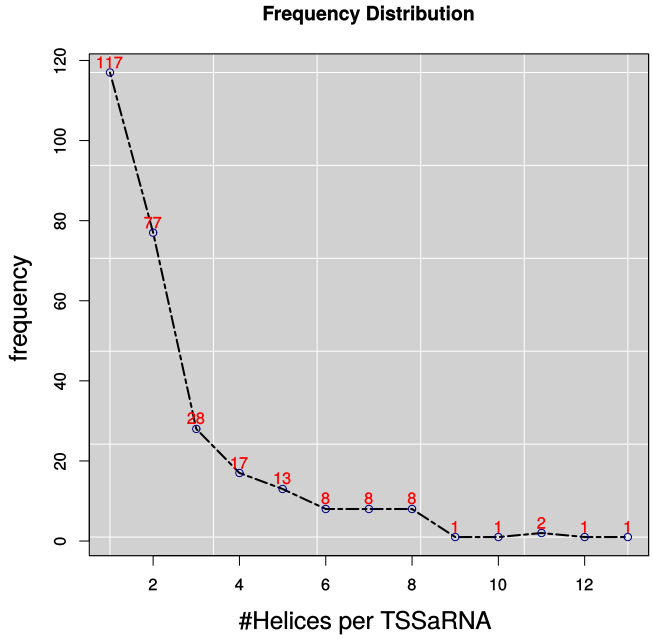
Figure 3.12: Example of predicted TSSaRNAs secondary structures topologies.
To see all the secondary structures topologies of TSSaRNAs refer to the Digital Appendix 1: TSSaRNAs 2D topologies. The colored bars between the paired nucleotides give more annotation of RNA secondary structure. The color annotation is ranged from the red color to indicate the existence of extreme well-determined base pairing to the black color to indicate poor base pairing. For more details of the Mfold color annotation refer to the link `http://unafold.rna.albany.edu/ref/s-annotate/node3.php`



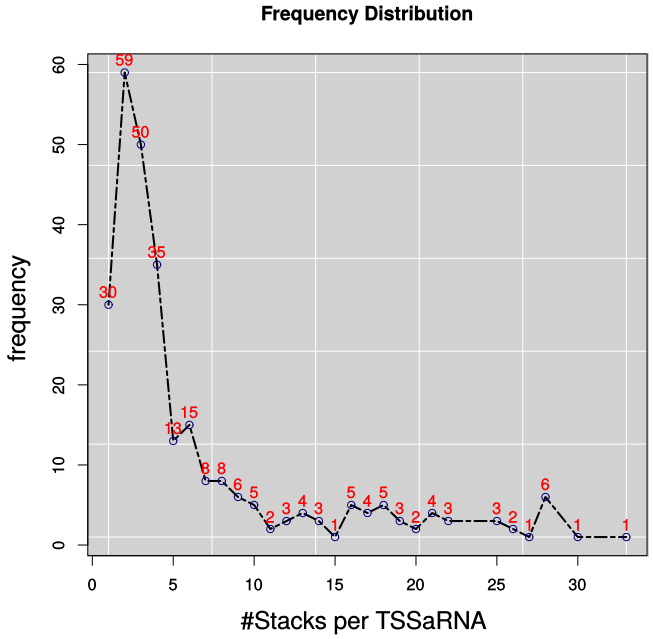
(a) Number of Hairpins Per TSSaRNA.



(b) Number of Bulges Per TSSaRNA.

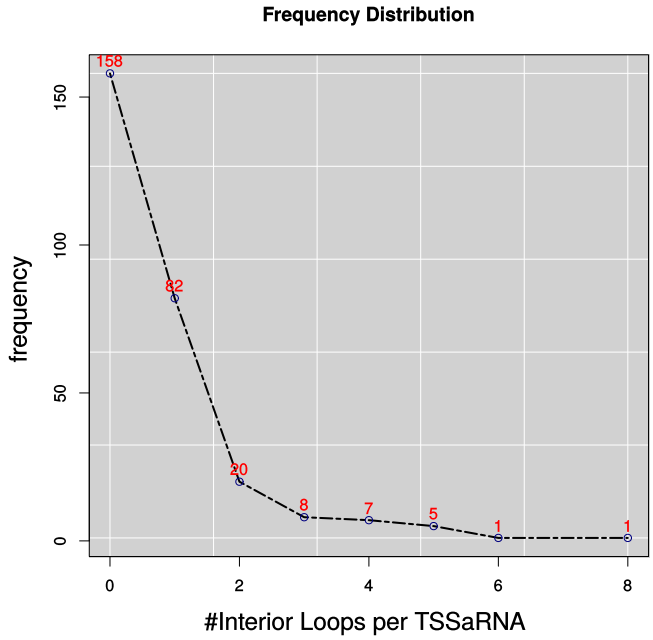


(c) Number of Helices Per TSSaRNA.

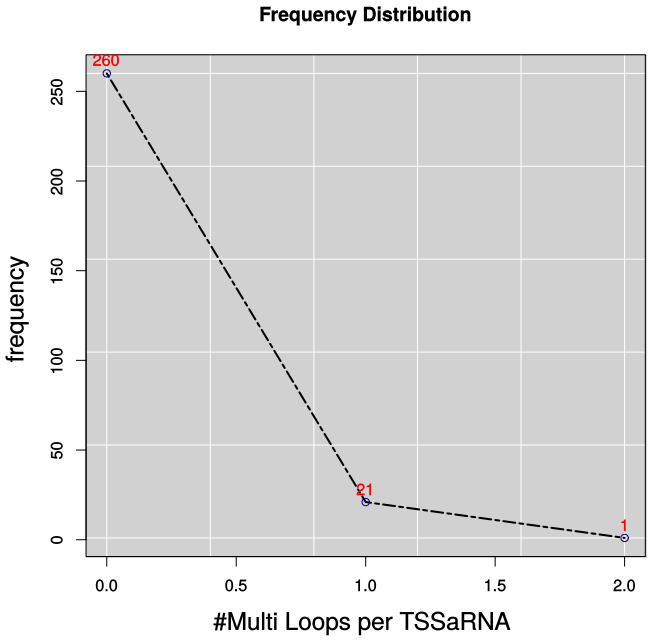


(d) Number of Stacks Per TSSaRNA.

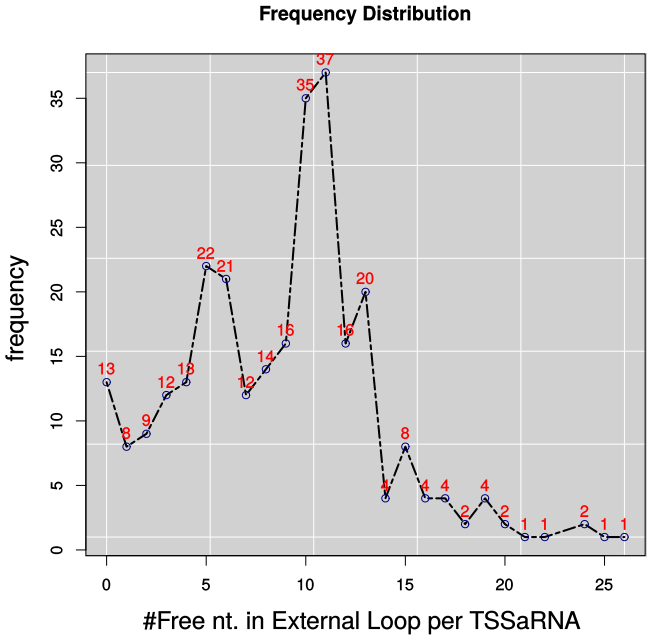
Figure 3.13: Distribution of TSSaRNAs secondary structure motifs I.



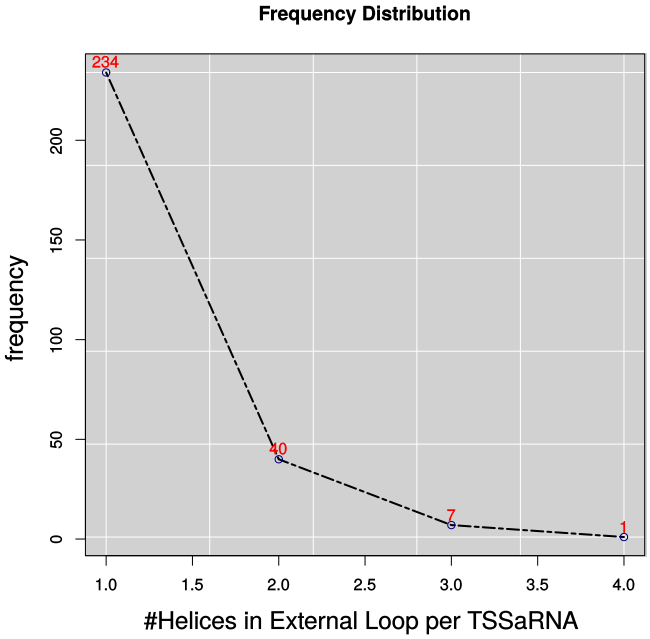
(a) Numbers of Interior loops TSSaRNA.



(b) Numbers of MultiLoops per TSSaRNA.



(c) Numbers of Free nucleotides in external Ends.



(d) Numbers of Helices in external Ends.

Figure 3.14: Distribution of TSSaRNAs secondary structure motifs II.

3.3.2 Predicting TSSaRNAs tertiary structures

By using Rosetta-Commons RNA tools, we succeeded to model the tertiary structures for all TSSaRNAs. Rosetta algorithm minimized the thermodynamic energy using Monte Carlo sampling process. We used the secondary structures as constraints to model the tertiary structures (Fig. 3.16 & 3.17). The overall weighted free energy for all TSSaRNAs showed highly favorable free energy except two TSSaRNAs which had associated with positive free energy. The minimum free energy was -702.379 in the Rosetta energy unit (REU). The energy median value for overall weighted energy after excluding the two positive outliers was -116.90 REU with an average of -181.40 Rosetta energy unit (Fig. 3.15).

Alongside the value of the overall weighted free energy of TSSaRNAs tertiary structures, we have paid more attention to some specific energy terms. These terms contribute significantly in the stabilization of the RNA tertiary structures or could participate in triggering some critical biological function. The energy terms with the particular interest include but are not limited to: (a) radius of gyration; (b) Van der Waals interactions;(c) Lennard-Jones; (d) stacking energy; (e) hydrogen bond; (f) solvation energy; (g) electrostatic repulsion between phosphates. The detailed result of these essential energy terms has presented in the following.

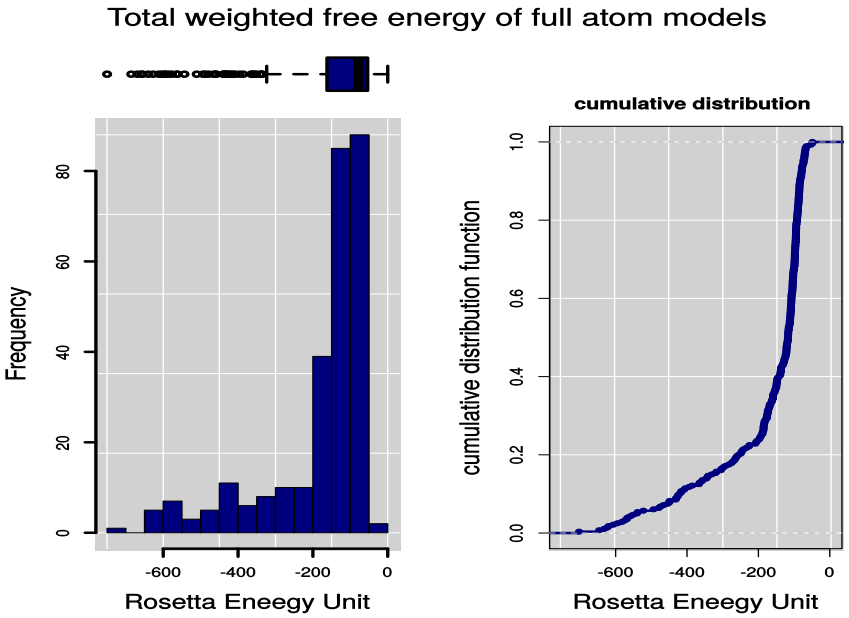
Touching on the radius of gyration term, the results showed a median energy value of 12.89 with an average of 16.19 in Rosetta energy unit. In the case of the Van der Waals interactions energy, the result showed a median energy value of 12.68 with a mean of 15.67 in Rosetta energy unit. On the other hand, the median value of the free energy that arose from the Lennard-Jones the attractive part was -60.02 with an average of -92.04 Rosetta energy unit.

Regarding the stacking energy which contribute to the thermal stability of RNA tertiary structure, the result showed favorable negatives energies of a median of -7.576 and average of -10.220 Rosetta energy unit. It was also noticeable that in the case of the hydrogen bond all the hydrogen bond components in the predicted TSSaRNAs tertiary structures

were associated with favorable negative values.

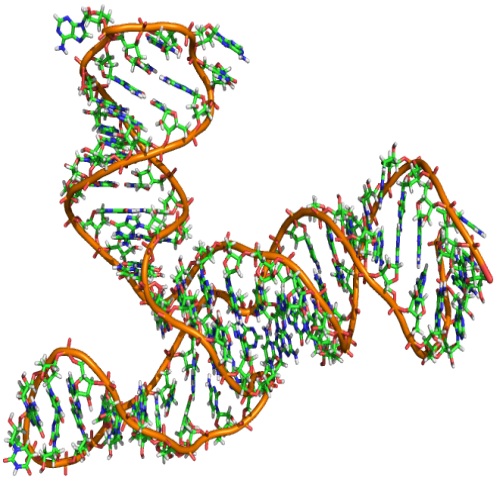
In respect of the solvation energy which helps to interpret the stability of biomolecules interaction, the result showed energy values of a median of 3.710 and average of 5.073 in Rosetta energy unit for the non-polar contribution of the solvation energy. On the other hand, the polar contribution of the solvation energy showed median energy of 58.60 with an average of 92.69 in Rosetta energy unit. About the electrostatic repulsion between phosphates, the results showed a weak electrostatic repulsion between phosphates of a median of 0.7865 and average of 0.9857 in Rosetta energy unit.

In summary, the values of the overall weighted free energy which had associated with the tertiary structures of TSSaRNAs showed highly negative values. These results imply that TSSaRNAs tertiary structures are thermodynamically stable structures.

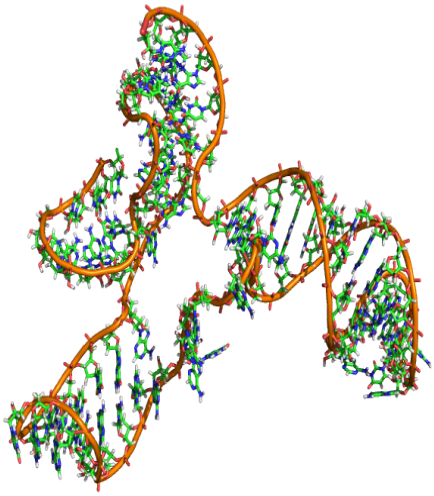


(a) Total free energy in TSSaRNAs tertiary structures

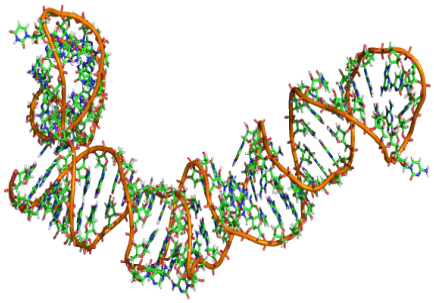
Figure 3.15: The overall free energy after excluding the two positive outliers. The values of energy are in Rosetta energy unit



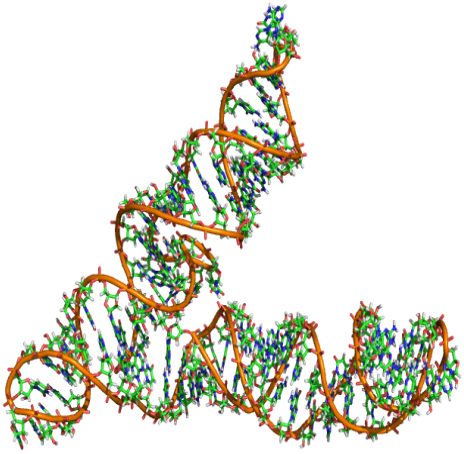
(a) chr_255575_255653_VNG0324G.



(b) chr_860255_860335_VNG1132G.

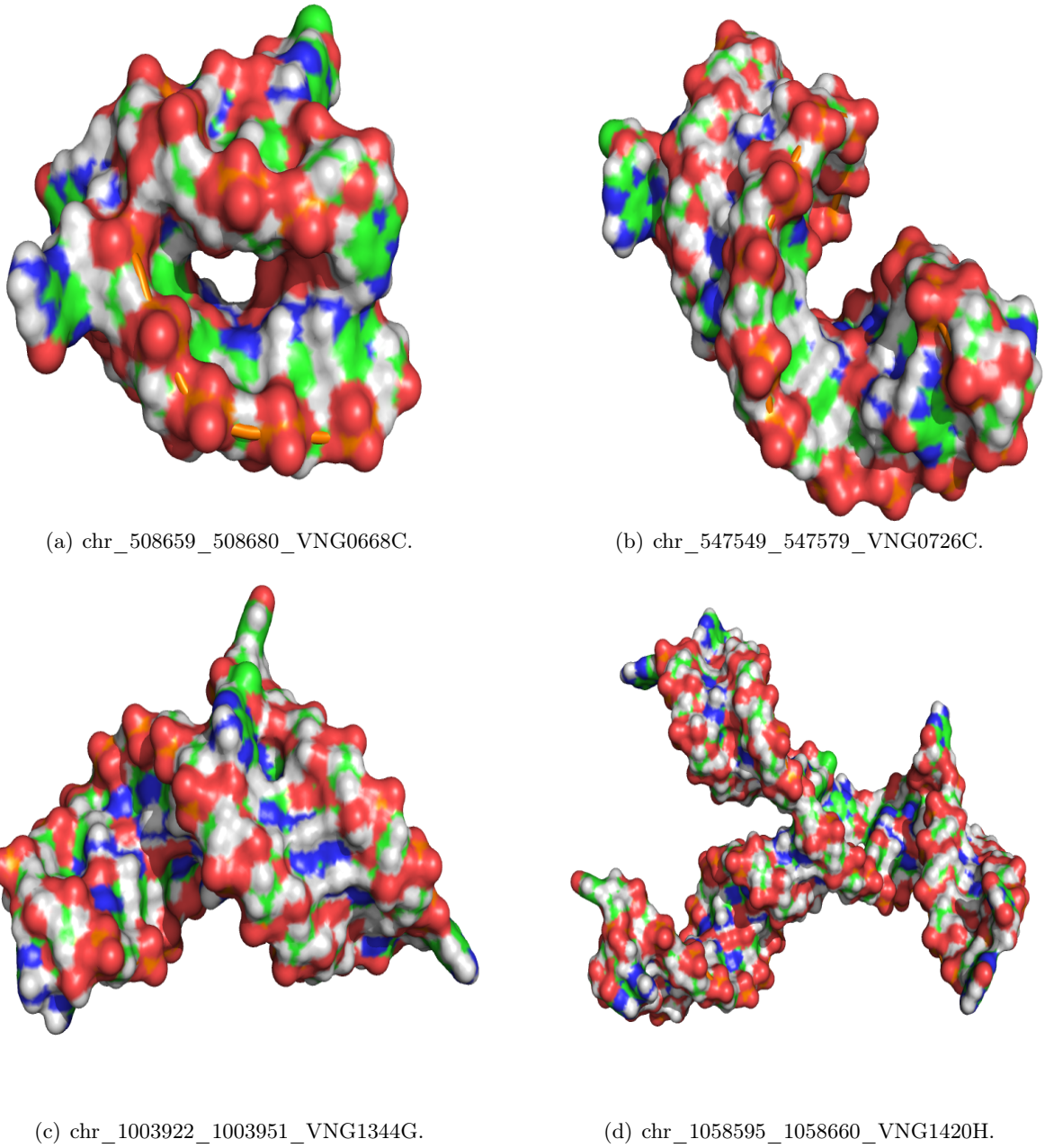


(c) chr_1977576_1977676_VNG2639G.



(d) plasmid_pNRC100_31821_31926_VNG7032.

Figure 3.16: Example of TSSaRNAs tertiary structures topologies.
To see all the tertiary structures of TSSaRNAs in cartoon representation refer to the Digital Appendix 1: TSSaRNAs 3D cartoon representation



(a) chr_508659_508680_VNG0668C.

(b) chr_547549_547579_VNG0726C.

(c) chr_1003922_1003951_VNG1344G.

(d) chr_1058595_1058660_VNG1420H.

Figure 3.17: Example of TSSaRNAs tertiary structures topologies (Surface View). To see all the tertiary structures of TSSaRNAs in surface representation refer to the Digital Appendix 1: TSSaRNAs 3D surface representation.

3.3.3 Predicting TSSaRNAs higher-order structures

3.3.3.1 TSSaRNA cognate gene hybridization

The approach to investigate the interaction of TSSaRNAs with its corresponding mRNA using IntaRNA tool has revealed that almost all TSSaRNAs (92.2%) were capable of hybridizing into their corresponding cognate genes except 22 of TSSaRNAs.The values of the TSSaRNAs-mRNA interaction energy showed a median of -9.22 kcal mole^{-1} and an average of -10.06 kcal mole^{-1} (Fig. 3.18).

Considering the mode of interaction between TSSaRNAs and their corresponding cognate genes, we have defined three modes of interactions depending on the interaction location: (1) 5 prime interactions, when a TSSaRNA hybridizes at the first one third of the cognate gene; (2) an embedded mode of interaction when TSSaRNA hybridize at the second one third of the cognate gene; and (3) the final mode of interaction as a 3 prime mode of interaction when TSSaRNA hybridize at the third one third of the cognate gene. The result obtained from investigating the mode of interaction showed that around 132 (50.77%) of the interacting TSSaRNAs hybridized on the of their corresponding cognate genes while 128 (49.23%) 5' of the interacting TSSaRNAs hybridized on the 3' of their corresponding cognate genes (Fig.3.19). The result of TSSaRNA-cognate gene hybridizations showed the existence of some interesting functional motifs such as (AUCA) sequence motif which is highly conserved among the three domain of life [113] and (AUGA) sequence motif which regulates the translation process [114]. For more detailed results of TSSaRNAs-Cognate genes, interactions refer to Digital Appendix 1: Table of TSSaRNA Cognate genes interactions.

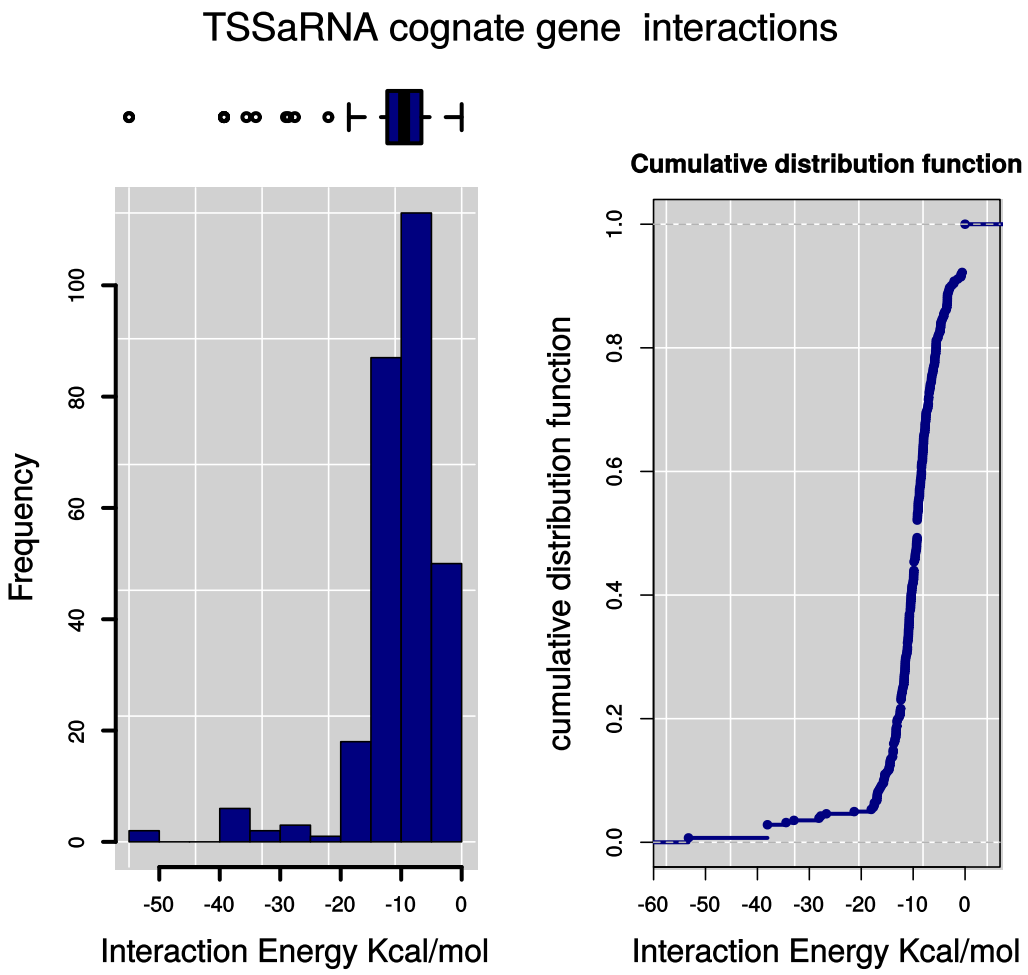
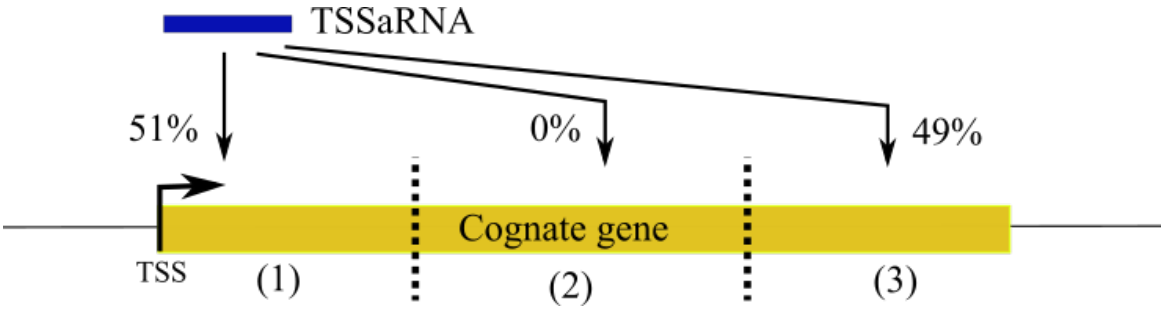


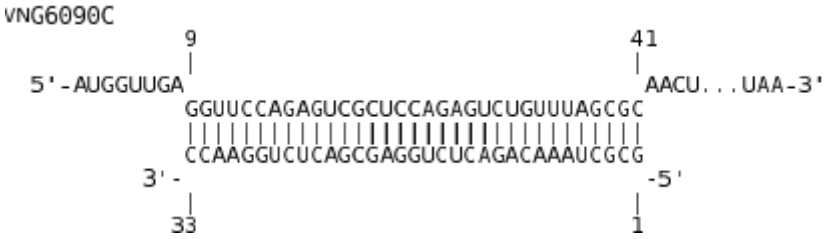
Figure 3.18: TSSaRNA-Cognate Genes Interactions.

Figure shows the potential interaction between TSSaRNA with cognate genes. Free Energy is in units of kcal $mole^{-1}$

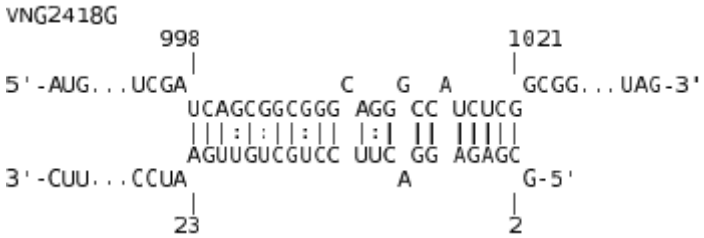
.



(a) TSSaRNA-Cognate gene mode of interactions.



(b) 5 Prime Mode of interactions.



(c) 3 Prime Mode of interactions.

Figure 3.19: Putative TSSaRNA-cognate gene hybridization.
Figure shows putative TSSaRNA-cognate gene interactions. Sub-figure (a) demonstrates TSSaRNA-cognate gene mode of hybridization where (1) indicates 5 prime interaction, (2) an embedded mode of interaction, and (3) 3 prime mode of interaction. Sub-figure (b) shows 5 prime mode of interactions. Sub-figure (c) shows 5 prime mode of interactions. To see all topologies of TSSaRNA-Cognate gene hybridization refer to the Digital Appendix 1: TSSaRNA Cognate genes interactions.

3.3.3.2 TSSaRNA Lsm binding

The approach to investigate the interaction of TSSaRNAs with the Lsm protein by mining the experimental data of Lsm-RNA interaction has revealed the existence of putative interactions between Lsm protein and TSSaRNAs molecules. The raw data generated by the scientist in the LaBiSisMi[1]. We have categorized the putative interactions into three types of interactions depending on the possibility of Lsm to interact with TSSaRNAs and the potential co-regulation of TSSaRNAs and their cognate genes by Lsm protein (Fig. 3.20). The three types of interactions are: (I) Lsm signals are more likely to indicate interaction of Lsm with TSSaRNA. We have considered Lsm to be more likely to indicate interaction of Lsm with TSSaRNA when there is a single Lsm signal across the cognate genes. This unique signal should cover the whole genomic coordinates of TSSaRNAs. Alternatively, the Lsm signal should be covered by the genomic coordinates of TSSaRNA. The first option of this type of interaction will show that the Lsm signal is extended to interact with the cognate genes. However, we have expected that Lsm may have the capability to interact with all RNAs that expressed within the interaction signal including TSSaRNAs. As there is a single Lsm signal that might interact with both TSSaRNA and its cognate gene, therefore, we have expected that in this type of interaction Lsm might be required to regulate both TSSaRNA and its cognate gene. (II) Lsm interacts with less degree of specificity with the genomic coordinates of TSSaRNAs. We have considered the interaction type (II) when there are more than one Lsm signal across the cognate genes. Also, we have considered the interaction type (II) when the interaction between Lsm and the genomic coordinates of TSSaRNAs is not a complete interaction, i.e., there is an overlapped interaction between Lsm signal and the genomic coordinates of TSSaRNAs rather than coverage interaction. The existence of more than one Lsm signal across the cognate genes adds a layer of regulatory complexity. Therefore, in this type of interaction, we have expected that Lsm might co-regulate TSSaRNA and its cognate gene with less degree of specificity as the other interactions might regulate the cognate genes as well. (III) Lsm is less likely to interact with TSSaRNAs. We have

[1] http://labisismi.fmrp.usp.br/index.php/en/

considered this type of interaction when the Lsm signal is leaked out to interact with a large part of the cognate genes. In this type of interaction, we have expected Lsm to interact mainly with the cognate genes rather than TSSaRNAs.

The result indicated that Lsm interacted with the peaks of 26 TSSaRNAs molecules as an interaction of type I (Fig. 3.22 a). Also, the results indicated that Lsm protein interacted will less degree of specificity (overlapped) with 12 TSSaRNAs molecules. Referring to the rest of interacted TSSaRNAs (17), the result indicated that Lsm interacted non specifically where Lsm protein leaked out to interact with a large part of the cognate genes, and even in some cases, Lsm interacted with the whole cognate gene. More detailed results of TSSaRNA-LSm interactions is provided in Digital Appendix 1: TSSaRNAs-Lsm table.

The result obtained from the computational docking experiment suggested that most of TSSaRNAs bind into the middle of the ring of Lsm structure (Fig. 3.21 a & b). Furthermore, we succeeded to calculate the binding energy of 47 (85.45%) of TSSaRNAs based on the APBS (Adaptive Poisson-Boltzmann Solver). The result of binding energies showed high stable thermodynamic energies with a median of -542900 kcal $mole^{-1}$ and mean of -575800 kcal $mole^{-1}$ (Fig. 3.22 b). The binding energy of the remaining 8 TSSaRNAs requires a high grid dimension for calculation due to the size of TSSaRNAs.

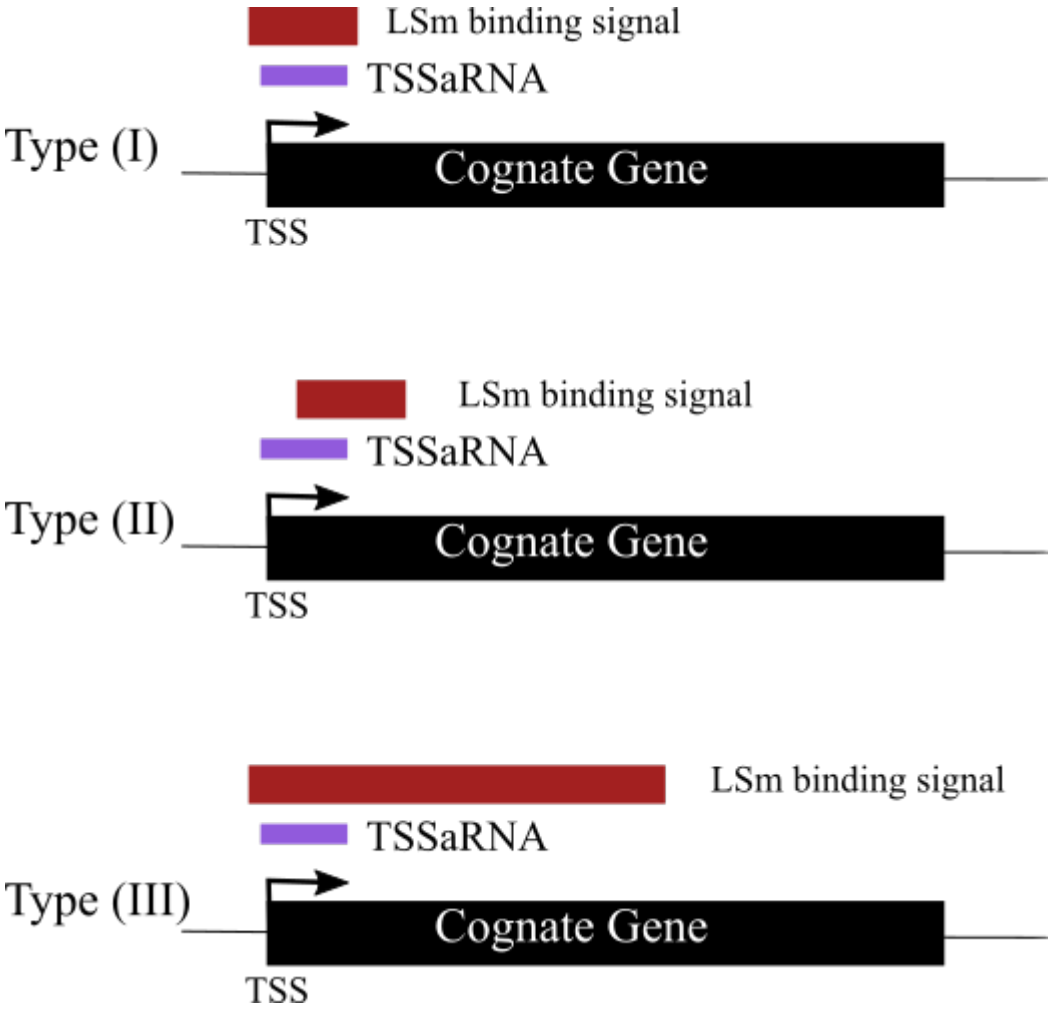
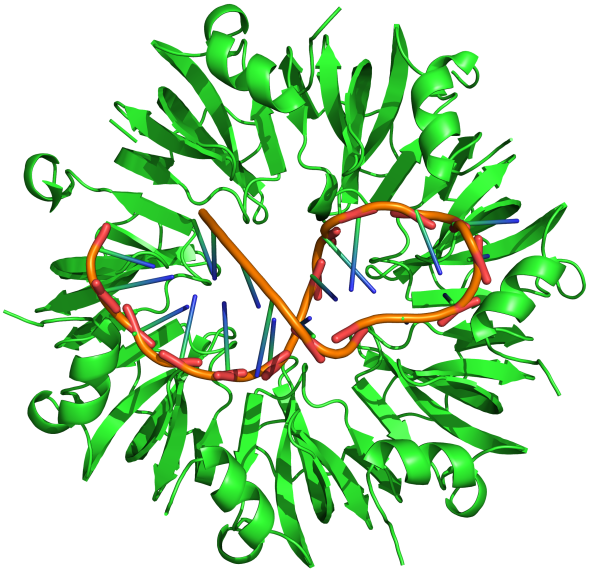


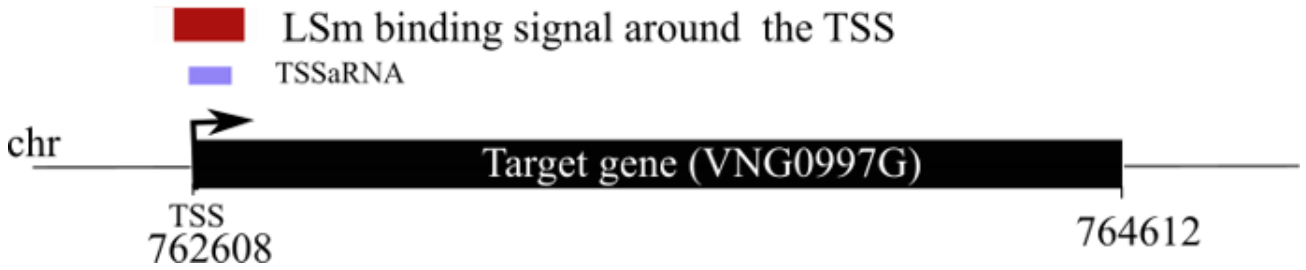
Figure 3.20: A figure demonstrates the three types of TSSaRNAs-Lsm interactions . The three types of the interactions are: (I) Lsm signal is more likely to indicate interaction of Lsm with TSSaRNA, (II) Lsm interacts with less degree of specificity with the genomic coordinates of TSSaRNAs (Overlapped interactions or wthere are more than one Lsm signal), (III) Lsm is less likely to interact with TSSaRNA (Lsm signal is leaked out to interact with a large part of the cognate gene).



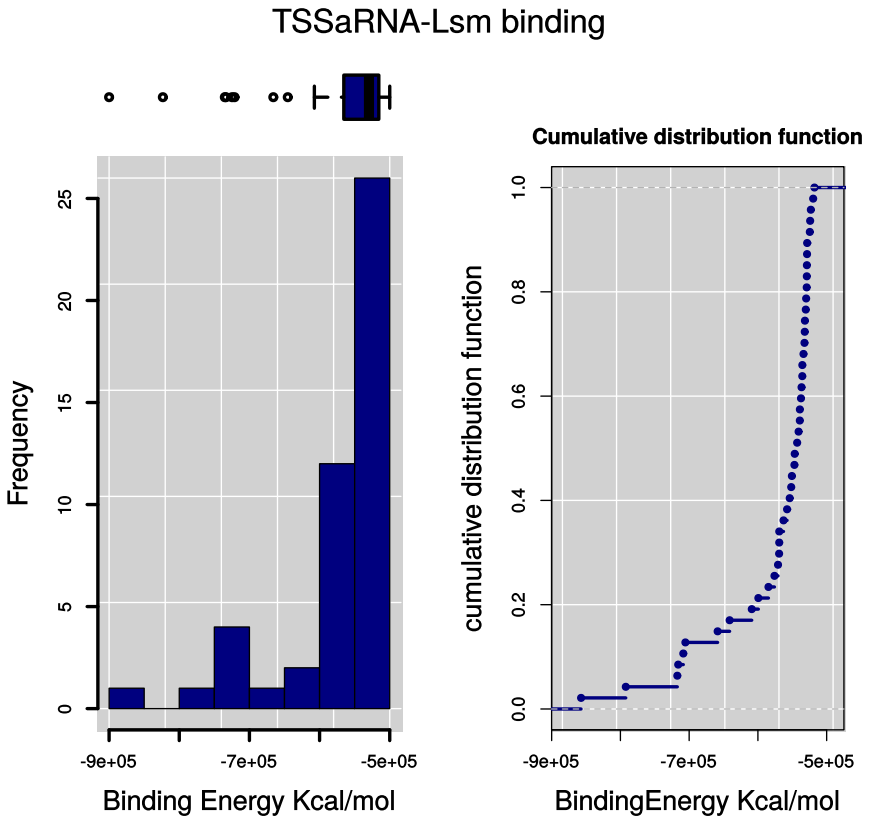
(a) TSSaRNA-Lsm interaction example 1.

(b) TSSaRNA-Lsm interaction example 2.

Figure 3.21: A figure demonstrates the result of potential TSSaRNAs-Lsm interactions. This figure demonstrates interested result of the docking experiment where two TSSaRNAs fit into the pocket at LSm protein, i.e. TSSaRNAs bind to the middle of Lsm ring structure.



(a) Potential interaction between Lsm and TSSaRNA



(b) TSSaRNA-Lsm binding energy.

Figure 3.22: A figure demonstrates the result of potential TSSaRNAs-Lsm interactions. Sub-figure (a) shows Lsm protein - the red color bar- binds into TSSaRNA region without leaking out into other regions on the gene.
Sub-figure (b) demonstrates a histogram of Lsm-TSSaRNAs bindings energies calculated using Apbs program. The result shows high stable binding energies.

3.4 TSSaRNAs annotation based on Rfam classifications

Our pipeline to annotate TSSaRNAs to the corresponding Rfam families succeeded to assign 147 (52.12%) of TSSaRNAs to their potential Rfam families. This classification is done either based on the approach of single molecule information or based on the consensus secondary structures information. Below is the detailed results.

3.4.1 Annotations of TSSaRNAs based on single molecule information

The first approach to annotate TSSaRNAs based on individual molecule information by querying TSSaRNAs sequences against covariance models of the Rfam families has succeeded to annotate 137 of TSSaRNAs. The orientation of annotated TSSaRNAs was as follows: 113 TSSaRNA molecules were sense TSSaRNAs and 24 molecules were antisense TSSaRNAs (Fig. 3.23 & Fig. 3.24). The primary result showed that a single TSSaRNA could be capable of hitting multiple Rfam families, for instance, TSSaRNA could hit multiple Rfam families that belong to the same ncRNAs class such as a particular TSSaRNA hit multiple families belong to the sRNA class. In some cases, TSSaRNA could hit more than one ncRNAs class such as a specific TSSaRNA hit a family of sRNA and another family of cis-regulatory. These classes could be functionally closely related classes or even in more complex case these classes will not be functionally associated with each other. To overcome the challenge that arose from the possibility of the multiple annotations of a single TSSaRNA molecule, we have annotated TSSaRNA to only one Rfam family that associated with the lowest E-value. The final Rfam annotation of TSSaRNAs is provided in Digital Appendix 1: Rfam annotation -only the minimum E-value-. However, to see all possible Rfam annotation of each annotated TSSaRNA refer to the Digital Appendix 1: Rfam annotation -all the possible E-values-.

Regarding the Rfam classes that encoded by TSSaRNAs, the results showed that more than 35% of annotated TSSaRNAs had annotated to the sRNA class where about 38 of sense TSSaRNAs and 11 of antisense TSSaRNAs have been identified as potential sRNAs. The second major Rfam class that encoded by TSSaRNAs was the cis-reg (cis-regulatory

elements) where 37 of sense TSSaRNAs and 8 of antisense TSSaRNAs had identified as putative cis-reg elements. Also, there were a considerable number of TSSaRNAs which have been annotated into putative snoRNAs particularly the sense TSSaRNAs, where 14 TSSaRNAs have been annotated as putative CD-box snoRNAs and seven sequences, have been annotated as putative HACA-box TSSaRNAs.

Furthermore, the results showed that six sense TSSaRNAs and one antisense TSSaRNA had annotated as potential CRISPR molecules (Table 3.3). These 7 TSSaRNAs have been annotated into distinct four CRISPR families. These CRISPR families are CRISPR-DR13, CRISPR-DR23, CRISPR-DR29, and CRISPR-DR36. We performed a local alignment to characterize further TSSaRNAs that had annotated as putative CRISPR molecules. To perform the local alignment, we retrieved all the seeds sequences of the four CRISPR families such as four seeds of CRISPR-DR13, three seeds of CRISPR-DR23, two seeds of CRISPR-DR29, and the only one seed sequence of CRISPR-DR36. The local alignment showed the existence of a minimum of 11 conserved nucleotides between TSSaRNAs sequences and CRISPR seeds. Moreover, the local alignment showed the existence of conserved secondary structures between CRISPR and 6 TSSaRNAs, i.e., except for one TSSaRNA. (Fig. 3.26). The TSSaRNA which had associated with the gene VNG1006 was the only one putative CRISPR that was without any conversed secondary structure (Fig 3.25). We have observed the existence of all alignment possibilities such as insertion, deletion and the substitutions on the result of the local alignment experiments (Fig 3.25, and Fig. 3.26).

The overall results of the Rfam annotation revealed the existence of some other Rfam classes with low abundance. The low abundance classes are including miRNAs and riboswitch (Fig. 3.23 & Fig. 3.24).

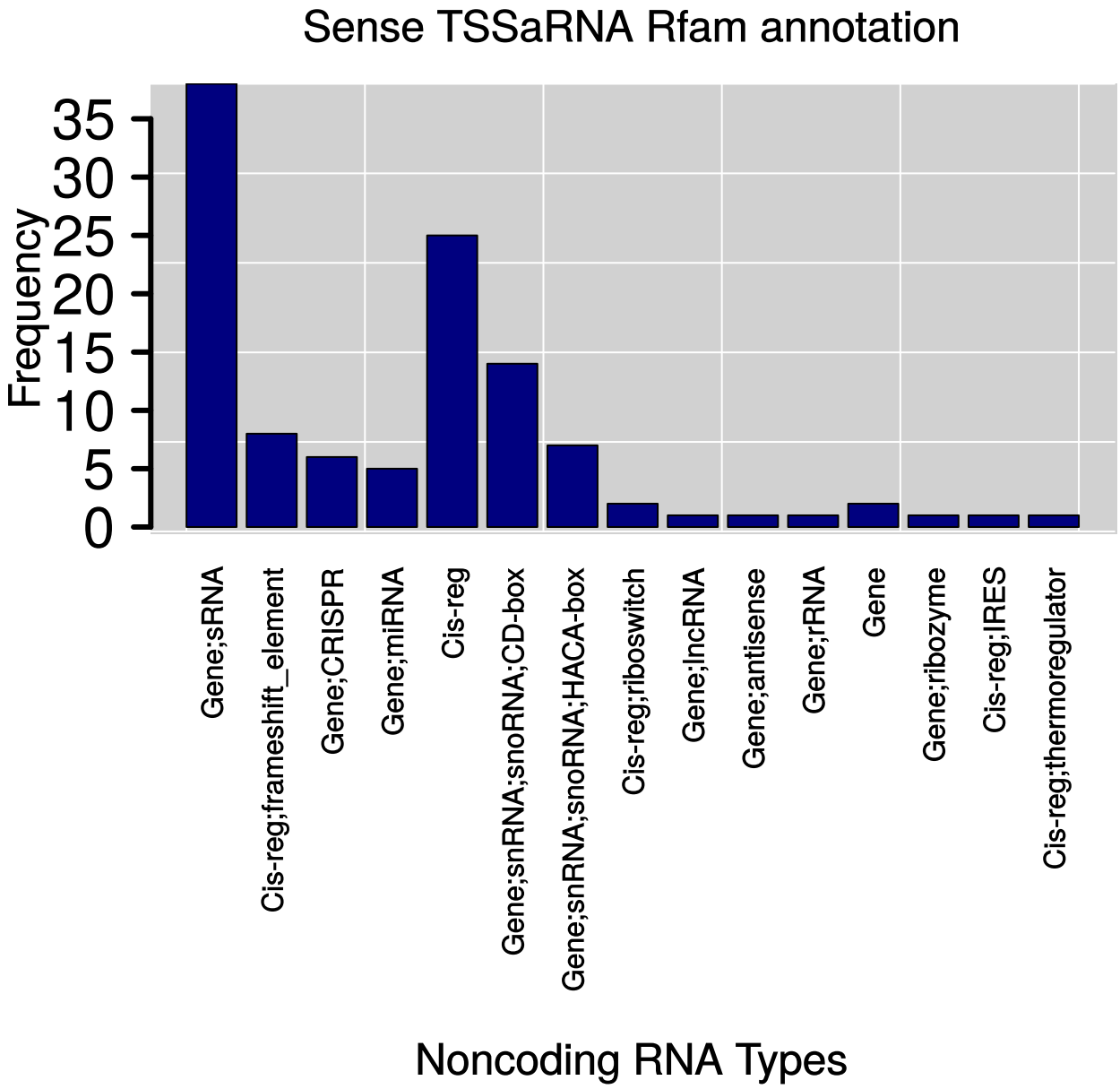


Figure 3.23: Rfam annotation of sense TSSaRNAs

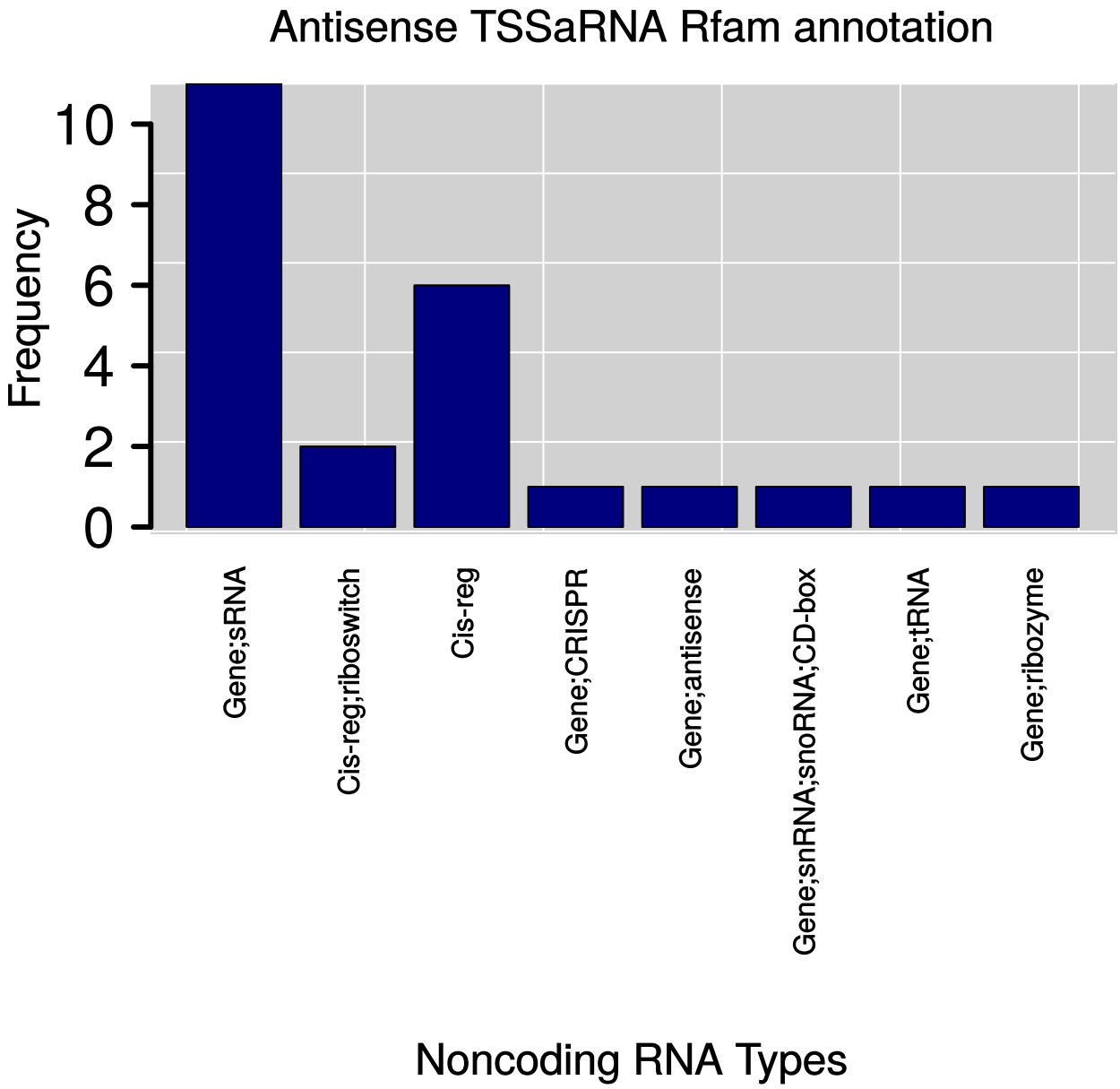


Figure 3.24: Rfam annotation of antisense TSSaRNAs

Table 3.3: Annotation of TSSaRNAs CRISPR molecules.

CRISPR name	Rfam accession	TSSaRNAs	GC %	E-value
CRISPR-DR13	RF01326	chr-770712-770821-VNG1006H	0.35	0.00037
CRISPR-DR23	RF01336	chr-1812831-1812934-VNG2418G	0.74	$2.4 * 10^{-5}$
CRISPR-DR36	RF01346	chr-1508836-1508887-VNG2048G	0.59	0.007
CRISPR-DR29	RF01340	chr-590829-590853-VNG0786G	0.68	0.00036
CRISPR-DR23	RF01336	chr-1676536-1676557-VNG2253H	0.83	0.0015
CRISPR-DR36	RF01346	chr-1277692-1277713-VNG1727G	0.56	0.0017
CRISPR-DR29	RF01340	chr-729356-729445-VNG0955G	0.77	$2.2 * 10^{-5}$

TSSaRNAs named as a combination of the chromosome id , TSSaRNA start, TSSaRNA end and cognate gene id
The orientation of the first TSSaRNA is antisense while the rest are sense TSSaRNAs.

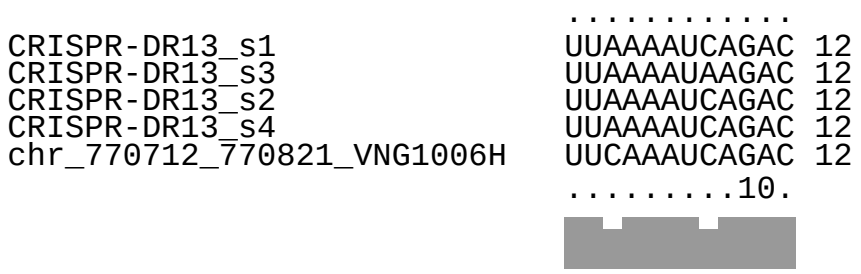


Figure 3.25: CRISPR seeds-TSSaRNA local Alignments (A).
A figure shows the local alignment of CRISPR seeds and TSSaRNA. The alignment shows a subsequence of length 12 nucleotides conserved among the four seeds sequences of CRISPR and TSSaRNA. This figure generated using LocARNA on-line tool (v1.9.1).

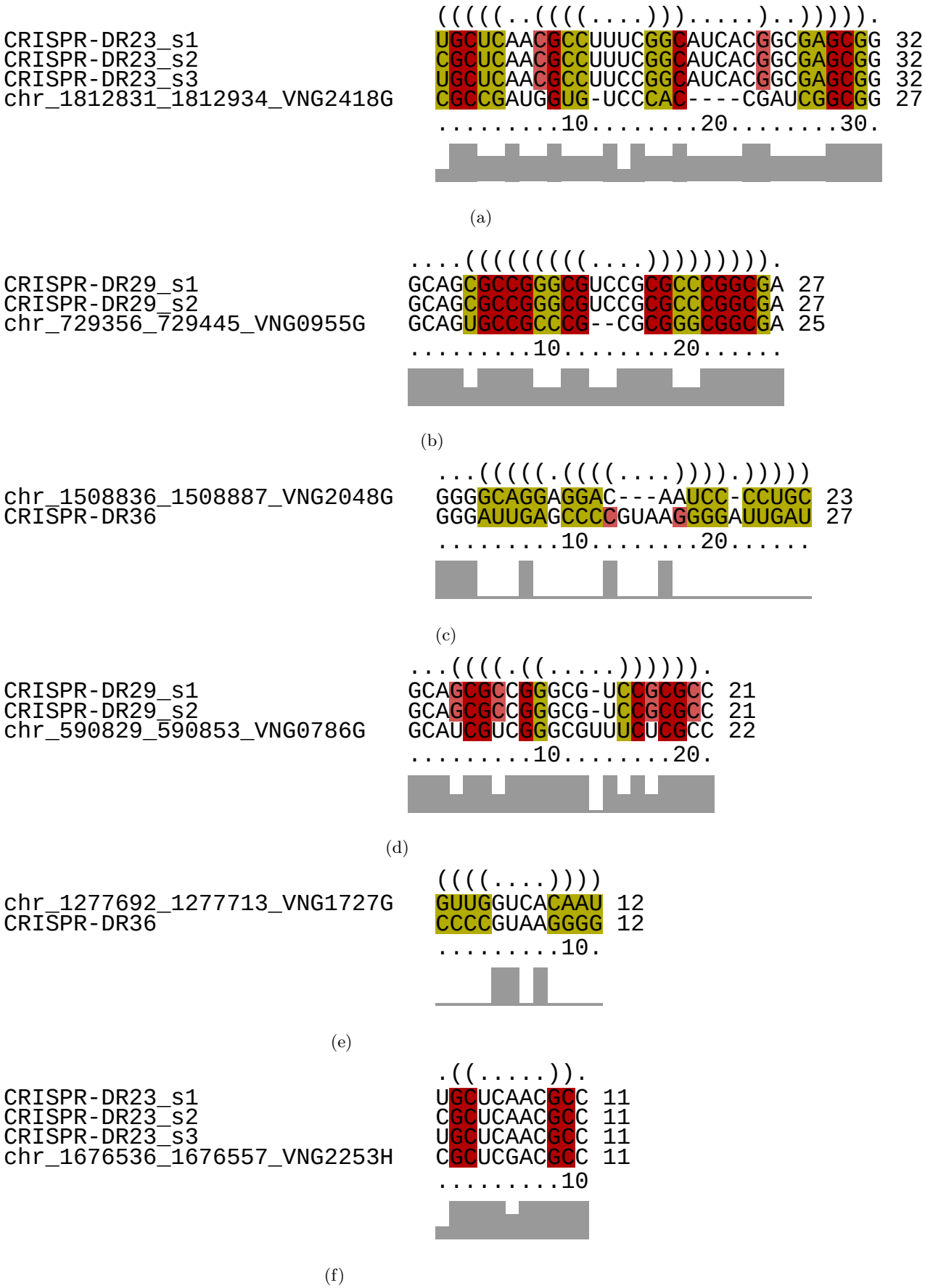


Figure 3.26: CRISPR seeds-TSSaRNA local Alignments (B).
A figure shows the local alignment of CRISPR seeds and TSSaRNA, alignment shows the existence of a conversed subsequences as well as the existence of a conversed secondary structures. These figures generated using LocARNA on-line tool (v1.9.1).

3.4.2 Annotations of TSSaRNAs based on the consensus secondary structures information

The second approach to annotate TSSaRNAs which had based on the consensus structure information of TSSaRNAs has succeeded to annotate 10 sense TSSaRNAs into their corresponding Rfam families. To query the covariance models of TSSaRNAs against the Rfam families, first, we had divided the cluster tree of the unannotated TSSaRNAs to groups of notes such that each node has a consensus secondary structure associated with stable thermodynamic energy. Consequently, we had grouped the sense TSSaRNAs into 17 nodes and the antisense TSSaRNAs into 4 nodes. We queried the covariance models of the selected nodes against Rfam sequences with E-value of 0.001 as a cutoff value. The results showed that 10 of sense TSSaRNAs in 5 covariance models had been assigned into corresponding Rfam families. The 10 of sense TSSaRNAs and their corresponding Rfam families are listed on the table (Table 3.4).

Table 3.4: Annotated sense TSSaRNA based on consensus secondary structures.

CM ID	TSSaRNAs	Rfam accession	Rfam Class	E-value
3	chr_395270_395293_VNG0510G chr_1667801_1667825_VNG2239C	RF02666	Gene;snRNA; snoRNA; scaRNA	0.00092
68	chr_413374_413396_VNG0536G chr_284266_284287_VNG0360C	RF01529	Gene;sRNA; snoRNA; CD-box	0.00062
65	pNRC200_357948_357968_VNG6478H pNRC200_39499_39519_VNG6048H	RF00608	Gene;snRNA; snoRNA; CD-box	0.0005
101	pNRC200_22463_22510_VNG6029G pNRC100_22463_22510_VNG7025	RF02349	Gene; sRNA	0.00018
4	chr_442209_442232_VNG0574C pNRC200_220232_220255_VNG6288C	RF01129	Gene;snRNA; snoRNA; CD-box	0.00038

TSSaRNAs named as a combination of chromosome name, TSSaRNA start, TSSaRNA end, and cognate gene name.

3.5 The functional annotation of the cognate genes

The results of the functional classification of the cognate genes based on the PANTHER classification system result in a total of 87 biological process hits, 78 molecular function hits and 54 cellular component hits. The biological process of the cognate genes showed an indication of TSSaRNAs-associated genes could be involved in various essential processes such as metabolic process (GO:0008152) and cellular process (GO:0009987)(Fig. 3.27). Regarding the molecular function of the cognate genes the results revealed that the cognate genes had incorporated into a larger network of catalytic activity (GO:0003824) or binding to other biomolecules (GO:0005488) (Fig. 3.28). In respect to the significant cellular components, the results revealed that the cognate genes of the annotated TSSaRNAs have belonged to four main terms that are: cell part (GO:0044464), macromolecular complex (GO:0032991), organelle (GO:0043226) and membrane (GO:0016020) (Fig. 3.29).

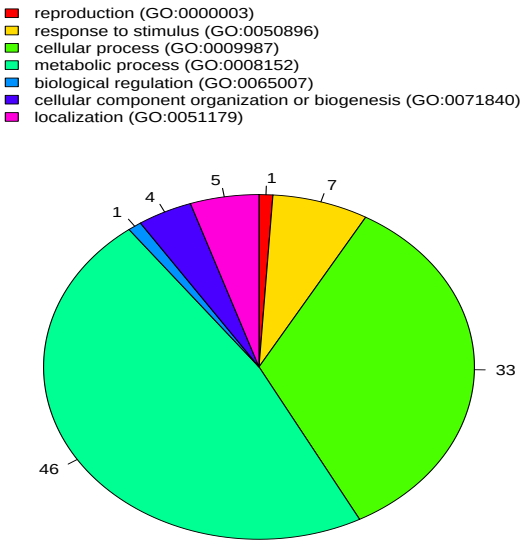


Figure 3.27: The functional annotation of cognate genes (the biological process terms in the list of cognate genes).

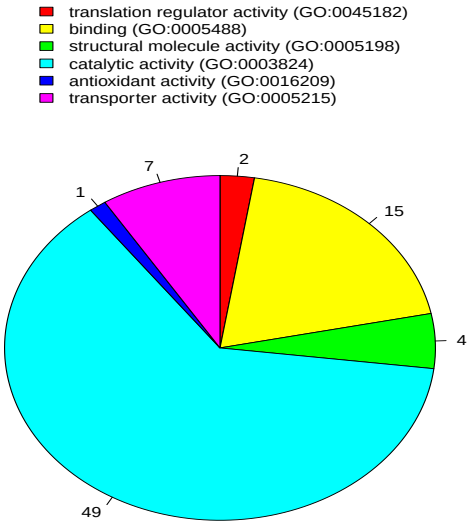


Figure 3.28: The functional annotation of cognate genes (the molecular function terms in the list of cognate genes).

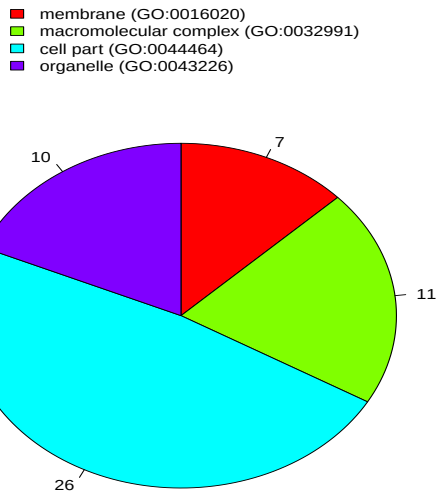


Figure 3.29: The functional annotation of cognate genes (the cellular components terms in the list of cognate genes).

CHAPTER 4
Discussion

4.1 TSSaRNAs

In this work we have identified and characterized the transcription start site associated RNAs in *Halobacterium salinarum* NRC-1 based on the statistical distribution of the RNA-seq reads. The raw data has generated from genome-wide scale strand-specific sequencing experiments. We were able to identify a set of 224 sense TSSaRNAs and 58 antisense TSSaRNAs that enriched near to the transcription start sites of cognate genes. In comparison with our groups own previous work establishing TSSaRNAs in *Halobacterium salinarum* NRC-1, We were more restrictive in defining TSSaRNAs regarding the size of the expressed peaks. Therefore, we considered TSSaRNA if only its expression peak is greater than or equal to the average of reads' expression within the predefined window for each cognate gene. However, by applying such restrictive criteria, the results showed 87.5% of predicted sense TSSaRNAs in concordance with the previously published TSSaRNAs by Zaramela et al. [46]. Our antisense TSSaRNA dataset is, on the other hand, totally original.

Referring to the characteristics of TSSaRNAs length size distribution, the results showed that the median size of TSSaRNAs is 25 nucleotides. This median size is quite similar to what Zaramel and her coauthors stated in their publication as they noted the median size of TSSaRNAs is 27 nucleotides. The obtained median size of TSSaRNAs is also in

the range of TSSaRNAs median size in other organisms. The other organisms which have included in this comparison are human, chicken, drosophila, and bacteria [45, 49, 55, 56]. The distribution of TSSaRNAs length size reveals that there are variations in TSSaRNAs size. The overall TSSaRNAs length sizes ranged from 20 to 145 nucleotides. Furthermore, by observing the histogram of the length size distribution of TSSaRNAs, we could suggest that TSSaRNAs categorized into three groups such as small-sized TSSaRNAs, medium-sized TSSaRNAs, and reasonable long TSSaRNAs. The length size of the small-sized TSSaRNAs is in a range of 20 to 50. On the other hand, the length size of the medium-sized TSSaRNAs is in range of 45 to 90 nucleotides. The group of the reasonable long TSSaRNA consists of TSSaRNAs with a length size greater than 90 nucleotides.

We observed that most of TSSaRNAs initiated exactly from the same point of the transcript start site signal of their cognate genes. However, in some cases, there were deviations from the transcription start sites, which was not even considered by Zaramela et al. [46]. This finding provides supportive evidence about the alternative transcription start sites of some annotated genes. The recent studies of the alternative transcription sites have proposed that this phenomenon plays as a transcriptional regulator in Eukaryote [115]. Moreover, the alternative transcription start sites have correlated with the expression of ncRNAs in bacteria [116]. In the Archaeal domain, the association of several ncRNAs classes with the existence of the alternative transcription sites has been investigated in the haloarchaeon *Haloferax volcanii* by Babski and his colleagues in 2016 [117]. Putting all together in connection with these findings, we could postulate that some TSSaRNAs could be involved in the RNA-based regulation mechanism that initiated from the alternative transcription site of some genes. Therefore, as a prediction for the future of genes annotation, we think that the process of re-annotation of genes of many organisms will take place in the near future. The prospective annotation will take into account the dynamical phenomena of transcription initiation signal as many genes posses alternatives transcription start site. This annotation is expected to provide information about the most dominant transcription starting site in a particular condition. One of the interesting findings in this project is that TSSaRNAs prediction pipeline

has succeeded to identify for the first time the antisense TSSaRNAs besides the sense TSSaRNAs in Archaea. The previous attempts of prediction TSSaRNAs in the other domains revealed the existence of both sense and antisense TSSaRNAs in Eukaryote while only sense TSSaRNAs that have identified in Bacterial domain. Linking these finding together, we could consider this finding as supportive evidence of the hypothesis of biological taxonomy that considers the Archaea as a distinct evolutionary domain, but in many aspects, Archaea is more close to Eukaryote domain than the Bacterial domain.

4.2 TSSaRNAs structures

Accurate prediction of TSSaRNAs structure provides insights into their potential functional role by considering that RNA secondary structure is stable enough such that it is capable of triggering its function. Therefore, in this study, we investigated the secondary structures of TSSaRNAs based on the most common prediction algorithm that relies on the minimization the thermodynamic free energy.

The obtained free energies of TSSaRNAs secondary structures suggested that almost all of TSSaRNAs were capable of folding into thermodynamically stable secondary structures with various topologies. The majority of the folded structures contained at least one stable hairpin secondary structure motif. The existence of the hairpins associated with favorable free energy would suggest that TSSaRNAs could trigger potential regulatory roles. The review [18] gives more details about the possible regulatory functions that could be triggered by the hairpin motifs. The expected regulatory functions of TSSaRNAs due to its hairpin motifs are including: protecting cognate gene degradation, recognition binding motifs in RNA-binding proteins, or acting as a regulatory element.

We have considered the dynamical phenomena in TSSaRNAs secondary structures where a particular TSSaRNA could possess multiple suboptimal secondary structures. However, we used only the most stable structures for each TSSaRNA as modeling constrain to model the tertiary structures because in the normal physiological condition the optimal structures supposed to be the most dominant one. The overall weighted energies obtained

from TSSaRNAs tertiary structures suggest that these structures are stable due to the highly negative values of their free energy. In respect to the low values of the radius of gyration energy, we could imply that there was no much conflict between TSSaRNAs residues in their tertiary structures. The values of the energy obtained from the terms of Van der Waals interactions & Lennard-Jones interactions suggested the existence of a repulsive force in TSSaRNAs tertiary structures. The repulsive force in TSSaRNAs tertiary structures expected as it has known that RNA molecules are highly charged molecules. The existence of the repulsive force in TSSaRNAs tertiary structures does not imply that the process of the unfolding of these structures is a spontaneous process as there are counterions within the cells to reduce this repulsive force (counter ions effect). Due to the highly negative values of the energies arising from the hydrogen bonds, we could suggest that TSSaRNAs are favorable in binding with other biomolecules to form complex higher order structures. However, due to the values of the solvation energies, we expect energy is needed to trigger the formation of such complexes. The variation in the topologies of folded tertiary structures suggests that TSSaRNAs could be capable of triggering many biological functions as many pockets motifs have observed in their tertiary structures; however, a further experimental investigation is needed to validate their roles.

By considering TSSaRNAs to function as part of the RNA-based regulation system, we have investigated the possibility of the TSSaRNAs to bind with the cognate genes to regulate their transcription process or regulate the translation machinery. The fundamental mode of action in the RNA-based regulation system is different among the Bacteria and Eukaryote. In Bacteria small regulatory RNAs bind into the 5' (five prime) of mRNA while in Eukaryote sRNAs bind to the 3' (three prime) of mRNA [36]. Therefore, here in this study, we have hypothesized three types of TSSaRNAs-cognate genes binding patterns. These three types of interactions include: (a) the hybridization with five prime regions of cognate genes. This type of binding expected to act as a fine-tuning of gene expression [36] or it could be responsible for the inhibition of the translation process; (b) in the second pattern of the interaction we hypothesized that TSSaRNAs could be

capable of binding to the 3' regions of the cognate genes. This type of interactions is expected to increase the half-life time of cognate genes [36]; (c) in the third type of interaction we expected that TSSaRNAs bind within the coding region to trigger function somehow similar to siRNA system [36]. Based on the results of TSSaRNAs-cognate gene hybridization we could consider that most of TSSaRNAs are involved into RNA-based regulation system to play as putative cis-acting small RNAs as the results suggested that around 92% of TSSaRNAs were capable of interacting with their cognate genes. However, based on this computational experiment we can not affirm that the rest of the 22 TSSaRNAs molecules could not be involved in the RNA-based regulation system as there is alternative hypothesis where TSSaRNAs could be capable of hybridizing with other genes rather than their cognate genes to function as trans-acting regulators.
Based on the result of TSSaRNAs-cognate gene hybridization we observed that there was a balance between the number of TSSaRNAs that bind to the 5' regions of the cognate genes (51%) and those bind to the 3' regions of their cognate genes (49%). This finding suggests the idea that the RNA-based regulations system in Archaea is a mixture of both Bacteria and Eukaryote regulations systems.

4.3 TSSaRNAs-Lsm intercation

Our findings suggest the existence of potential interactions between Lsm protein and TSSaRNAs. Lsm protein has the capability to interact with various RNAs within the cell [38]. Therefore, we speculate Lsm to bind to all TSSaRNAs that expressed at the interaction signals including TSSaRNAs and it cognate mRNA. By considering the existence of putative interaction of Lsm protein with both of TSSaRNA and its cognate gene we could speculate that TSSaRNA regulates its cognate gene by either competitive or cooperative interactions with Lsm protein. Hence, we could postulate that some TSSaRNAs could trigger their potential functions in a way similar to the mode of action of some bacterial sRNA that regulates their targeted mRNAs by competitive or cooperative interactions with Hfq protein [118]. Putting all together, we could sum up this

section by suggesting the existence of potential interactions between some TSSaRNAs and Lsm protein to elucidate their regulatory function in the similar way of small regulatory RNAs in bacteria which require to bind to Hfq protein and small regulatory RNAs in Eukaryote which require to bind to the Ago protein. As the mode of the action of most Archaeal small regulatory RNAs is not well characterized yet then based on this finding, we could speculate that the potential mechanism of some regulatory RNAs in Archaea including some TSSaRNAs might require binding to Lsm protein to mediate their function. Although, more experimental investigations are needed to ensure that Lsm interacts with both TSSaRNA and its cognate gene. Also, experimental investigations will help to well characterize the importance of such interaction in their potential regulatory roles and the mechanism of the mode of action.

4.4 Rfam annotation

By considering the Rfam database as a repository of a comprehensive collection of ncRNAs families represented by ncRNAs sequence and conserved structures, we have attempted to annotated TSSaRNAs into their corresponding families.

In theory, a single TSSaRNA could hit multiple Rfam families that belong into the same ncRNA class, and even the result could yield a more complex situation where a single TSSaRNA could be annotated into multiple ncRNA classes. However, in this work for the purpose of simplification, we have annotated TSSaRNAs into only one Rfam family that associated with the minimum E-value. The results of annotating TSSaRNAs into their corresponding Rfam families revealed that the majority of TSSaRNAs have considered as putative cis-acting regulators such as cis-regulatory element, sRNAs, and microRNAs. However, there was a possibility of TSSaRNAs to function as potential trans-acting regulators to regulate distant molecules such as CRISPR and Antisense RNA. Moreover, the results suggested that some TSSaRNAs could trigger more complex functions as catalytic functions such as Riboswitch or play a potential role in the process of defense against a virus such as CRISPR.

We have observed that most of TSSaRNAs have annotated as potential regulatory ncRNAs where the majority of TSSaRNAs have interpreted as putative sRNA or putative cis-regulatory elements. Taken together this finding in connection with the result that obtained from TSSaRNAs-cognate genes hybridization, we could conclude that it is reasonable to speculate that most of TSSaRNAs function as fine-tuning of cognate gene expression by direct interaction into target mRNA molecule. Also, we could consider that some TSSaRNAs could function as sRNA that regulates several biological processes and trigger the process of adaptation to stressed conditions [36]. The result showed the existence of a specific class of cis-regulatory called frame-shift element. Frame-shift elements regulate the gene expressions by changing the open reading frame to allow target mRNAs to code for different protein [119]. Taken the finding of the existence of TSSaRNAs as a potential frame-shift element with reference to the hypothesis of the existence of alternative transcription sites, we could suggest that some TSSaRNAs could be capable of regulating gene expression by providing an alternative transcript.

The second major ncRNA class which has predicted in TSSaRNAs (mainly sense TSSaRNAs) is the snRNAs class. The existence of snRNA (mainly snoRNAs) in Archaeal domain has been proved by both experimental studies and computational approach [120–122]. snoRNAs guide the modifications of other non-coding RNAs such as ribosomal, and transfer RNAs. As almost all of the snoRNAs are not transcribed from their own genes (few exceptions such as U17A and U17B6 snoRNA) therefore, prediction of the snoRNAs as part of various genes are expected. In Eukaryote there are two major types of snoRNAs. The types of Eukaryote snoRNAs are H/A snoRNAs and C/D snoRNAs. The consensus sequences in H/A snoRNAs consist of H-box motif (ANANNA, where N = A, C, G or U) and A-box motif (ACA). In the case of H/A snoRNAs, the consensus sequences consist of C-box motif (RUGAUGA, where R = A or G) and D-box motif (CUGA). However, in this study, we did not succeed to find these motifs on TSSaRNAs that have annotated as potential snoRNAs. Failing in finding the consensus motifs of snoRNAs on annotated TSSaRNAs of *Halobacterium salinarum* NRC-1 expected as a previous study concluded that snoRNAs in halophilic species are very divergent and without the canonical sequence

motifs compared to other organisms [123].

The result showed few TSSaRNAs that have annotated as putative CRISPR molecules. The class of CRISPR has considered as an adaptive immune system of Archaea against viral infection and foreign nucleic acids [124]. Regarding our best knowledge, the CRISPR class has not been identified yet in the model organism *Halobacterium salinarum* NRC-1 however, we can not totally ignore the annotated TSSaRNAs as CRISPR molecules even if the associated E-value is not very much extreme value. The existence of insertion and deletion in the local alignment of TSSaRNAs with the CRISPR seeds had expected as the insertion and deletions had existed even when we performed the local alignment of CRISPR seed sequences of the same CRISPR family. Although, there are no any wet-lab results that proof the existence of the CRISPR in *Halobacterium salinarum* NRC-1 but we consider this finding as a motivative finding for the experimental biologist to validate it.

Furthermore, the result interestingly showed some of TSSaRNAs that have been annotated as putative microRNAs to consider some TSSaRNAs as gene silencing elements. We have paid some attention for this class because the existence of this type of ncRNAs in Archaea has been predicted computationally [125] yet no experimental evidence that proves the existence of microRNAs in Archaea. Therefore, we consider this finding also as another motivative finding for the experimental biologist to validate it.

We have considered the possibility of TSSaRNAs to trigger enzymatic functions such as riboswitch or ribozyme. However, the pipeline succeeded to annotate only three TSSaRNAs as putative riboswitch and only two TSSaRNAs as a putative ribozyme. Therefore, we suggest that the enzymatic activity of TSSaRNAs may exist, but it is not a major function of TSSaRNAs molecules.

4.5 TSSaRNAs potential functions in respect to cognate genes functional classification

By assuming that TSSaRNAs potential functions might be connected to the functions

of the cognate genes, we performed the functional classification of the cognate genes based on the Panther database of gene ontology. The functional classification of the cognate genes showed that the genes are involved in many essential and fundamental functions (Fig. 3.27, Fig. 3.28, and Fig. 3.29). For instance, the results showed that the cognate genes have involved in the following functions: forming macromolecular complexes (GO:0032991), structural activity(GO:0005198), regulator activity (GO:00045182, GO:0065007), catalytic activity (GO:0003824), binding to other biomolecules (GO:0005488), besides many other functions. The results of the cognate genes classification gave indirect evidence about the variability of TSSaRNAs potential functions. With reference to TSSaRNAs functional annotations, we hypothesize that TSSaRNAs could control their cognate genes as Rfam annotations showed many TSSaRNAs as putative cis regulators. Also, the results showed potential interactions between many TSSaRNAs and their cognate genes. However, still many bench experiments are needed to clarify this hypothesis by investigating the roles of TSSaRNAs in controlling the functions of their cognate genes.

CHAPTER 5
Conclusion

During this doctoral studies, I have attempted to predict TSSaRNAs accurately using data generated from the model organism *Halobacterium salinarum* NRC-1 and then functionally annotate the predicted TSSaRNAs. I conducted and performed *in silico* experiments using bioinformatics tools to gain the objectives of this project. This project aims to contribute in the knowledge discovery by further our understanding about the potential functions of TSSaRNAs molecules and their regulatory roles in *Halobacterium salinarum* NRC-1 as well as giving an interpretation of their potential molecular mode of action. To summarize:

- The strand-specific RNA-seq is suitable to predict TSSaRNAs however, without considering checking the quality control of the RNA-seq data as a crucial stage the result will be misleading. As the mapping process has considered as the most time-consuming stage in any traditional RNA-seq analysis, therefore, bioinformatician should make a trade-off between the acceleration of the mapping process by filtering out very short reads and losing the short TSSaRNAs.
- The algorithm which was used to predict TSSaRNAs depended on the distribution of the mapped reads within a predefined window (local distribution). However, a more sophisticated algorithm by using a global distribution of reads (chromosomal, or among the entire genome) or hybrid of local and global could be considered for future prediction. Also, it evident that a bigger predefined window size will increase

the false positive TSSaRNAs. On the other hand, a short length of predefined window size will improve the accuracy of prediction, but it will result in higher false negatives.

- We could conclude from the predicted TSSaRNAs secondary and tertiary structures that TSSaRNAs are thermodynamically stable molecules associated with high free energy. The significance of this result is that many scientists consider the structure of the minimum free energy as the dominant structures. Consequently, many scientists believe the structure of the minimum free energy as the biologically active structure. However, it is certain that the structures of bio-molecules are dynamic when performing their biological functions. Therefore, multiple active forms of a single molecule exited. To investigate the dynamical nature of TSSaRNAs structures thus predict all the active form for each TSSaRNAs molecule all the suboptimal structures should be predicted using a method that considers both thermodynamics and statistical mechanics.
- The significance of predicting TSSaRNAs secondary structures is that the RNA secondary structure has considered as stable enough to trigger a biological function.
- Although predicting RNA tertiary structures is an essential step in annotating their biological functions, but it is still challenging to define all functional motifs in their structures.
- Investigating the higher order structures of TSSaRNAs is an essential step towards understanding their potential regulatory roles. Thus, by examining TSSaRNAs cognate genes interactions, this study has demonstrated that the majority of TSSaRNAs are capable of playing regulatory roles as part of RNA-based regulation system.
- By considering TSSaRNAs-Lsm binding, we could conclude that some of TSSaRNAs might require Lsm protein (the chaperone protein) to trigger their functions. Moreover, the computational docking experiment suggested that most of the interacted TSSaRNAs are capable of binding to the middle of the ring in the Lsm structure with favorable binding energy. Based on the computational docking experiment,

we could comment that the molecular docking techniques provide a window to model the interaction between two molecules even without prior knowledge about the binding sites. This kind of the molecular docking approach is technically called blind docking. Blind docking could provide a considerable amount of information to demonstrate the potential biomolecular interactions and enhance our understanding of the higher order structures in the complex biological system.

- From the functional annotation of TSSaRNAs based on the Rfam classification, we could consider TSSaRNAs as versatile molecules that could be capable of triggering various potential functions. The majority of the TSSaRNAs could be viewed as potential cis-acting regulators to regulate near molecules such as cis-regulatory element, sRNAs and microRNAs. On the other hand, some potential trans-acting TSSaRNAs regulating exist such as CRISPR and Antisense RNA.
- The majority of TSSaRNAs are predicted to be involved in a putative RNA based regulation system to regulate their cognate genes.
- TSSaRNAs could have more complex functions such as catalytic function as Riboswitch Ribozyme or as a defense against a virus such as CRISPR.

In conclusion, this work addressed some unanswered questions on the function of TSSaRNAs providing insights into their potential roles inside the cell where they could regulate cognate genes or act as potential trans-acting molecules. Although our functional annotation approach has succeeded to annotate many TSSaRNAs still more effort is needed to understand their functional mechanism deeply and to emerge in their full complexity.
In this project, we have reported many exciting insights and promising results that have gained throughout this *in silico* study. Although, we think that any task that is depending only on bioinformatics (*in silico*) and relies solely on the computational experiments is a challenging task with many obvious limitations in many situations, however, the *in silico* experiment is an absolutely indispensable step for many wet-lab experiments. According to the current paradigm in experimental designing, many scientists consider the computational works as supplementary to the bench-work trials. However, we think

that in the future the bioinformatics studies and the computational biology hypotheses will be more fundamental part to address any scientific question in the field of system biology and other related fields. Moreover, we expect that bioinformatics will play a central role in opening many sophisticated questions.

Bibliography

[1] C R Woese and G E Fox. Phylogenetic structure of the prokaryotic domains: the primary kingdoms. *Proceedings of the National Academy of Sciences, USA*, 74(11):5088–5090, 1977.

[2] C R Woese, O Kandler, and M L Wheelis. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, 87(12):4576–4579, 1990.

[3] S Winker and C R R Woese. A Definition of the Domains Archaea, Bacteria and Eucarya in Terms of Small Subunit Ribosomal RNA Characteristics. *Systematic and applied microbiology*, 14(4):305–310, 1991.

[4] Idan Bodaker, Itai Sharon, Roi Feingersch, and Mira Rosenberg. Archaeal diversity in the Dead Sea : Microbial survival under increasingly harsh conditions. *Natural Resources and Environmental Issues*, 15:137–143, 2009.

[5] S Leuko, C Domingos, A Parpart, G Reitz, and P Rettberg. The Survival and Resistance of Halobacterium salinarum NRC-1, Halococcus hamelinensis, and Halococcus morrhuae to Simulated Outer Space Solar Radiation. *Astrobiology*, 15(11):987–997, 2015.

[6] W V Ng, S P Kennedy, G G Mahairas, B Berquist, M Pan, H D Shukla, S R Lasky, N S Baliga, V Thorsson, J Sbrogna, S Swartzell, D Weir, J Hall, T A Dahl, R Welti, Y A Goo, B Leithauser, K Keller, R Cruz, M J Danson, D W Hough, D G Maddocks, P E Jablonski, M P Krebs, C M Angevine, H Dale, T A Isenbarger, R F Peck, M Pohlschroder, J L Spudich, K W Jung, M Alam, T Freitas, S Hou, C J Daniels, P P Dennis, A D Omer, H Ebhardt, T M Lowe, P Liang, M Riley, L Hood, and S DasSarma. Genome sequence of Halobacterium species NRC-1. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):12176–12181, 2000.

[7] Ronald F Peck, Shiladitya DasSarma, and Mark P Krebs. Homologous gene knockout in the archaeon Halobacterium salinarum with ura3 as a counterselectable marker. *Molecular Microbiology*, 35(3):667–676, 2000.

[8] James a Coker, Priya DasSarma, Jeffrey Kumar, Jochen a Müller, and Shiladitya DasSarma. Transcriptional profiling of the model Archaeon Halobacterium sp. NRC-1: responses to changes in salinity and temperature. *Saline systems*, 3:6, 2007.

[9] Patrick E Gygli and Linda C DeVeaux. Adaptation of the Halobacterium salinarum ssp. NRC-1 gene deletion system for modification of chromosomal loci. *Journal of Microbiological Methods*, 99(1):22–26, 2014.

[10] Scott A Strobel and Jesse C Cochrane. RNA catalysis: ribozymes, ribosomes, and riboswitches, 2007.

[11] Arati Ramesh and Wade C Winkler. Metabolite-binding ribozymes, 2014.

[12] Zhichao Miao and Eric Westhof. RNA Structure: Advances and Assessment of 3D Structure Prediction. *Annual Review of Biophysics*, 2017.

[13] Jörg Fallmann, Sebastian Will, Jan Engelhardt, Björn Grüning, Rolf Backofen, and Peter F. Stadler. Recent advances in RNA folding. *Journal of Biotechnology*, 2017.

[14] Nilay Chheda and Manish K Gupta. RNA as a Permutation. *arXiv.org*, 2014.

[15] Marilyn Kozak. Regulation of translation via mRNA structure in prokaryotes and eukaryotes, 2005.

[16] Sébastien Lemieux and François Major. RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic acids research*, 30(19):4250–4263, 2002.

[17] Avinash Achar and Pål Sætrom. RNA motif discovery: a computational overview. *Biol Direct*, 10(1):61, 2015.

[18] P. Svoboda and A. Di Cara. Hairpin RNA: A secondary structure of primary importance, 2006.

[19] I Tinoco, O C Uhlenbeck, and M D Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230(5293):362–367, 1971.

[20] M Zuker. On finding all suboptimal foldings of an RNA molecule. *Science (New York, N.Y.)*, 244(4900):48–52, 1989.

[21] E Rivas and S R Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of molecular biology*, 285(5):2053–2068, 1999.

[22] M Zuker and A B Jacobson. Using reliability information to annotate RNA secondary structures. *RNA (New York, NY)*, 4(6):669–679, 1998.

[23] D H Mathews, J Sabina, M Zuker, and D H Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of molecular biology*, 288(5):911–940, 1999.

[24] Ye Ding and Charles E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 2003.

[25] Martin Mann, Patrick R. Wright, and Rolf Backofen. IntaRNA 2.0: Enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Research*, 45(W1):W435–W439, 2017.

[26] Junichi Iwakiri, Michiaki Hamada, Kiyoshi Asai, and Tomoshi Kameda. Improved Accuracy in RNA-Protein Rigid Body Docking by Incorporating Force Field for Molecular Dynamics Simulation into the Scoring Function. *Journal of Chemical Theory and Computation*, 12(9):4688–4697, 2016.

[27] Samuel C Flores, Yaqi Wan, Rick Russell, and Russ B Altman. Predicting RNA structure by multiple template homology modeling. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 2010.

[28] Rhiju Das, John Karanicolas, and David Baker. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature methods*, 7(4):291–294, 2010.

[29] Rhiju Das and David Baker. Macromolecular modeling with rosetta. *Annual review of biochemistry*, 77:363–82, 2008.

[30] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of chemical theory and computation*, 13(6):3031–3048, jun 2017.

[31] David Dufour and Marc A. Marti-Renom. Software for predicting the 3D structure of RNA molecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 5(1):56–61, jan 2015.

[32] Jörg Fohrer, Mirko Hennig, and Teresa Carlomagno. Influence of the 2-hydroxyl group conformation on the stability of A-form helices in RNA. *Journal of Molecular Biology*, 356(2):280–287, 2006.

[33] Manja Marz and Peter F. Stadler. RNA interactions. *Advances in Experimental Medicine and Biology*, 722:20–38, 2011.

[34] Frédéric Chédin. Nascent Connections: R-Loops and Chromatin Patterning, 2016.

[35] Giedrius Gasiunas, Rodolphe Barrangou, Philippe Horvath, and Virginijus Siksnys. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 2012.

[36] Julia Babski, Lisa-Katharina Maier, Ruth Heyer, Katharina Jaschinski, Daniela Prasse, Dominik Jäger, Lennart Randau, Ruth A Schmitz, Anita Marchfelder, and Jörg Soppa. Small regulatory RNAs in Archaea. *RNA Biology*, 11(5):484–493, 2014.

[37] Angela Re, Tejal Joshi, Eleonora Kulberkyte, Quaid Morris, and Christopher T. Workman. RNA–protein interactions: An overview. *Methods in Molecular Biology*, 1097:491–521, 2014.

[38] Meghna Sobti Jens M. Moll and Bridget C. Mabbutt. *The Lsm Proteins: Ring Architectures for RNA Capture, RNA Processing.* Prof. Paula Grabowski (Ed.), InTech, 2011.

[39] William R. Pearson. An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, 2013.

[40] Eugene W. Myers Stephen F. Altschul, Warren Gish, Webb Miller and David J. Lipman. BLAST, 1990.

[41] Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna, and Sean R. Eddy. Rfam: An RNA family database, 2003.

[42] Michael Brown, Kimmen Sjolander, Rebecca C Underwood, David Haussler, and Santa Cruz. Stochastic context-free grammars. 22(23):5112–5120, 1994.

[43] Sean R Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic acids research*, 22(11):2079–2088, jun 1994.

[44] Ryan A Flynn, Albert E Almada, Jesse R Zamudio, and Phillip A Sharp. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proceedings of the National Academy of Sciences of the United States of America*, 108(26):10460–10465, 2011.

[45] Eva Yus, Marc Güell, Ana P Vivancos, Wei-Hua Chen, María Lluch-Senar, Javier Delgado, Anne-Claude Gavin, Peer Bork, and Luis Serrano. Transcription start site associated RNAs in bacteria. *Mol Syst Biol*, 8(585):585, 2012.

[46] Livia S Zaramela, Ricardo Z N Vêncio, Felipe Ten-Caten, Nitin S Baliga, and Tie Koide. Transcription start site associated RNAs (TSSaRNAs) are ubiquitous in all domains of life. *PloS one*, 9(9):e107680, 2014.

[47] Dongliang Yu, Xiaoxia Ma, Ziwei Zuo, Huizhong Wang, and Yijun Meng. Classification of Transcription Boundary-Associated RNAs (TBARs) in Animals and Plants. *Frontiers in Genetics*, 9(MAY):1–10, may 2018.

[48] Ryan J. Taft, Craig D. Kaplan, Cas Simons, and John S. Mattick. Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle*, 8(15):2332–2338, aug 2009.

[49] R J Taft, E A Glazov, N Cloonan, C Simons, S Stephen, G J Faulkner, T Lassmann, A R Forrest, S M Grimmond, K Schroder, K Irvine, T Arakawa, M Nakamura, A Kubosaki, K Hayashida, C Kawazu, M Murata, H Nishiyori, S Fukuda, J Kawai, C O Daub, D A Hume, H Suzuki, V Orlando, P Carninci, Y Hayashizaki, and J S Mattick. Tiny RNAs associated with transcription start sites in animals. *Nature Genetics*, 41(5):572–578, 2009.

[50] Nuankanya Sathira, Riu Yamashita, Kousuke Tanimoto, Akinori Kanai, Takako Arauchi, Soutaro Kanematsu, Kenta Nakai, Yutaka Suzuki, and Sumio Sugano.

Characterization of Transcription Start Sites of Putative Non-coding RNAs by Multifaceted Use of Massively Paralleled Sequencer. *DNA research an international journal for rapid publication of reports on genes and genomes*, 17(3):169–183, 2010.

[51] Jiang Han, Daniel Kim, and Kevin V Morris. Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells. *Proc Natl Acad Sci U S A*, 104(30):12422–12427, 2007.

[52] Ryan J Taft, Cas Simons, Satu Nahkuri, Harald Oey, Darren J Korbie, Timothy R Mercer, Jeff Holst, William Ritchie, Justin J-L Wong, John E J Rasko, Daniel S Rokhsar, Bernard M Degnan, and John S Mattick. Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nature Structural and Molecular Biology*, 17(8):1030–1034, 2010.

[53] Justin Layer and Anthony Weil. Ubiquitous antisense transcription in eukaryotes: novel regulatory mechanism or byproduct of opportunistic RNA polymerase? *F1000 Biology Reports*, 2009.

[54] Matthew G. Guenther, Stuart S. Levine, Laurie A. Boyer, Rudolf Jaenisch, and Richard A. Young. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell*, 130(1):77–88, 2007.

[55] Amy C. Seila, J. Mauro Calabrese, Stuart S. Levine, Gene W. Yeo, Peter B. Rahl, Ryan A. Flynn, Richard A. Young, and Phillip A. Sharp. Divergent transcription from active promoters. *Science*, 322(5909):1849–1851, 2008.

[56] Y. S. Choi, W. Patena, A. D. Leavitt, and M. T. McManus. Widespread RNA 3'-end oligouridylation in mammals. *RNA*, 18(3):394–401, 2012.

[57] Xiangfeng Wang, John D. Laurie, Tao Liu, Jacqueline Wentz, and X. Shirley Liu. Computational dissection of Arabidopsis smRNAome leads to discovery of novel microRNAs and short interfering RNAs associated with transcription start sites. *Genomics*, 97(4):235–243, 2011.

[58] Dmitry Belostotsky. Exosome complex and pervasive transcription in eukaryotic genomes, 2009.

[59] Karen Adelman and John T. Lis. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Reviews Genetics*, 13(10):720–731, 2012.

[60] Diego Duchi, David L.V. Bauer, Laurent Fernandez, Geraint Evans, Nicole Robb, Ling Chin Hwang, Kristofer Gryte, Alexandra Tomescu, Pawel Zawadzki, Zakia Morichaud, Konstantin Brodolin, and Achillefs N. Kapanidis. RNA Polymerase Pausing during Initial Transcription. *Molecular Cell*, 63(6):939–950, 2016.

[61] Xiuli Liu, W. Lee Kraus, and Xiaoying Bai. Ready, pause, go: Regulation of RNA polymerase II pausing and release by cellular signaling pathways. *Trends in Biochemical Sciences*, 40(9):516–525, 2015.

[62] Pedro Rafael Costa, Marcio Luis Acencio, and Ney Lemke. Cooperative RNA Polymerase Molecules Behavior on a Stochastic Sequence-Dependent Model for Transcription Elongation. *PLoS ONE*, 2013.

[63] Daniel S. Day, Bing Zhang, Sean M. Stevens, Francesco Ferrari, Erica N. Larschan, Peter J. Park, and William T. Pu. Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types. *Genome Biology*, 17(1):120, 2016.

[64] Irene M. Min, Joshua J. Waterfall, Leighton J. Core, Robert J. Munroe, John Schimenti, and John T. Lis. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes and Development*, 25(7):742–754, 2011.

[65] Daniel A. Gilchrist, Gilberto Dos Santos, David C. Fargo, Bin Xie, Yuan Gao, Leping Li, and Karen Adelman. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell*, 143(4):540–551, 2010.

[66] Michael Levine. Paused RNA polymerase II as a developmental checkpoint, 2011.

[67] Riki Kurokawa. Promoter-associated long noncoding RNAs repress transcription through a RNA binding protein TLS. *Advances in Experimental Medicine and Biology*, 722:196–208, 2011.

[68] Andreas Mayer, Heather M. Landry, and L. Stirling Churchman. Pause & go: from the discovery of RNA polymerase pausing to its functional implications, 2017.

[69] Stefan Klumpp. Pausing and Backtracking in Transcription Under Dense Traffic Conditions. *Journal of Statistical Physics*, 142(6):1252–1267, 2011.

[70] Evgeny Nudler. RNA polymerase backtracking in gene regulation and genome instability, 2012.

[71] Alan C M Cheung and Patrick Cramer. Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature*, 471(7337):249–253, 2011.

[72] Eivind Valen, Pascal Preker, Peter Refsing Andersen, Xiaobei Zhao, Yun Chen, Christine Ender, Anne Dueck, Gunter Meister, Albin Sandelin, and Torben Heick Jensen. Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes. *Nature structural & molecular biology*, 18(9):1075–82, aug 2011.

[73] Andreas P M Weber. Discovering New Biology through Sequencing of RNA. *Plant physiology*, 169(3):1524–31, 2015.

[74] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: A revolutionary tool for transcriptomics, 2009.

[75] Stephan C. Schuster. Next-generation sequencing transforms today's biology, 2008.

[76] Eric E. Schadt, Steve Turner, and Andrew Kasarskis. A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2), 2010.

[77] James Dominic Mills, Yoshihiro Kawahara, and Michael Janitz. Strand-Specific RNA-Seq Provides Greater Resolution of Transcriptome Profiling. *Current Genomics*, 14:173–181, 2013.

[78] Terrence S. Furey. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions, 2012.

[79] Adam Roberts, Cole Trapnell, Julie Donaghey, John L Rinn, and Lior Pachter. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology*, 12(3):R22, 2011.

[80] Xing Li, Asha Nair, Shengqin Wang, and Liguo Wang. Quality control of RNA-seq experiments. In *RNA Bioinformatics*, pages 137–146. 2015.

[81] Paul Flicek and Ewan Birney. Sense from sequence reads: methods for alignment and assembly. *Nature methods*, 6(11 Suppl):S6–S12, 2009.

[82] Cole Trapnell and Steven L Salzberg. How to map billions of short reads onto genomes. *Nature Biotechnology*, 27(5):455–457, 2009.

[83] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments. *U.C. Berkeley Div. Biostat. Pap. Ser.*, 11(1):94, 2009.

[84] Sabrina Moretti. In Silico Experiments in Scientific Papers on Molecular Biology. *Science Studies*, 24(2):23–42, 2011.

[85] David R. Burgess. Cytokinesis: LET-ting the Asters Signal. *Current Biology*, 17(4):R130–R132, feb 2007.

[86] D Karolchik, R Baertsch, M Diekhans, T S Furey, A Hinrichs, Y T Lu, K M Roskin, M Schwartz, C W Sugnet, D J Thomas, R J Weber, D Haussler, and W J Kent. The UCSC Genome Browser Database, 2003.

[87] Donna Karolchik, Angela S Hinrichs, Terrence S Furey, Krishna M Roskin, Charles W Sugnet, David Haussler, and W James Kent. The UCSC Table Browser data retrieval tool. *Nucleic acids research*, 32(Database issue):D493–6, 2004.

[88] Patricia P Chan, Andrew D Holmes, Andrew M Smith, Danny Tran, and Todd M Lowe. The UCSC Archaeal Genome Browser: 2012 update. *Nucleic Acids Research*, 40(D1), 2012.

[89] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. The sequence read archive. *Nucleic Acids Research*, 39(SUPPL. 1), 2011.

[90] Yuichi Kodama, Martin Shumway, and Rasko Leinonen. The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Research*, 40(D1), 2012.

[91] S Andrews. FastQC: A quality control tool for high throughput sequence data, 2010.

[92] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10:R25, 2009.

[93] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–359, 2012.

[94] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[95] Aaron R Quinlan and Ira M Hall. The BEDTools manual. *Genome*, 16(6):1–77, 2010.

[96] Aaron R Quinlan and Ira M Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[97] J H Christian and J A Waltho. Solute concentrations within cells of halophilic and non-halophilic bacteria. *Biochimica et biophysica acta*, 65:506–508, dec 1962.

[98] Michael T. Madigan, John M. Martinko, David A. Stahl, and David P. Clark. *Brock Biology of Microorganisms, 13th Edition.* 2012.

[99] Rhiju Das and David Baker. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences of the United States of America*, 104(37):14664–14669, 2007.

[100] L. S. ZARAMELA. *Mapping of non-coding RNAs and their interactions with chaperone Lsm in archaea Halobacterium salinarum.* PhD thesis, Faculdade de Medicina de Ribeirão Preto – Universidade de São Paulo, Ribeirão Preto, y, 2015.

[101] Brian G. Pierce, Yuichiro Hourai, and Zhiping Weng. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS ONE*, 6(9), 2011.

[102] Nikulin A.D., Lekontseva N.V., and Tishchenko S.V.

[103] Elizabeth Jurrus, Dave Engel, Keith Star, Kyle Monson, Juan Brandi, Lisa E. Felberg, David H. Brookes, Leighton Wilson, Jiahui Chen, Karina Liles, Minju Chun, Peter Li, David W. Gohara, Todd Dolinsky, Robert Konecny, David R. Koes, Jens Erik Nielsen, Teresa Head-Gordon, Weihua Geng, Robert Krasny, Guo Wei Wei, Michael J. Holst, J. Andrew McCammon, and Nathan A. Baker. Improvements to the APBS biomolecular solvation software suite. *Protein Science*, 27(1):112–128, 2018.

[104] Eric P. Nawrocki and Sean R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

[105] Cameron Smith, Steffen Heyne, Andreas S. Richter, Sebastian Will, and Rolf Backofen. Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA. *Nucleic acids research*, 38(Web Server issue):W373–7, jul 2010.

[106] Sebastian Will, Tejal Joshi, Ivo L Hofacker, Peter F Stadler, and Rolf Backofen. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA (New York, N.Y.)*, 18(5):900–14, 2012.

[107] Sebastian Will, Kristin Reiche, Ivo L. Hofacker, Peter F. Stadler, and Rolf Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS computational biology*, 3(4):e65, 2007.

[108] Jan Engelhardt, Steffen Heyne, Sebastian Will, and Kristin Reiche. RNAclust . pl Documentation. *Therapy*, pages 1–9, 2010.

[109] Ronny Lorenz, Stephan H Bernhart, Christian zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[110] Kristin Reiche. RNAsoup Documentation. (December 2008):1–7, 2008.

[111] C Notredame, DG Higgins, and Jaap Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217, 2000.

[112] Paul D. Thomas, Michael J. Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research*, 13(9):2129–2141, 2003.

[113] C. Tu, X. Zhou, J. E. Tropea, B. P. Austin, D. S. Waugh, D. L. Court, and X. Ji. Structure of ERA in complex with the 3' end of 16S rRNA: Implications for ribosome biogenesis. *Proceedings of the National Academy of Sciences*, 2009.

[114] Marco Mariotti, Alexei V. Lobanov, Bruno Manta, Didac Santesmasses, Andreu Bofill, Roderic Guigó, Toni Gabaldón, and Vadim N. Gladyshev. Lokiarchaeota Marks the Transition between the Archaeal and Eukaryotic Selenocysteine Encoding Systems. *Molecular Biology and Evolution*, 2016.

[115] Peter Zhang, Emmanuel Dimont, Thomas Ha, Douglas J. Swanson, Winston Hide, and Dan Goldowitz. Relatively frequent switching of transcription start sites during cerebellar development. *BMC Genomics*, 18(1), 2017.

[116] Magali Boutard, Laurence Ettwiller, Tristan Cerisy, Adriana Alberti, Karine Labadie, Marcel Salanoubat, Ira Schildkraut, and Andrew C. Tolonen. Global repositioning of transcription start sites in a plant-fermenting bacterium. *Nature Communications*, 7:13783, 2016.

[117] Julia Babski, Karina A. Haas, Daniela Näther-Schindler, Friedhelm Pfeiffer, Konrad U. Förstner, Matthias Hammelmann, Rolf Hilker, Anke Becker, Cynthia M. Sharma, Anita Marchfelder, and Jörg Soppa. Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). *BMC Genomics*, 17(1), 2016.

[118] E. Sauer, S. Schmidt, and O. Weichenrieder. Small RNA binding to the lateral surface of Hfq hexamers and structural rearrangements upon mRNA target recognition. *Proceedings of the National Academy of Sciences*, 2012.

[119] B Cobucci-Ponzano, B Cobucci-Ponzano, M Rossi, M Rossi, M Moracci, and M Moracci. Recoding in archaea. *Mol Microbiol*, 55(2):339–348, 2005.

[120] Thean-Hock Tang, Jean-Pierre Bachellerie, Timofey Rozhdestvensky, Marie-Line Bortolin, Harald Huber, Mario Drungowski, Thorsten Elge, Jürgen Brosius, and Alexander Hüttenhofer. Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. *Proceedings of the National Academy of Sciences of the United States of America*, 99(11):7536–41, 2002.

[121] Thean Hock Tang, Timofey S Rozhdestvensky, Béatrice Clouet D'Orval, Marie-Line Bortolin, Harald Huber, Bruno Charpentier, Christiane Branlant, Jean-Pierre Bachellerie, Jürgen Brosius, and Alexander Hüttenhofer. RNomics in Archaea reveals a further link between splicing of archaeal introns and rRNA processing. *Nucleic acids research*, 30(4):921–30, 2002.

[122] Sean R. Eddy. Computational genomics of noncoding RNA genes. *Cell*, 109(2):137–140, 2002.

[123] P P Dennis, a Omer, and T Lowe. A guided tour: small RNA function in Archaea. *Molecular microbiology*, 40(3):509–519, 2001.

[124] Gisle Vestergaard, Roger a Garrett, and Shiraz a Shah. CRISPR adaptive immune systems of Archaea. *RNA biology*, 11(2):156–67, 2014.

[125] Shengqin Wang, Yuming Xu, and Zuhong Lu. Genome-wide miRNA seeds prediction in archaea. *Archaea*, 2014, 2014.

Appendices

Digital Appendix: 1. A table provides links for the detailed results

Appendix	Link
Predicted TSSaRNAs	https://www.goo.gl/afFa3v
TSSaRNAs 2D topologies	https://www.goo.gl/Z2NmA6
TSSaRNAs 3D cartoon representation	https://goo.gl/Va8U5D
TSSaRNAs 3D surface representation	https://goo.gl/NwnCY2
TSSaRNAs-Lsm table	https://www.goo.gl/BU6d91
Mode of interaction TSSaRNA-cognate gene	https://www.goo.gl/H6TDYf
Rfam annotation -only the minimum E-value-	https://www.goo.gl/2cvSF3
Rfam annotation -all the possible E-values-	https://goo.gl/GGyDt2
TSSaRNA Cognate genes interactions	https://goo.gl/H6TDYf
Table of TSSaRNA Cognate genes interactions	https://goo.gl/ia6vxQ
TSSaRNAs mfold table	https://goo.gl/KCuSyk

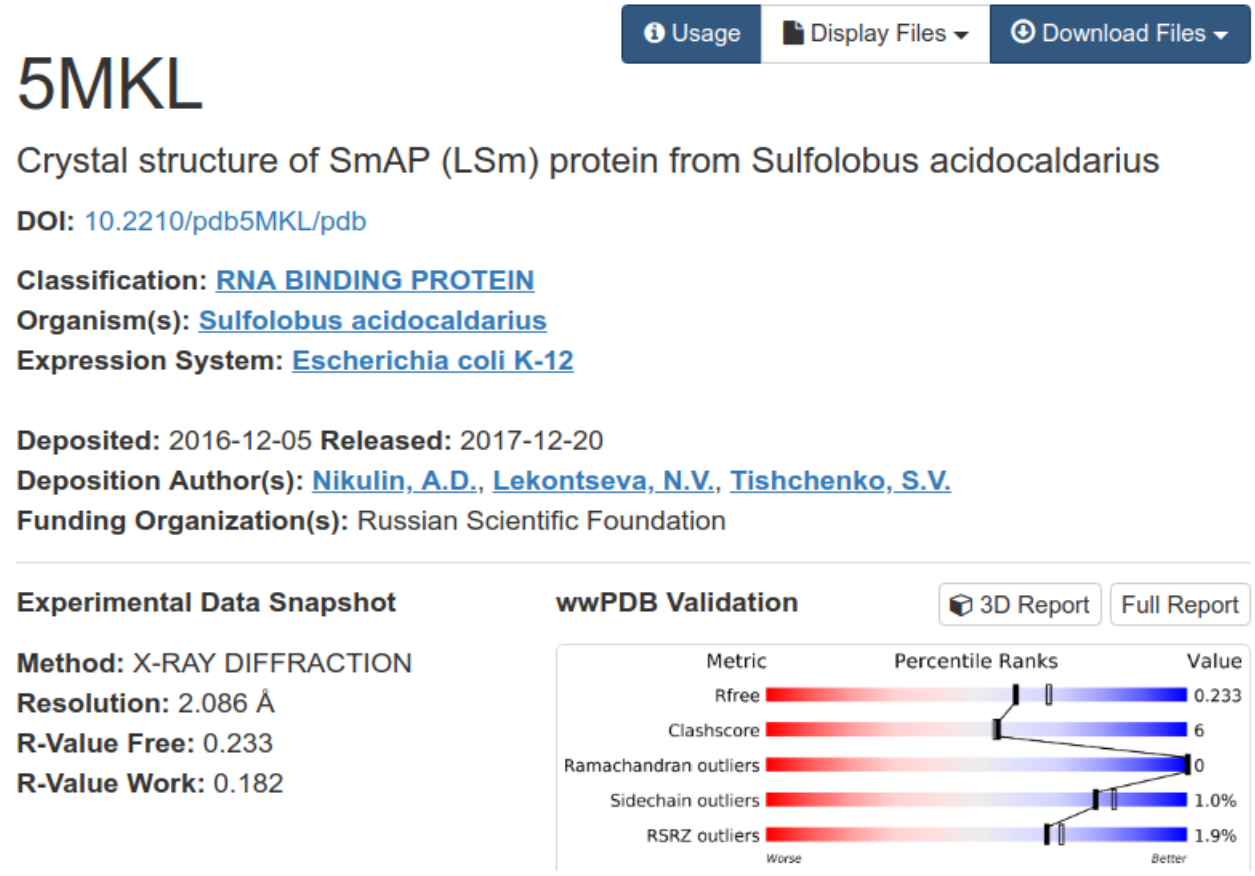
Digital Appendix: 2. A table shows some programming environments and tools that I used intensively during my Ph.D. without mention them explicitly in the main text.

Tool / programming environment	Purpose
Perl programming language	I used Perl language in this project for many purposes which are include but not limited to: a) to automate the query process when downloading the raw data sets; b) to Parallel Computing the computation processes (Parallel Computing) mainly to paralleling the docking experiments; c) we use the power of Perl in editing text for the text mining automation process.
R programming language	I used the R environment for all the calculations and statistical analysis also, I applied the power of The R in the data exploratory analysis.
C++	We used C++11 to overcome the slowness of R, mainly in computation of TSSaRNAs prediction.
Bourne shell (sh)	As the main scripting language
Archaeopteryx v 0.9	I used it to visualize, analysis, and editing of TSSaRNAs cluster trees
Pymol	I used it to visualize TSSaRNAs secondary, tertiary, and higher order structures.
UCSF Chimera	I used it to visualize TSSaRNAs secondary, tertiary, and higher order structures.
Visual studio	As the main text editor and Integrated development environment (IDE) for C++
Geany	As the main text editor and Integrated development environment (IDE) for Perl and R
Textmaker	As the main LaTeX editor
LibreOffice	As the main office suite
Inkscape	As the main program for drawing structured diagrams.

Digital Appendix: 3. A table shows examples of some open source computational tools for RNA-seq processing

Tool	Processing stage	Link
NGS QC Toolkit	Quality control	http://www.nipgr.res.in/ngsqctoolkit.html
QuaCRS	Quality control	http://bioserv.mps.ohio-state.edu/QuaCRS/
RSeQC	Quality control	http://rseqc.sourceforge.net/
FastQC	Quality control	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
RNA-SeQC	Quality control	https://www.osc.edu/book/export/html/4379
HTSeq	Quality control	https://htseq.readthedocs.io/en/
dupRadar	Quality control	https://bioconductor.org
AfterQC	Quality control	https://github.com/OpenGene/AfterQC
SAMStat	Quality control	http://samstat.sourceforge.net/
FASTX	Quality control	http://hannonlab.cshl.edu/fastx_toolkit/
RSeQC	Quality control	http://rseqc.sourceforge.net/
SolexaQA	Quality control	http://solexaqa.sourceforge.net/
Bowtie	Mapping	http://bowtie-bio.sourceforge.net/index.shtml
Bowtie2	Mapping	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
maQ	Mapping (for Illumina-Solexa)	http://maq.sourceforge.net/
tophat	Full pipeline to process RNA-seq Data	https://ccb.jhu.edu/software/tophat/index.shtml
soap2	Mapping	http://soap.genomics.org.cn/soapaligner.html

Digital Appendix: 4. A screenshot of `https://www.rcsb.org` website shows the parameters of 5MKL PDB entery. 5MKL contains crystal structure of LSm protein from *Sulfolobus acidocaldarius*. In this study we used 5MKL version 1.1; to see complete history refer to the link `https://www.rcsb.org/structure/5MKL#entry-history`.



Digital Appendix: 5. Coursework: Academic Achievement and Activities

95131 - 8913607/1 - Yagoub Ali Ibrahim Adam

Sigla	Nome da Disciplina	Início	Término	Carga Horária	Cred.	Freq.	Conc.	Exc.	Situação
IBI5029-1/2	Fundamentos Algébricos para Biologia Sintética e Bioinformática	07/05/2014	29/07/2014	120	8	100	A	N	Concluída
IBI5026-1/3	Biofísico-Química Computacional	04/08/2014	07/11/2014	120	8	100	B	N	Concluída
IBI5044-1/1	Modelos Dinâmicos em Biologia	05/08/2014	27/10/2014	120	8	100	A	N	Concluída
IBI5070-2/1	Seminários em Tópicos Avançados de Bioinformática	14/08/2014	05/11/2014	60	4	100	A	N	Concluída
QBQ5884-1/1	Introdução à Bioinformática – Análise de Dados de Microarranjos de DNA e Sequenciamento de nova Geração (Instituto de Química - Universidade de São Paulo)	15/09/2014	21/09/2014	30	2	100	A	N	Concluída
IBI5045-1/1	Seminários Integrados em Bioinformática	02/03/2015	24/05/2015	60	4	100	A	N	Concluída
5915776-2/3	Fundamentos de Física Estatística e Enovelamento de Proteínas (Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - Universidade de São Paulo)	03/03/2015	05/05/2015	90	6	100	A	N	Concluída
IBI5015-2/1	Processamento e Recuperação de Informação Textual para Bioinformática	12/03/2015	03/06/2015	120	8	100	A	N	Concluída
5935922-3/2	Docência no Ensino Superior: Aspectos Didáticos e Pedagógicos (Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - Universidade de São Paulo)	19/03/2015	28/05/2015	60	4	90	B	N	Concluída
IBI5000-3/1	Seminários em Bioinformática	03/06/2015	25/08/2015	60	4	100	A	N	Concluída
Atividade do Programa	Participou da Etapa de Estágio Supervisionado em Docência do Programa de Aperfeiçoamento de Ensino junto à disciplina - Álgebra Booleana e Aplicações, ministrada aos alunos de Graduação do Departamento de informática Biomédica de Ribeirão Preto - SP - sob a supervisão do prof. Dr. Ricardo Zorzetto Nicolliello Vencio. (2)	22/02/2016	10/06/2016	-	2	100	-	-	-
6025851-2/1	Tópicos Avançados de Pesquisa (Faculdade de Ciências Farmacêuticas de Ribeirão Preto - Universidade de São Paulo) (3)	05/03/2018	11/03/2018	30	2	100	A	N	Concluída
5915792-1/2	Tópicos em Física Médica (Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - Universidade de São Paulo) (3)	27/08/2018	02/09/2018	15	1	100	A	N	Concluída