BIOCOMPUTATIONAL GENOME-WIDE IDENTIFICATION OF sRNA IN *Leptospira interrogans* serovar Lai

TAN XINQ YUAN

UNIVERSITI SAINS MALAYSIA

2019

BIOCOMPUTATIONAL GENOME-WIDE IDENTIFICATION OF sRNA IN *Leptospira interrogans* serovar Lai

by

TAN XINQ YUAN

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

August 2019

ACKNOWLEDGEMENT

First, I would like to express my deepest and heartfelt gratitude to my supervisor, Prof. Dr. Tang Thean Hock for his endless support, guidance and patience throughout this research. His valuable advices have always been an encouragement to me in completing this work. Next, I would like to place my sincere gratitude to my co-supervisor, Dr. Citartan Marimuthu for his continuous guidance and patience throughout this journey. I am indebted to Universiti Sains Malaysia and Advanced Medical and Dental Institute (AMDI) for providing excellent facilities for conducting my research. I also wanted to thank Ministry of Higher Education (MOHE) for providing me financial assistance via MyMasters scholarship under MyBrain15 scheme.

I would also like to place my appreciation to all the members of RNA-Bio Group, Infectomics Cluster, AMDI, especially, my labmates; Madam Siti Aminah Ahmed, Mr. Cheah Hong Leong, Dr. Ang Kai Cheen, Dr. Lee Li Pin, Ms. Presela Ravinderan, Mr. Prabu Siva Sankar, Ms. Rogini Sivalingam, Mr. Yeoh Tzi Shien, Mr. Alvin Paul and Ms. Tan Lee Lee for their great assistance, endless support and heartwarming encouragement throughout this journey. I wish all of them great luck in their future undertakings.

I owe my sincere most appreciation to my loving parents, Mr. Tan Tang Toong and Mrs. Wong Joo Kin who have been continuously motivates me with their unconditional love and patience. Words cannot describe my profound gratitude towards their endless support and encouragement. I wouldn’t have reached this far without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS AND SYMBOLS

A	Adenine
APS	Ammonium persulfate
ATP	Adenosine 5’-triphosphate
bp	Base pair(s)
C	Cytosine
°C	Degrees Celsius
cDNA	Complementary DNA
ddH_2O	Double-distilled water
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	Deoxyribonucleotide triphosphate
dsDNA	Double-stranded DNA
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic Acid
EMJH	Ellinghausen-McCullough-Johnson-Harris
et al.	and others
EtBr	3,8-diamino-5-Ethyl-6-phenyl phenanthridinium Bromide
g	Gravitational acceleration
g	Gram
G	Guanine
HCl	Hydrochloric acid
Hr	Hour(s)
IGR	Intergenic Region
IPTG	Isopropyl-β-D-thiogalactopyranoside
KCl	Potassium chloride
kV	Kilovolts
LB	Luria Bertani medium
M	Mol/Liter, molar
Mg2+	Magnesium ion
Min	Minute(s)
ml	Milliliter
mM	Millimolar
mRNA	Messenger RNA
NaCl	Sodium chloride
NaOH	Sodium hydroxide
ng	Nanogram

nM	Nanomolar
nt	Nucleotide(s)
OD460	Optical density at 460nm wavelength
OH	Hydroxyl
OMP	Outer Membrane Protein
ORF(s)	Open Reading Frames
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
QS	Quorum sensing
RBS	Ribosomal Binding Site
RNA	Ribonucleic acid
RNase	Ribonuclease
RNA-Seq	RNA Sequencing
RNP	Ribonucleoprotein
rpm	Rotations per minute
rRNA	Ribosomal RNA
RT	Room temperature
RT-PCR	Reverse transcription-PCR
s	Second(s)
SD	Shine Dalgarno sequence
sRNA	Regulatory small RNA
S. aureus	*Staphylococcus aureus*
T	Thymine
TAE	Tris–Acetic Acid–EDTA
TBE	Tris-Boric Acid-EDTA
TIR	Translation Initiation Region
Tris	Tris-(Hydroxymethyl)-Aminomethane
tRNA	Transfer RNA
U	Unit
UTP	Uridine 5’-triphosphate
UTR	Untranslated region
utRNA	Untranslated RNA
u.v.	Ultraviolet
V	Volt (s)
v/v	Volume per volume
w/v	Weight per volume
µg	Microgram
µl	micro liter
µM	micro molar

32P	Phosphorus-32
Ω	Ohm
α	Alpha
β	Beta
γ	Gamma
δ	Delta
σ	Sigma
λ	Lamda

PENGENALPASTIKAN sRNA KESELURUHAN GENOM *Leptospira interrogans* serovar Lai MENGGUNAKAN KAEDAH BIOINFORMATIK

ABSTRAK

Regulatory small RNA (sRNA) adalah transkrip RNA yang tidak diterjemahkan kepada protein tetapi bertindak mempunyai fungsi sel. Dalam bakteria, mereka terlibat dalam pelbagai proses modulasi sel biologi, misalnya sebagai tindak balas tekanan umum, penderiaan kuorum dan virulensi. Kepentingan sRNA dalam pelbagai kawalan biologi telah membawa kepada penemuan mereka dalam spesies bakteria seperti *Salmonella enterica, Mycobacterium tuberculosis* dan *Vibrio cholera*, sepanjang dekad yang lalu. Patogenik Leptospira, penyakit zoonosis yang spirochaetal menyebabkan Leptospirosis, satu wabak yang menyumbang kepada kematian dan mobiliti di serata dunia, terutama di Asia Tenggara. Bakteria ini dapat menyesuai-diri di persekitaran tekanan yang pelbagai, contohnya di dalam host yang dijangkiti ataupun di keadaannya nutrien yang rendah. sRNA di jangka memainkan peranan kawal selia dalam fisiologi pada Leptospira patogenik. sRNA telah berjaya ditemui dan diarkibkan dengan kos efektif dan berkesan melalui konsensus serta kemajuan teknologi pengkomputeran. Dalam kajian ini, RNAz-nocoRNAc, satu kejayaan yang terbukti program ramalan sRNA telah dipilih untuk menyaring genom *Leptospira interrogans* serovar Lai. Hasilnya, sebanyak 126 sRNA telah diramalkan. Keputusan dibandingkan dengan data RNA-seq telah dijana dalam kajian terdahulu dan mendapati bahawa 7 sRNA telah ekspres dalam fasa mid-log, pegun, tekanan serum, tekanan suhu dan tekanan besi dan 2 sRNA ekspres dalam semua fasa kecuali tekanan besi. RT-PCR telah mengesahkan semua sRNA ini ekspres dalam fasa mid-log yang dipelihara dalam *Leptospira spp.* fungsi mereka telah diramalkan oleh TargetRNA2 dan salah satu calon telah diramalkan untuk menyasarkan pelbagai ubat rintangan

mRNA. Kajian berfungsi eksperimen adalah wajar pada masa akan datang untuk memastikan peranan pengawalseliaan sRNAs.

BIOCOMPUTATIONAL GENOME-WIDE IDENTIFICATION OF sRNA IN *Leptospira interrogans* serovar Lai

ABSTRACT

Regulatory small RNAs (sRNA) are RNA transcripts that are not translated into protein but act as functional RNAs. In bacteria, they are majorly involved in diverse biological modulating process such as general stress response, quorum sensing and virulence. A myriad of sRNAs were discovered in pathogenic bacteria species such as *Salmonella enterica, Mycobacterium tuberculosis* and *Vibrio cholera.* Pathogenic Leptospira is a widespread spirochaetal zoonosis that causes Leptospirosis, an epidemic disease that accounts for deaths and mobility around the globe, especially in South East Asia. These bacteria are known to adapt to stressful environment such as in the vicinity of infected host or in low nutrient condition to retain their survivability. It is speculated that sRNAs that have pivotal regulatory roles are involved in orchestrating the pathogenicity of Leptospira. In this study, biocomputational-based strategy was adopted to identify sRNAs from the annotated genome of Leptospira. In fact, this is the first study to apply integrated computational-based bacterial sRNA prediction to identify sRNAs in *Leptospira interrogans* serovar Lai, a common pathogenic Leptospira in South East Asia. In this study, RNAz-nocoRNAc, a proven success of sRNA prediction program was selected to screen the genome of *Leptospira interrogans* serovar Lai. As a result, a total of 126 sRNA was predicted. The results were compared to the RNA-seq data generated in previous study and found that 7 sRNA were expressed in mid-log, stationary, serum stress, temperature stress and iron stress phases and 2 sRNA expressed in all these phases except iron stress. RT-PCR has further confirmed all these sRNA was expressed in mid-log phase which were highly conserved in *Leptospira spp*. Their

prospected function has been predicted by TargetRNA2 and one of the candidates were predicted to target multi-drug resistance mRNA. Experimental functional study is warranted in future to ascertain the regulatory roles of the sRNAs.

CHAPTER 1

INTRODUCTION

1.1 Chronology of bacterial sRNA discovery

Regulatory small RNAs (sRNA) occur in all kingdoms of life; archaea, eukaryotes and prokaryotes (Hüttenhofer et al., 2002, Ambros, 2004, Zhang et al., 2006, Babski et al., 2014). Heterogeneous molecules with various sizes and structures, sRNA generally does not encode for protein but are able to perform physiological control of the cells. In bacteria, sRNAs were found to be involved in a myriad of gene regulation such as in glucose metabolism, outer membrane regulation, quorum sensing, virulence and antibiotic resistance (Hoe et al., 2013).

sRNAs were first reported in prokaryotes as early as in 1967, when Hindley (1967) first observed an abundance of unidentified low molecular weight RNA species in their study of *E.coli* tRNA. Besides a single band that constitutes tRNA, they detected another three other components, one of which was found to be ribosomal RNA (rRNA). It was after 4 years, the second components was identified and named 6S RNA (Brownlee, 1971). The function of 6S RNA was reported in 2000 (29 years later), to be involved in the regulation of RNA polymerase activity (Wassarman and Storz, 2000). The third component was later discovered as Spot42 by Ikemura et al. (1973). The function of the Spot42 sRNA was revealed to mediate gal operon, responsible for the carbon metabolism (Møller et al., 2002).

In 1981, first *cis*-antisense sRNA was discovered to be derived from the plasmid and named as RNAI (108nts). RNAI functions to control the copy number of

the ColE1 plasmid by preventing the maturation of ColE1 replication primer from RNAII pre-primer (Tomizawa and Itoh, 1981).

Subsequently, antisense RNAs of the mobile elements that control the copy number of plasmids, transposon and bacteriophages were discovered (Stougaard et al., 1981, Simons and Kleckner, 1983, Krinke et al., 1991). One example is the OOP antisense sRNA in bacteriophage λ, whereby the overexpression of the 77nts sRNA causes two cleavage at two different sites on the *cII* mRNA. As a result, cII protein expression is reduced and consequently bacteriophage λ lytic life cycle was prevented (Krinke et al., 1991).

Chromosomal encoded sRNA (MicF, 93 nts) was discovered in 1989, MicF target *ompF* mRNA and inhibits the translation of major outer membrane, mechanism of which is important in the osmosis regulation of *E.coli* (Andersen and Delihas, 1990, Mizuno et al., 1984).

Majority of the regulatory sRNAs detected in the early phase are housekeeping RNA (tmRNA and RNase P and 4.5 S RNA) which participate in ribosomal rescue, tRNA maturation and protein translocation in *E.*coli (Wassarman et al., 1999). They were initially identified on basis of their high abundance.

1.2 Challenges in the identification of sRNA-based on the physical properties

The major challenge in the identification of sRNAs is the lack of common characteristics that are present in the transcripts. Moreover, the size of the sRNAs that is relatively shorter (40-500 nts) compared to mRNAs, making them almost impossible to be identified from phenotype-derived-based mutational screening. The

secondary structures of the sRNAs also vary from one to another, which lack specific modules or common motifs for the identification (Altuvia, 2007, Hershberg et al., 2003). Their identification is complicated further by the less conservation of the sRNAs among members of the same species compared to protein sequences which have higher level of conservation. Apart from that, due to the absence of open reading frame, sRNAs often escape the identification by the conventional Open Reading Frame (ORF) searching tools, which is based on the codon-based model. Systematic identification of sRNAs relies on two major strategies; the bioinformatics and the experimental approaches.

1.2.1 Experimental sRNA identification

sRNA were first identified in the radiolabeled total RNA from *E.coli*. The labeling of sRNAs enables determination of the expression of sRNA in the cells from the relative band intensities from the autoradiography (Hindley, 1967). However, such labeling method has some drawbacks, mainly due to concerns of the potential health risk associated with handling radioactive cultures (Vogel and Sharma, 2005). Hence more efficacious RNA experimental techniques that could achieve higher detection sensitivity were introduced. They are short-gun cloning, microarray and RNA-seq.

1.2.1(a) Short-gun cloning

Short-gun cloning of RNAs has led to the successful identification of many sRNAs in all domains of life (Marker et al., 2002, Tang et al., 2002, Hüttenhofer et al., 2002, Vogel et al., 2003, Yuan et al., 2003, Tang et al., 2005, Chinni et al., 2010). The main idea of a short-gun cloning protocol involves the directional cDNA cloning and the subsequent sequencing of the cDNA library, which is derived from the size-fractionated total RNA on polyacrylamide gel. This method also allows identification of sRNAs that are expressed under certain environmental culture or growth stages of the bacteria without any prior knowledge of the transcript. However this approach is labor intensive, plus it also requires intensive computational cDNA library assembly (Nielsen et al., 2010).

1.2.1(a) Microarray

Microarray is a probe-based technique that can be used to identify sRNAs on a genome-wide scale. This technique had enabled successful identification of sRNAs from intergenic regions which are often referred to as the hotspot for sRNAs (Tjaden et al., 2002). It has successfully identified sRNAs in *E.coli* (Wassarman et al., 2001) and other bacteria (Valverde et al., 2008, Ahmad et al., 2012, Gierga et al., 2012). However, the most significant disadvantages of microarray are the financial cost and the off-target effect of the probes that does not guarantee detection of sRNA (Jaksik et al., 2015).

1.2.1(b) RNA-seq

RNA-seq (RNA-sequencing) was first pioneered for the transcriptomic profiling (RNA expression) of yeast (Nagalakshmi et al., 2008). In brief, this approach consists of few critical steps; (1) isolation of total RNA (2) RNA species enrichment (removal of highly abundant rRNA) (3) fragmentation (4) conversion of RNA to first-strand cDNAs and adaptor ligation (commercial adaptor). Subsequently the ligated cDNAs are amplified and sequenced via high-throughput sequencing with the adaptor specific primers (Wang et al., 2009) . This method can generates millions (ranging between 25 to 300 bps) of reads (Shendure and Ji, 2008, Oshlack et al., 2010).

Though it provides mapping of RNA at a single nucleotide resolution, identification of RNA of low abundance remains a challenge as the RNA could be lost during library preparation or the due to the sequencing bias that is characteristics of (G+C) rich transcripts (Shendure and Ji, 2008, Raabe et al., 2014). Moreover, RNA-seq is very expensive considering the cost required to manage, process and analyze the data (Schadt et al., 2010).

1.2.2 Biocomputational-based screening of sRNA candidates- a much more cost effective and less labor-intensive approach

Alternatively, sRNAs could be identified by genome-wide scale sRNA computational screen; a much more cost effective, less labor-intensive approach. With the currently available complete genome and improved annotation such as Ref-seq (reference sequence) deposited in NCBI (National Center for Biotechnology), *in silico* sRNA prediction has become much more expedient.

There are four major features of sRNAs that can be capitalized for the prediction of sRNAs from the genomic sequence of the bacteria, which include (a) comparative genomics (b) secondary structure and thermodynamic stability (c) 'Orphan' transcriptional signals in intergenic regions and (d) ab initio method (Argaman et al., 2001, Rivas et al., 2001, Wassarman et al., 2001, Sridhar and Gunasekaran, 2013).

1.2.2(a) Comparative genomics-based approach

Comparative genomics is grounded on the genomes which are highly conserved among several bacterial families (Lindgreen et al., 2014). This is combined with the prediction of sRNAs based on consensus secondary structures, for instance QRNA program (Rivas and Eddy, 2001). This program employs probabilistic model to detect RNA region and protein coding region by pairwise alignment analysis. Furthermore, it also computes sRNAs by covariance-based mutational analysis. Apart from that, another extensive comparative genomic method such as "Infernal" was developed based on Rfam, a database of RNA alignment and covariance model (Griffiths-Jones et al., 2005, Nawrocki et al., 2009). This program works by building a consensus secondary structure covariance model from an 'RNA family' and searching across the Rfam database for sequence and structure similarities. This results in identification of putative homology sRNA family from the Rfam database. A variety of sRNA species has been successfully identified by comparatives genomics of several bacteria genomes (Wassarman et al., 2001, Pánek et al., 2008, Chen et al., 2011).

1.2.2(b) Computational Thermodynamic-based approach

sRNAs or functional RNAs form favorable minimal free energy (MFE) –based secondary structures, which have a lower thermodynamic free energy (Washietl et al., 2005). Thermodynamic-based approach is often applied in conjunction with structure conservation screening to separate reliable sequence from random sequence. One of the prominent tools such as RNAz, adopts both secondary structural conservation and thermodynamic stability of sRNAs for the efficient detection of sRNA in alignment of few genomes (Gruber et al., 2010). This tool has successfully aided in the discovery of sRNAs in *M. tuberculosis* (Wang et al., 2016), *Shigella* (Peng et al., 2011) and *Pseudomomas aeruginosa* (Sonnleitner et al., 2008).

1.2.2(c) Transcription signal-based method

Intergenic regions are the hotspots for identification of sRNA which is believed to harbor stand-alone genes, 'Orphan' transcription factor binding site/ promoter signals and terminator signals (Livny et al., 2008, Sridhar et al., 2010, Herbig and Nieselt, 2011). SIPHT is the one of the computational tools that predicts sRNAs that are derived from the intergenic regions. Putative sRNAs are predicted based on the promoter signals, transcriptional factor binding sites and rho-independent termination signals screened by TRANSTERMHP and BLAST (Livny et al., 2008). Similar strategy was also applied by another sRNA prediction program, nocoRNAc. However instead of predicting the transcription factor binding sites, this program computes SIDD (Stress Induced Duplex Destabilization) of the genome to identify potential promoter signals (Herbig and Nieselt, 2011).

1.3 Classifications of Bacteria sRNA

In general sRNA can be classified into *cis*-encoded or *trans*-encoded sRNA based on their location to the target (Figure1.1).

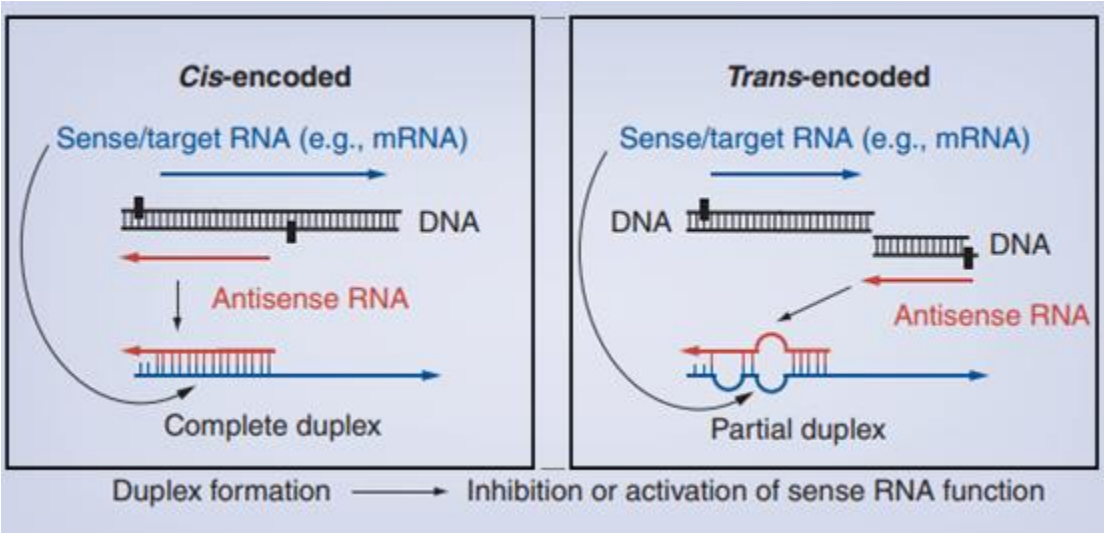

Figure 1.1: Mechanism of *cis*- and *trans*-encoded sRNA. Antisense RNAs are drawn in red and sense RNAs are drawn in blue. The left panel shows perfect complementary base pairing between *cis*-encoded sRNA with its target. Non-contiguous *trans*-encoded sRNA possess partial complementary with its target shown in the right panel. (adopted from Brantl, 2012)

1.3.1 *trans*-encoded RNA

In bacteria, *trans*-encoded sRNAs are usually distantly located from their target mRNAs. These sRNAs usually form imperfect base pairing to their complementary mRNA. Consequently, one *trans*-encoded sRNA could potential have more than one target mRNA (Waters and Storz, 2009). Due to the imperfect base pairing their target mRNAs, these sRNAs generally require the aid of RNA chaperons such as Hfq, S1, StpA or Ro, ProQ (Pichon and Felden, 2007, Smirnov et al., 2016). Among them, Hfq is the most prominent and well-studied RNA chaperone that strengthen the base pairing interaction by remodeling the RNA structures (Storz

et al., 2004). The interaction may result in positive or negative regulation of mRNA stability (Waters and Storz, 2009). The diverse roles of *trans*-encoded sRNA are discussed below.

1.3.1(a) Inhibition of translation

Translational repression has been generally recognized as the main mode of action of *trans*-encoded sRNA. Regulation is achieved by complementary base pairing of the sRNA to the target mRNA, particularly targeting the ribosomal binding site at the 5' untranslated region (UTR) that leads to the inhibition of the translation. One example is manifested by OxyS, which targets *fhla* mRNA, by overlapping the Shine-Dalgarno sequence, leading to the translational repression (Argaman and Altuvia, 2000). Some others sRNAs such as MicF and RyhB also perform translational repression at the upstream of (AUG) codon of their respective target mRNA (Figure 1.2) (Andersen and Delihas, 1990, Večerek et al., 2007).

1.3.1(b) Activation of translation

Apart from translational repression, *trans*-encoded RNA also performs their function as a translational activator. In the inhibitory state, 5' UTR of some mRNA folds into a secondary structure that sequesters ribosomal binding site of the mRNA, inhibiting the translational process. The complementary base pairing of the *trans*-encoded sRNA converts the translational-inhibitory secondary structure to a translational-permissive secondary structure that allows the binding of the ribosomal RNA to the Shine-Dalgarno sequence that fuels translation. For instance, DsrA and RprA are known to serve as a translational activator of rpoS, a sigma factor

responsible for the adaptive response of the bacteria in *E.coli* (Majdalani et al., 2002, McCullen et al., 2010). rpoS mRNA has a long 5' UTR (~600 nts) which folds into a self-inhibitory translational inactive structure. The binding of the sRNA at the upstream region of the *rpoS* mRNA leads to the collapse of the self-inhibitory stem-loop structure. This allows the binding of ribosome to the RBS, driving the translation (Figure 1.2) (Majdalani et al., 1998).

1.3.1(c) Occlusion of protein activity

Some *trans*-encoded sRNAs interact with cellular protein. These sRNAs bind to their target protein, this in turn blocked the binding of the protein to their accociate targets. A typical example is 6S RNA, first found in bacteria, specifically associate with RNA polymerase holoenzyme that contains sigma 70 factor (Wassarman and Storz, 2000). 6S RNA mimics B-form DNA, a form DNA helices whereby major groove is wider than the minor groove (Chen et al., 2017). This sRNA-protein interaction results in the occlusion of the binding of the downstream sigma70 RNA polymerase holoenzyme dependent genes in stationary phase. Another example is CrsB/ CrsC sRNA that antagonize protein CrsA. CrsA protein is a repressor of glycogen synthesis and catabolism, biofilm synthesis, motility and cell surface-protein (Romeo, 1998). As CrsB/ CrsC accumulates, these sRNAs binds to CrsA, thus reversing its role as a repressor (Liu et al., 1997). This results in the upregulation of the downstream genes, which are otherwise repressed by CrsA protein (Figure 1.2).

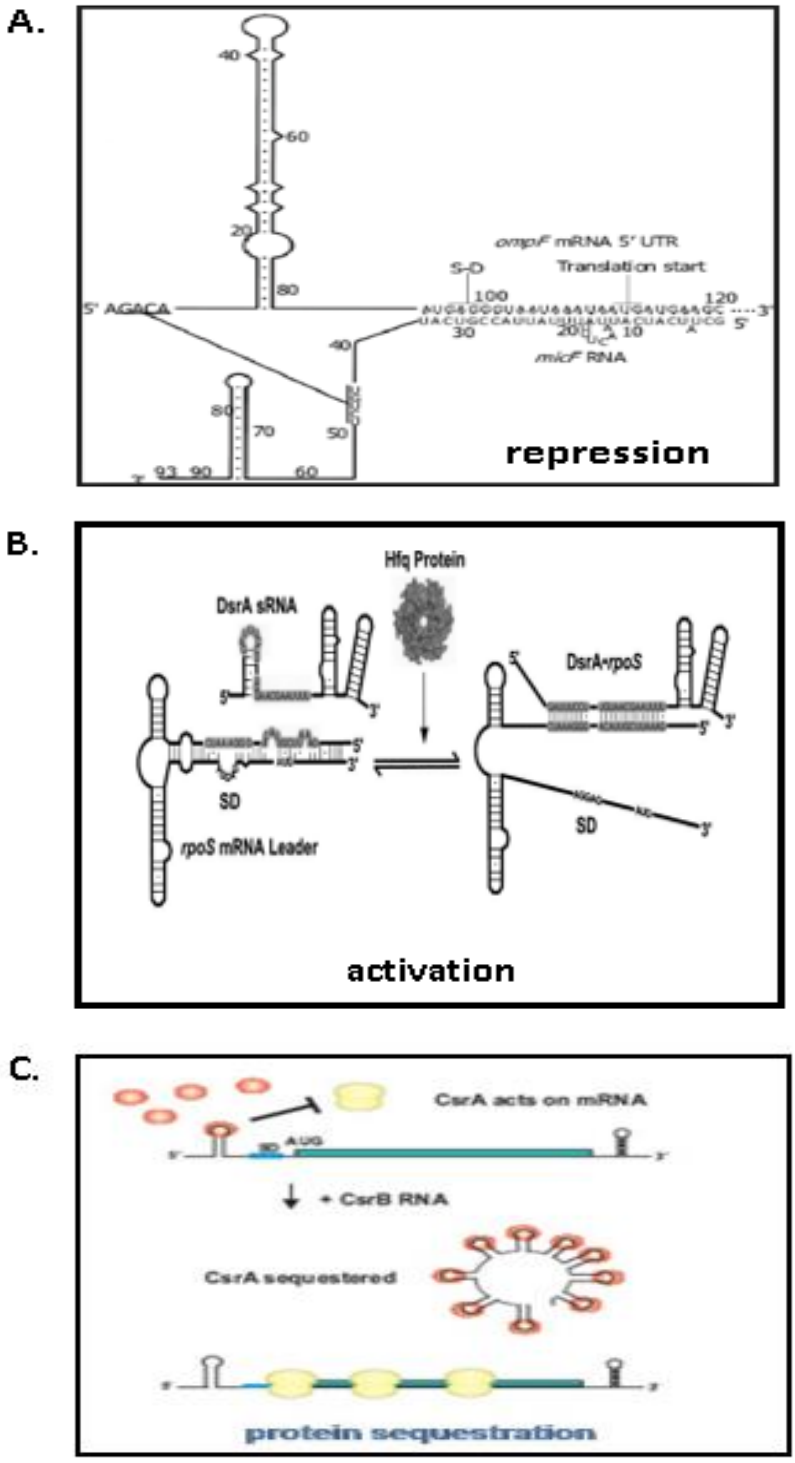

Figure 1.2: Schematic presentation of gene regulation by *trans*-encoded sRNA. **(A)** sRNA mediated translation repression. **(B)** sRNA mediated translational activation (adopted from Delihas & Forst, 2001). **(C)** Sequestration of CsrA proteinby CsrB. The binding of CsrA (red circles) to the hairpin of the mRNA leads to translational repression. (adopted from Wassarman, 2007).

1.3.2 *Cis*-encoded sRNA

Located at the opposite strand of the target mRNAs, *cis*-encoded sRNAs are fully complementary to their target mRNAs. They do not require helper protein such as Hfq. Although increasing numbers of *cis*-encoded sRNA being reported, their mode of mechanism still remains to be elucidated (Georg et al., 2009).

1.3.2(a) Alternation of target stability

Majority of antisense *cis*-encoded sRNAs with known functions negatively regulates translation or mRNA degradation (Opdyke et al., 2004). For instance, in *E.coli*, a 105 nts sRNA GadY, act as an activator of the glutamate dependent acid response system (*gadXW*). GadY is located at the intergenic region between GadX and GadW bis*cis*tronic operon. Interestingly GadW that is located immediate downstream of GadX has its own promoter and act as an independent transcript (Ma et al., 2002, Tramonti et al., 2008). GadY sRNA binds at the 3'-UTR of the *GadX* mRNA and also to the transcribed intergenic region between *GadX* and *GadW* mRNA. The resulting duplex is subject to cleavage by RNAse III or RNAse E which enhanced the stability of the *GadX* as they are protected from cleavage by GadY (Opdyke et al., 2004). As a result, more GadX protein is accumulated. Study has shown that significance reduction survival rate and decrease amount of GadY and GadX transcript under RNAse E knockout strain in acid stress (Takada et al., 2007). However, in contrast, another study shown that only RNAse III is involve in GadY dependent cleavage, suggesting that involvement of RNAse E in cleavage machinery could be growth-stage dependent manner (Opdyke et al., 2011).

1.3.2(b) Transcription termination

Other than posttranscriptional mechanism, *cis*-encoded sRNA is also shown to influence the transcriptional process of the target gene. The best example for this mechanism is the regulation of iron transport biosynthesis operon in *Vibrio anguillarum* that resides in plasmid pJM1. This operon consists of four ferric siderophore transport genes (*fatDCBA*) and two siderophore synthesis gene *(angR and angT).* The opposite strand of the operon gives rise to two *cis*-encoded antisense RNA (RNAα and RNAβ). Under iron-rich condition, a 650 nts *cis*-encoded antisense sRNA, RNAα shown to repress the complementary gene *fatA* and *fatB* by modification of the secondary structure of the poly*cis*tronic (*fatDCBA*) mRNA (Stork et al., 2007). However, under iron-deficiency, RNAβ is expressed and the interaction of RNAβ to the 3' region of the *fatA* leads to transcription termination, causes different transcript of the full-length operon (*fatDCBA-angRT*). This results in accumulation of short form transcript *fatDCBA* about 17 times more abundant than *(angR and angT)* (Waldbeser et al., 1993). Similarly in *Shigella flexneri*, a *cis*-encoded antisense sRNA RnaG also shown to facilitate premature termination of *icsA* mRNA which encodes virulence protein requires in host invasion (Giangrossi et al., 2010).

1.3.2(c) Modulation of translation

Cis-encoded antisense sRNA can regulate the translation of the expressed gene. For example, SymE, a *cis*-encoded antisense sRNA was found responsible in regulating Sos response induced protein *SymR*. SymE complementary binds the 5' UTR of the SymR which also occludes the ribosome binding site (RBS) and the start codon (AUG). The duplex formation of SymR and SymE disrupts the access of the

30s ribosome; thereby inhibits the translation process (Kawano et al., 2007).

1.4 Riboswitches and RNA thermometers

Riboswitches are generally located at the 5' UTR region of the mRNA, it serves as the modulator of gene expression by forming complex structure fold. This structure is highly selective receptors for many class organic compounds (Henkin, 2008). A well-known example for riboswitches is TPP riboswitch, also known as THI element and Thi-box riboswitch. It was first described at the 5' UTR of *thiC* which responsible for thiamin biosynthesis (Miranda-Ríos et al., 2001). At low TPP concentration, the riboswitches at the 5' UTR is not fully folded, thus gene is expressed, whereas at high TPP concentration TPP binds and induces the clustering of Shine-Dalgarno sequence, ultimately resulting inhibition of gene (Figure 1.3) (Guedich et al., 2016).

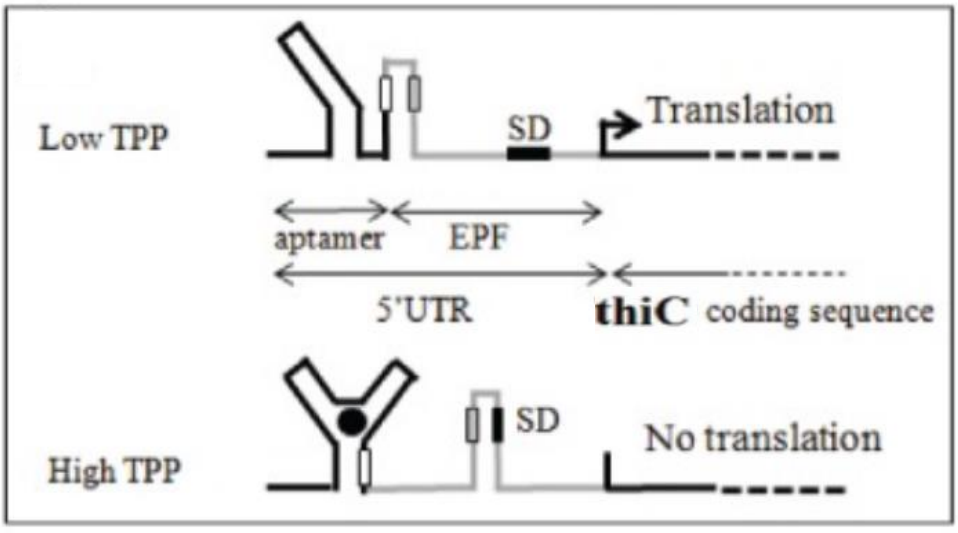

Figure 1.3: Regulation and mechanisms of thiamine pyrophosphate riboswitches. Riboswitches are located in the 5'-UTR and control transcription (thiC) of the downstream genes; EPF (in gray): expression platform, Black circle: TPP (adopted from Guedich et al., 2016).

RNA thermometers (RNAT) also generally found at 5’ UTR of genes, but unlike riboswitches, they act on the level of mRNA stability caused by temperature change (Roßmanith and Narberhaus, 2016). It was found important to regulate gene that is involve in respond to heat shock and cold shock in respond to temperature change. It mask the ribosome binding site at low temperature and the structure melts at high temperature (Narberhaus et al., 2006).

1.5 sRNA involved in stress response and virulence mechanism

Recently, sRNAs were implicated as a vital modulator of gene expression in response to an ever-changing environment. They have been shown to be involved in orchestrating different physiological stress responses such as oxidative stress, pH stress, temperature fluctuation, iron homeostasis and outer membrane protein (OMP) stress (Hoe et al., 2013). For instance, OxyS, a109 nts sRNA serve as a regulator involves in the adaptation of hydrogen peroxide by protecting cells against oxidative damage. OxyS believed to regulate at least 40 genes, including negative regulation of fhlA, a transcriptional activator for hyp operon encoding necessary proteins for the maturation of [NiFe] hydrogenase (oxidative enzyme) by forming two kissing complexes at the 5’ UTR of fhlA mRNA masking the RBS (Argaman and Altuvia, 2000).

Besides that, sRNAs are also known to participate in regulating the virulence genes expression in several pathogenic bacteria, such as *Staphylococcus aureus*, *Streptococcus pyogenes*, *Clostridium perfringens* and *Vibrio cholera* (Novick et al., 1993, Carter et al., 2014, Pérez-Reytor et al., 2017). For example, RNAIII, a 514 nts sRNA in *S. aureus* serves as a sensor for population density and virulence to animal

models. It is a transcription unit in the *agr* system, that controls the early expression of surface protein and late expression of endotoxin, but also encodes for δ-hemolysin (hld) (Novick et al., 1993, Morfeldt et al., 1995). The base paring of the 5' end of the RNAIII to the 5' UTR of the *hla* mRNA that encodes for α-hemolysin, promotes it translation. On the other hand, the 3' end and the central domain of this sRNA repress the translation of the mRNA that encodes for the major cell surface protein (fibrinogen-binding protein/ adhesin, cell-surface protein A) and major pleiotropic transcription factor, Rot by binding to the translation initiation region (TIR) (Boisset et al., 2007, Geisinger et al., 2006).

The 514 nts RNAIII sRNA, has revealed its importance in regulatory mechanism of gene expression involved in response to the environment signals and to control virulence in pathogenic *S. aureus*. In spite of this, pathogenic bacteria, *Leptospira spp.* could adopt a similar strategy to regulate its virulence mechanism and retain their viability in harsh environment.

1.6 Pathogenic *Leptospira spp.*

Leptospira spp. are gram-negative bacteria that belong to Spirochetes family. They have a thin, helically coiled morphology and are usually 6–20 μm in length. This bacteria was first observed in 1907 by silver stained kidney tissues which was misdiagnosed as a case of yellow fever (Stimson, 1907). During that time Stimson named the bacteria as *Siprochetes interrograns* because their resemblance of hook ends. The etiology of Leptospira is then demonstrated independently in German and Japan. In Japan, specific antibodies and spirohetes (*Siprochaeta icterohaemorrhagiae*) were detected by Inada and Ido (1916) in the blood of Japanese miners with

infectious jaundice, whereby two groups of German physicians studied German soldiers with similar symptom (Flügge, 2009). The genus Leptospira was later proposed by another Japanese scientist which differed from other Spirochetes, meaning ‘slender coil’(Noguchi, 1918).

Traditionally *Leptospira spp.* can be classified into saprophytic species (*Leptospira biflexa*) and pathogenic species (*Leptospira interrogans*). DNA hybridization analyses have shown that there are at least 19 species (13 pathogenic and 6 saprophytic) (Mohammed et al., 2011). These species of Leptospira can be further classified into 24 serogroups and 250 serovars based on the agglutination test of the surface-exposed lipopolysaccharide (LPS) (Bharti et al., 2003, Adler and de la Pena Moctezuma, 2010).

Pathogenic Leptospira causes Leptospirosis and each year there is an estimated of over 853,000 cases of leptospirosis, responsible for about 48,000 deaths (Bandara et al., 2014). Clinical symptoms of leptospirosis include fever, abdominal pain, jaundice, myalgia and meningitis, and even death (Taylor et al., 2015).

Pathogenic Leptospira is transmitted to human from the water contaminated by urine of infected rodent reservoirs (Adler et al., 2011). Pathogenic Leptospira cause asymptomatic infection in maintenance host animals. After being shed in the urine, most of the pathogenic Leptospira species are able to adapt in a poor nutrient conditions, such as moist soil, natural bodies and even capable to avoid or counteract innate immunity as soon as they overcome the initial barrier of the human host (Crawford et al., 1971).

1.6.1 Adaptation of pathogenic Leptospira in stressful environment

Pathogenic Leptospira is known to be frequently exposed to a drastically changing environment and is able to adapt to harsh environment to retain its viability. Adaptive response of the pathogenic *Leptospira spp.* has been analyzed by microarray to determine the changes of gene expression at the transcriptomic level in a response to osmolarity, temperature, iron depletion and interaction with phagocytic cells. These parameters are relevant during host invasion or under low nutrient environment (Lo et al., 2006, Qin et al., 2006, Matsunaga et al., 2007, Xue et al., 2010).

The complete genome of *L. interrogans* serovar Lai has been available in NCBI since 2003 (NC_004342.2 and NC_004343.2). In general, the genome of *L. interrogans* serovar Lai consists of two chromosomes; chromosome I and chromosome II. These two chromosome were used as the model for sRNA discovery, in which both chromosomes harbors total of 3683 protein coding genes including proteins related to adhesion, invasion and the haematological changes that characterize leptospirosis and could be regulated by sRNAs (Ren et al., 2003).

The transcriptomic analyses have revealed that the pathogenic Leptospira is capable to modulate transcriptomic signals in response to diverse array of environment condition. Currently, only 366 sRNAs were identified in pathogenic *L. interrogans* serovar Manilae, among them only 8 sRNAs have been experimental validated by northern blot analysis, whereby 2 sRNAs were proposed to prevent the translation of the mRNAs that encode for surface exposed lipoprotein of Lipl21 and Lipl32 (Zhukova et al., 2017). A host of other sRNAs that are potentially involved in orchestrating biological pathway remain to be discovered. As a result, a much more exhaustive identification of sRNAs in pathogenic Leptospira is required.

1.7 Research Objective

This study aims to discover sRNAs in *Leptospira interrogans* serovar Lai via biocomputational approach.

The objectives of this study:

1) To identify *Leptospira interrogans* serovar Lai sRNAs via computational approach

2) To experimentally expression validate the predicted sRNA candidates

3) To characterize the validated sRNA candidates

CHAPTER 2

MATERIALS AND METHODS

2.1 Chemicals and reagents

Chemicals and reagents utilized in this study are listed below in the alphabetical order according to the manufacturer's name. All chemicals were of analytical grade.

Table 2.1 Chemicals and reagents

Manufacture	**Chemicals**
Ambion ® (Austin, USA)	RNase Away ™ reagent
BD (New Jersey, USA)	Difco™ Leptospira Enrichment EMJH; Difco™ Leptospira Medium Base EMJH
Biotools (Madrid, Spain)	*Taq* DNA Polymerase; PCR buffer, MgCl
Bio Basic Inc (Markham, Canada)	EDTA, Ammonium persulphate (APS)
Bio-Rad (Hercules, USA)	Ethidium bromide solution, 10mg/ml, 30% Acrylamide and bis-acrylamide solution 29:1; TEMED
Invitrogen® (Carlsbad, CA)	UltraPure ™ Tris
Molecular Research Center (MRC), Inc	TRI Reagent ®

(Ohio, USA)	
Promega (Madison, USA)	Agarose LE; 100bp DNA ladder; 25bp DNA step ladder
Roche (Mannhiem, Germany)	DNAse 1 Recombinant; Transcriptor First Strand cDNA Synthesis Kit
Sigma-Aldrich (Missouri, USA)	Chelex®100

2.2 Buffers/ Solutions

All buffers used in this study are listed below in the alphabetical order.

Table 2.2 Buffers and Solutions

Buffer/ Solutions	**Components**
50X TAE	242g Tris base; 57.1ml acetic acid; 100ml 0.5M EDTA (pH 8.0); top up with ddH_2O to a final volume of 1L
DNA loading dye (6x)	30% (v/v) glycerol.; 0.25% (w/v) bromophenol blue top up with ddH_2O to a final volume of 10ml

2.3 Bacterial Strains

Table 2.3 Bacterial Strains

Strain	**Source/ reference**
Leptospira interrogans serovar Lai	Makmal Kesihatan Awam Ipoh (MKAI), Ipoh, Malaysia

2.4 Culture Media

Table 2.4 Culture Media

	Ingredients	**Source**
EMJH medium	2.3g of Difco™ Leptospira Medium Base EMJH in 900ml of ddH20 (autoclaved); 100ml of Difco™ Leptospira Enrichment EMJH	BD (USA)

2.5 Primers used for RT-PCR analysis

Online software Primer-BLAST

(http://www.ncbi.nlm.nih.gov/tools/primer-blast/) was used for the design of the primers. The primers were purchased from 1st BASE Pte Ltd (Singapore). Listed below are the PCR primers.

Table 2.5 Primer used for RT-PCR

Primer **(F=forward primer;** **R=reverse primer)**	**Sequence (5' to 3')**	**sRNA**	**Amplicon size**
Lai1_1365_F	TGAGCATACCCGATCGGAAC	Lai1_1365	60 bps
Lai1_1365 R	AATCACCGCGGTTCAAGATG		
Lai1_1526_F	GTTGGGGGTGGGATTCACTG	Lai1_1526	80 bps
Lai1_1526_R	CCGTTTTCCTACGCAGGTCT		
Lai1_1763_F	AACACGGGACCGGGTAATTC	Lai1_1763	167 bps
Lai1_1763_R	AGGGTTTGCCATTCCCGATT		
Lai1_3021_F	ACCTGGTTCGAAGGTATGGA	Lai1_302	92 bps
Lai1_3021_R	TCACTTCCATCTCCAAGGCG		
Lai1_3635_F	TGGAATCAGTTAATCCCTTGAGAA	Lai1_3635	122 bps
Lai1_3635_R	GATCGATGAGGCGGGAAGAG		
Lai1_3741_F	GAATCGAACCCCCGACCTTT	Lai1_3741	71 bps
Lai1_3741_R	CTTCTCAAGCGGGGACGTAG		

Lai1_5029_F	TCTGGATGAATCCCCAGATCG	Lai1_5029	61 bps
Lai1_5029_R	TCGGATTTTCCGAAGCGACT		
Lai1_6129_F	ACACGGAGGAATAGTCAATGGT	Lai1_6129	166 bps
Lai1_6129_R	AGCGGGTTTTAAAATTTCATCCCT		
Lai2_382_F	ATTCTGTGAAATTGCAAGCGAGT	Lai2_382	60 bps
Lai2_382_R	CGAGGTTTAATCCCGATTTCTTCCT		
Lai1_16s_F	GCGTAGGCGGACATGTAAGT	16s_rRNA	211 bps
Lai1_16s_R	AATCCCGTTCACTACCCACG		

2.6 Hardware analysis tools

Analysis was carried out by using laptop. Specification and details of laptop are:

- Manufacture: Asus
- Model: Asus Laptop A43S
- Processor: Intel Core i5-2450M @2.5 GHz
- RAM: 4GB
- System type : 64 bit operating system
- Operating system: Ubuntu GNOME 16.04

2.7 Databases

Rfam

http://rfam.xfam.org/

Multiple Sequence Alignment

https://www.ebi.ac.uk/Tools/msa/

BLASTn Non-Redundance Nucleotide Database

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_PROGRAMS=megablast&PAGE_TYPE=BlastSearch

Bedtools

http://bedtools.readthedocs.io/en/latest/

Primer-BLAST

http://www.ncbi.nlm.nih.gov/tools/primer-blast/

NCBI database

ftp://ftp.ncbi.nlm.nih.gov/

RNA Secondary Structure Alignment (LocaRNA)

(http://rna.informatik.uni-freiburg.de:8080/LocARNA/Input.jsp)

TargetRNA2

http://cs.wellesley.edu/~btjaden/TargetRNA2/

Mauve

http://darlinglab.org/mauve/mauve.html

RNAz

https://www.tbi.univie.ac.at/software/RNAz/