Identification and Characterization of small RNAs in *Escherichia Coli*

By

Rebecca Dan Zhu

A Thesis

Submitted to the School of Graduate Studies

In partial Fulfillment of the Requirements

For the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2008) **McMaster University**

(Biochemistry and Biomedical Sciences) **Hamilton, Ontario**

TITLE: **Identification and Characterization of Non-coding RNAs in *Escherichia coli***

AUTHOR: **Rebecca Dan Zhu, B. Sc. (University of Western Ontario)**

SUPERVISOR: **Professor Yingfu Li**

NUMBER OF PAGES: ix, 111

ABSTRACT

Until a little over a decade ago, the regulatory roles of small RNAs (sRNAs) in prokaryotes were largely undetected. Since then, there has been an explosion in the discovery of novel sRNA sequences and we have begun to understand their functions and mechanisms of regulation. The identification and characterization of sRNAs from different organisms have largely been achieved through computational and experimental approaches that focus on sequence elements in intergenic regions. Based on these previously established techniques, we have developed and applied a new bioinformatics approach to search for highly conserved sequences in unannotated intergenic regions from several bacterial genomes, which may contain new sRNA sequences. Through this search, we have identified seven candidate sequences that are conserved at the primary sequence level, and some of the secondary structure motifs are also conserved among multiple bacteria genomes. When we examined those seven candidates experimentally, it was found that when the expression of one mutated candidate (rUIG0803_4D) was induced at the RNA level, minor morphological changes and a delayed lethal phenotype was elicited. The expression of the RNA also may result in the altered expression of kanamycin kinase and glycerol kinase, as indicated by the mass spectrometry data. Experimental characterizations of eight previously identified sRNAs from literature with functions unknown have also been performed but no apparent phenotypic phenomenon was observed in this project, which indicated that all or some of those 8 sRNAs might not play any regulatory roles in cells, or their roles need to be characterized through other genetic screens. To further search for RNA sequences with regulatory functions, we created a library of random DNA transcript using the Lambda Phage genomic DNA. Preliminary screening efforts show that three of the 192 clones screened could trigger reduced cell growth when their RNA was overexpressed. This study marks the first use of a bioinformatics approach that uses primary sequence and secondary structure information to search for sRNAs in the unannotated intergenic region. Moreover it also marks the first time that the effects of introducing random lambda phage RNA in an *E. coli* host.

ACKNOWLEGEMENTS

First, I would like to express my sincere appreciation to Professor Yingfu Li, for his supervision on this thesis. The guidance, support, and encouragement he gave me have become the foundation of my work. I am also grateful to Dr. Li for giving me invaluable advice on improving my writing and presentation skills and encouraging me to think creatively. The knowledge and advice will be extremely useful for me in my career development.

I also wish to acknowledge Professor John A. Hassell for giving me tremendous help on my experimental design and scientific thinking and for allowing me access to the equipment in his lab. My appreciation goes to my committee member Professor Zhou Xing for his encouragement and suggestions. I would also like to thank Professor David Andrews and Professor Daniel Yang for allowing my access to their equipment.

I am indebted to many co-workers during the completion of my thesis. Particularly, I would like to thank Dr. Naveen Kumar for sharing his knowledge of microbiology and helping me setting up experiment as well as data interpretation, Weian and William in DNA related experiments, Simon and Wendy for very helpful suggestions on my writing, Casey, Dr. Ali and Kacper for many helpful discussions, Michael, Victor and Katherine with help for SDS-PAGE gel, Jeff and Jim in all the help related to computer issues, Courtney and Bobbijo and Brigitte for their assistance on my experiments.

I would like to acknowledge my colleagues in the Department of biochemistry for their friendship, help, and sharing many great times. Many thanks go to Ms. Dale Tomlinson, Ms Lisa Kush, Ms Mary Margret for their help during my study.

I also want to thank my brother, Lin Zhu for his great help in the bioinformatics part. I also appreciate the funding from Dr. Yingfu Li and Dr. John Hassell supporting my study in Bioinformatics area.

Finally, this work is dedicated to my parents and my dearest brother for their support and encouragement during the past years and to my husband, Shunxing Su, for his love, encouragement, and understanding.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

Ara	Arabinose
aTc	anhydrotetracycline
Cm	chloromphenical
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleoside 5′-triphosphate
E. coli	*Escherichia coli*
Hfq	Host factor I protein
hr	hour
LB	Luria-Bertani broth
MCS	multiple cloning site
min	minutes
miRNA	microRNA
ml	mililiter
mRNA	messenger RNA
ncRNA	non-coding RNA
ng	nanogram
OD	optical density
ORF	open reading frame
PBS	phosphate buffered saline
PCR	polymerase chain reaction
RBS	ribosomal binding site
RNA	ribonucleic acid
rRNA	ribosomal RNA
SRP	signal recognition particle
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
snoRNA	small nucleolar RNA
sRNA	small RNA
tRNA	transfer RNA
UIG	Undefined Intergenic Gene
μg	microgram

Chapter One

General Introduction

1.1 Development in Non-coding RNAs Research

The central dogma of biology dictates that genetic information flows from DNA to RNA to protein. As a consequence, it has usually been assumed that proteins not only fulfill most structural and catalytic roles, but also mediate most regulatory functions in cells. However, in recent years, various families of non-coding RNAs (ncRNAs) have been found to play catalytic and regulatory roles in cells as well. Non-coding RNAs have been found both in prokaryotes and eukaryotes (Storz, Altuvia et al. 2005). The potential importance of ncRNAs is suggested by the observation that the complexity of an organism is poorly correlated with the number of protein coding genes it contains. Rather, it is highly correlated with the number of ncRNAs it possesses. In fact, in the human genome, only a small fraction of genetic transcripts (2-3%) are actually translated into proteins (Volinia, Calin et al. 2006).

Non-coding RNAs differ from coding RNAs (messenger RNAs, mRNAs) in that they lack an open reading frame (ORF) and therefore are not translated into protein (Mattick and Makunin 2006). Some classes of ncRNAs have been known for quite some time (Mattick and Makunin 2006), and these include transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs) as well as small nucleolar RNAs (snoRNAs). Other classes of non-coding RNAs with roles in regulation have only been identified relatively recently. In eukaryotes, the new ncRNAs include microRNAs (miRNAs) and small interfering RNAs (siRNAs); in prokaryotes, new ncRNAs are generally referred to as small RNAs (sRNAs) (Urban and Vogel 2007).

In terms of the functions of non-coding RNAs, many studies have convincingly shown that ncRNAs play crucial roles in the regulation of chromatin structure, gene expression, mRNA processing and splicing, mRNA stability and translational control (Storz, Altuvia et al. 2005; Sevignani, Calin et al. 2006). For instance, sRNAs tend to use base-paring interactions to bind various mRNAs and regulate gene expression post-transcriptionally. This mode of regulation has been found in both prokaryotes and eukaryotes (Storz, Altuvia et al. 2005). Despite of the differences in the characteristics of the eukaryotic and prokaryotic regulatory RNAs and in the fine details of their mechanism of action, sRNAs and ncRNAs can both exert their regulatory function by base pairing with the mRNA to influence translation or mRNA stability (Shimoni, Friedlander et al. 2007).

1.2 Small RNAs in Bacteria: Discovery

The early discoveries of sRNAs in bacterial systems were rather accidental. In fact, many of the early sRNAs were discovered during studies of the transcriptional regulation of neighboring protein-coding genes (Ikemura and Dahlberg 1973; Mizuno, Chou et al. 1984; Andersen, Forst et al. 1989; Coyer, Andersen et al. 1990; Schmidt, Zheng et al. 1995; Delihas and Forst 2001). The first sRNA was discovered in 1967, and ~140 sRNAs have been identified since then. However, the cellular functions of most sRNAs have yet to be elucidated (Argaman, Hershberg et al. 2001; Rivas, Klein et al. 2001; Wassarman, Repoila et al. 2001; Chen, Lesnik et al. 2002; Tjaden, Saxena et al. 2002; Vogel, Bartels et al. 2003; Zhang, Wassarman et al. 2003; Kawano, Reynolds et al. 2005). The majority of these sRNAs were discovered in *E. coli*, and a smaller subset was characterized in other (mostly pathogenic) bacteria. sRNAs are defined as untranslated RNA species ranging in size from 50nt to 500 nt, and they have various housekeeping or regulatory roles instead of functioning as rRNA or tRNA (Hershberg, Altuvia et al. 2003; Storz, Opdyke et al. 2004).

Genome-wide searches had begun once it was realized that sRNAs play a significant role in bacterial regulation. In fact, there has been a growing interest in sRNAs within bacteria due to their role in regulating many aspects of gene expression (Urban and Vogel 2007). In 2001, systematic genome-wide searches for new sRNAs in *E. coli* were first performed by several laboratories (Argaman, Hershberg et al. 2001; Rivas, Klein et al. 2001; Wassarman, Repoila et al. 2001). More than 60 sRNA candidates were generated from these searches. Other sRNAs were also identified recently in cyanobacteria and other bacterial genomes through similar efforts (Axmann, Kensche et al. 2005).

The sRNAs whose cellular functions have been elucidated were used and are being used to define common characteristics of sRNAs in hopes of discovering more such molecules. In many cases, it has been found that sRNAs are encoded within the intergenic regions (non-protein coding regions) of the genome and are terminated by a rho-independent terminator (Gottesman 2004). This information, along with the fact that many of the known sRNAs are conserved within related bacteria species such as *E. coli*, *Salmonella* and *Shigella*, has been utilized in many computational searches of potential sRNAs (Gottesman 2004).

1.2.1 Small RNAs in Bacteria: Classifications and Functions

While a large number of sRNA candidates discovered so far have unknown functions, most sRNAs with identified functions bind various mRNAs via imperfect sequence complementarities (Masse, Escorcia et al. 2003; Storz, Opdyke et al. 2004; Vogel, Argaman et al. 2004). In addition, it has been found that the activities of some cellular proteins are controlled directly by sRNAs in *E. coli* (Romeo 1998; Wassarman and Storz 2000; Weilbacher, Suzuki et al. 2003; Trotochaud and Wassarman 2004; Barrick, Sudarsan et al. 2005). Small RNAs with housekeeping functions have also been shown to be present in virtually all bacterial genomes sequenced to date (Stark, Kole et al. 1978; Poritz, Bernstein et al. 1990; Ribes, Romisch et al. 1990; Keiler, Waller et al. 1996).

Detailed characterization of some sRNAs has revealed that sRNAs are involved in homeostasis, sugar metabolism, transcriptional regulation and growth-dependant outer membrane protein expression (Masse and Gottesman 2002; Chen, Zhang et al. 2004; Vanderpool and Gottesman 2004; Storz, Opdyke et al. 2006). Researchers have also found that about one third of known *E. coli* sRNAs bind to Hfq with high affinity. Since Hfq (Host factor I protein) functions as a polyfunctional translational regulator of numerous bacterial mRNAs, this finding suggests that these sRNAs may play potential regulatory roles in translation of many mRNAs (Zhang, Wassarman et al. 2003; Gottesman, McCullen et al. 2006).

The known sRNAs can be divided into three general classes depending on their mechanisms of action. The fist class belongs to sRNAs that possess catalytic activity or belongs to an RNA-protein complex (Tjaden, Goodwin et al. 2006). There are two well-characterized sRNAs which fall into this category: 4.5S RNA and RNase P (Gottesman 2004). 4.5S RNA was one of the first sRNAs identified and is the RNA component of the signal recognition particle (SRP) in *E. coli* which is involved in recognizing and transporting proteins to the plasma membrane (Wassarman, Zhang et al. 1999). It is thought that this RNA is involved in stabilizing one of the protein components (Ffh protein) of the SRP complex (Jensen and Pedersen 1994). RNase P is also part of a ribonucleotide complex that is involved in tRNA and rRNA processing (Gopalan, Vioque et al. 2002). These two sRNAs are the only essential sRNAs known to date (Gottesman 2004). The second class of sRNAs includes the ones that affect protein activity by mimicking the structure of nucleic acids (Tjaden, Goodwin et al. 2006). These sRNAs are well conserved and include 6S RNA, CsrB and CsrC (Storz, Opdyke et al. 2004). 6S RNA binds and

inhibits RNA polymerase containing σ^{70} subunit, as it is thought that this RNA resembles the σ^{70} promoter (Wassarman and Storz 2000). CsrB and CsrC, on the other hand, are involved in regulating carbon storage by binding to CsrA (the carbon storage regulatory protein), which is known to bind and inhibit its target mRNAs when CsrB and CsrC are not bound (Romeo 1998). This class of sRNAs seems to play a large role in regulating protein activity in bacteria through protein-RNA interactions.

The final class belongs to those sRNAs which post-transcriptionally regulate mRNA through RNA-RNA interactions affecting the stability or translation of the transcript (Tjaden, Goodwin et al. 2006). This class is the most common and as a result is the best characterized (Storz, Opdyke et al. 2004). These sRNAs are either encoded in *cis* (which is on the opposite strand of the target mRNA but at the same genetic location resulting in perfect complementarity), or encoded in *trans* (which is located at a different chromosomal locus from the target mRNA and usually results in non-perfect base pairing) (Storz, Altuvia et al. 2005). *Cis*-acting sRNAs have been found to be involved in regulating some aspects of plasmid and bacteriophage functions (Storz, Opdyke et al. 2004). For example, OxyS RNA plays a role in plasmid stability through a toxin/antitoxin system, where the RNA is the antitoxin which inhibits the translation of the toxin (Gottesman 2004). However, most of the sRNAs found in this final class are encoded in *trans* and require the Sm-like protein Hfq (Tjaden, Goodwin et al. 2006). Hfq is an RNA chaperone that has been shown to have similar functions as the Sm-like proteins of eukaryotes that are involved in splicing and mRNA degradation complexes (Moller, Franch et al. 2002). Hfq is thought to bind to AU-rich sequences on the sRNA facilitating base-pairing with their target mRNA, and thus the structural changes in the RNA will disrupt the protection from digestion by RNase E (Moller, Franch et al. 2002; Storz, Altuvia et al. 2005). The first *trans*-encoded sRNA discovered is MicF RNA (mRNA-interfering complementary RNA) (Mizuno, Chou et al. 1984; Storz, Opdyke et al. 2004). MicF has been shown to bind to *ompF* mRNA blocking the synthesis of OmpF, an essential protein that acts as a pore allowing for passive diffusion of small molecules (Mizuno, Chou et al. 1984). MicF binding occurs in response to stress stimuli and induces the degradation of *ompF* mRNA, the resulting down-regulation of OmpF can lead to multiple antibiotic resistance (Delihas and Forst 2001).

The above are just a few examples of some of the characterized sRNAs. It can be seen that they are involved in a variety of biological functions, but most have an underlying regulatory

role. There are still many sRNAs to be characterized, suggesting that perhaps bacterial sRNAs may have a more significant role in regulating gene expression than previously thought. Knowing how known sRNAs function can be important in characterizing and identifying other unknown sRNAs. For example, the interaction of Hfq with *trans*-encoded sRNAs has been utilized to help identify and characterize new sRNAs (Wassarman, Repoila et al. 2001). Gathering more information on sRNAs and on how they function will allow us to obtain a better understanding of gene regulation in bacteria, as sRNAs may be providing a new level of genetic control. Findings in bacteria may also help elucidate related aspects of eukaryotic gene expression as non-coding RNAs have also been discovered in eukaryotes.

1.2.2 Small RNAs Discovery: Two Major Approaches

As it becomes more apparent that sRNAs are playing critical physiological roles in diverse prokaryotes, the discovery of these RNA species and the elucidation of their functions has become important areas of research. However, identifying sRNA-encoding genes has been no trivial matter, as classical genetic approaches used to identify regulatory proteins cannot be applied. This is mainly due to the small sizes of sRNAs and their immunity to the effects of frame shift and nonsense mutations (Hershberg, Altuvia et al. 2003; Gottesman 2004). The earliest sRNAs were discovered by chance; others were identified based on abundance in the cell after metabolic labeling, while the more recent attempts used computational approaches. For example, some studies used sequence conservation in intergenic regions between related bacterial species as an identification method (Gottesman 2004). Experimental based approaches, including microarray and shotgun cloning, have also been successfully used in sRNA discoveries. In this section, I will briefly discuss both bioinformatics approaches and some key experimental approaches that have been used in the identification of potential sRNAs. Table 1-1 (taken from (Vogel and Sharma 2005)) summarizes the advantages and disadvantages of each approach.

1.2.2.1 Bioinformatics Approaches

Previous computational studies have been successful in uncovering families of functional RNAs with well-defined sequence characteristics, such as snoRNAs in eukaryotes (Yang, Zhang et al. 2006). However, only limited groups of regulatory RNA families contain such defined

elements. In spite of the lack of defined sequence or structural characteristics, a number of computational approaches have been shown to be effective in discovering a large number of non-coding RNA genes in both prokaryotes and eukaryotes (Vogel and Sharma 2005).

Table 1-1: Overview of strategies for discovering sRNAs in bacteria (J. Vogel and C.M. Sharma, 2005). This table contains the most commonly used sRNAs discovery methods and their advantages and disadvantages.

Strategy		Advantages (⊕) and disadvantages (⊖)
RNA labeling and staining	⊕	Most abundant sRNAs and/or sRNAs with highest synthesis rate under a given growth condition are readily visualized; does not require prior knowledge of sRNA characteristics in the organism of interest; allows detection of species-specific sRNAs; points to the mature form of the sRNA identified
	⊖	Does not distinguish between sRNAs and abundant processed fragments of rRNAs or tRNAs; can require handling of highly radiolabeled bacterial cultures (orthophosphate labeling)
Functional genetic screens	⊕	May immediately pinpoint a functional role of the identified sRNA; could build on mutant strains and methods already established in genetic studies
	⊖	Difficult if sRNA is either essential or toxic when overexpressed; sRNAs acting under special conditions may not be identified; labor-intensive
Biocomputational searches	⊕	Rapidly generates a list of many potential sRNA candidates; allows phylogenetic comparison with genomes of related bacteria
	⊖	Requires prior knowledge of sRNA characteristics and validation of many candidate loci
Microarray detection	⊕	Yields transcriptional profiles for many sRNA genes in parallel; rapid detection of condition-dependent sRNA expression patterns; allows detection of species-specific sRNA transcripts
	⊖	Requires microarrays that cover intergenic regions; expensive; often yields inconsistent sRNA detection results compared to Northern blot signals
Shotgun cloning (RNomics)	⊕	Should allow detection of all RNAs of a certain size range that are expressed at a given time point; does not require prior knowledge of sRNA characteristics; can be automated; can detect processed, species-specific and non-canonical sRNAs; permits detection of primary transcripts
	⊖	Expensive (sequencing); labor-intensive (screening and evaluation of non-canonical candidates); cDNA synthesis may be biased against highly structured sRNAs
Co-purification with proteins	⊕	Could indicate specific interactions with proteins and the active form of the sRNA
	⊖	RNA has to remain tightly associated with the protein throughout purification; co-immunoprecipitation requires highly specific antibodies; limited to a subclass of sRNAs

As more genome sequences become available, genome-wide annotation of sRNAs becomes more realistic. The prediction of sRNA genes in *E. coli* has been done by comparing sequence and structural conservation between related species in non-protein coding regions or the intergenic regions, as discussed earlier (Rivas, Klein et al. 2001). In fact, several groups have based their predictions on sequence homology in intergenic regions from closely related microbial genomes. In these studies, criteria for the identification of sRNAs were derived based on a machine learning strategy, which builds models from conserved motifs in known sRNAs (Wassarman, Zhang et al. 1999). For instance, the presence of binding sites for specific DNA-binding proteins, the promoter and terminator sequences in non-protein coding regions in *E. coli*

provided another criterion to predict possible sRNA genes (Argaman, Hershberg et al. 2001; Eddy 2002; Zhang, Wassarman et al. 2003). Many of the discovered sRNAs candidates, ranging in size from 50 to 500 nucleotides, are conserved and located in intergenic regions (between two open reading frames). The expression of many of these genes is growth-phase dependent or stress related. Because each search employed specific parameters (sequence elements), sRNAs candidates with distinct characteristics are often identified. Consequently, unique sRNAs such as those that are species-specific, those that are transcribed under unique conditions or those that are located on the antisense strand of protein-encoding genes, were probably missed. Meanwhile, these computationally predicted sRNAs require further experimental verification of their expression through experiments such as Northern Blotting (Wassarman, Repoila et al. 2001).

The computational annotations have led to the identification of over 100 sRNAs in *E. coli*. However, the number of sRNAs that can be identified by this method is still limited due to several factors. First, all the computational annotations have focused on defined intergenic regions. They do not account for sRNAs in the coding regions of the genome or undefined intergenic regions. Second, these algorithms may fail to detect certain sRNAs without the recognized sequence element, while these sRNAs might be highly conserved across different genomes and possibly serve certain functions. Third, most of the algorithms for sRNA search rely heavily on sequence conservation between genomes. Species-specific sRNAs will be disregarded from these screens. Finally, some recently discovered sRNAs are not annotated because the start and end point of sRNAs in the genome are usually defined in relation to the actual coding sequence, thus the relative length of predicted sRNAs can be variable in different annotation and might not be annotated (Trotochaud and Wassarman 2004).

1.2.2.2 Experimental Approaches

Many sRNAs were first experimentally and systematically discovered by size fractionation of total RNA isolated from cells (Huttenhofer, Brosius et al. 2002; Vogel, Bartels et al. 2003; Huttenhofer, Cavaille et al. 2004). Direct cloning after size selection (the so-called 'RNomics approach') has also been employed in sRNA discovery. However, only RNA species that are present in high quantity will be detected due to the low sensitivity of this method. As a consequence, sRNAs in low or extremely low abundance will fail to be detected and recognized. Interestingly, the RNomics approach has identified several novel sRNA candidates, which were

not identified by computational annotations. This suggests that the sRNAs definition is far more complex than previously assumed. This method has also been applied in the screening of microRNAs (miRNAs) in eukaryotes (Huttenhofer, Kiefmann et al. 2001; Huttenhofer, Brosius et al. 2002; Tang, Bachellerie et al. 2002; Tang, Rozhdestvensky et al. 2002; Vogel, Bartels et al. 2003; Huttenhofer, Cavaille et al. 2004).

As a powerful tool for simultaneous monitoring of gene expression on a genome-wide scale, microarray technology has also been applied in sRNAs discovery, and total RNA extracts or RNAs isolated by co-immunoprecipitation with Hfq has been used in the experiments (Tjaden, Saxena et al. 2002; Zhang, Wassarman et al. 2003). However, the detection of sRNAs requires probes that are specific to the Intergenic Regions (IGRs) from both strands where most of the newly identified sRNAs reside, as well as strand-specific probes for all ORFs (open reading frames), tRNAs and rRNAs of a specific genome in detecting sRNAs (Selinger, Cheung et al. 2000). This is different from the standard microarrays, which are designed to detect RNAs expressed from ORFs only. Although DNA microarrays are a valuable tool for both identification and transcription profiling of sRNAs, it still faces some challenges in sRNA detection. One of the most challenging aspects in the use of this technique is the preparation and labeling of RNA samples. The small size and relatively stable secondary structure of the sRNAs make these transcripts poor substrates for amplification and labeling. In addition, shorter sRNAs (<50 nt) might be harder to detect, especially if the probes are not closely spaced. Therefore, short sRNAs and sRNAs that are highly structured and/or modified are likely to be missed. In addition, independent validation of the microarray data are still recommended, which also applies to the data from regular microarray experiments.

Some sRNAs are also found to be in complexes with proteins, either because these sRNAs require proteins for their activity or because they act on and modify the activity of their target proteins (Montzka and Steitz 1988). Hfq is required for the function of a great number of sRNAs that act as antisense regulators (Wassarman, Repoila et al. 2001). As a consequence, many of the sRNAs are bound to Hfq, and can therefore be co-immunoprecipitated by the use of Hfq antibodies (Wassarman, Repoila et al. 2001; Zhang, Wassarman et al. 2003). A small subset of sRNAs has also been found in complex with their target proteins in modulating their activity. CsrB of *E. coli* and RsmZ of *P. fluorescens* were identified by co-purification with their corresponding target proteins, CsrA and RsmA (Liu, Gui et al. 1997; Heeb, Blumer et al. 2002).

The genomic SELEX approach, which is based on the binding of sRNAs to Hfq, has recently been adopted to identify new *E. coli* Hfq binding sRNAs (Lorenz, von Pelchrzim et al. 2006). In this method, a library of random sequences that are 50–500 nt long from the *E. coli* genome is transcribed *in vitro* and then incubated with Hfq protein. Hfq-binding RNAs are isolated, converted to cDNA and subjected to additional rounds of selection and amplification. Specific Hfq interaction is then determined *in vivo* using the yeast three-hybrid system (Lorenz, von Pelchrzim et al. 2006). Unlike other experimental approaches where the sRNAs identification relies on RNA expression under certain growth conditions and growth stages, this SELEX approach generates RNAs *in vitro*, and thus is independent of growth conditions and stages. However, it should be noted that the newly identified sRNAs still need to be physiologically verified.

The detection of antisense transcripts and transcripts outside the known transcription units has also been performed through mRNA analysis and analyzing expressed sequence tags (EST) libraries, which will not be discussed in detail here (Lavorgna, Dahary et al. 2004).

1.3 My Research Objectives

As discussed above, the discovery of sRNAs and the elucidation of their functions has become an important area of research. My thesis has been designed to lay some groundwork in our lab to establish a new research interest in the area of discovering new non-coding RNA genes and examining their biological functions. Four interlinked projects will be pursued in my thesis. In the first project, we will perform a new bioinformatics study to search for new sRNA genes in *E. coli*. Since all the systematic methods that have been used in sRNA screening are focused on the intergenic regions with similar sequence elements, there are still many sequence elements within intergenic regions that may encode for some unique sRNAs. Therefore, the first objective of my thesis is to develop a bioinformatics search algorithm to look for DNA elements that only exhibit sequence conservation among a small number of genomes. These sequences cannot be discovered in all previous bioinformatics efforts because of the limited sequence conservation and the fact that we will not use any promoter or terminator sequences as confinements. After this bioinformatics search, I will use commonly used RNA secondary structure prediction program such as 'Mfold' (http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html) to see if these

DNA elements are predicted to have a potential RNA sequence with a relatively stable secondary structure. Following these computational analyses, in the second part of my thesis, I will insert some candidate genes into an RNA expression system and examine the effect of RNA expression on the physiology of *E. coli* cells. I will use a tetracycline inducible system to clone each of the candidate genes and examine if its expression into RNA can result in a lethal phenotype (cell death or reduced cell growth). A specific plasmid, named pNYL9-MCS11, that does not contain a ribosome-binding site (RBS) will be used, and the RNA expression in this vector is under the control of a tetracycline promoter.

In addition to the expression of the putative sRNA genes identified from the bioinformatics search above, I am also interested in applying the same cloning strategy to express some previously identified sRNA genes in *E coli* whose functions have not been elucidated. Finally, I will attempt to utilize the same expression vector to screen for random expression of DNA fragments from lambda phage. Since its first discovery in 1951, enterobacteria phage λ (lambda phage) has been intensively studied. The mechanism of infection of lambda phage in bacterial host has been also well understood (Rybakov, Shestakov et al. 1976; Narajczyk, Baranska et al. 2007; Osterhout, Figueroa et al. 2007). However, because no specific studies have been conducted to determine effects of controlled expression of RNA molecules from lambda phage in a bacterial host, I am interested in carrying out a simple experiment to screen for potential genes from lambda phage, when over-expressed into RNA in a bacterial host (*E. coli* for instance), can result in retarded growth of the host. This may lead to the discovery of new RNA-based regulation of gene expression employed by lambda phage against the bacterial hosts.

Chapter Two

Materials and Methods

2.1 Bioinformatics Analysis

The genomic sequences of all 551 microbial genomes (the names of these genomes are given in Appendix 1) were obtained from NCBI database (through following web address ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Protein-coding and intergenic regions in *E. coli* MG1655 were created based on the gene annotations using RegCompareI, a program I created to detect and eliminate all the protein-coding regions (the code for RegCompare I is given in Appendix 2). Both strands from the bacteria genomes are annotated separately in all the annotations in this project. A protein-coding region was defined as a genomic sequence containing an open reading frame (ORF) on either of the two DNA strands, whereas other parts of genome were defined as intergenic regions. A true ORF in this study was defined as the longest possible reading frame that begins with a start codon and ends with a stop codon. Intergenic regions of our particular interest in this study are all the intergenic sequences that do not contain tRNA genes, rRNA genes listed in the EcoGene database (Rudd 2000) and 80 sRNA candidates listed in the sRNAs dataset (Hershberg, Altuvia et al. 2003). Using RegCompareI, the protein-coding and known RNA genes were eliminated from the *E. coli* MG1655 genome, and the remaining intergenic regions with 50-500 nucleotides in length were collected. This data set was termed '*E. coli* MG1655 DataSet-I'. This sequence set was also scanned with Riboswitch Finder (Bengert and Dandekar 2004); no potential riboswitch was detected.

A sequence alignment was then performed with RegCompareII (a program I wrote as well; the code for RegCompareII is given in Appendix 3) using *E. coli* MG1655 DataSet-I as the query to compare against the genomic sequences of all other 550 genomes. A minimum of 50% identity was defined in RegCompareII; this alignment produced RegCompare DataSet-II, which was organized in the order based on their positions in *E. coli* MG1655 genome. The sequences in RegCompare Dataset-II were ranked based on the number of genomes in which each sequence was conserved. The one appeared in a highest number of genomes was ranked as #1. Because these sequences lie in the Unannotated Intergenic Region (UIG), they were named as UIG####. The top 7 most conserved sequences (their sequences and other relevant information are given in Appendix 4; the names of conserved genomes are given in Appendix 5) were chosen for some experimental characterization. A flowchart in search of highly conserved inter-intergenic elements is shown in Figure 2-1. The 7 top candidates were subjected to secondary structure prediction using 'Mfold' (Zuker 2003). The adjacent gene information of each of these 7 candidates was also obtained using BLAST (Altschul, Madden et al. 1997).

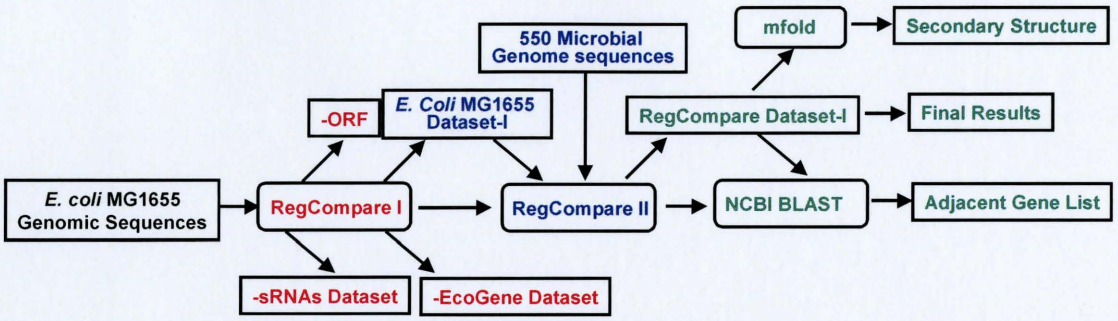


Figure 2-1: Flow Chart in Search of Potential sRNA Candidates. *E. coli* MG1655 genomic sequence is analyzed by RegCompareI to eliminate ORFs and known RNA genes in the published sRNA Dataset and EcoGene Dataset with two strands annotated separately. The resulting *E. coli* MG1655 Dataset-I is analyzed with RegCompareII in which *E. coli* MG1655 Dataset I is used as query sequences and each sequence is aligned against all 550 microbial genomic sequences to yield some candidate genes with sequence conservation in limited genomes as the RegCompare Dataset-I. A few top-ranked candidates (i.e., those occurred in most genomes) are subjected to (1) Mfold for secondary structure prediction and (2) NCBI BLAST to obtain adjacent gene information.

2.2 Bacterial Strains and Growth Conditions

E. coli cells were grown overnight at 37 °C and 250 rpm in Luria Broth (LB) (Sigma, St Louis, MO) supplemented with the appropriate antibiotics at final concentration of 50 μg/mL spectinomycin or 25 μg/mL kanamycin with or without various concentrations of anhydrous tetracycline (aTc). All antibiotics used for this study were purchased from Sigma (St. Louis, MO), unless otherwise noted. *E. coli* DH5α-Z' or *E. coli* top10 cells were used for plasmid propagation. *E. coli* MG1655 as well as *E. coli* DH5α-Z' cells were used for genomic DNA extraction. All restriction enzymes were purchased from MBI-Fermentas (Burlington, ON) unless otherwise indicated.

2.3 Cell Lines and Vectors

A plasmid named 'pZE21-MCS-1' that was used to create the tetracycline inducible system were obtained as a gift from Herman Bujard in Germany (Lutz and Bujard 1997). The

experimental design was adapted from a previous study on identifying essential *Staphylococcal* genes (Ji, Zhang et al. 2001).

The modification of pZE21-MCS-1 to remove the ribosomal binding site (RBS) was performed previously in our lab by Dr. Naveen Kumar Navani. pZE21-MCS-1 was digested with *EcoRI* to remove the RBS along with part of the multiple cloning sites (MCS). The MCS site was subsequently restored and the modified plasmid was termed pNYL9-MCS11 (Figure 2-2). The sequences of both pZE21-MCS-1 and pNYL9-MCS11 are provided in Appendix 6. Selection for pNYL9-MCS11 was performed in the presence of 25 μg/mL of kanamycin. All screening experiments described later were performed using either chemically competent DH5α-Z' *E. coli* cells (rubidium chloride treated) or electro-competent DH5α-Z' *E. coli* cells in LB containing 50 μg/mL of spectinomycin.

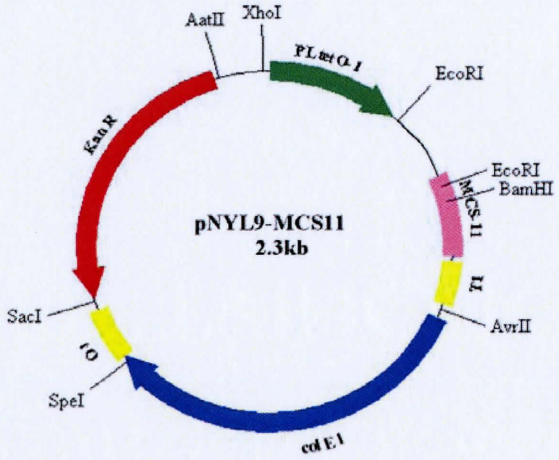


Figure 2-2: Map of the vector pNYL9-MCS11. It is derived from pZE21-MCS-1 with the ribosomal binding site removed and the MCS sites restored after the removal of the RBS. It carries the PLtet-O1 promoter, the tetracycline controlled promoter. All the genes of interest in this study are inserted into this vector and the constructs are subsequently transformed into *DH5αZ' E. coli* competent cell, which expresses a tetracycline repressor. Thus upon the addition of tetracycline or aTc, the PLtet-O1 promoter will be freed and the transcription of downstream insert can proceed.

2.4 Generation of DNA Templates of sRNA Genes from *E. coli*

Genomic DNA used to create the DNA templates of the 7 candidate sRNAs was isolated from *E. coli* MG1655 using the Wizard Genomic DNA Purification Kit from Promega (Nepean,

ON). Polymerase chain reaction (PCR) was performed using the genomic DNA as the template and primers complementary to the 5′ and 3′ ends of the following candidates: UIG0242, UIG0803, UIG0985, UIG1195, UIG1259, UIG1354, and UIG1585 (for sequences of these primers, see Appendix 7). Primers to each candidate for the PCR reaction were designed in the manner that the *BamHI* and *SalI* restriction enzyme recognition sites were incorporated into the forward primer and the reverse primer carries the *HindIII* restriction enzyme recognition site. Primers were ordered from IDT (Coralville, IA) in the 25-nmole scale and purified by denaturing PAGE (polyacrylamide gel electrophoresis). The PCR mixture for each candidate contains 12.5 μL of 10× dNTP (2 mM each, MBI-Fermentas), 12.5 μL of 10× ThermoPol Reaction Buffer (New England Biolabs, Ipswich, MA), 2.5 μL of Vent DNA Polymerase (New England Biolabs), 3 μL of forward primer (20 μM), 3 μL of reverse primer (20 μM) and 88.5 μL of ddH_2O. 2 μL of genomic DNA (0.002 μg/mL, *E. coli* MG1655) was added to 98 μL of the reaction mixture. Another PCR mixture at ¼ scale was made to which 1 μL ddH_2O was added. This was used as the negative control. Thermal cycles were set as follows on Robocyler Gradient 96 (Stratagene, La Jolla, CA): 1) 3 min at 92 °C, 2) 40 s at 92 °C, 30 s at 52 °C and 30 s at 72 °C for 10 cycles, 3) 30 s at 92 °C, 30 s at 55 °C and 30 s at 72 °C for 20 cycles, and 4) 10 min at 72 °C. The PCR products were loaded onto a 1.5% agrose gel with 0.0005% v/v SYBR Safe DNA Gel Stain from Invitrogen (Burlington, ON) and visualized on Typhoon Variable Mode Imager (GE Healthcare). Corresponding bands with right size was cut out and DNA was extracted and dissolved in 50 μL of ddH_2O using QIAquick Gel Extraction Kit from Qiagen (Mississauga, ON).

We also created clones for 8 previously identified sRNAs with unknown functions (Hershberg, Altuvia et al. 2003) (see Table 2-1 below for more information). The same genomic DNA was used for generation of the DNA fragments using the same PCR procedure described above. The sequences of each primer pairs are also provided in Appendix 7.

2.5 Construction of Inducible RNA Expression System

E. coli DH5α-Z' cells carrying pNYL9-MCS11 vector was recovered from glycerol stock by streaking onto a fresh agar plate supplied with 25 μg/mL of kanamycin and 50 μg/mL of spectinomycin. 100 mL of LB with kanamycin (25 μg/mL) and spectinomycin (50 μg/mL) was used for large scale inoculation of these cells. Large-scale preparation of pNYL9-MCS11 vector was performed using QIAprep Spin Miniprep Kit (Qiagen) to obtain 500 μL of concentrated plasmid. 5 μL of this solution was loaded onto 0.8% agrose gel with 0.0005% v/v SYBR Safe

DNA Gel Stain and visualized on Typhoon Imager. Double digestion of the vector was performed using the combination of *SalI* and *HindIII* as well as *BamHI* and *HindIII*. The double digestion reaction mixture was prepared by using 100 μL of purified plasmid mixed with 76 μL of ddH_2O, 20 μL of 10× *BamHI* buffer, 2 μL of *SalI* and 2 μL of *HindIII*. Another double digestion was performed using 100 μL of purified plasmid mixed with 76 μL of ddH_2O, 20 μL of 10× *BamHI* buffer, 2 μL of *BamHI* and 2 μL of *HindIII*. Both double digestion reactions were incubated at 37 °C for 2 hr and loaded onto 0.8% agrose gel with 0.0005% v/v SYBR Safe DNA Gel Stain followed by visualization on Typhoon Imager. Corresponding bands with right size were cut out and extracted in 60 μL of ddH_2O for each double digestion using QIAquick Gel Extraction Kit. The double digestions of the PCR products for our 7 candidates were performed using the combination of *SalI/HindIII* enzymes or *BamHI/HindIII* enzymes for inserting the templates in a specific orientation. The double digestion was prepared by mixing 20 μL of purified PCR product with 5 μL of ddH_2O, 3 μL of 10× *BamHI* buffer, 1 μL of *SalI* and 1 μL of *HindIII*. Another double digestion using 20 μL of purified PCR product mixed with 5 μL of ddH_2O, 3 μL of 10× *BamHI* buffer, 1 μL of *BamHI* and 1 μL of *HindIII* was also prepared. Each mixture was incubated at 37 °C for 1 hr and then loaded onto 1.5% agrose gel with 0.0005% v/v SYBR Safe DNA Gel Stain followed by visualization on Typhoon Imager. Corresponding bands with right size were cut out and extracted in 30 μL of ddH_2O using QIAquick Gel Extraction Kit. The PCR products for the 8 known sRNAs were also subjected to the double digestion in the same manner.

Table 2-1: List of sRNAs Selected from Literature

sRNA candidates	Adjacent Genes	Length
tpke11	dnaK/dnaJ	370
C0293	icd/ymfD	72
C0299	hlyE/umuD	78
c0343	ydaN/dbpA	74
sraD	ygaG/gshA	70
sraI	yhhX/yhhY	94
sraL	soxR/yjcD	140
ssrA	smpB/intA	363

Each digested PCR product was ligated with the matching vector for 20 hr at 16 °C. The ligation mixture contained 14 μL of the PCR product, 2 μL of the vector, 1 μL of T4 DNA ligase, 2 μL of 10× ligation buffer, and 1 μL of PEG6000 (all from MBI-Fermentas). A negative control containing 14 μL of ddH_2O to replace the PCR product was also prepared.

The ligation mixture were then electroporated and transformed into electro-competent DH5αZ' competent cells. Colonies were then selected and sent for sequencing to confirm the identities of the inserts. Each clone was named after its insert followed by orientation designation and whether it contains mutation. For example, the clone has candidate UIG0803 in pNYL9-MCS11 in forward orientation with no mutation was termed as fUIG0803, while the clone that has candidate UIG0803 in pNYL9-MCS11 in reverse orientation with 4 base pair deletion was termed as rUIG0803_4D.

2.6 Activity Screens

Each bacterial colony obtained above was grown overnight at 37 °C in 5 mL of LB containing 25 μg/mL kanamycin and 50 μg/mL of spectinomycin. Replica plates of each colony were obtained in the presence and absence of 400 ng/mL aTc on 0.5% LB agar plates containing the same two antibiotics. These plates were grown at 37 °C for 6 hr and screened for colonies that exhibit significantly reduced growth in the presence of aTc.

2.7 Cell Permeability Assay on Expression of RNAs

A chosen colony was grown overnight at 37 °C in 5 mL of LB containing 25 μg/mL of kanamycin and 50 μg/mL of spectinomycin. After induction with aTc for 4 hr, replica plates of each colony were obtained in the presence and absence of 400 ng/mL of aTc onto 0.5% LB agar MacConkey plates containing the same two antibiotics. These plates were grown at 37°C overnight and examined for colonies that show white or colorless morphology in the presence of aTc.

2.8 Growth Curve

Three clones were selected for the growth curve assay and they were: rUIG0803_4D (candidate UIG0803 in pNYL9-MCS11 in reverse orientation with 4 nt deletion), RRygC (a

positive control with a lethal RygC gene in pNYL9-MCS11 in reverse orientation) and MCS11 (the vector itself as a negative control clone). These clones were grown overnight in LB containing the two antibiotics. Cultures were then inoculated in duplicate at 1% into a 96-well plate containing 200 μL of LB broth with the two antibiotics. Each culture was grown both in the presence and absence of aTc, three different concentrations of aTc: 200 ng/mL, 400 ng/mL and 800 ng/mL. The plates were incubated at 37 °C at 250 rpm. The optical density at 600 nm was measured every 30 min using EnVision plate reader.

2.9 Fluorescence Microscope Study of Cell Morphology

Cells with rUIG0803_4D, rUIG0803, rRygC, MCS11, and fRygC (Rygc in forward orientation in pNYL9-MCS11) were chosen for this assay. These cells were inoculated in 5 mL of LB supplied with the two antibiotics and grew overnight. Fresh inoculation of each clone was then made the following morning from the overnight culture. 600 μL of the overnight culture was added to 6 mL of kanamycin/spectinomycin-containing LB with or without 400 ng/mL of aTc. Cells were harvested at 4, 8, 12 and 24 hr. After 4 hr of growth, 500 μL of cells were pelleted at 8,000 rpm for 4 min, the cell pellet was then washed in 800 μL of PBS (phosphate buffered saline: 137 mM NaCl, 10 mM Phosphate, 2.7 mM KCl, pH 7.4) and centrifuged at 13,000 rpm for 1 min. The liquid was then decanted and the pellet was resuspended in 100 μL of the dye mix containing Syto9 and propidium iodide (1:1 in ratio, LIVE/DEAD cell staining kit, Invitrogen). The cells in dye mix were incubated in dark for 15 min and 1.5 μL of dye-containing cells were dropped onto a slide and covered with cover-slip. Pictures of the slides were taken using a fluorescence microscope (Axiovert 100 from Zeiss, Toronto, ON) at two different magnifications (40× and 100×).

2.10 Protein Profiling Using SDS-PAGE

Cells with rUIG0803_4D, rUIG0803, rRygC, or MCS11 were inoculated freshly from glycerol stock and grew overnight. Fresh inoculation of each clone was then made the next morning from the overnight culture, from which 600 μL was taken and added into 6 mL of kanamycin/spectinomycin-containing LB with or without 400 ng/mL of aTc. Cells were harvested at 4, 8, 12 and 24 hr. After 4 hr of growth, 500 μL of cells were pelleted and washed with 800 μL of 1× PBS and resuspended in 50 μL of 1× PBS for protein profiling. 10% and 15%

SDS-PAGE gels were casted in advance. 10 μL of the cell suspension was mixed with 6× protein-loading buffer and loaded into each well. Every sample was analyzed on SDS-PAGE. The gels were then stained with freshly made Commassie Blue (0.1% Commassie Brilliant Blue R-250, 50% methanol and 10% glacial acetic acid in ddH_2O) for 2 hr at room temperature with gentle shaking followed by de-staining in de-staining solution (40% methanol and 10% glacial acetic acid in ddH_2O) for 45 min at room temperature with gentle shaking. The de-stained gels were washed with ddH_2O for 2 hr and gel images were scanned using the HP Scanjet 5370C scanner.

The size of the band was determined using the protein ladder (SM4401, MBI-Fermentas). Bands that showed difference in protein expression in any clone in the presence or the absence of aTc were cut out and sent for protein sequencing in Central Mass Spectrometry Facility in McMaster University.

2.11 Generation of Genomic Library from Lambda Phage

Genomic DNA used for creation of the genomic library was obtained from Invitrogen. The DNA was digested with *Sau3A1* (New England Biolabs) for 60 min. The reaction was stopped with the addition of 0.5 M EDTA (pH8.0) and analyzed on a 1% agarose gel. The digested fragments were excised and extracted using QIAquick Gel Extraction Kit.

The vector pNYL9-MCS11 was then digested with *BamHI* for 60 min and purified on a 1% agarose gel. The band corresponding to digested pNYL9-MCS11 was excised and extracted using QIAquick Gel Extraction Kit. The vector was subsequently treated with calf intestinal alkaline phosphotase (CIAP, from MBI-Fermentas) for 30 min to remove the 5′ phosphates to prevent self-ligation. The digested vector was cleaned up using the QIAquick Gel Extraction Kit.

Prior to ligation, the genomic DNA fragments and dephosphorylated pNYL-MCS11 were precipitated together in 100% ethanol and subsequently washed with 70% ethanol. Then the samples were dried and re-hydrated in 15 μL of ddH_2O. Ligation using T4 DNA ligase in the presence of PEG 6000 was performed. Ligation mixtures were incubated at 16 °C for 16 hr and subsequently transformed into the *E. coli* DH5α-Z' competent cells. Colonies were then picked and plasmids were sent for sequencing to reveal the identities of the inserts. 192 clones were selected for aTc induced screens below. All the clones were named numerically from clone 001L to clone 192L, with L denoting Lambda phage genomic clones.

2.12 Screening Lambda Phage Genomic Library

All 192 clones were screened using the same protocol described in Section 2.6 above. The clones showed reduced growth were selected for further characterization.

2.13 Growth Curve of Selected Clones

Cultures for clones 140L, 141L, 152L, RRygC and clone MCS11 were subjected to a growth curve analysis as described for clone rUIG0803_4D (see Section 2.8 above).

Chapter Three

Bioinformatics Search and Experimental Characterization of Unannotated DNA fragments from *E. coli* Genome

3.1 Results

3.1.1 Bioinformatics Analysis

All protein-coding regions in *E. coli* MG1655 were created based on the gene annotations in 'RegCompareI', a program I wrote for the detection and elimination of all of the protein-coding regions. Two strands were annotated separately in all annotation. In addition, we also eliminated 80 known sRNA genes (Hershberg, Altuvia et al. 2003), and the tRNA and rRNA genes annotated in the EcoGene database (Rudd 2000). As a result, we obtained 2457 unannotated intergenic DNA sequences between 50- and 500-nt (nucleotides) in length, collectively termed 'RegCompare DataSet-I'. These sequences were named 'UIG0001 to UIG3202' according to the order they occurred in our bioinformatics analysis (UIG: Unannotated Intergenic Genes), 745 sequences were either longer than 500 nt or shorter than 50 nt from initial analysis and were eliminated. The remaining qualified sequences are provided as Appendix 8 and can be downloaded from the following URL: http://www.flynature.com/Appendix_8.doc. It should be noted that we did not find any riboswitch sequences in RegCompare DataSet I when it was analyzed with the Riboswitch Finder (Bengert and Dandekar 2004).

We then performed a multiple sequence alignment using RegCompare DataSet-I against the genomic sequences of all 551 microbial genomes that are currently available from NCBI (National Center for Biotechnology Information). We decided to search for any sequence in our dataset that exhibits at least 50% identity to a sequence that occurs in at least two additional genomes. This search was performed using an in-house program named 'RegCompare II'. This effort generated a new dataset of 82 sequences, which we named the 'RegCompare DataSet II' (see Appendix 9 listed at http://www.flynature.com/Appendix_9.xls). The Top 7 sequences in terms of genome occurrence were selected for further analysis (UIG0242, UIG0803, UIG0985, UIG1195, UIG1259, UIG1354 and UIG1585, see Appendix 4). It is interesting to note that these sequences are highly conserved in more than 3 bacterial genomes that are phylogenetically distant from each other, rather than three different strains of the same microorganism. For example, UIG0242 is conserved in *E. coli*, *Shigella flexneri*, *Salmonella enterica* as well as other bacterial genomes. The detailed conservation information for each candidate was listed in Appendix 5. Additionally, the secondary structures of these candidates are also conserved among multiple genomes.

The names of the adjacent genes for each candidate were also obtained using BLAST (Table 3-1) and the conservation information of these adjacent genes was also examined by a

published method (Altschul, Madden et al. 1997) and the information is provided in Table 3-1 (as well as in Appendix 5).

Table 3-1: List of 7 most conserved UIG candidates

UIG Candidate	Adjacent Gene	Strand	Length (nt)	3' end position
UIG0242	ybhi / ybhj	→	84	802628
UIG0803	predicted transposase	→	107	1529841
UIG0985	ydhQ/ydhR	→	124	1744573
UIG1195	yecL/yecR	→	87	1986131
UIG1259	yeeN/adhesin	→	118	2057989
UIG1354	tRNA/tRNA	→	93	2192314
UIG1585	tRNA/DNA binding activator	→	117	2519140

Interestingly, the adjacent genes of each candidate also showed some level of sequence conservation. In general, the adjacent genes of UIG0242, UIG0985, UIG1195 and UIG1259 contain some hypothetical proteins that are believed to be involved in cell regulation (Riley, Abe et al. 2006). UIG0803 is flanked by genes that encode for transposases. In addition, the adjacent genes for UIG1354 are tRNA genes. Finally, UIG1585 has a tRNA gene located to its 5′ side and a gene for a DNA binding activator to its 3′ end.

Finally, the secondary structures of these 7 candidates were predicted using the Mfold program (Zuker 2003). The secondary structure of UIG0803 is depicted in Figure 3-1 and the secondary structures of all 7 UIGs are provided in Appendix 10. All these candidates appear to have a very stable secondary structure, pointing to a possibility that they may be transcribed into RNA transcript likely to be stable.

3.1.2 Amplification of Top UIGs from Genomic DNA

Using the genomic DNA isolated from *E. coli* MG1655 and specific primers complementary to the 5′ and 3′ ends of each of the 7 candidate genes, we performed polymerase chain reactions to obtain DNA materials for the cloning experiment. It should be specially noted that the *SalI* and *BamHI* restriction enzyme recognition sites were designed into the sequence of each forward primer, while the *HindIII* restriction enzyme recognition site was engineered into each reverse primer. The existence of these restriction sites will facilitate the cloning of each

candidate gene into pNYL9-MCS11 bi-directionally. All the primers sequences are listed in Appendix 7 with restriction enzyme recognition sites highlighted.

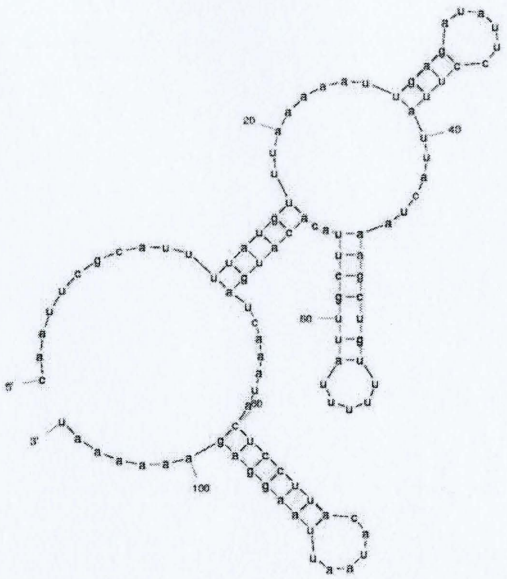


Figure 3-1: **A Predicted Secondary Structure of UIG0803.** The structure is predicted using Mfold program and the putative structure with lowest free energy (and thus are most stable) is shown above.

PCR was successfully conducted for the following six genes: UIG0242, UIG0803, UIG0985, UIG1259, UIG1354, and UIG1585 (Figure 3-2). The bands shown on 2% TAE-agrose gels are consistent with the predicted size of each gene. However, we failed to amplify the DNA for UIG1195 after several trials and thus were forced to exclude this candidate from experimental tests described in following sections.

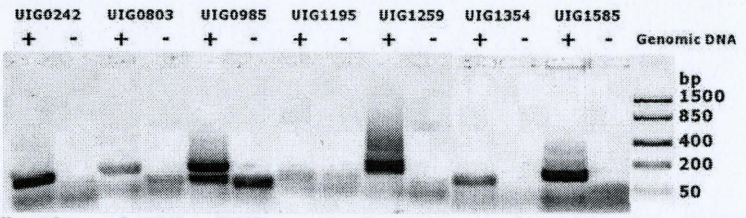


Figure 3-2: Amplification of UIGs. PCR products were analyzed on 2% TAE-agrose gels. The predicted size of each DNA fragment: UIG0242, 84 nt; UIG0803, 107 nt; UIG0985, 124 nt; UIG1259; 118 nt; UIG1354, 93 nt; UIG1585, 117 nt.

3.1.3 Sub-cloning and Expression of UIGs

We next attempted to clone each candidate into pNYL9-MCS11 (Figure 2-2 in Chapter 2). This plasmid has a tetracycline-controlled promoter but lacks a ribosome binding site (RBS). This configuration allows for the expression of the DNA insert into RNA but not protein. All 6 UIGs were successfully inserted into pNYL9-MCS11 in both directions. The cloning of each gene was verified by DNA sequencing. No mutations were found in almost all the sequenced clones, with the exception of one clone containing UIG0803 in the reverse orientation with 4-nt deletion. This clone was termed as rUIG0803_4D (the clone without mutation was named 'rUIG0803').

It is noteworthy that each vector carrying a candidate in a given orientation was transformed into the *E. coli* cell line DH5α-Z'. This cell line constitutively expresses a tetracycline repressor, thus each DNA insert will only be expressed as RNA in the presence of tetracycline. Anhydrous tetracycline (aTc) was used in our system because of its higher affinity to the tetracycline repressor and therefore a small concentration of this compound is required for induction, producing minimal toxic effect.

3.1.4 Identification of Clones with Abnormal Growth Phenotype

For the inducible RNA expression experiment, a positive control clone rRygC (carrying the sRNA gene known as 'RygC' inserted into pNYL9-MCS11 in the reverse orientation), which was isolated by a previous lab member from a gene screening experiment and was found to inhibit cell growth upon aTc induction. The negative control clone MCS11 (pNYL9-MCS11 without any DNA insert) was used to ensure that neither the addition of aTc or expression of a random RNA sequence (MCS11 sequence itself for instance) is not toxic to the cells. It is worth mentioning all the assays were done in 6 hr period of time. However, pictures were taken both at the end of the assay and 12 hr post assay time. It is because that the positive control clone rRygC exhibits a lethal phenotype around 6 hr of induction, and this clone start to recover after 6 hr of induction and thus might not be a good enough positive control. Upon aTc induction, a lethal phenotype should be seen with the rRygC clone, while the MCS11 clone should exhibit normal cell growth. As expected, the MCS11 clone showed normal growth while the rRygC clone had a lethal phenotype (Figure 3-3). To our disappointment, all 12 clones, each with one of the 6 candidates inserted into pNYL9-MCS11 vector in either forward direction or reverse orientation, did not show any lethal phenotype during the assay. Interestingly, however, the rUIG0803_4D

clone, which carries candidate UIG0803 in the reverse direction but with a 4-nt deletion, showed significant lethality. However, this phenotype was not observed at the end of the 6 hr screening period (data not shown) but only appeared 12 hr post assay when the positive control clone rRygC started to recover from lethal phenotype. Thus all the pictures shown will be the ones 12 hr post assay to better show the phenotypic change in clones other than rRygC.

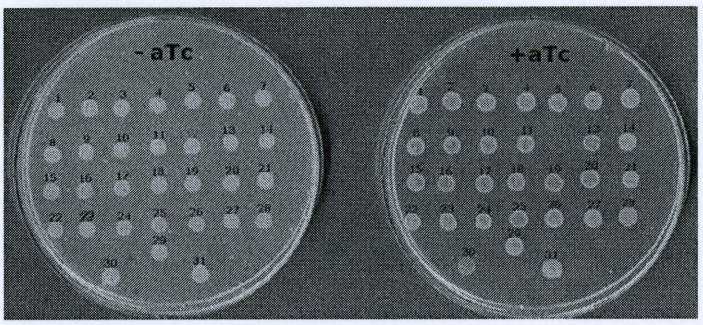


Figure 3-3: Phenotypic assay on LB agar plates. Cells were grown overnight in LB containing 25 μg/mL of kanamycin and 50 μg/mL of spectinomycin. 1% sub-cloning was incubated for 4 hr before 1.5 μL of each culture was spotted on 1.5% LB agar plates (containing the same antibiotics) in the absence (left) or presence (right) of 400 ng/mL of aTc. The pictures were all taken at 12 hr post assay. Clones 1, 2, 3 are rUIG1585 (r: reverse orientation); 4, 5 and 6 are rUIG1354; 7 and 8 are rUIG0242; 9 is fUIG0242 (f: forward orientation); 10 and 11 are rUIG1259; 12 is rUIG0803_4D; 13 is rUIG0803; 14 and 15 are fUIG0803; 16 and 17 are fUIG1259; 18 and 19 are fUIG1354; 20, 21 and 22 are fUIG985; 23, 24, 25 and 26 are fUIG1585; 27, 28 and 29 are rUIG0985; 30 and 31 are rRygC (the positive control) and MCS11 (the negative control), respectively. Clone 12 (rUIG0803_4D) is shown in pink to highlight the observation of the reduction of growth upon aTc induction.

3.1.5 Cell Permeability Assay

All 13 clones along with the positive and negative controls were also subjected to an assay to examine the effects of RNA expression on cell permeability. All 15 clones including rUIG0803_4D were induced by 400 ng/mL of aTc for 4 hr in LB supplied with 50 μg/mL of spectinomycin and 25 μg/mL of kanamycin, and then transferred onto the designated MacConkey Agar plate supplied with the same antibiotics with or without 400 ng /mL of aTc (Figure 3-4). The cells were incubated in aTc for 4 hrs so that there will be enough cells growing on the MacConkey plates. No clones showed any change in cell permeability (as no observation of white or colorless colonies). rUIG0803_4D, labeled as '11', showed a reduction of growth when

growing on aTc plates. As expected, cell growth was inhibited in clone rRygC (labeled as '+') upon aTc induction. It is interesting to notice that the reduction of cell growth in rUIG0803_4D (clone 11) requires the presence of aTc beyond 4 hr, as clone 11 only showed lethal phenotype on the plate with aTc but the clone that grew on the –aTc plate after 4 hr aTc liquid incubation doesn't show lethal phenotype. In contrast, cell growth in rRygC (clone15) was inhibited within 4 hr of aTc induction, which is shown by the lethal phenotype on both aTc and –aTc plate after 4 hr of aTc liquid incubation.

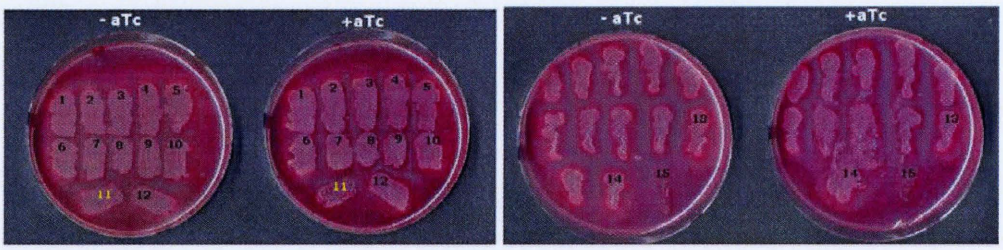


Figure 3-4: Phenotypic assay on a MacConkey plate. Cells were grown overnight in LB containing 25 μg/mL of kanamycin and 50 μg/mL of spectinomycin. After addition of 400 ng/mL aTc and further incubation for 4 hr, each culture was streaked on a 1.5% MacConkey agar plate in the absence (left) or presence (right) of 400 ng/mL of aTc. 1-13: fUIG0242, rUIG0242, fUIG0803, rUIG0803, fUIG0985, rUIG0985, fUIG1259, rUIG1259, fUIG1354, rUIG1354, rUIG0803_4D, fUIG1585, fUIG1585. 14 and 15: MCS11 and rRygC. The reduction of cell growth in rUIG0803_4D (clone 11) requires the presence of aTc for more than 4 hr. In contrast, cell growth in rRygC (clone15) was inhibited within 4 hr of aTc induction.

3.1.6 Growth Curve

Growth curves were obtained for clones rUIG0803_4D, rRygC and MCS11. Cells were grown overnight in LB containing 25 μg/mL of kanamycin and 50 μg/mL of spectinomycin. 1% of each overnight culture was freshly re-inoculated in LB supplemented with 25 μg/mL of kanamycin and 50 μg/mL of spectinomycin. All 3 clones were then grown either in the absence of aTc or presence of 200, 400 and 800 ng/mL of aTc in 96 well plates. OD_{600} readings were taken and recorded every 30 min for 14 hr and the data were plotted in Figure 3-5. Note that in this figure, '0 aTc', '0.5 aTc', '1 aTc' and '2 aTc' denote cell cultures treated with 0, 200, 400 and 800 ng/mL of aTc.

For rRygC (Figure 3-5A), a reduction in cell growth was observed within 1 hr of aTc induction; the cells seemed to grow slowly thereafter (OD_{600} was increasing, particularly after 8 hr). With rUIG0803_4D (Figure 3-5B) however, a reduction in cell growth was apparent only after ~6 hr of aTc induction with cells appearing to die shortly thereafter (OD_{600} decreased between 8-14 hr). Relatively small dose-response was observed for both rRygC and rUIG0803_4D, indicating the lowest concentration of aTc (200 ng/mL) was sufficient to induce the observed phenotype. As expected, little difference in cell growth was seen with MCS11 with or without aTc induction (Figure 3-5C).

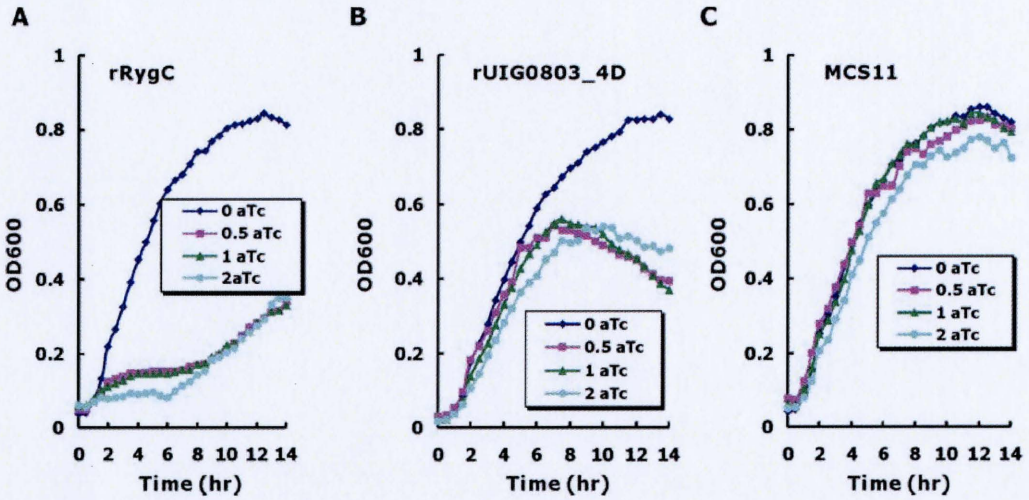


Figure 3-5: Growth curve of selected clones at various aTc concentrations. Cultures were grown overnight in LB containing kanamycin and spectinomycin. 1% of each culture was freshly inoculated with the same two antibiotics in the absence of aTc (0 aTc) and presence of 200 ng/mL (0.5 aTc), 400 ng/mL (1 aTc) and 800 ng/mL of aTc (2 aTc). OD600 was taken every 30 min up to 14 hr. (A) rRygC. (B) rUIG0803_4D. (C) MCS11.

3.1.7 Fluorescence Microscopy Analysis

Four clones were selected for microscopy study: rUIG0803_4D, rUIG0803, rRygC, and MCS11. Cells were grown overnight in LB containing 25 µg/mL of kanamycin and 50 µg/mL of spectinomycin. Each cell culture was then further grown in the absence or presence of 400 ng/mL of aTc, and harvested at 4, 8, 12 and 24 hr. After washing, the cells were re-suspended in 1× PBS to 1× 10^8 cells/mL, incubated in SYTO9 and propidium iodide stains (Invitrogen) for 15 min, and

studied under a microscope. Images were photographed in an Axiovert 100 microscope (Carl Zeiss, Inc.) at 100× and 40× optical magnifications. After staining, live bacteria with intact cell membranes show green fluorescence and dead bacteria with compromised membranes exhibit red fluorescence.

Pictures taken with rUIG0803_4D with (left panel) and without (right panel) aTc are shown in Figure 3-6. No significant morphology changes were observed in the cells at all four incubation times. However, two points are noteworthy: (1) cells became slightly elongated at 4-hr time point with aTc, and (2) more cell death was observed for cells treated with aTc for 24 hr.

For rRygC (Figure 3-7), elongated cells were observed at 8, 12 and 24 hr in the sample treated with aTc. In addition, significant cell death was seen in the sample incubated with aTc for 24 hr.

In contrast, no significant changes in cell morphology or cell death were seen with clones MCS11 (Figure 3-8) or rUIG0308 (Figure 3-9) upon aTc induction when compared to un-induced cells.

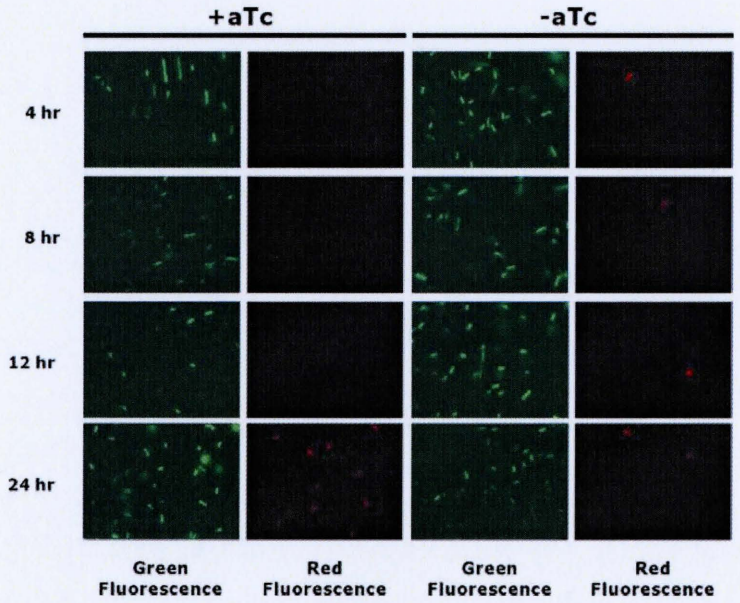


Figure 3-6: Fluorescent cell images of rUIG0803_4D. Cells were grown overnight in LB containing kanamycin and spectinomycin and re-cultured under the same conditions in the absence or presence of 400 ng/mL of aTc. Cells were then harvested after 4, 8, 12 and 24 hr and re-suspended in 1× PBS at 1×10^8 cells/mL. Cells in suspension were then incubated in SYTO9 and propidium iodide stains for 15 min and fluorescence microscopy images were taken at 100× magnification.

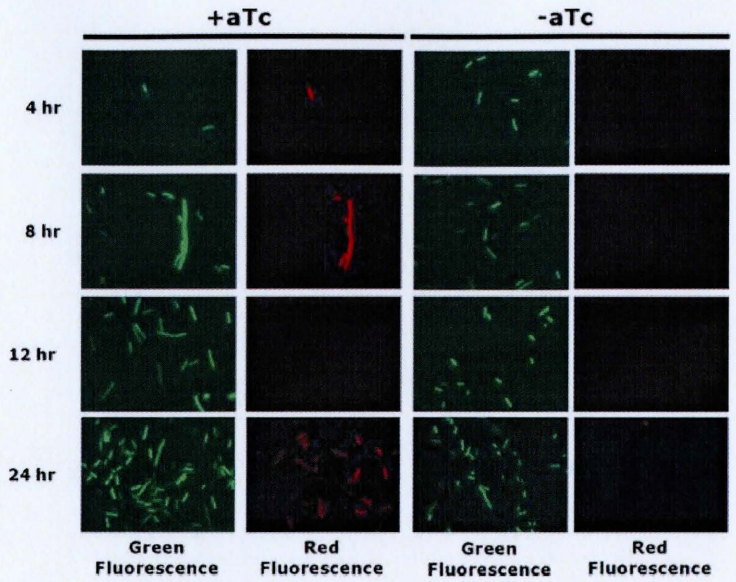


Figure 3-7: Fluorescent cell images of rRygC. See Figure 3-6 for legend.

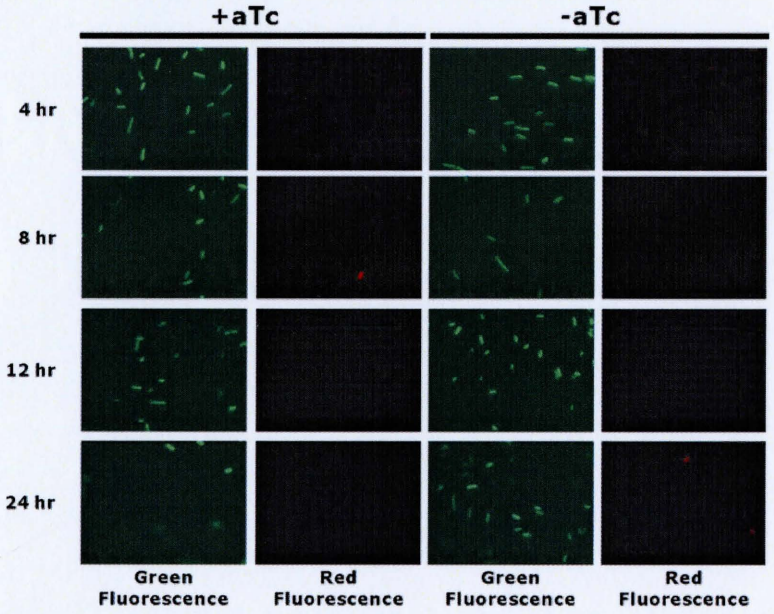


Figure 3-8: Fluorescent cell images of MCS11. See Figure 3-6 for legend.

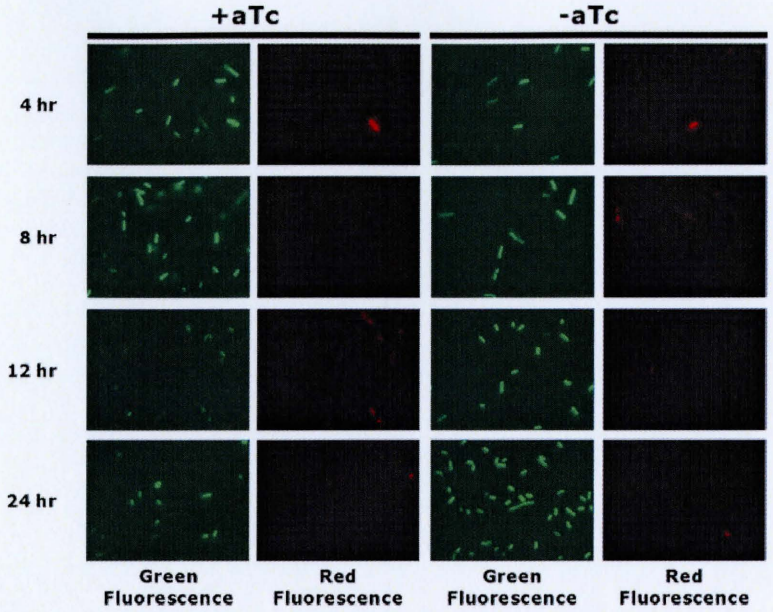


Figure 3-9: Fluorescent cell images of rUIG0803. See Figure 3-6 for legend.

3.1.8 Differential Protein Expression

The four clones, rUIG0803_4D, rUIG0803, rRygC and MCS11, were then subjected to a preliminary analysis of differential protein expression with and without aTc. Each cell line was grown overnight in LB containing 25 µg/mL of kanamycin and 50 µg/mL of spectinomycin; 1% of each cell culture was transferred into a fresh batch of LB with the same antibiotics in the absence or presence of 400 ng/mL of aTc. Cells were harvested and washed after 4, 8, 12 and 24 hr of growth and re-suspended in 1× PBS to a concentration of 1×10^8 cells/mL. Each cell suspension was then mixed with 6× protein loading buffer and subjected to total protein analysis on 10% and 15% SDS-PAGE, followed by staining with Commassie blue gel staining solution. The images of these gels are provided in Figure 3-10.

Differential protein expression was observed for both rRygC and rUIG0803_4D expressing cells at two locations, one at ~50 KDa (indicated by an arrow in Figure 3-10) and another at ~30 KDa (arrowhead) while the negative control clones MCS11 and clone rUIG0803 showed no significant difference in band intensity (thus protein expression) at these two locations. For rUIG0803_4D, reduced expression at 50 KDa location was observed at 8, 12 and 24 hr; at the 30 KDa location, increased expression was evident at 12 and 24 hr. For rRygC, the reduced expression at both 50 KDa and 30 kDa locations was most evident at 8 hr.

3.1.9 Protein Sequencing

The six protein bands (boxed in Figure 3-10) that showed differential expression upon aTc induction were excised and sent for Mass Spectrometry in the McMaster Regional Centre for Mass Spectrometry. The results indicated two possible proteins were likely involved, kanamycin kinase (~30 KDa) and glycerol kinase (~56 KDa).

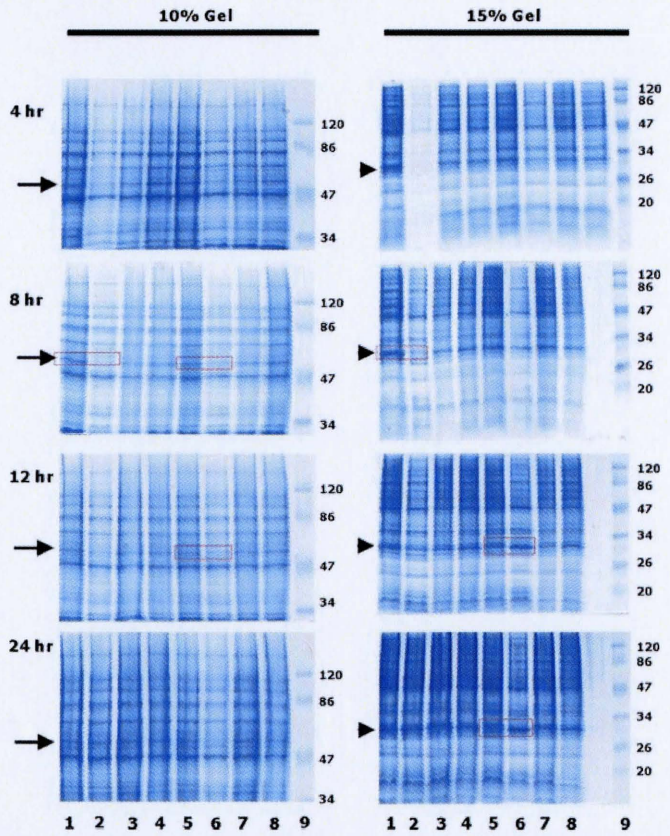


Figure 3-10: Protein expression in rRygC, MCS11, rUIG0803_4D and rUIG0803. All cultures were grown overnight in LB containing 25 µg/mL of kanamycin and 50 µg/mL of spectinomycin. 1% of each culture was re-inoculated in LB with the same two antibiotics. All clones were then growing in the absence or presence of 400 ng/mL of aTc. Cells were harvested and washed at 4, 8, 12 and 24 hr and re-suspended in 1× PBS in 1 × 10^8 cells/mL. Cells in suspension were then mixed with 6× protein loading buffer in 5:1 ratio and subject to 10% and 15% SDS-PAGE. Gels were stained with Commassie blue. The left gel on each row is a 10% gel while the gels on the right are 15% gels. Lanes 1 and 2: rRygC without and with aTc; 3 and 4: MCS11 without and with aTc; 5 and 6: UIG0803_4D without and with aTc; 7 and 8: UIG0803 without and with aTc. Lane 9 contains the protein markers.

3.2 Discussion

3.2.1 Bioinformatics Analysis

All the protein-coding and intergenic regions in *E. coli* MG1655 genome were predicted from RegCompareI. After the exclusion of all the protein-coding regions, 99% of the most conserved intergenic regions were found to be rRNAs or partial rRNA genes (the most conserved genes that comprised over 70% of most genomes). Thus, another exclusion step was applied to remove the rRNA genes as well as tRNA genes listed in the EcoGene database (Rudd 2000).

Eighty more previously identified sRNA genes in the published sRNA dataset were also excluded from the unannotated intergenic regions, the focus of this project (Hershberg, Altuvia et al. 2003). These manipulations created 3203 sequence entries, and they were named from UIG0001 to UIG3203. Since almost all the previously identified sRNAs are between 50 and 500 nucleotides in length, we removed 745 sequences that are beyond this size range in this effort. In the end, we established the RegCompare Dataset I with 2457 entries. Furthermore, no potential riboswitch sequences were detected using the Riboswitch finder in RegCompare DataSet I (Bengert and Dandekar 2004), supported by the fact that no riboswitch known to date is in the unannotated intergenic regions, where we conducted our bioinformatics analysis from.

We are interested in determining whether these UIG candidates exist in many microbial genomes and how well they are conserved across multiple genomes. Thus, we performed multiple sequence alignments using RegCompare II and found 82 UIGs that do occur in multiple bacterial genomes. Additionally, functional RNAs are known to have very defined or stable secondary structures. Thus, we also predicted the potential secondary structures of the 7 most conserved UIGs using Mfold (Figure 3-1 and Appendix 10). Interestingly, all seven UIGs indeed show interesting and relatively stable secondary structures, and this is consistent with their possible cellular regulatory role.

We also ran BLASTN search of all the UIGs candidates (Table 3-1) and obtained the information of their genomic locations as well as their adjacent gene information (Appendix 5). We found that most of the adjacent genes have interesting features that might guide us in the experimental testing of the predicted candidates in the future. For instance, the adjacent genes of UIG0242, UIG0985, UIG1195 and UIG1259 are hypothetical proteins that are believed to be involved in cell regulation (Riley, Abe et al. 2006), indicating that these UIGs might be involved in cellular regulation pathways. Interestingly, the adjacent genes for UIG0803 on both side of the candidate are predicted to be transposases. This may point to the possibility that the location of

UIG0803 could be mobile and thus its function is flexible. UIG1354 has tRNA genes both upstream and downstream of it, suggesting that UIG1354 might be involved in tRNA related cell regulation and possibly translational control. A tRNA gene is also present upstream of UIG1585, which has the DNA binding activator downstream. This may suggest that UIG1585 might be involved in translational control via DNA binding related mechanisms. We have also found that the more closely related the bacterial genomes are the higher level of conservation for the adjacent genes of the candidates are. To be clear, the adjacent gene of UIG0242 in *E. coli* is more identical to that in *Shigella* than the one in *Bacillus*. This could be explained by the fact that more closely related species have more similar genomic sequence content.

However, it should be noted that these seven candidates from the primary bioinformatics search are not experimentally verified yet. For instance, the existences of their RNA products in cells are not validated yet. One verification method would be the Northern Blotting experiment, which uses probes specific to each candidate. However, this method is relatively challenging and risky because the time and conditions of the RNA expression for any candidate could be variable and are not predictable at all. Due to the time constraint in this project and the amount of work that is required to perform Northern Blotting for all 7 candidates, we have decided to perform other simpler experimental testing first.

Conveniently, a tetracycline inducible system has been created in our lab by previous lab member Dr. Naveen Kumar. This system can identify a lethal cell phenotype from RNA expression of a candidate gene in *E. coli*. A vector, named pNYL9-MCS11, can express RNA under the control of a tetracycline promoter but it does not have a ribosome binding site (RBS) for protein expression. As a positive control, we used a previously identified clone called rRygC (small RNA RygC in pNYL9-MCS11 in reverse orientation). Upon induction by aTc, this clone results in a lethal phenotype. The empty vector pNYL-MCS11 itself was used as the negative control.

3.2.2 Amplification of 7 Candidate UIGs

The seven most well conserved UIGs were amplified from *E. coli* MG1655 genomic DNA. It is known that the orientation of inserted sRNA genes is important to its function. For instance, RygC in the antisense orientation can induce cell death while RygC in sense orientation cannot (data not shown). Considering this, we designed PCR primers to clone each candidate into the expression vector in both orientations. By making use of the multiple cloning sites in the

vector, we have incorporated *SalI* and *BamHI* recognition sites into the forward primer and *HindIII* site into the reverse primer (See Appendix 6 for the map of the multiple cloning sites in the vector). Thus if we digest the PCR products and the vector with *BamHI* and *HindIII* followed by ligation, we will obtain clones that express antisense RNA only. Conversely, the ligation product of the *SalI* and *HindIII* digested PCR products with vector will produce clones with sense RNA expressed only. As shown in Figure 3-2, we were able to amplify 6 of the 7 UIGs but UIG1195. The primers for this latter UIG contained many G residues, and G-rich sequences are known to form stable guanine quartet based structure (Shafer and Smirnov 2000), which might be a reason for the PCR failure. After several attempts, we decided to abandon this UIG candidate in this project.

3.2.3 Cloning and Sequencing

The amplified DNA fragments were successfully cloned into pNYL9-MCS11 vector and transformed into *E. coli* DH5αZ', which constitutively expresses a tetracycline repressor. The sequencing results showed that we obtained 13 different clones: 6 UIGs in both forward and reverse orientations and a deletion mutant of 1 UIG. The mutant clone had 4 nucleotide deletions and was named 'rUIG0803_4D' (the prefix r indicates a clone that is inserted in the reverse orientation, and 4D indicates 4 nt deletion).

3.2.4 aTc Induction

Upon aTc induction, a lethal phenotype was observed in clone rRygC, while clone MCS11 showed normal cell growth, indicating no detectable toxic effect of aTc at the given concentration. From this screen, clone rUIG0803_4D, which has a 4 nt deletion in rUIG0803, showed a lethal phenotype observed only 12 hr after 6 hr of aTc induction. The rest 12 clones, including rUIG0803, showed no lethal phenotype upon aTc induction. We repeated this experiment 3 times and similar lethal phenotype was always observed for clone rUIG0803_4D at least 12 hr following the induction. It should be noted that by the time the lethal phenotype of rUIG0803_4D was detectable, the rRygC colony started to recover from the lethal effect, demonstrated by clone 30 in Figure 3-3.

The delayed onset of lethality for rUIG0803_4D may suggest that the mechanism by which rUIG0803_4D regulates cell growth might be different from that of rRygC. However, this

assay does not show if the regulatory effect is growth phase dependent or cell population dependent.

Surprisingly, clone rUIG0803 showed no lethal phenotype. It begs the question why that 4 nt difference produces such a dramatic difference in cell growth upon RNA expression? Initially, we thought it could be due to the use of the cell line DH5α-Z', an *E. coli* cell line that is slightly different from MG1655. We used the genomic DNA from MG1655 because we used its published genomic sequence in our bioinformatics analysis. To test for sequence differences, we have also extracted genomic DNA from *E. coli* cell line DH5α-Z' and sent for sequencing (data not shown). The sequencing result showed that the sequence of UIG0803 region is exactly the same in both DH5α-Z' and MG1655.

There could be several possible explanations why only the clone with 4 nt deletion has a delayed lethal effect upon RNA expression. Firstly, it is possible that rUIG0803 is required as a binding component in a non-essential pathway, the binding partner of rUIG0803 could be either protein or RNA or both. And the affinity of the rUIG0803 RNA to its target might be a bit less than that of rUIG0803_4D due to the deletions. As a result, rUIG0803_4D may eventually replace rUIG0803 from their target by competitive binding and thus interfere with the regular binding of rUIG0803. The cells thus may eventually die out slowly because rUIG0803 could not bind to targets and carry out its normal regulatory function. It should be noticed that the difference in affinity between rUIG0803 and rUIG0803_4D might be trivial, which may explain why it took long time (12 hr) for clone rUIG0803_4D to show a visible lethal phenotype.

Secondly, rUIG0803 might not have any function in cells and thus do not interact with any target. The 4 nt deletions, however, might have changed the secondary structure of this RNA, allowing it to interact with the protein synthesis machinery (ribosomal binding, for example). Thus the binding of rUIG0803_4D might eventually disrupt the normal protein synthesis and lead to cell death because of protein deficiency. Lastly, rUIG0803 might be able to bind to the cell membrane after that 4 nt deletion and make the cells permeable to their environment thus lead the cells to eventual death. There are also other possible explanations why only the rUIG0803_4D RNA expression triggers delayed cell death.

Additionally, all 12 clones with 6 candidates expressed in both orientations didn't exert lethal effect on cell growth upon aTc induction. This might indicate that these RNAs are actually not expressed in cells and thus exert no effect on cell when expressed. If they are indeed expressed in cells, it still doesn't mean they have to kill cells when expressed. Thus, other assay

will be useful in assessing their functions. In conclusion, these clones should still be tested experimentally in the future at different aspects other than cell death.

3.2.5 Cell Permeability Assay of All 13 Clones

To further understand the effect of RNA expression from those 13 clones with two control clones, more assays, including the cell permeability assay using MacConkey agar, were conducted. MacConkey agar is the culture medium for growing Gram-negative bacteria including *E. coli* and for staining them for lactose fermentation. By utilizing the lactose available in the medium, *E. coli* with intact membranes will produce acid, which lowers the pH of the agar to below 6.8 and results in the appearance of red/pink colonies. However, if the cell membranes are not intact, the cells are permeable to peptone in the agar. Thus ammonia will form and raises the pH of the agar, resulting in the formation of white/colorless colonies.

All 13 clones were treated by aTc for 4 hr to ensure sufficient RNA expression before being transferred to the designated MacConkey Agar plate supplied with 50 μg/mL of spectinomycin and 25 μg/mL of kanamycin with or without 400 ng /mL of aTc (Figure 3-4). No white or colorless colonies were observed in any clones, suggesting that the membrane integrity of the clones were not disrupted upon RNA expression. Clone rUIG0803_4D only had lethal effect on cells after 4hr of aTc pre-incubation plus 6 hr on +aTc plate (clone 11 in Figure 3-4), which is consistent with the RNA expression assay data in Figure 3-3. On the other hand, cell growth was inhibited as soon as the rRygC clone was induced (clone 15 in Figure 3-4), which is also consistent with previous observation. In conclusion, the RNA expression of neither rUIG0803_4D nor rRygC changed the cell permeability and thus the membranes under both conditions were still intact. However, a proper control containing cells with permeable membrane should be included as the positive control. Since permeable membrane of cells will cause cell lysis and thus cell death, another assay should be implied to further test the cell permeability of candidates.

3.2.6 Growth Curve of Selected Clones

The RNA expression assay has been a great method in screening for clones that showed lethal phenotype upon aTc induction. However, this method could only tell us if there is detectable lethal phenotype occurs during the assay and the approximate time that occurs. Thus in

assessing the detailed effect of RNA expression of each clone on cell growth, a growth curve assay was performed. Clone rUIG0803_4D (the reverse complement of clone UIG0803_4D) and the positive control clone rRygC were subject to a growth curve assay. The negative control clone MCS11 was also examined in this experiment.

Three different aTc concentrations, 200 ng/mL, 400 ng/mL and 800 ng/mL, were used in this assay. It should be noted that in a similar experiment previously carried out by another lab member, only 100 ng/ mL of aTc was used. However, it was found that 100 ng/ mL of aTc may not induce the lethal phenotype at times due to its sensitivity to light and heat, thus cells were tested using different aTc concentrations with both positive and negative control clones for optimal aTc concentration in RNA expression assays. All 3 of the aforementioned aTc concentrations were able to induce RNA expression without exerting significant toxic effects on normal cells when growing the clones on agar plates. Control experiment showing RNA expression was done using the cells survived after rRygC RNA expression induction. The cells survived were shown to immune aTc induction afterwards, showing that small portion of cells survived lack rRygC RNA expression machinery (Figure 3-11).

For clone rRygC, a reduction in cell growth was observed within 1 hr of aTc induction. This effect was observed up to 14 hr of growth as the OD_{600} was much lower in aTc induced clone rRygC cells compared with un-induced cells (Figure 3-5A). No obvious dose dependent growth inhibition of clone rRygC was observed during the assay. For instance, the OD_{600} for 200 ng/mL and 400 ng/mL aTc induction is exactly the same during the assay. However, the OD_{600} with 800 ng/mL of aTc was slightly lower than that with other two aTc concentrations between 1.5 hr and 8 hr of growth. While dose dependent growth inhibition of clone rRygC RNA expression was observed in previous lab member's result. This is because the concentrations of aTc used in previous experiment were between 0 and 100 ng/mL, while we used 200 ng/mL, 400 ng/mL and 800 ng/mL in this assay. The effect of rRygC RNA expression might have been saturated between 100 and 200 ng/mL of aTc induction, thus higher aTc concentration wouldn't change the growth inhibition dose dependently any more. The reason OD_{600} with 800 ng/mL of aTc induction was slightly lower than other two aTc induction concentrations between 1.5 hr and 8 hr might because the high aTc concentration is slightly toxic to cells and thus start to kill cell in addition to the rRygC RNA expression. This is possible because we used agar plates with or without aTc to test for the optimal concentration in inducing RNA expression. While aTc did not exert toxic effects on normal cells, slight OD changes within 0.1 would not be obvious enough on

agar plate to be observed. Thus in the future, 800 ng/mL of aTc should not be used in assays because of its possible toxic effect.

As for clone rUIG0803_4D, a reduction in cell growth was observed after 4 hr of aTc induction and this effect was noticeable up to 14 hr of growth as lower OD_{600} was observed in aTc induced cells compared to the un-induced ones (Figure 3-5B). For instance, the growth inhibition becomes more apparent as the cells grow from 4 hr to 14 hr. However, the OD_{600} difference between the aTc induced cells and un-induced cells was smaller than 0.1 up to 8 hr of growth. After 8 hr of growth, the difference in cell density between induced and un-induced cells have become at least 0.2 units. This data is consistent with the previous RNA expression assay results, where apparent growth inhibition was only observed after more than 4 hr of growth. This growth curve has given us a more thorough view of cell growth upon aTc induction of rUIG0803_4D. It suggests the inhibition of cell growth started after around 4 hr and the inhibitory effect increases as the cell grows, thus the log phase has been shortened as compared to the un-induced cells. No cell growth difference was apparent among all cells with or without MCS11 RNA expression induction since no significant OD_{600} difference was shown at any time point of growth (Figure 3-5C). However, the OD_{600} was slightly lower in the clone induced by 800 ng/mL aTc when compared to three other clones by less than 0.1 starts from 1.5 hr of growth.

For future work, it will be interesting to see the growth inhibition of rRygC or rUIG0803_4D RNA over expression is growth phase dependent or cell population dependent. Thus, serial dilution of starting cell population should be tested in the same manner as the experiment here for both rRygC and rUIG0803_4D. If growth inhibition was observed around the same time as here, it should be time dependent otherwise it should be cell population dependent, and this will further open up our knowledge on the functions of these clones. In terms of optimal aTc concentration, 400 ng/mL of aTc should be used in future work.

Since we now have obtained detailed information of the RNA expression effect on cell growth, we also want to know if the cell morphology has changed, fluorescence microscopy study of these clones will be a great method of judging that.

3.2.7 Fluorescence Microscopy Study of Selected Clones

The positive control clone rRygC, and clone fRygC (RygC in forward direction), the negative control clone MCS11 as well as clone rUIG0803_4D were selected in this assay. Images

were photographed at 100× and 40× optical magnification and only pictures at 100× were presented. When incubated with the SYTO 9 and propidium iodide nucleic acid stains provided, live bacteria with intact cell membranes fluorescence green and dead bacteria with compromised membranes fluorescence red (Figure 3-6, 3-7, 3-8 and 3-9).

For clone rUIG0803_4D, cells were slightly elongated compared to un-induced cells at 4 hr of growth (Figure 3-6). No significant cell morphology differences were observed between cells with and the ones without clone rUIG0803_4D RNA expression at 8, 12 or 24 hr of growth. On the other hand, for clone rUIG0803, no apparent cell morphology or cell growth differences were observed between cells with and without RNA expression (Figure 3-9). It suggests that the expression of rUIG0803_4D RNA has caused abnormal cell elongation at 4 hr, which subsequently resulted in the growth inhibition. Cell elongation might indicate problems in cell division, thus rUIG0803_4D expression might have prevented cell division around 4 hr of growth for a short time period. This problem in cell division at 4 hr might have decreased the cell numbers in the media and as cells continue to propagate, a small change in cell number at 4 hr has been magnified leading to a high OD_{600} reading after 8 hr. In addition, when we compare the green panels on the left to the red panels from the same field on the right, more red cells appeared at 24 hr of growth than that at other time points. This indicates that there were large amount of dead cells at 24 hr, when the cells were going to the exponential phase, and thus more cell death was expected. However, even though only cells with compromised membrane will stain red, cells might be lysed and thus do not stain any more, thus the portion of red cell to green cell does not represent the actual portion of dead to live cells and hence were not studied in this case.

Interestingly, the positive control clone rRygC illustrated elongated cell morphology at 8 hr, 12hr and 24 hr of growth upon aTc induction (Figure 3-7). This suggests that cell division was interrupted at 8 hr, 12 hr and 24 hr of growth. However, the fact that the lethal phenotype was observed within 2 hr of aTc induction seems to suggest that rRygC RNA expression disrupted the cell division indirectly.

No cell morphology changes were observed during the assay for clone MCS11, indicating that the changes in cell morphology in clones rUIG0803_4D and rRygC were due to RNA expression from a specific DNA insert. This has further demonstrated that 400ng/mL of aTc is safe to the cells because no change in cell morphology was detected in the negative control clone. Furthermore, over-expression of random sequence didn't affect cell morphology, as shown by the result of the negative control clone MCS11. In addition, the fRygC clone didn't change cell

morphology within 24 hr, suggesting that the orientation of RNA expression for RygC is essential in its regulating roles.

In the future, cells with the RNA of these clones expressed in the pSAD system should also be examined to confirm their RNA expression effect. In addition, more clones from the bioinformatics screening should be examined in this assay. For instance, the rest 12 clones should also be studied to see the effects of their RNA expression on cell morphology.

3.2.8 Protein Profiling of Selected Clones

Up to this point, all the experimental characterizations of selected clones were only on cell growth or morphology. Thus we decided to examine the protein expression. SDS-PAGE was used to see the total protein distribution on a PAGE gel with amount of each protein proportional to its corresponding band intensity. 10% and 15% PAGE gels were used to ensure best resolution for proteins in both high and low size ranges.

Clones rUIG0803_4D and rUIG0803 were chosen in this study. The positive control clone rRygC and negative control clone MCS11 were also examined at the same time. In general, only two proteins, at around 50 kDa and 30 kDa locations, have shown unusual expression upon aTc induction. Abnormal protein expression has only occurred in either rRygC or rUIG0803_4D expressing cells. In addition, the cells with the negative control or rUIG0803 RNA showed no difference in protein expression between aTc induced and un-induced cells. This has confirmed that 400 ng/mL of aTc is not toxic to cells because no abnormal protein expression was observed. It also has suggested that over-expression of random sequence does not change the protein expression in cells, as supported by the data on MCS11 and rUIG0803 clones (Figure 3-10).

After 8 hr of growth, both the 50 KDa and the 30 KDa proteins showed decrease in expression when rRygC RNA was expressed, as the intensity of those two bands were much weaker in aTc induced lane than that in the un-induced lanes (Figure 3-10). This is the only time point where band intensity change between aTc induced and un-induced cells has occurred for clone rRygC. This has indicated that those two proteins have either lower synthesis rate or higher degradation rate after 8 hr of aTc induction of clone rRygC. The identity of those two proteins will further increase our knowledge on how rRygC RNA over-expression has caused the lethal phenotype. Thus those two bands were excised and sent for protein sequencing to identify possible protein candidates.

For clone rUIG0803_4D, abnormal protein expression occurred at 8, 12 and 24 hr of growth when induced with aTc. For instance, the 50 KDa protein showed decreased expression in aTc induced cells when compared to that of un-induced cells at 8 hr of growth. At 12 hr of growth, the same protein's expression was still lower in aTc induced cells while the 30 KDa protein was higher in induced cells compared to the un-induced cells. Furthermore, there was also an increase in the 30 KDa protein expressions at 24 hr of growth in aTc induced cells as compared to the un-induced cells. These results suggest that the 50 kDa protein has either lower synthesis rate or higher degradation rate around 8 to 12 hr of growth with aTc induction, while the 30 KDa protein has either increased synthesis rate or decreased degradation rate after 12 hr of aTc induction. These two bands have also been excised and sent for protein sequencing.

It should be noted that no change in protein expression was observed at 4 hr in both rRygC and rUIG0803_4D clones regardless of aTc induction. This suggests that the effect of RNA over-expression for these two clones on protein expression was not obvious at 4 hr of growth. However, the total proteins that were loaded into each well are not exactly the same because no protein concentration assay was performed, as seen on the PAGE gels. We might have missed out some protein bands that showed a slight change in expression when induced with aTc.

Thus for the future work, this experiment should be repeated with an additional protein concentration determination step before loading the gel so that equal amount of total protein will be examined. Two-dimensional gel electrophoresis may be utilized in obtaining a more comprehensive view on protein expression.

3.2.9 Protein Sequencing

The protein sequencing results showed that the 30 KDa is most likely to be aminoglycoside 3'-phosphotransferase (kanamycin kinase) and glycerol kinase is the most possible candidate to be the protein around 56 KDa, which is consistent with Figure 3-10.

Aminoglycoside 3'-phosphotransferase (kanamycin kinase) expression was down-regulated at 8 hr of growth in rRygC RNA induced cells. It was up-regulated at 12 and 24 hr of growth in rUIG0803_4D RNA expressed cells. This protein functions in conferring resistance to the antibiotic kanamycin. However, kanamycin resistance gene was encoded in the vector we used for inducible RNA expression, pNYL9-MCS11 and was used as a selectable marker to select for cells expressing this vector. Thus only cells with this vector will survive in the presence

of kanamycin. It is not clear to us at this moment why there is expression alteration of this protein at various growth phases when rRygC or rUIG0803_4D RNA was expressed, and two possible answers to that are listed as follows. It might because the expression rRygC or rUIG0803_4D RNA has interfered with the kanamycin resistance gene expression. It is possible that rRygC RNA binds to its target with a higher affinity while rUIG0803_4D binds with much lower affinity. However, a control experiment with rRygC RNA expressed in pSAD under chloramphenichol and arabinose control was studied and similar lethal phenotype was observed (data not shown). This suggests that the expression of rRygC RNA might have disrupted the expression of the kanamycin resistance gene or chloramphenichol gene expression, leading to cell death. Without conducting the protein profiling experiment using a pSAD expression vector with rRygC, however, it is difficult to make a precise conclusion.

Glycerol kinase expression was down-regulated at 8 hr of growth in both rRygC and rUIG0803_4D RNA expressed cells. It was also down-regulated at 12 hr of growth in rUI0803_4D RNA expressed cells. This protein is involved in glycerol utilization of glycerol as carbon source and thus is the key enzyme in the regulation of glycerol uptake and metabolism. It catalyzes the transfer of a phosphate from ATP to glycerol forming glycerol phosphate. The decreased expression of this glycerol kinase might have caused problem in glycerol uptake thus eventually leading to cell death. However, since no total protein determination was performed before the SDS-PAGE analysis, there might be more growth phase or point showing differential expression of glycerol kinase. In addition, not only differential protein expression could cause decreased cell growth, proteins with no differential expression but altered activity may also lead to cell death. Thus, co-immunoprecipitation experiment should be done in the future to find the potential binding partners of rRygC and rUIG0803_4D RNA. However, a proper positive control should be included in the assay to show the white/colorless colonies in the future. Any gram positive bacteria will be a good positive control because they can take up peptone and ammonia thus raising the pH of the agar, and white/colorless colonies will form.

For the future work from the bioinformatics prospect, more candidates could be selected from the RegCompare DataSet II by expanding the parameters. For instance, lowering the homology requirement from 50% to 45% or changing the required size range from 50-500 to 50-700 nt would generate more candidates. With the constant discovery of more and more functional sRNAs, more sequences could be eliminated from RegCompare DataSet I as well. Finally, Northern Blotting experiments should be performed on those 7 UIG candidates we have selected above for preliminary verification of these sRNAs candidates.

For those top ranked UIG candidates, first of all, effort should be made to amplify UIG1195 using a different method, even though it is not optimal to change the length of the candidate, new primer sets with longer or shorter amplification product than the candidate should be attempted. In addition to that, another RNA expression system should be utilized using different antibiotic choices to exclude the potential effect of the antibiotic used in the assay. One alternative to the pNYL9-MCS11 system is the pSAD one, which uses chloramphenichol as the antibiotic selection marker and arabinose as the RNA expression inducing reagents. This system has actually been used in the case of rRygC and same lethal phenotype was observed (data not shown). The 13 clones mentioned above should also be tested in pSAD with chloramphenichol and arabinose to examine the RNA expression effect in future work.

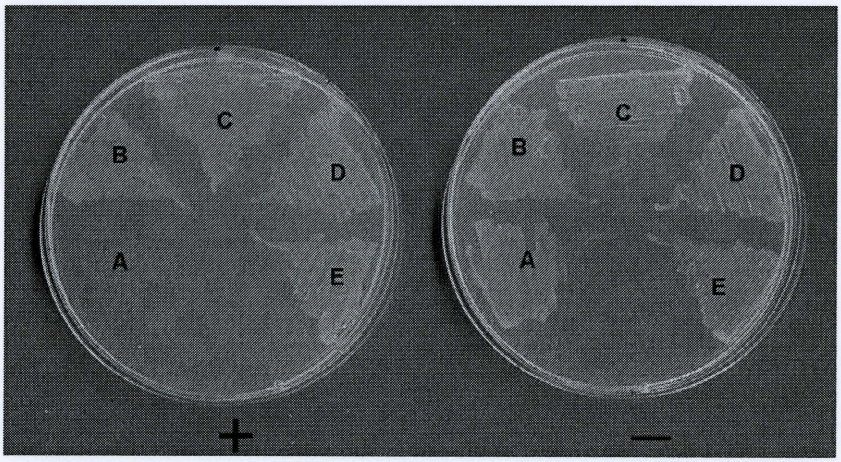


Figure 3-11: rRygC RNA induction assay. All cultures were grown overnight in LB containing 25 μg/mL of kanamycin and 50 μg/mL of spectinomycin. 1% of each culture was re-inoculated in LB with the same two antibiotics. All clones were then growing in the absence (-) or presence (+) of 400 ng/mL of aTc. A: Fresh rRygC in pNYL9-MCS11 with OD ~0.3. B: rRygC survived from A on aTc treated plate after 6 hours of incubation. C: rRygC from two series of sub-inoculation with Kan, Spec, aTc of A. D: rRygC from three series of sub-inoculation with Kan, Spec, aTc of A. E: rRygC from four series of sub-inoculation with Kan, Spec, aTc of A. Cells survived from A didn't have the plasmid with rRygC, which lead to no lethal phenotype observed when they were treated with aTc because lack of RNA expression in these cells.

Chapter Four

Experimental Characterization of Known sRNAs

4.1 Results

4.1.1 Amplification of Selected sRNAs

Following the protocol in section 3.1.2, we amplified the following 8 sRNA genes from the *E. coli* MG1655 genomic DNA: C0293, C0299, C0343, sraD, sraI, sraL, tpke11 and ssrA. The corresponding bands on 2% TAE-agarose gels (Figure 4-1) are consistent with the predicted sizes of these 8 candidates. All the primer sequences are provided in Appendix 7.

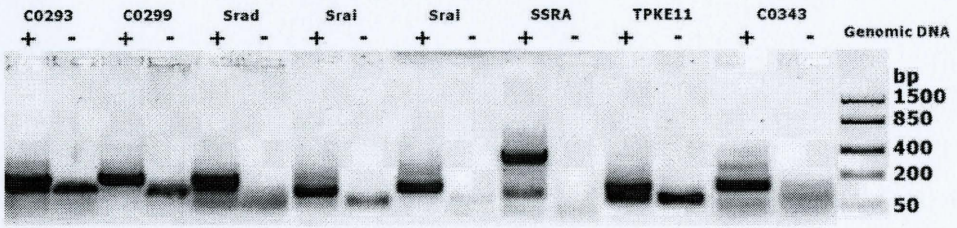


Figure 4-1: Amplification of 8 sRNAs. The PCR products were analyzed on a 2% TAE-agarose gel. The lane on the right is a DNA ladder. The expected sizes of the PCR products: C0293, 73 nt; C0299, 79 nt; sraD, 75 nt; sraI, 94 nt; sraL, 140 nt; ssrA, 363 nt; tpke11, 89 nt; C0343, 75 nt.

4.1.2 Sub-cloning and Expression of 8 sRNA Genes

The PCR amplified sRNA genes were cloned into pNYL9-MCS11 in both orientations and were transformed into the *E. coli* DH5α-Z' cell line using the method described in section 3.1.3. In total, 16 clones were established and each contains a given sRNA gene in one orientation. The DNA sequence of each clone was verified by sequencing and no mutation was observed in any clone.

4.1.3 aTc Induction

The aTc induction experiment was performed on all 16 clones using the procedure described in section 3.1.4 (rRygC and MCS11 were used as positive and negative controls, respectively), and the data are given in Figure 4-2. There was no clone showing detectable lethal phenotype upon aTc induction.

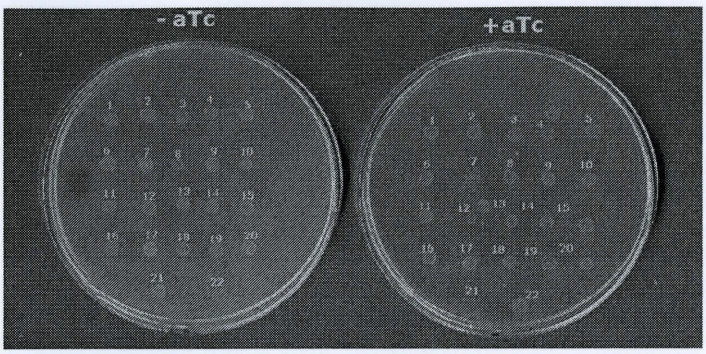


Figure 4-2: Phenotypic assay on LB agar plates. Cells were grown overnight in LB containing 25 µg/mL of kanamycin and 50 µg/mL of spectinomycin. 1.5 µL of each culture was spotted on 1.5% LB agar plates (containing the same antibiotics) in the absence (left) or presence (right) of 400 ng/mL of aTc. Clone 1: fC0293 (f, forward orientation); 2: rC0293 (r, reverse orientation); 3: fC0299; 4: rC0299; 5: fC0343; 6: rC0343; 7: fsraD; 8: rsraD; 9: fsraI; 10: rsraI; 11: fsraL; 12: rsraL; 13: fssrA; 14: rssrA; 15: ftpke11; 16: rtpke11. 17-20: clones from other known sRNAs (not discussed here). 21 and 22: rRygC (a positive control) and MCS11 (a negative control).

4.1.4 Cell Permeability Assay

Cell permeability assay was also performed on all 16 clones utilizing the protocol in section 2.7. rRygC and MCS11 were included in this assay as controls. All 18 clones were induced by aTc for 4 hr before being transferred to MacConkey agar plate supplied with 50 µg/mL of spectinomycin and 25 µg/mL of kanamycin in the absence or presence of 400 ng/mL of aTc (Figure 4-3). No visible white or colorless colonies were observed in any clones that were tested upon induction. However, the rRygC clone exhibited a lethal phenotype (Figure 4-3).

Figure 4-3: Assay on MacConkey plates. Cells were grown overnight in LB containing 25 μg/mL of kanamycin and 50 μg/mL of spectinomycin. After 4 hr of sub-inoculation of 1% overnight cells with 400 ng/mL of aTc, each culture was streaked on 1.5% MacConkey agar plates in the absence (left) or presence (right) of 400 ng/mL of aTc. 1-16: fC0293, rC0293, fC0299, rC0299, fC0343, rC0343, fsraD, rsraD, fsraI, rsraI, fsraL, rsraL, fssrA, rssrA, ftpke11, rtpke11; 17 and 19: MCS11; 18 and 20: rRygC.

4.2 Discussion

Eight known sRNAs, C0293, C0299, C0343, sraD, sraI, sraL, tpke11 and ssrA, were successfully amplified and cloned into pNYL9-MCS11 in two orientations, generating 16 clones.

In the aTc induction experiment, rRygC (the positive control) and MCS11 (the negative control) were tested in addition to those 16 clones. Upon aTc induction, only rRygC showed a noticeable lethal phenotype while none of the 16 test clones or MCS11 showed detectable lethal phenotype. This data shows that the over-expression of these small RNAs or their antisense counterpart do not inhibit cell growth. Therefore, these sRNAs might function in a mechanism that is different from the one by rRygC or their RNA expression may not be essential to cell growth. However, our results do not mean that these sRNAs do not have a function in cells. Thus a different functional assay will have to be developed to assess their biological roles.

All 16 clones were also subjected to the cell permeability assay. No change in color was observed in all the clones. Thus the over-expression of these 8 small RNAs in any orientation does not appear to change cell permeability. However, a proper control containing cells with permeable membrane should be included as the positive control. Since permeable membrane of cells will cause cell lysis and thus cell death. Another assay should be implied to further test the cell permeability of candidates.

Chapter Five

Initial Characterization of a Random Lambda Phage Genomic Library

5.1 Results

5.1.1 Generation of a Random Genomic Library from Lambda Phage

The genomic DNA from Lambda phage was digested with *Sau3A1* for 60 min, and the reaction was stopped by addition of 0.5M EDTA (pH 8.0). The DNA mixture was analyzed on a 1% agarose gel, along with the undigested Lambda phage genomic DNA as a control (Figure 5-1). The DNA fragments in the highlighted region from Figure 5-1 were excised and purified.

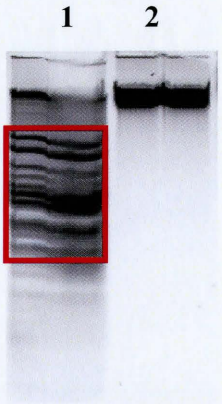


Figure 5-1: Digestion of Lambda phage genomic DNA using *Sau3AI* analyzed on 1% agarose gel. Lane 1 - Lambda phage genomic DNA digested with *Sau3AI* for 60 minutes. The DNA bands in red box were excised and purified. Lane 2 - Lambda phage genomic DNA.

The pNYL9-MCS11 vector was digested with *BamHI* and purified on a 1% agarose gel. The band corresponding to the digested pNYL9-MCS11 was excised and extracted. The 5′ phosphates of the digested vector were removed using calf intestinal alkaline phosphotase (CIAP) to prevent self-ligation.

The Lambda genomic DNA and the linearized vector were ligated and then transformed into *E. coli* DH5αZ' competent cells. 192 colonies were selected for initial characterization. Twenty random colonies were sent for sequencing and the results revealed that all twenty clones contained a fraction of Lambda genomic DNA.

5.1.2 Activity Screening

All 192 clones were assayed on their RNA expression effect following the protocol described above for the 7 UIG candidates (sections 2.6 and 2.7). Clones 140L, 141L and 152L

showed significant growth inhibition upon aTc induction, while the rest 189 clones didn't show detectable growth changes (Figure 5-2).

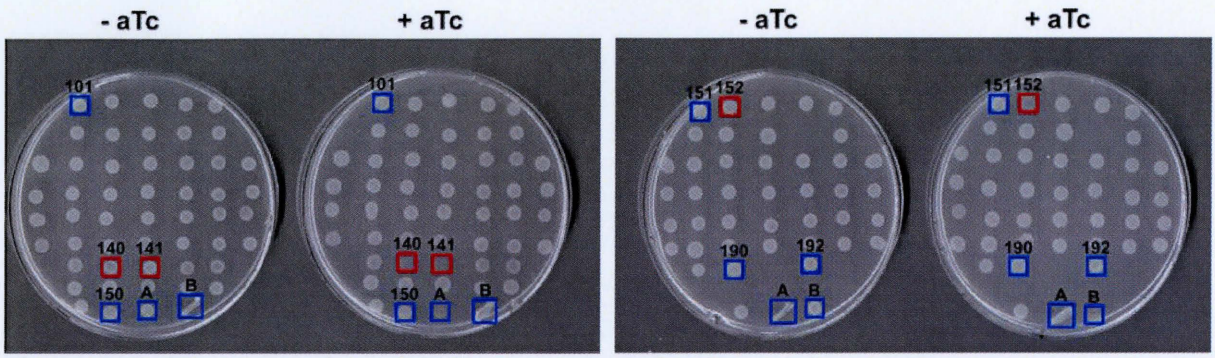


Figure 5-2: The expression of RNA upon aTc induction. Cultures were grown overnight in LB containing 25 μg/mL of kanamycin and 50 μg/mL of spectinomycin. 1.5 μL of each culture was spotted on 1.5% LB agar plates with and without 400 ng/mL of aTc. Clones 140L, 141L and 152L showed significant lethality. Clone A was the positive control clone rRygC and clone B was the negative control clone MCS11.

5.1.3 Growth Curve

Clones 140L, 141L, 152L, as well as rRygC and MCS11 were selected in the growth curve assay using the protocol previously described (section 2.8). The results of this assay are presented in Figure 5-3. After 2 hr, clones 140L and 152L started to grow slower with aTc induction. While for clone 141L, apparent growth inhibition was also observed with aTc induction past 2 hr, and cells appeared to die out after 4 hr (the optical density decreased significantly). It should be noted that clone rRygC grew in an aTc dependent pattern after 2 hr. No significant growth inhibition was observed for MCS11. It is also noteworthy that there was only minor growth differences (by optical density) among various aTc concentrations, which have suggested that 400 ng/mL of aTc is appropriate to induce RNA expression without toxic effect.

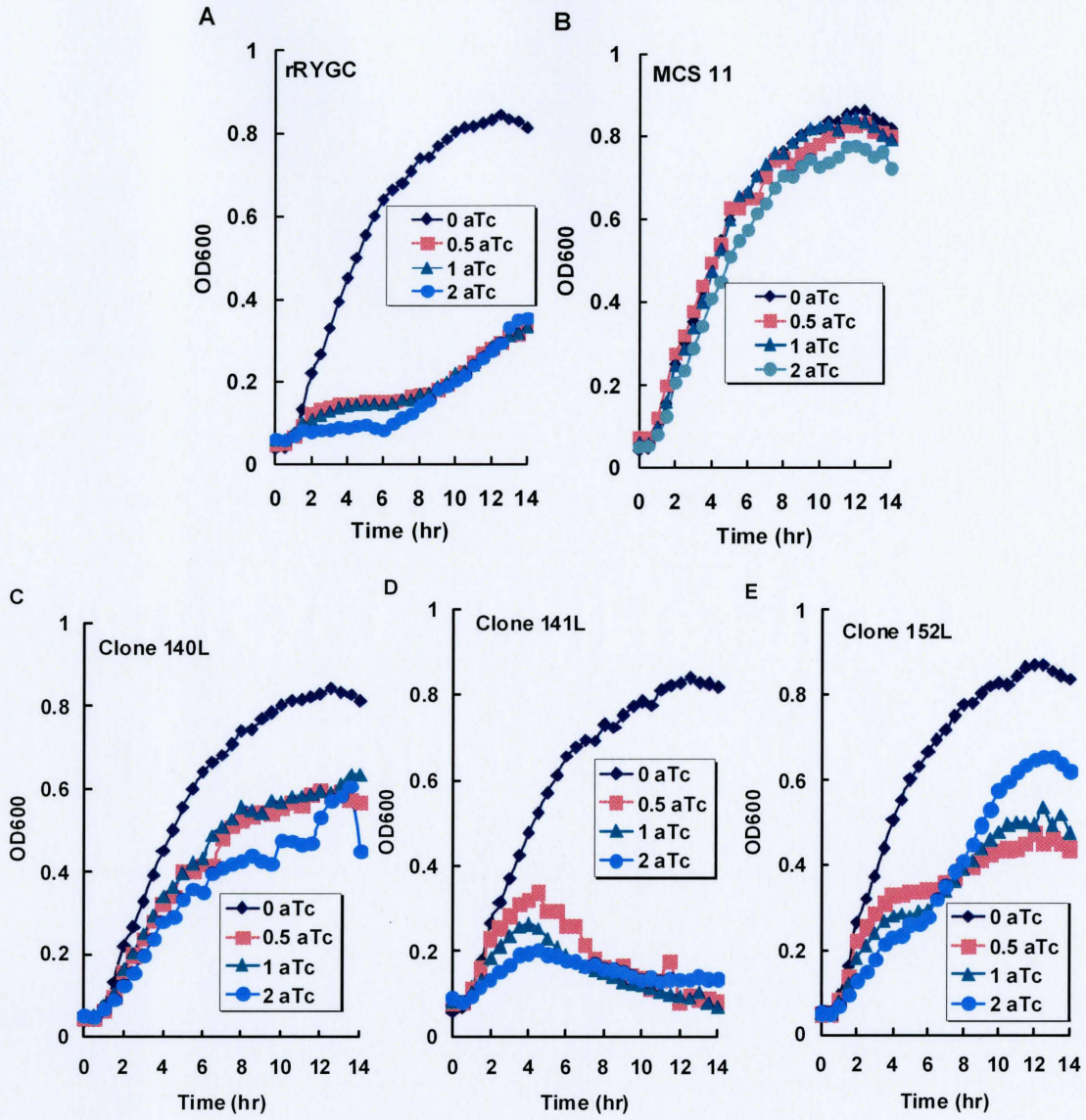


Figure 5-3: Growth curve of selected clones at various aTc concentrations. Cultures were grown overnight in LB broth containing kanamycin and spectinomycin. 1% of each culture was freshly inoculated with the same two antibiotics in the absence of aTc (0 aTc) and presence of 200 ng/mL (0.5 aTc), 400 ng/mL (1 aTc) and 800 ng/mL of aTc (2 aTc). OD600 was taken every 30 min up to 14 hr. (A) rRygC, (B) MCS11, (C) 140L, (D) 141L, (E) 152L.

5.2 Discussion, Conclusions and Future Work

5.2.1 Initial Examination of the Selected Clones

Following successful cloning of the DNA fragments into pNYL9-MCS11, 192 clones were assayed using the same protocol previously discussed (section 2.5). Clones 140L, 141L and

152L showed growth inhibition upon aTc induction, while the rest didn't show detectable growth change. Clones 140L, 141L and 152L were sequenced and the sequence information can be found in Appendix 11. BLAST search has been performed and the results are highlighted and presented in Figure 5-4.

Clone 140L contains a DNA motif of 669 nt in length (Figure 5-3A). DNA fragment $140L_{19\text{-}669}$ (the sequence from 19 to 669 nt in 140L) shows 97% identity with the antisense strand of entire *E. coli* putative single-stranded DNA binding protein of prophage gene, and this same sequence is also 97% identical to antisense strand of the full lambda ant-restriction protein gene. $140L_{118\text{-}440}$ is found to be 95% identical to the antisense strand of both the whole lambda ant-restriction protein N gene and the complete lambda restriction inhibitor protein ral gene. $140L_{118\text{-}394}$ is 98% identical to the entire Bacteriophage phi-21 transcription antitermination protein (N) gene in the sense direction. $140L_{396\text{-}669}$ is found also 95% identical to antisense strand of the whole Stx2 converting phage II DNA gene.

Clone 141L contains a DNA sequence of 960 nt long (Figure 5-3B). The BLAST results showed that three different sections of this DNA motif showed high level of identity to various complete lambda genes all in sense direction. $141L_{65\text{-}765}$ shows 100% identity with the outer membrane protein Lom precursor gene. $141L_{130\text{-}750}$ is found to be 90% identical to the conserved bacterial internalization gene protein (cig) gene. $141L_{623\text{-}960}$ is also found to be 87% identical to the putative tail fiber protein.

Clone 152L, which is 953 nt in length, contains a DNA fragment $152L_{624\text{-}853}$ that is 88% identical to entire putative outer host membrane protein precursor / putative fiber protein gene in both directions at multiple locations in *E. coli*. $152L_{66\text{-}766}$ is 100% identical to the complete outer host membrane of Enterobacteria phage lambda protein gene in sense direction. $152L_{131\text{-}751}$ is also 100% identical to the whole sense Bacteriophage lambda conserved bacterial internalization gene protein (cig) gene. There are two sections in 152L that are 100% identical to the partial lambda phage protein genes in sense direction. For instance, $152L_{197\text{-}748}$ is identical to the partial Bacteriophage mep503 Cig5-like protein gene, and $152L_{175\text{-}748}$ is identical to the Bacteriophage mep123 cig12-like gene.

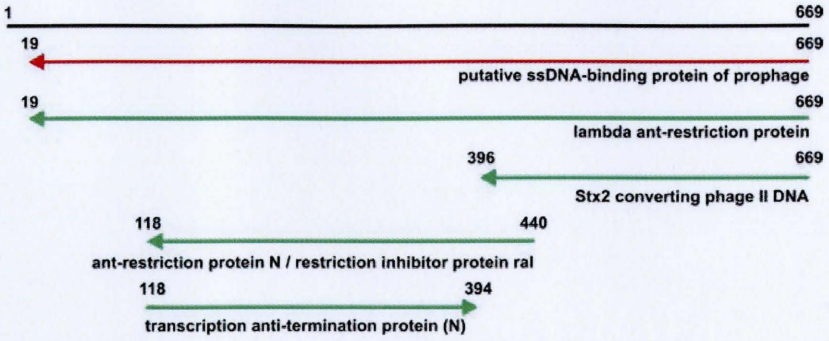


Figure 5-4A

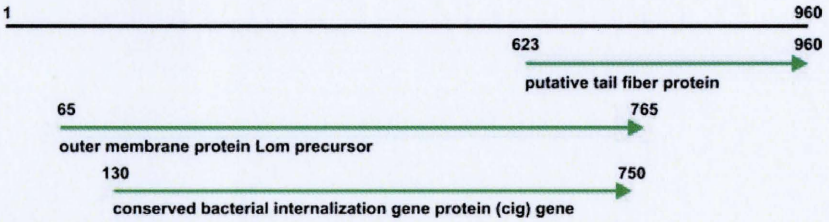


Figure 5-4B

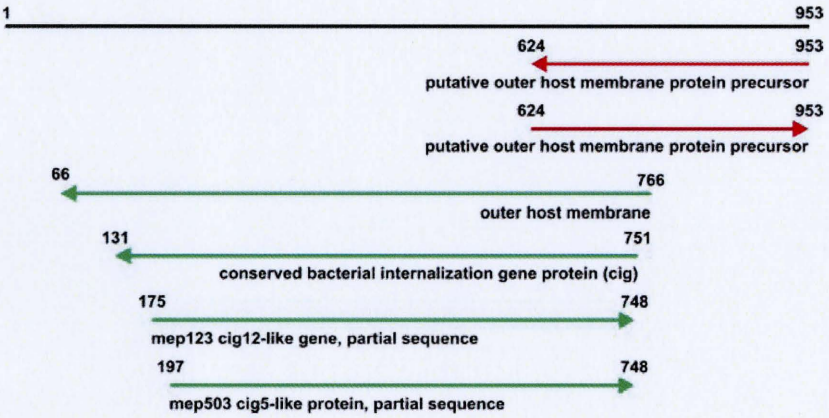


Figure 5-4C

Figure 5-4: Summary of the BLAST results of three candidates from Lambda phage library screening. The green arrows represent proteins in lambda phage, red in *E. coli*, arrow to the right indicates sense direction, arrow to the left meaning antisense direction, proteins are whole proteins unless specified at the end, there are two partial proteins in clone 152L. 5-3A: 140L. 5-3B: 141L. 5-3C: 152L.

5.2.2 Growth Curve

Clones 140L, 141L, 152L, as well as the positive control clone rRygC and negative control clone MCS11 were assayed following the protocol of generating the growth curve of

clone rUIG0803_4D (section 2.8). Each point on the plot represents the average of two duplicate samples with background reading deducted. Since the same growth curve was plotted for clones rRygC and MCS11 and was discussed previously in Chapter 3, the results of these two clones are not discussed in detail in this section.

For clone 140L, a slight decrease in growth was observed in aTc induced cells from 2 hr of growth compared to the un-induced cells. However, the cells in both conditions still grew in the same pattern as both have gradual increase in OD during the assay. In addition, there is slight growth decrease in 800 ng/mL of aTc induced clones compared to 200 ng/mL and 400 ng/mL of aTc induced cells, which may suggest that 800 ng/mL of aTc is slightly toxic to cells and thus start to kill cells. More importantly, no dose-dependent growth inhibition was observed, suggesting that effect of the 140L RNA expression on cell growth was saturated with 200 ng/mL of aTc.

While for clone 141L, growth inhibition was observed in aTc induced cells after 2 hr and no detectable growth was observed after 4 hr compared to the un-induced cells. Interestingly, a dose-dependent growth inhibition was observed up to 7 hr, this suggests that the effect of 141L RNA over-expression was not saturated with 200 ng/mL of aTc induction.

For clone 152L, a general decrease in growth was observed in aTc induced cells after 2 hr compared to the un-induced cells. However, a dose-dependent growth inhibition was observed up to 7 hr of growth with the cells with 800 ng/mL of aTc induction showing lowest cell density. On the other hand, starting from 7 hr, reverse dose-dependent growth inhibition was observed with the 800 ng/mL of aTc induction showing highest cell density. We cannot explain this interesting growth response.

Despite the large size of these three sequences, none of them contain the prokaryotic consensus ribosomal binding site, the essential element in protein synthesis that has been removed from pNYL9-MCS11. Thus, there is little chance for protein translation upon aTc-induced transcription. So the change in cell growth of three clones is less likely caused by protein synthesis. However, the synthesis of large RNAs of these three clones might have disrupted the regular transcription or translation machinery and thus affected cell growth by affecting protein synthesis.

We have yet to determine molecular mechanism(s) behind the inhibitory effect for the three clones identified. However, the BLAST search result, together with the growth curve, should provide priming information for future investigations. The growth curve assay has also

indicated that the optimal aTc induction concentration is clone-dependant and thus should be optimized for each clone (400 ng/mL of aTc is a good starting concentration).

Chapter Six

Summary and Contribution

Firstly, I developed the algorithms that can be used to perform bioinformatics search for conserved sequence motifs in unannotated intergenic regions in bacterial genomes. By applying these algorithms, I created a set of ~70 such candidates in *E. coli* genome that show significant level of conservation to other bacterial genomes. To the best of my knowledge, this is the first time that a bioinformatics analysis was performed to the unannotated intergenic regions in bacterial genomes. This effort opens up a new research direction in our lab and will help expand sRNA research in general.

Secondly, I performed some preliminary experiments to probe the possible biological functions of the top seven candidates from the bioinformatics search. The key component of these experiments was the use of a plasmid system that can induce RNA expression from a DNA insert of interest. The data from the solid agar assay, the growth curve experiment and the microscope study provide the preliminary characterization that should assist future, more in-depth analysis of these candidate genes.

Thirdly, I cloned 8 previously identified sRNA genes into the same plasmid system. These genes were reported by other researchers from bioinformatics search of sRNAs and their cellular expression was experimentally validated; however, their functions have not been elucidated. I found that that the RNA expression of these 8 candidates is not lethal to the cells. This result means that we will have to develop other assays to assess the biological functions of these and other sRNA genes.

Finally, I have obtained a Lambda phage DNA library by random digestion of the Lambda phage genome. Using the above plasmid system, I discovered three DNA fragments, which, upon inducible RNA expression, can slow down the growth of the host *E. coli*. Future experiments are required to analyze the mechanisms behind the observed phenotype.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25, 3389-3402.

Andersen, J., Forst, S. A., Zhao, K., Inouye, M. & Delihas, N. (1989). The function of micF RNA. micF RNA is a major factor in the thermal regulation of OmpF protein in Escherichia coli. *J Biol Chem* 264, 17961-17970.

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G., Margalit, H. & Altuvia, S. (2001). Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr Biol* 11, 941-950.

Axmann, I. M., Kensche, P., Vogel, J., Kohl, S., Herzel, H. & Hess, W. R. (2005). Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biol* 6, R73.

Barrick, J. E., Sudarsan, N., Weinberg, Z., Ruzzo, W. L. & Breaker, R. R. (2005). 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *Rna* 11, 774-784.

Bengert, P. & Dandekar, T. (2004). Riboswitch finder--a tool for identification of riboswitch RNAs. *Nucleic acids research* 32, W154-159.

Chen, S., Lesnik, E. A., Hall, T. A., Sampath, R., Griffey, R. H., Ecker, D. J. & Blyn, L. B. (2002). A bioinformatics based approach to discover small RNA genes in the Escherichia coli genome. *Bio Systems* 65, 157-177.

Chen, S., Zhang, A., Blyn, L. B. & Storz, G. (2004). MicC, a second small-RNA regulator of Omp protein expression in Escherichia coli. *J Bacteriol* 186, 6689-6697.

Coyer, J., Andersen, J., Forst, S. A., Inouye, M. & Delihas, N. (1990). micF RNA in ompB mutants of Escherichia coli: different pathways regulate micF RNA levels in response to osmolarity and temperature change. *J Bacteriol* 172, 4143-4150.

Delihas, N. & Forst, S. (2001). MicF: an antisense RNA gene involved in response of Escherichia coli to global stress factors. *J Mol Biol* 313, 1-12.

Eddy, S. R. (2002). Computational genomics of noncoding RNA genes. *Cell* 109, 137-140.

Gish, W. & States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nat Genet* 3, 266-272.

Gopalan, V., Vioque, A. & Altman, S. (2002). RNase P: variations and uses. *The Journal of biological chemistry* 277, 6759-6762.

Gottesman, S. (2004). The small RNA regulators of Escherichia coli: roles and mechanisms*. *Annual review of microbiology* 58, 303-328.

Gottesman, S., McCullen, C. A., Guillier, M. & other authors (2006). Small RNA Regulators and the Bacterial Response to Stress. *Cold Spring Harb Symp Quant Biol* 71, 1-11.

Griffin, B. E. (1971). Separation of 32P-labelled ribonucleic acid components. The use of polyethylenimine-cellulose (TLC) as a second dimension in separating oligoribonucleotides of '4.5 S' and 5 S from *E. coli*. *FEBS Lett* 15, 165-168.

Heeb, S., Blumer, C. & Haas, D. (2002). Regulatory RNA as mediator in GacA/RsmA-dependent global control of exoproduct formation in Pseudomonas fluorescens CHA0. *Journal of bacteriology* 184, 1046-1056.

Hershberg, R., Altuvia, S. & Margalit, H. (2003). A survey of small RNA-encoding genes in Escherichia coli. *Nucleic acids research* 31, 1813-1820.

Huttenhofer, A., Kiefmann, M., Meier-Ewert, S., O'Brien, J., Lehrach, H., Bachellerie, J. P. & Brosius, J. (2001). RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *Embo J* 20, 2943-2953.

Huttenhofer, A., Brosius, J. & Bachellerie, J. P. (2002). RNomics: identification and function of small, non-messenger RNAs. *Curr Opin Chem Biol* 6, 835-843.

Huttenhofer, A., Cavaille, J. & Bachellerie, J. P. (2004). Experimental RNomics: a global approach to identifying small nuclear RNAs and their targets in different model organisms. *Methods Mol Biol* 265, 409-428.

Ikemura, T. & Dahlberg, J. E. (1973). Small ribonucleic acids of Escherichia coli. I. Characterization by polyacrylamide gel electrophoresis and fingerprint analysis. *J Biol Chem* 248, 5024-5032.

Jensen, C. G. & Pedersen, S. (1994). Concentrations of 4.5S RNA and Ffh protein in Escherichia coli: the stability of Ffh protein is dependent on the concentration of 4.5S RNA. *Journal of bacteriology* 176, 7148-7154.

Kawano, M., Reynolds, A. A., Miranda-Rios, J. & Storz, G. (2005). Detection of 5'- and 3'-UTR-derived small RNAs and cis-encoded antisense RNAs in Escherichia coli. *Nucleic acids research* 33, 1040-1050.

Keiler, K. C., Waller, P. R. & Sauer, R. T. (1996). Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science* 271, 990-993.

Lavorgna, G., Dahary, D., Lehner, B., Sorek, R., Sanderson, C. M. & Casari, G. (2004). In search of antisense. *Trends Biochem Sci* 29, 88-94.

Liu, M. Y., Gui, G., Wei, B., Preston, J. F., 3rd, Oakford, L., Yuksel, U., Giedroc, D. P. & Romeo, T. (1997). The RNA molecule CsrB binds to the global regulatory protein CsrA and

antagonizes its activity in Escherichia coli. *The Journal of biological chemistry* 272, 17502-17510.

Lorenz, C., von Pelchrzim, F. & Schroeder, R. (2006). Genomic systematic evolution of ligands by exponential enrichment (Genomic SELEX) for the identification of protein-binding RNAs independent of their expression levels. *Nature protocols* 1, 2204-2212.

Madden, T. L., Tatusov, R. L. & Zhang, J. (1996). Applications of network BLAST server. *Methods Enzymol* 266, 131-141.

Masse, E. & Gottesman, S. (2002). A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc Natl Acad Sci U S A* 99, 4620-4625.

Masse, E., Escorcia, F. E. & Gottesman, S. (2003). Coupled degradation of a small regulatory RNA and its mRNA targets in Escherichia coli. *Genes Dev* 17, 2374-2383.

Mattick, J. S. & Makunin, I. V. (2006). Non-coding RNA. *Human molecular genetics* 15 Spec No 1, R17-29.

Mizuno, T., Chou, M. Y. & Inouye, M. (1984). A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). *Proc Natl Acad Sci U S A* 81, 1966-1970.

Moller, T., Franch, T., Hojrup, P., Keene, D. R., Bachinger, H. P., Brennan, R. G. & Valentin-Hansen, P. (2002). Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Molecular cell* 9, 23-30.

Montzka, K. A. & Steitz, J. A. (1988). Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc. *Proc Natl Acad Sci U S A* 85, 8885-8889.

Narajczyk, M., Baranska, S., Szambowska, A., Glinkowska, M., Wegrzyn, A. & Wegrzyn, G. (2007). Modulation of lambda plasmid and phage DNA replication by Escherichia coli SeqA protein. *Microbiology (Reading, England)* 153, 1653-1663.

Osterhout, R. E., Figueroa, I. A., Keasling, J. D. & Arkin, A. P. (2007). Global analysis of host response to induction of a latent bacteriophage. *BMC Microbiol* 7, 82.

Poritz, M. A., Bernstein, H. D., Strub, K., Zopf, D., Wilhelm, H. & Walter, P. (1990). An *E. coli* ribonucleoprotein containing 4.5S RNA resembles mammalian signal recognition particle. *Science* 250, 1111-1117.

Ribes, V., Romisch, K., Giner, A., Dobberstein, B. & Tollervey, D. (1990). *E. coli* 4.5S RNA is part of a ribonucleoprotein particle that has properties related to signal recognition particle. *Cell* 63, 591-600.

Riley, M., Abe, T., Arnaud, M. B. & other authors (2006). Escherichia coli K-12: a cooperatively developed annotation snapshot--2005. *Nucleic acids research* 34, 1-9.

Rivas, E., Klein, R. J., Jones, T. A. & Eddy, S. R. (2001). Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr Biol* 11, 1369-1373.

Romeo, T. (1998). Global regulation by the small RNA-binding protein CsrA and the non-coding RNA molecule CsrB. *Mol Microbiol* 29, 1321-1330.

Rudd, K. E. (2000). EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic acids research* 28, 60-64.

Rybakov, N. I., Shestakov, V. A. & Aniskin, E. D. (1976). [Study of the virustatic action of rimantadine based on a phage-bacterium model]. *Biulleten' eksperimental'noi biologii i meditsiny* 81, 564-566.

Schmidt, M., Zheng, P. & Delihas, N. (1995). Secondary structures of Escherichia coli antisense micF RNA, the 5'-end of the target ompF mRNA, and the RNA/RNA duplex. *Biochemistry* 34, 3621-3631.

Selinger, D. W., Cheung, K. J., Mei, R., Johansson, E. M., Richmond, C. S., Blattner, F. R., Lockhart, D. J. & Church, G. M. (2000). RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. *Nature biotechnology* 18, 1262-1268.

Sevignani, C., Calin, G. A., Siracusa, L. D. & Croce, C. M. (2006). Mammalian microRNAs: a small world for fine-tuning gene expression. *Mamm Genome* 17, 189-202.

Shafer, R. H. and I. Smirnov (2000). Biological aspects of DNA/RNA quadruplexes. *Biopolymers* 56(3), 209-27.

Shimoni, Y., Friedlander, G., Hetzroni, G., Niv, G., Altuvia, S., Biham, O. & Margalit, H. (2007). Regulation of gene expression by small non-coding RNAs: a quantitative view. *Molecular systems biology* 3, 138.

Stark, B. C., Kole, R., Bowman, E. J. & Altman, S. (1978). Ribonuclease P: an enzyme with an essential RNA component. *Proc Natl Acad Sci U S A* 75, 3717-3721.

Storz, G., Opdyke, J. A. & Zhang, A. (2004). Controlling mRNA stability and translation with small, noncoding RNAs. *Curr Opin Microbiol* 7, 140-144.

Storz, G., Altuvia, S. & Wassarman, K. M. (2005). An abundance of RNA regulators. *Annual review of biochemistry* 74, 199-217.

Storz, G., Opdyke, J. A. & Wassarman, K. M. (2006). Regulating Bacterial Transcription with Small RNAs. *Cold Spring Harb Symp Quant Biol* 71, 269-273.

Tang, T. H., Bachellerie, J. P., Rozhdestvensky, T., Bortolin, M. L., Huber, H., Drungowski, M., Elge, T., Brosius, J. & Huttenhofer, A. (2002a). Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. *Proc Natl Acad Sci U S A* 99, 7536-7541.

Tang, T. H., Rozhdestvensky, T. S., d'Orval, B. C. & other authors (2002b). RNomics in Archaea reveals a further link between splicing of archaeal introns and rRNA processing. *Nucleic Acids Res* 30, 921-930.

Tjaden, B., Saxena, R. M., Stolyar, S., Haynor, D. R., Kolker, E. & Rosenow, C. (2002). Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays. *Nucleic acids research* 30, 3732-3738.

Tjaden, B., Goodwin, S. S., Opdyke, J. A., Guillier, M., Fu, D. X., Gottesman, S. & Storz, G. (2006). Target prediction for small, noncoding RNAs in bacteria. *Nucleic acids research* 34, 2791-2802.

Trotochaud, A. E. & Wassarman, K. M. (2004). 6S RNA function enhances long-term cell survival. *J Bacteriol* 186, 4978-4985.

Urban, J. H. & Vogel, J. (2007). Translational control and target recognition by Escherichia coli small RNAs in vivo. *Nucleic acids research* 35, 1018-1037.

Vanderpool, C. K. & Gottesman, S. (2004). Involvement of a novel transcriptional activator and small RNA in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system. *Mol Microbiol* 54, 1076-1089.

Vogel, J., Bartels, V., Tang, T. H., Churakov, G., Slagter-Jager, J. G., Huttenhofer, A. & Wagner, E. G. (2003). RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res* 31, 6435-6443.

Vogel, J., Argaman, L., Wagner, E. G. & Altuvia, S. (2004). The small RNA IstR inhibits synthesis of an SOS-induced toxic peptide. *Curr Biol* 14, 2271-2276.

Vogel, J. & Sharma, C. M. (2005). How to find small non-coding RNAs in bacteria. *Biol Chem* 386, 1219-1238.

Volinia, S., Calin, G. A., Liu, C. G. & other authors (2006). A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci U S A* 103, 2257-2261.

Wassarman, K. M., Zhang, A. & Storz, G. (1999). Small RNAs in Escherichia coli. *Trends Microbiol* 7, 37-45.

Wassarman, K. M. & Storz, G. (2000). 6S RNA regulates *E. coli* RNA polymerase activity. *Cell* 101, 613-623.

Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G. & Gottesman, S. (2001). Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev* 15, 1637-1651.

Weilbacher, T., Suzuki, K., Dubey, A. K. & other authors (2003). A novel sRNA component of the carbon storage regulatory system of Escherichia coli. *Mol Microbiol* 48, 657-670.

Yang, J. H., Zhang, X. C., Huang, Z. P., Zhou, H., Huang, M. B., Zhang, S., Chen, Y. Q. & Qu, L. H. (2006). snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res* 34, 5112-5123.

Zhang, A., Wassarman, K. M., Rosenow, C., Tjaden, B. C., Storz, G. & Gottesman, S. (2003). Global analysis of small RNA and mRNA targets of Hfq. *Mol Microbiol* 50, 1111-1124.

Zhang, J. & Madden, T. L. (1997). PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res* 7, 649-656.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31, 3406-3415.

Appendix 1 List of Completed Genomes Used in Bioinformatics Analysis

Acaryochloris marina	Alkaliphilus metalliredigens QYMF
Acidiphilium cryptum JF-5	Alkaliphilus oremlandii OhILAs
Acidobacteria bacterium Ellin345	Anabaena variabilis ATCC 29413
Acidothermus cellulolyticus 11B	Anaeromyxobacter dehalogenans 2CP-C
Acidovorax avenae subsp. citrulli AAC00-1	Anaeromyxobacter sp. Fw109-5
Acidovorax sp. JS42	Anaplasma marginale str. St. Maries
Acinetobacter baumannii ATCC 17978	Anaplasma phagocytophilum HZ
Acinetobacter sp. ADP1	Aquifex aeolicus VF5
Actinobacillus pleuropneumoniae L20	Arcobacter butzleri RM4018
Actinobacillus succinogenes 130Z	Arthrobacter aurescens TC1
Aeromonas hydrophila subsp. hydrophila ATCC 7966	Arthrobacter sp. FB24
Aeromonas salmonicida subsp. salmonicida A449	Aster yellows witches'-broom phytoplasma AYWB
Agrobacterium tumefaciens str. C58	Azoarcus sp. BH72
Agrobacterium tumefaciens str. C58	Azoarcus sp. EbN1
Alcanivorax borkumensis SK2	Azorhizobium caulinodans ORS 571
Alkalilimnicola ehrlichei MLHE-1	
Bacillus amyloliquefaciens FZB42	Borrelia afzelii PKo
Bacillus anthracis str. 'Ames Ancestor'	Borrelia burgdorferi B31
Bacillus anthracis str. Ames	Borrelia garinii PBi
Bacillus anthracis str. Sterne	Bradyrhizobium japonicum USDA 110
Bacillus cereus ATCC 10987	Bradyrhizobium sp. BTAi1
Bacillus cereus ATCC 14579	Bradyrhizobium sp. ORS278
Bacillus cereus E33L	Brucella abortus biovar 1 str. 9-941
Bacillus cereus subsp. cytotoxis NVH 391-98	Brucella melitensis 16M
Bacillus clausii KSM-K16	Brucella melitensis biovar Abortus 2308
Bacillus halodurans C-125	Brucella ovis ATCC 25840
Bacillus licheniformis ATCC 14580	Brucella suis 1330
Bacillus licheniformis ATCC 14580	Buchnera aphidicola str. APS (Acyrthosiphon pisum)
Bacillus pumilus SAFR-032	Buchnera aphidicola str. Bp (Baizongia pistaciae)
Bacillus subtilis subsp. subtilis str. 168	Buchnera aphidicola str. Cc (Cinara cedri)
Bacillus thuringiensis serovar konkukian str. 97-27	Buchnera aphidicola str. Sg (Schizaphis graminum)
Bacillus thuringiensis str. Al Hakam	Burkholderia ambifaria AMMD
Bacteroides fragilis NCTC 9343	Burkholderia cenocepacia AU 1054
Bacteroides fragilis YCH46	Burkholderia cenocepacia HI2424
Bacteroides thetaiotaomicron VPI-5482	Burkholderia mallei ATCC 23344
Bacteroides vulgatus ATCC 8482	Burkholderia mallei NCTC 10229
Bartonella bacilliformis KC583	Burkholderia mallei NCTC 10247
Bartonella henselae str. Houston-1	Burkholderia mallei SAVP1
Bartonella quintana str. Toulouse	Burkholderia pseudomallei 1106a
Baumannia cicadellinicola str. Hc (Homalodisca coagulata)	Burkholderia pseudomallei 1710b
Bdellovibrio bacteriovorus HD100	Burkholderia pseudomallei 668
Bifidobacterium adolescentis ATCC 15703	Burkholderia pseudomallei K96243

Bifidobacterium longum NCC2705	Burkholderia sp. 383
Bordetella bronchiseptica RB50	Burkholderia thailandensis E264
Bordetella parapertussis 12822	Burkholderia vietnamiensis G4
Bordetella pertussis Tohama I	Burkholderia xenovorans LB400
Caldicellulosiruptor saccharolyticus DSM 8903	Chlorobium chlorochromatii CaD3
Campylobacter concisus 13826	Chlorobium phaeobacteroides DSM 266
Campylobacter curvus 525.92	Chlorobium tepidum TLS
Campylobacter fetus subsp. fetus 82-40	Chromobacterium violaceum ATCC 12472
Campylobacter hominis ATCC BAA-381	Chromohalobacter salexigens DSM 3043
Campylobacter jejuni RM1221	Citrobacter koseri ATCC BAA-895
Campylobacter jejuni subsp. doylei 269.97	Clavibacter michiganensis subsp. michiganensis NCPPB 382
Campylobacter jejuni subsp. jejuni 81-176	Clostridium acetobutylicum ATCC 824
Campylobacter jejuni subsp. jejuni 81116	Clostridium beijerinckii NCIMB 8052
Campylobacter jejuni subsp. jejuni NCTC 11168	Clostridium botulinum A str. ATCC 19397
Candidatus Blochmannia floridanus	Clostridium botulinum A str. ATCC 3502
Candidatus Blochmannia pennsylvanicus str. BPEN	Clostridium botulinum A str. Hall
Candidatus Carsonella ruddii PV	Clostridium botulinum F str. Langeland
Candidatus Desulfococcus oleovorans Hxd3	Clostridium difficile 630
Candidatus Pelagibacter ubique HTCC1062	Clostridium kluyveri DSM 555
Candidatus Protochlamydia amoebophila UWE25	Clostridium novyi NT
Candidatus Ruthia magnifica str. Cm (Calyptogena magnifica)	Clostridium perfringens ATCC 13124
Candidatus Vesicomyosocius okutanii HA	Clostridium perfringens SM101
Carboxydothermus hydrogenoformans Z-2901	Clostridium perfringens str. 13
Caulobacter crescentus CB15	Clostridium tetani E88
Chlamydia muridarum Nigg	Clostridium thermocellum ATCC 27405
Chlamydia trachomatis A/HAR-13	Colwellia psychrerythraea 34H
Chlamydia trachomatis D/UW-3/CX	Corynebacterium diphtheriae NCTC 13129
Chlamydophila abortus S26/3	Corynebacterium efficiens YS-314
Chlamydophila caviae GPIC	Corynebacterium glutamicum ATCC 13032
Chlamydophila felis Fe/C-56	Corynebacterium glutamicum ATCC 13032
Chlamydophila pneumoniae AR39	Corynebacterium glutamicum R
Chlamydophila pneumoniae CWL029	Corynebacterium jeikeium K411
Chlamydophila pneumoniae J138	Coxiella burnetii Dugway 7E9-12
Chlamydophila pneumoniae TW-183	Coxiella burnetii RSA 493
	Cytophaga hutchinsonii ATCC 33406
Dechloromonas aromatica RCB	Desulfotalea psychrophila LSv54
Dehalococcoides ethenogenes 195	Desulfotomaculum reducens MI-1
Dehalococcoides sp. BAV1	Desulfovibrio desulfuricans G20
Dehalococcoides sp. CBDB1	Desulfovibrio vulgaris subsp. vulgaris DP4
Deinococcus geothermalis DSM 11300	Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough
Deinococcus radiodurans R1	Dichelobacter nodosus VCS1703A
Desulfitobacterium hafniense Y51	
Ehrlichia canis str. Jake	Escherichia coli 536
Ehrlichia chaffeensis str. Arkansas	Escherichia coli APEC O1
Ehrlichia ruminantium str. Gardel	Escherichia coli CFT073

Ehrlichia ruminantium str. Welgevonden	Escherichia coli E24377A
Ehrlichia ruminantium str. Welgevonden	Escherichia coli HS
Enterobacter sakazakii ATCC BAA-894	Escherichia coli K12
Enterobacter sp. 638	Escherichia coli O157:H7 EDL933
Enterococcus faecalis V583	Escherichia coli O157:H7 str. Sakai
Erwinia carotovora subsp. atroseptica SCRI1043	Escherichia coli UTI89
Erythrobacter litoralis HTCC2594	Escherichia coli W3110
Fervidobacterium nodosum Rt17-B1	Francisella tularensis subsp. tularensis FSC198
Flavobacterium johnsoniae UW101	Francisella tularensis subsp. tularensis SCHU S4
Flavobacterium psychrophilum JIP02/86	Francisella tularensis subsp. tularensis WY96-3418
Francisella tularensis subsp. holarctica	Frankia alni ACN14a
Francisella tularensis subsp. holarctica FTA	Frankia sp. CcI3
Francisella tularensis subsp. holarctica OSU18	Frankia sp. EAN1pec
Francisella tularensis subsp. novicida U112	Fusobacterium nucleatum subsp. nucleatum ATCC 25586
Geobacillus kaustophilus HTA426	Gloeobacter violaceus PCC 7421
Geobacillus thermodenitrificans NG80-2	Gluconobacter oxydans 621H
Geobacter metallireducens GS-15	Gramella forsetii KT0803
Geobacter sulfurreducens PCA	Granulibacter bethesdensis CGDNIH1
Geobacter uraniumreducens Rf4	
Haemophilus ducreyi 35000HP	Helicobacter acinonychis str. Sheeba
Haemophilus influenzae 86-028NP	Helicobacter hepaticus ATCC 51449
Haemophilus influenzae PittEE	Helicobacter pylori 26695
Haemophilus influenzae PittGG	Helicobacter pylori HPAG1
Haemophilus influenzae Rd KW20	Helicobacter pylori J99
Haemophilus somnus 129PT	Herminiimonas arsenicoxydans
Hahella chejuensis KCTC 2396	Hyphomonas neptunium ATCC 15444
Halorhodospira halophila SL1	
Idiomarina loihiensis L2TR	
Jannaschia sp. CCS1	Janthinobacterium sp. Marseille
Kineococcus radiotolerans SRS30216	Klebsiella pneumoniae subsp. pneumoniae MGH 78578
Lactobacillus acidophilus NCFM	Legionella pneumophila str. Corby
Lactobacillus brevis ATCC 367	Legionella pneumophila str. Lens
Lactobacillus casei ATCC 334	Legionella pneumophila str. Paris
Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	Legionella pneumophila subsp. pneumophila str. Philadelphia 1
Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365	Leifsonia xyli subsp. xyli str. CTCB07
Lactobacillus gasseri ATCC 33323	Leptospira borgpetersenii serovar Hardjo-bovis JB197
Lactobacillus johnsonii NCC 533	Leptospira borgpetersenii serovar Hardjo-bovis L550
Lactobacillus plantarum WCFS1	Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130
Lactobacillus reuteri F275	Leptospira interrogans serovar Lai str. 56601
Lactobacillus sakei subsp. sakei 23K	Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293
Lactobacillus salivarius subsp. salivarius UCC118	Listeria innocua Clip11262

Lactococcus lactis subsp. cremoris MG1363	Listeria monocytogenes EGD-e
Lactococcus lactis subsp. cremoris SK11	Listeria monocytogenes str. 4b F2365
Lactococcus lactis subsp. lactis Il1403	Listeria welshimeri serovar 6b str. SLCC5334
Lawsonia intracellularis PHE/MN1-00	
Magnetococcus sp. MC-1	Mycobacterium sp. MCS
Magnetospirillum magneticum AMB-1	Mycobacterium tuberculosis CDC1551
Mannheimia succiniciproducens MBEL55E	Mycobacterium tuberculosis F11
Maricaulis maris MCS10	Mycobacterium tuberculosis H37Ra
Marinobacter aquaeolei VT8	Mycobacterium tuberculosis H37Rv
Marinomonas sp. MWYL1	Mycobacterium ulcerans Agy99
Mesoplasma florum L1	Mycobacterium vanbaalenii PYR-1
Mesorhizobium loti MAFF303099	Mycoplasma agalactiae PG2
Mesorhizobium sp. BNC1	Mycoplasma capricolum subsp. capricolum ATCC 27343
Methylibium petroleiphilum PM1	Mycoplasma gallisepticum R
Methylobacillus flagellatus KT	Mycoplasma genitalium G37
Methylococcus capsulatus str. Bath	Mycoplasma hyopneumoniae 232
Moorella thermoacetica ATCC 39073	Mycoplasma hyopneumoniae 7448
Mycobacterium avium 104	Mycoplasma hyopneumoniae J
Mycobacterium avium subsp. paratuberculosis K-10	Mycoplasma mobile 163K
Mycobacterium bovis AF2122/97	Mycoplasma mycoides subsp. mycoides SC str. PG1
Mycobacterium bovis BCG str. Pasteur 1173P2	Mycoplasma penetrans HF-2
Mycobacterium gilvum PYR-GCK	Mycoplasma pneumoniae M129
Mycobacterium leprae TN	Mycoplasma pulmonis UAB CTIP
Mycobacterium smegmatis str. MC2 155	Mycoplasma synoviae 53
Mycobacterium sp. JLS	Myxococcus xanthus DK 1622
Mycobacterium sp. KMS	
Neisseria gonorrhoeae FA 1090	Nitrosococcus oceani ATCC 19707
Neisseria meningitidis FAM18	Nitrosomonas europaea ATCC 19718
Neisseria meningitidis MC58	Nitrosomonas eutropha C91
Neisseria meningitidis Z2491	Nitrosospira multiformis ATCC 25196
Neorickettsia sennetsu str. Miyayama	Nocardia farcinica IFM 10152
Nitratiruptor sp. SB155-2	Nocardioides sp. JS614
Nitrobacter hamburgensis X14	Nostoc sp. PCC 7120
Nitrobacter winogradskyi Nb-255	Novosphingobium aromaticivorans DSM 12444
Oceanobacillus iheyensis HTE831	Onion yellows phytoplasma OY-M
Ochrobactrum anthropi ATCC 49188	Orientia tsutsugamushi Boryong
Oenococcus oeni PSU-1	
Parabacteroides distasonis ATCC 8503	Prochlorococcus marinus subsp. marinus str. CCMP1375
Paracoccus denitrificans PD1222	Prochlorococcus marinus subsp. pastoris str. CCMP1986
Parvibaculum lavamentivorans DS-1	Propionibacterium acnes KPA171202
Pasteurella multocida subsp. multocida str. Pm70	Prosthecochloris vibrioformis DSM 265
Pediococcus pentosaceus ATCC 25745	Pseudoalteromonas atlantica T6c
Pelobacter carbinolicus DSM 2380	Pseudoalteromonas haloplanktis TAC125
Pelobacter propionicus DSM 2379	Pseudomonas aeruginosa PA7

Pelodictyon luteolum DSM 273	Pseudomonas aeruginosa PAO1
Pelotomaculum thermopropionicum SI	Pseudomonas aeruginosa UCBPP-PA14
Photobacterium profundum SS9	Pseudomonas entomophila L48
Photorhabdus luminescens subsp. laumondii TTO1	Pseudomonas fluorescens Pf-5
Polaromonas naphthalenivorans CJ2	Pseudomonas fluorescens PfO-1
Polaromonas sp. JS666	Pseudomonas mendocina ymp
Polynucleobacter sp. QLW-P1DMWA-1	Pseudomonas putida F1
Porphyromonas gingivalis W83	Pseudomonas putida KT2440
Prochlorococcus marinus str. AS9601	Pseudomonas stutzeri A1501
Prochlorococcus marinus str. MIT 9215	Pseudomonas syringae pv. phaseolicola 1448A
Prochlorococcus marinus str. MIT 9301	Pseudomonas syringae pv. syringae B728a
Prochlorococcus marinus str. MIT 9303	Pseudomonas syringae pv. tomato str. DC3000
Prochlorococcus marinus str. MIT 9312	Psychrobacter arcticus 273-4
Prochlorococcus marinus str. MIT 9313	Psychrobacter cryohalolentis K5
Prochlorococcus marinus str. MIT 9515	Psychrobacter sp. PRwf-1
Prochlorococcus marinus str. NATL1A	Psychromonas ingrahamii 37
Prochlorococcus marinus str. NATL2A	
Ralstonia eutropha H16	Rhodopseudomonas palustris HaA2
Ralstonia eutropha JMP134	Rhodospirillum rubrum ATCC 11170
Ralstonia metallidurans CH34	Rickettsia akari str. Hartford
Ralstonia solanacearum GMI1000	Rickettsia bellii OSU 85-389
Rhizobium etli CFN 42	Rickettsia bellii RML369-C
Rhizobium leguminosarum bv. viciae 3841	Rickettsia canadensis str. McKiel
Rhodobacter sphaeroides 2.4.1	Rickettsia conorii str. Malish 7
Rhodobacter sphaeroides ATCC 17025	Rickettsia felis URRWXCal2
Rhodobacter sphaeroides ATCC 17029	Rickettsia massiliae MTU5
Rhodococcus sp. RHA1	Rickettsia prowazekii str. Madrid E
Rhodoferax ferrireducens T118	Rickettsia rickettsii str. 'Sheila Smith'
Rhodopirellula baltica SH 1	Rickettsia typhi str. Wilmington
Rhodopseudomonas palustris BisA53	Roseiflexus castenholzii DSM 13941
Rhodopseudomonas palustris BisB18	Roseiflexus sp. RS-1
Rhodopseudomonas palustris BisB5	Roseobacter denitrificans OCh 114
Rhodopseudomonas palustris CGA009	Rubrobacter xylanophilus DSM 9941
Saccharophagus degradans 2-40	Staphylococcus aureus subsp. aureus str. Newman
Saccharopolyspora erythraea NRRL 2338	Staphylococcus epidermidis ATCC 12228
Salinibacter ruber DSM 13855	Staphylococcus epidermidis RP62A
Salinispora tropica CNB-440	Staphylococcus haemolyticus JCSC1435
Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67	Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305
Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150	Streptococcus agalactiae 2603V/R
Salmonella enterica subsp. enterica serovar Typhi Ty2	Streptococcus agalactiae A909
Salmonella enterica subsp. enterica serovar Typhi str. CT18	Streptococcus agalactiae NEM316
Salmonella typhimurium LT2	Streptococcus gordonii str. Challis substr. CH1
Serratia proteamaculans 568	Streptococcus mutans UA159
Shewanella amazonensis SB2B	Streptococcus pneumoniae D39
Shewanella baltica OS155	Streptococcus pneumoniae R6
Shewanella baltica OS185	Streptococcus pneumoniae TIGR4

Shewanella denitrificans OS217	Streptococcus pyogenes M1 GAS
Shewanella frigidimarina NCIMB 400	Streptococcus pyogenes MGAS10270
Shewanella loihica PV-4	Streptococcus pyogenes MGAS10394
Shewanella oneidensis MR-1	Streptococcus pyogenes MGAS10750
Shewanella pealeana ATCC 700345	Streptococcus pyogenes MGAS2096
Shewanella putrefaciens CN-32	Streptococcus pyogenes MGAS315
Shewanella sediminis HAW-EB3	Streptococcus pyogenes MGAS5005
Shewanella sp. ANA-3	Streptococcus pyogenes MGAS6180
Shewanella sp. MR-4	Streptococcus pyogenes MGAS8232
Shewanella sp. MR-7	Streptococcus pyogenes MGAS9429
Shewanella sp. W3-18-1	Streptococcus pyogenes SSI-1
Shigella boydii Sb227	Streptococcus pyogenes str. Manfredo
Shigella dysenteriae Sd197	Streptococcus sanguinis SK36
Shigella flexneri 2a str. 2457T	Streptococcus suis 05ZYH33
Shigella flexneri 2a str. 301	Streptococcus suis 98HAH33
Shigella flexneri 5 str. 8401	Streptococcus thermophilus CNRZ1066
Shigella sonnei Ss046	Streptococcus thermophilus LMD-9
Silicibacter pomeroyi DSS-3	Streptococcus thermophilus LMG 18311
Silicibacter sp. TM1040	Streptomyces avermitilis MA-4680
Sinorhizobium medicae WSM419	Streptomyces coelicolor A3(2)
Sinorhizobium meliloti 1021	Sulfurovum sp. NBC37-1
Sodalis glossinidius str. 'morsitans'	Symbiobacterium thermophilum IAM 14863
Solibacter usitatus Ellin6076	Synechococcus elongatus PCC 6301
Sphingomonas wittichii RW1	Synechococcus elongatus PCC 7942
Sphingopyxis alaskensis RB2256	Synechococcus sp. CC9311
Staphylococcus aureus RF122	Synechococcus sp. CC9605
Staphylococcus aureus subsp. aureus COL	Synechococcus sp. CC9902
Staphylococcus aureus subsp. aureus JH1	Synechococcus sp. JA-2-3B'a(2-13)
Staphylococcus aureus subsp. aureus JH9	Synechococcus sp. JA-3-3Ab
Staphylococcus aureus subsp. aureus MRSA252	Synechococcus sp. RCC307
Staphylococcus aureus subsp. aureus MSSA476	Synechococcus sp. WH 7803
Staphylococcus aureus subsp. aureus MW2	Synechococcus sp. WH 8102
Staphylococcus aureus subsp. aureus Mu3	Synechocystis sp. PCC 6803
Staphylococcus aureus subsp. aureus Mu50	Syntrophobacter fumaroxidans MPOB
Staphylococcus aureus subsp. aureus N315	Syntrophomonas wolfei subsp. wolfei str. Goettingen
Staphylococcus aureus subsp. aureus NCTC 8325	Syntrophus aciditrophicus SB
Staphylococcus aureus subsp. aureus USA300	
Thermoanaerobacter tengcongensis MB4	Thiobacillus denitrificans ATCC 25259
Thermobifida fusca YX	Thiomicrospira crunogena XCL-2
Thermosipho melanesiensis BI429	Thiomicrospira denitrificans ATCC 33889
Thermosynechococcus elongatus BP-1	Treponema denticola ATCC 35405
Thermotoga lettingae	Treponema pallidum subsp. pallidum str. Nichols
Thermotoga maritima MSB8	Trichodesmium erythraeum IMS101
Thermotoga petrophila RKU-1	Tropheryma whipplei TW08/27
Thermus thermophilus HB27	Tropheryma whipplei str. Twist
Thermus thermophilus HB8	

Ureaplasma parvum serovar 3 str. ATCC 700970	
Verminephrobacter eiseniae EF01-2	Vibrio harveyi ATCC BAA-1116
Vibrio cholerae O1 biovar eltor str. N16961	Vibrio parahaemolyticus RIMD 2210633
Vibrio cholerae O395	Vibrio vulnificus CMCP6
Vibrio fischeri ES114	Vibrio vulnificus YJ016
Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	Wolbachia endosymbiont strain TRS of Brugia malayi
Wolbachia endosymbiont of Drosophila melanogaster	Wolinella succinogenes DSM 1740
Xanthobacter autotrophicus Py2	Xanthomonas oryzae pv. oryzae KACC10331
Xanthomonas axonopodis pv. citri str. 306	Xanthomonas oryzae pv. oryzae MAFF 311018
Xanthomonas campestris pv. campestris str. 8004	Xylella fastidiosa 9a5c
Xanthomonas campestris pv. campestris str. ATCC 33913	Xylella fastidiosa Temecula1
Xanthomonas campestris pv. vesicatoria str. 85-10	
Yersinia enterocolitica subsp. enterocolitica 8081	Yersinia pestis Pestoides F
Yersinia pestis Antiqua	Yersinia pestis biovar Microtus str. 91001
Yersinia pestis CO92	Yersinia pseudotuberculosis IP 31758
Yersinia pestis KIM	Yersinia pseudotuberculosis IP 32953
Yersinia pestis Nepal516	
Zymomonas mobilis subsp. mobilis ZM4	

Appendix 2 Code for RegCompare I

```
#include "formatDB.hpp"

int main(int argc, char** argv)
{
  /*P INIT length of file TO ZERO */
  unsigned int fileLen = 0;

  /*P INIT LENGTH CUTTER TO 0 */
  unsigned int lengthCutter = 0;

  /*P INIT startpos of file TO ZERO */
  unsigned int fileStart = 0;

  /*P INIT length of sequences TO ZERO */
  unsigned int seqsLen = 0;

  /* INIT NAME BUFFER */
  char fdataOutputFile[256];
  char nameBuf[256];
  char nameBufStart[256];

  /*P INIT INPUT FILE HANDLER TO NULL */
  FILE* inputHandler = NULL;

  /*P INIT OUTPUT LIST FILE HANDLER TO NULL */
  FILE* listHandler = NULL;

  /*P INIT POINTER TO POINT TO DATA TO NULL */
  char* fileBuf = NULL;
  char* fileBufRev = NULL;

  /*P INIT FILE COUNT TO ZERO */
  int fileCount = 0;

  /*P INIT DIR DATASTRUCT TO NULL */
  DIR* dir = NULL;
  struct dirent* dent = NULL;

  /*P CREATE VECTOR FOR HOLD ORF */
  vector<ORF_t> orfVector;

  /*P CREATE VECTOR FOR HOLD RR */
  vector<RR_t> rrVector;

  /*P VERIFY INPUT PARAMS */
  if(argc<3)
  {
```

```
    cout<<"Usage: "<<argv[0]<<" "<<"[Min length of ORF] [Database directory] {length
cutter}"<<endl;
    exit(0);
  }

  fMinLenOrf = atoi(argv[1]);
  if(fMinLenOrf<=0)
  {
    cout<<"Min Length of ORF should great zero!"<<endl;
    exit(0);
  }

  /*P READ DIR INFO */
  dir = opendir(argv[2]);
  if(NULL == dir)
  {
    cout<<"Database directory ["<<argv[2]<<"] not exist!"<<endl;
    exit(0);
  }

  if(argc == 4)
  {
    if(atoi(argv[3]) <100)
    {
      cout<<"Cutter Length cannot less than 100. System will use default 200 instead! "<<endl;
    }else
    {
      lengthCutter = atoi(argv[3]);
    }

  }

  memset(nameBufStart, '\0', 256);
  while(strlen(argv[2])>0 && argv[2][strlen(argv[2])-1] == '/')
  {
    argv[2][strlen(argv[2])-1] = '\0';
  }

  if(strlen(argv[2]) == 1 && argv[2][0] != '.' );
  {
    strncpy(nameBufStart, argv[2],strlen(argv[2]));
    nameBufStart[strlen(nameBufStart)]='/';
  }

  /*P OPEN LIST OUTPUT FILE HANDLER*/
  memset(nameBuf, '\0', 256);
  strncpy(nameBuf, nameBufStart,strlen(nameBufStart));
  strncat(nameBuf,"list.txt",8);
  listHandler = fopen(nameBuf,"w");
```

```
  /*P READ EACH FILE UNDER DATABASE DIRECTORY */
  while(dent = readdir(dir))
  {
    /*P NOT A VALID DATABASE FILE */
    if(0!=strncmp(&dent->d_name[strlen(dent->d_name)-4],".gbk",4))
      continue;

    /*P OPEN INPUT FILE*/
    memset(nameBuf, '\0', 256);
    strncpy(nameBuf, nameBufStart,strlen(nameBufStart));
    strncat(nameBuf,dent->d_name,strlen(dent->d_name));
    inputHandler = fopen(nameBuf,"r");
    if(!inputHandler)
    {
      cout<<"Could not open database file:"<<nameBuf<<endl;
      continue;
    }

    /*P READ OUTPUT FILE NAME */
    memset(fdataOutputFile,'\0',256);
    fgets(fdataOutputFile,256,inputHandler);
    while((fdataOutputFile[strlen(fdataOutputFile)-1] ==' ' )||
          (fdataOutputFile[strlen(fdataOutputFile)-1] =='\n' )||
          (fdataOutputFile[strlen(fdataOutputFile)-1] ==10 )||
          (fdataOutputFile[strlen(fdataOutputFile)-1] ==13 ))
        fdataOutputFile[strlen(fdataOutputFile)-1] = '\0';
    /*P CONVERTING MSG TO USER*/
    cout<<"Converting database:"<<fdataOutputFile<<"......"<<endl;
    fileCount++;
    strncat(fdataOutputFile,".db",3);

    /*P ALLOCATE MEMORY FOR HOLD DATA */
    fileStart = ftell(inputHandler);
    /*P go to end of file */
    fseek(inputHandler,0,SEEK_END);
    /*P GET LENGTH OF FILE */
    fileLen = ftell(inputHandler);
    fileLen = fileLen-fileStart;
    /*P ALLOCATE SPACE */
    fileBuf = new char[fileLen+2];
    /*P READ FROM FILE */
    fseek(inputHandler, fileStart, SEEK_SET);
    fread((void*)fileBuf, fileLen+1, 1, inputHandler);

    /*P CLOSE INPUT FILE*/
    fclose(inputHandler);

    /*P PHRASE BUFFER, TAKE OFF ALL GARBAGER CHAR*/
    seqsLen=phraseSequence(fileBuf, fileLen);
//    cout<<"seq:"<<seqsLen<<endl<<"fileLen:"<<fileLen<<endl;
```

```
    /*P EXTRACT ORF */
    fileBufRev = new char[seqsLen + 32];
    revSeqs(fileBuf, fileBufRev, seqsLen);
    orfVector.clear();
    extractORF(fileBuf, fileBufRev, orfVector, seqsLen);

    //for(int jj=0;jj<orfVector.size();jj++)
    //{
    //  cout<<orfVector[jj].start<<"..."<<orfVector[jj].end<<"  "<<orfVector[jj].count<<endl;
    //}
    //
    /*P GENERATE RR FRAME */
    rrVector.clear();
    generateRRFrame(orfVector, rrVector, lengthCutter);

    /*P ENCODE RR FRAME TO DATABASE */
    /*P OPEN INPUT FILE*/
    memset(nameBuf, '\0', 256);
    strncpy(nameBuf, nameBufStart,strlen(nameBufStart));
    strncat(nameBuf,fdataOutputFile,strlen(fdataOutputFile));
    encodeRRFrame(fileBuf, fileBufRev, rrVector, nameBuf);

    /*P WRITE DATABASE NAME TO LIST */
    fputs(fdataOutputFile, listHandler);
    fputc('\n',listHandler);

    /*P RELEASE MEMORY ALLOCATION */
    delete [] fileBufRev;
    delete [] fileBuf;
  }

  /*P CLOSE LIST FILE HANDLER */
  fclose(listHandler);

  /*P CLOSE DIRECTORY */
  closedir(dir);

  /*P MSG TO USER*/
  cout<<fileCount<<" database files have been converted."<<endl;
}

//encode RR frame
void encodeRRFrame(const char* bufOrg, const char* bufRev, vector<RR_t> &rr, const char* output)
{
  /*P INITIAL len to zero*/
  unsigned int len = 0;

  /*P INITIAL readC to 0 */
  int readC = 0;  /* Temp variable to hold data read from inFile */
```

```
  /*P INITIAL int_buf TO 0 */
  unsigned int intBuf = 0;  /* buffer to hold up to 16 codon chars */

  /*P INITIAL int_code_count TO 0 */
  unsigned int intCodeCount = 0; /* code count to indecated if need swap buffer */

  /*P INITIAL int_code to 5 */
  unsigned int intCode = 5; /* temp variable to hold current code*/

  seqsHead_t tempSeqsHead;
  unsigned int idx = 1;
  unsigned int sizeofSeqsHead = sizeof(tempSeqsHead);
  bool previousRev = false;

  FILE* outputHandler =fopen(output, "wb");

  /*P GO OVER ALL SEQUENCE IN LIST */
  for(int i = 0; i < rr.size(); i++,idx++)
  {
    if(previousRev != rr[i].isReversed)
    {
      idx = 1;
      previousRev = rr[i].isReversed;
    }
    /*P CALC LENGTH OF FILE FOR CURRENT SEQUENCE */
    len = 1 + rr[i].end - rr[i].start;

    /*P SET CORRECT VALUE FOR SEQSHEAD_T */
    tempSeqsHead.id = idx;
    tempSeqsHead.start = rr[i].start;
    tempSeqsHead.end = rr[i].end;
    tempSeqsHead.len = len;
    tempSeqsHead.isReversed = rr[i].isReversed;
//    cout<<"LEN("<<rr[i].start<<"<->"<<rr[i].end<<"):"<<len<<endl;
    /*P WRITE SIZE OF CURRENT SEQUENCE TO FILE */
    fwrite((const void*)&tempSeqsHead, sizeofSeqsHead, 1, outputHandler);

    /*P INITIAL int_buf to 0 FOR EACH SEQUENCE */
    intBuf = 0;

    /*P INIT int_code_count to 0 at begin of each sequence */
    intCodeCount = 0;
    for(int j = rr[i].start; j <= rr[i].end; j++)
    {
      /*P READ ONE CHAR FROM inFile*/
      if(rr[i].isReversed)
        readC = bufRev[j];
      else
        readC=bufOrg[j];
```

```
      /*P INITIAL intCode to 5**/
      intCode=5;
      /*P GET CORRESPOND CODE */
      switch(readC)
      {
        case 'a':
          intCode = 0;
          break;
        case 'c':
          intCode = 1;
          break;
        case 'g':
          intCode = 2;
          break;
        case 't':
          intCode = 3;
          break;
        default:
          break;
      }
      /*P IF inCode GET VALID CODE CHAR*/
      if(intCode != 5)
      {
        intCodeCount++;
        intBuf|=SHIFT(intCode,intCodeCount);
        if(intCodeCount == 16)
        {
          fwrite((const void*)&intBuf, sizeof(intBuf), 1, outputHandler);
//          cout<<"__"<<intBuf<<"__";
          intCodeCount = 0;
          intBuf = 0;
        }
      }
    }//for
    /*P IF THERE IS SOME DATA LEFT IN BUFFER, WRITE TO FILE*/
    if(intCodeCount>0)
    {
      /*P WRITE INT_BUF TO FILE*/
      fwrite((const void*)&intBuf, sizeof(intBuf), 1, outputHandler);
//      cout<<"__"<<intBuf<<"__";
    }
//    cout<<endl;
  }//for
  fclose(outputHandler);
}

void generateRRFrame(vector<ORF_t> &orf, vector<RR_t> &rr, unsigned int lengthCutter)
{
  /*P new RR for hold temp value */
  RR_t rrTemp;
```

```
  /*P INIT MIN LENGTH TO 200 OR LENGTHCUTTER */
  unsigned int minLen = lengthCutter;
  int idx = 0;
  int restart = 0;

  /*P RESET MINLEN ACCORDING TO LENGTHCUTTER */
  if( 0 == minLen )
    minLen = 200;

  while(restart < 2)
  {//FIRST WHILE LOOP
    /*P IF THIS IS THE FIRST ORF OF ONE BUF SEQUENCE */
    if(orf.size() > idx)
    {
      rrTemp.start = 0;
      rrTemp.end = orf[idx].start - 1;
      rrTemp.isReversed = orf[idx].isReversed;
      if((rrTemp.end - rrTemp.start + 1) >= minLen )
      {
        if( 0 != lengthCutter )
        {
          rrTemp.start = rrTemp.end - minLen + 1;
        }
        rr.push_back(rrTemp);
      }
    }

    idx++;

    while(idx < orf.size() && orf[idx].isReversed == orf[idx-1].isReversed)
    {//SECOND WHILE LOOP
      rrTemp.start = orf[idx-1].end + 1;
      rrTemp.end = orf[idx].start - 1;
      rrTemp.isReversed = orf[idx].isReversed;
      if((rrTemp.end - rrTemp.start +1) >= minLen )
      {
        if( 0 != lengthCutter)
        {
          rrTemp.start = rrTemp.end - minLen + 1;
        }
        rr.push_back(rrTemp);
      }
      idx++;
    }//END SECOND WHILE LOOP
    restart++ ;
  }//END FIRST WHILE LOOP

}
```

```
//extract open reading frame
void extractORF(const char* bufOrg, const char* bufRev, vector<ORF_t>& orf, unsigned int len)
{
  /*P INITIAL TEMP BUF TO NULL */
  const char* buf = NULL;

  /*P INITIAL START POS FOR SEARCH */
  unsigned int startPos=0;

  /*P INITIAL NEXT START POS OF FRAME*/
  unsigned int nextStartPos = 0;

  /*P INITIAL STOP CODON'S POSITION*/
  unsigned int stopPos1 = 0, stopPos2 = 0, stopPos3 = 0;

  /*P INITIAL TEMP MAX ORF TO NULL*/
  ORF_t* maxOrf = NULL;

  /*P DEFINE ARRAY OF ORF */
  ORF_t orfArray[3];

  /*P INIT TEMP VARIABLE FOR FOR LOOP */
  int i=0;
  unsigned int lenMinus3 = len - 3;

  /*P INIT REVERSE INDICATOR TO FALSE */
  bool reversed = false;

  /*P GO ORIGNAL ORDER FIRSE, THEN REVERSED ORDER */
  for(int OrgToRev = 0; OrgToRev < 2; OrgToRev++)
  {//FIRST FOR LOOP
    /*P SET CURRECT BUF */
    if( 0 == OrgToRev)
    {
      buf = bufOrg;
      reversed = false;
    }else
    {
      buf = bufRev;
      reversed = true;
    } //IF - ELSE
    /*P RESET START POS  */
    startPos = 0;

    /*P WHILE THERE ARE DNA CHAR LEFT */
    while(startPos<len)
    {//FIRST WHILE LOOP
      /*P GET NEXT START CODON */
      nextStartPos = findStartCodon(buf,startPos, len-1);
      /*P GET NEXT VALID START CODON */
```

```
    if(nextStartPos < (len-1))
    {//IF
      /*P TRY TO FIND THREE FRAME */
      for(i=0; i<3; i++)
      {//SECOND FOR LOOP
        orfArray[i].start = nextStartPos+i;
        orfArray[i].end = findStopCodon( buf, orfArray[i].start, len-1);
        orfArray[i].isReversed = reversed;
        if(orfArray[i].end > (lenMinus3) )
        {
          orfArray[i].valid=false;
        }else
        {
          orfArray[i].count = orfArray[i].end - orfArray[i].start + 1;
          orfArray[i].valid = true;
        }//IF-ELSE
      }//END SECOND FOR LOOP

      /*P CHOICE THE ORF WITH LARGEST COUNT*/
      maxOrf = NULL;
      for(i=0;i<3;i++)
      {//THIRD FOR LOOP
        if(orfArray[i].valid)
          maxOrf = &orfArray[i];
      }//END THIRD FOR LOOP
      /*P NO VALID ORF*/
      if(!maxOrf)
      {
        startPos = nextStartPos + 1;
        continue;
      }

      /*P PICK LARGEST ORF AND ADD TO VECTOR*/
      for(i=0; i<3; i++)
      {//FOURTH FOR LOOP
        if(orfArray[i].valid && orfArray[i].count > maxOrf->count)
          maxOrf = &orfArray[i];
      }//END FOURTH FOR LOOP

      if(maxOrf->count >= (fMinLenOrf*3))
        orf.push_back(*maxOrf);
      startPos = maxOrf -> end + 1;
    }
    else
    {
      startPos = nextStartPos + 1;
    }//END IF - ELSE
  }//END FIRST WHILE LOOP
}//END FIRST FOR LOOP
}
```

```
//find next start codon
unsigned int findStartCodon(const char* buf, unsigned int start, unsigned int end)
{
  while(start < end)
  {
    if(strncmp(&buf[start], fstartCodon, 3) == 0)
    {
      break;
    }else
    {
      start += 1;
    }
  }
  return start;
}

//find stop codon
unsigned int findStopCodon(const char* buf, unsigned int start, unsigned int end)
{

  start += 3;
  while(start < end)
  {
    if( (strncmp(&buf[start],fstopTaa, 3) == 0) ||
        ((strncmp(&buf[start],fstopTag, 3) == 0) ||
         (strncmp(&buf[start],fstopTga, 3) == 0)))
      break;
    else
      start += 3;
  }

  return (start + 2);
}


//phrase sequence, take off all non-codon char
unsigned int phraseSequence(char* buf, unsigned int len)
{
  /*P INITIAL RETURN VALUE*/
  unsigned int rev = 0;
  for(unsigned int i=0;i<len;i++)
  {
    switch(buf[i])
    {
      case 'a':
      case 'c':
      case 'g':
      case 't':
        swapChars(buf[rev],buf[i]);
```

```
        rev++;
        break;
      default:
        break;
    }//switch
  }
  buf[rev]='\0';
  return rev;
}

void swapChars(char& c1, char& c2)
{
  char c=c1;
  c1 = c2;
  c2 = c;
}

//reverse DNA sequence file in order to get another direction's sequence
void revSeqs(const char* orignal, char* reversed, unsigned int len)
{
  /*P INIT LENMINUSONE TO LEN - 1 */
  unsigned lenMinusOne = len -1;

  /*P GO OVER ORIGNAL FROM START TO END,     */
  /*  AND WRITE CORRESPOND CHAR TO REVERSED */
  for(unsigned int i=0; i<len; i++)
  {
    switch(orignal[i])
    {
      case 'a':
        reversed[lenMinusOne - i]='t';
        break;
      case 't':
        reversed[lenMinusOne - i]='a';
        break;
      case 'c':
        reversed[lenMinusOne - i]='g';
        break;
      case 'g':
        reversed[lenMinusOne - i]='c';
        break;
      default:
        /*P NOTHING TO DO WITH INVALID CHAR*/
          ;
    }//end with switch
  }//end for
}

#include "RegCompareI.hpp"
```

```
int main(int argc, char** argv)
{

  /*P INITIAL PARAMS' FILE HANDLER TO NULL */
  FILE* RRHandler = NULL;
  FILE* databaseHandler = NULL;
  FILE* reportHandler = NULL;
  /*P INITIAL DATABASE FILE NAME BUFFER */
  char databaseFileNameBuf[256];
  char databaseFileName[256];
  char writeBuf[256];
  /*P INIT SIMLAR TO ZERO */
  int samePer = 0;

  /*P INIT DIR DATASTRUCT TO NULL */
  DIR* dir = NULL;
  struct dirent* dent = NULL;


  /*P re-calc MAPBUF */
  switchMap();

  /*P check argc */
  if(argc<4)
  {
    cout<<"Usage: "<<argv[0]<<" [Threshold] [RR_database] [Database Directory]"<<endl;
    exit(0);
  }

  /*P check threshold */
  fthreshold=atoi(argv[1]);
  if(fthreshold<=0 || fthreshold>100)
  {
    cout<<"ERROR: Threshold should between [0-100]"<<endl;
    exit(0);
  }

  /*P check if RR file exist */
  RRHandler=fopen(argv[2],"rb");
  if(!RRHandler)
  {
    cout<<"RR data file cannot open! Please check it!"<<endl;
    exit(0);
  }

  /*P READ DIR INFO */
  dir = opendir(argv[3]);
  if(NULL == dir)
  {
    cout<<"Database directory ["<<argv[2]<<"] not exist!"<<endl;
```

```
  fclose(RRHandler);
  exit(0);
}

memset(databaseFileNameBuf, '\0', 256);
while(strlen(argv[3])>0 && argv[3][strlen(argv[3])-1] == '/')
{
  argv[3][strlen(argv[3])-1] = '\0';
}

if(strlen(argv[3]) == 1 && argv[3][0] != '.' );
{
  strncpy(databaseFileNameBuf, argv[3],strlen(argv[3]));
  databaseFileNameBuf[strlen(databaseFileNameBuf)]='/';
}

/*P CREATE A REPORT FILE FOR OUTPUT RESULT */
reportHandler = fopen("compareReport.txt","w");
fputs("RR compare result file\n\n\nRR File:",reportHandler);
fputs(argv[2],reportHandler);
fputs("\nDatabase Directory: ",reportHandler);
fputs(argv[3],reportHandler);
fputs("\n\n",reportHandler);

/*P FOR EACH FILE paramsHandler LIST */
while( dent = readdir(dir))
{
  /*P NOT A VALID DATABASE FILE */
  if(0!=strncmp(&dent->d_name[strlen(dent->d_name)-3],".db",3))
    continue;

  /*P RESET NAME BUFFER */
  memset(databaseFileName, '\0', 256);
  strncpy(databaseFileName, databaseFileNameBuf, strlen(databaseFileNameBuf));
  strncat(databaseFileName, dent->d_name, strlen(dent->d_name));

  /*P OPEN DATABASE FILE */
  databaseHandler = fopen(databaseFileName,"rb");
  if( NULL == databaseHandler)
  {
    cout<<"Cannot open: "<<databaseFileNameBuf<<" ***"<<endl;
    continue;
  }

  /*P MSG TO USER */
  cout<<"Searching database file '"<<dent->d_name<<"'..... "<<flush;


  /*P GO TO COMPARE*/
  fOneToOneResult.empty();
```

```
    samePer = compareSequences(databaseHandler, RRHandler);

    /*P OUTPUT COMPARE RESULT */
    cout<<'\b'<<samePer << "% matched" <<endl;
    fputs("*****************************************************\n",reportHandler);
    fputs(dent->d_name, reportHandler);
    sprintf(writeBuf,"Max Matched %d\n\n", samePer);
    fputs("\n*****************************************************\n",reportHandler);
    fputs(writeBuf, reportHandler);
    printCutResult(reportHandler);

    /*P CLOSE DATABASE FILE*/
    fclose(databaseHandler);
  }//WHILE

  cout<<"Please check report file[compareReport.txt]. \nSearch Completed!"<<endl;

  /*P CLOSE DIRECTORY */
  closedir(dir);

  /*P CLOSE RR DATA FILE */
  fclose(reportHandler);
  fclose(RRHandler);
}

void printCutResult(FILE* reportHandler)
{
  //print something stuff
  char* RRBuf = NULL;
  char* DBBuf = NULL;
  char* ALBuf = NULL;
  char buf[256];
  char RRrev[] = {'\0', '\0'};  //if RR reversed
  char DBrev[] = {'\0', '\0'};  //if DB reversed
  unsigned int bufLen = 0;
  unsigned int seqsLen = 0;
  seqsCompareResultUnit_t unit;

  if(!fOneToOneResult.empty())
  {
    bufLen = fOneToOneResult.top().RRSeqsHead.len + 1;
    /*P ALLOCATE MEMORY FOR BUF */
    RRBuf = new char[bufLen];
    DBBuf = new char[bufLen];
    ALBuf = new char[bufLen];
  }

  while(!fOneToOneResult.empty())
  {
    /*P GET FIRST ONE ELE FROM QUEUE */
```

```
    unit = fOneToOneResult.top();
    fOneToOneResult.pop();
    /*P EMPTY BUFFER */
    memset(RRBuf, '\0', bufLen);
    memset(DBBuf, '\0', bufLen);
    memset(ALBuf, '\0', bufLen);
    memset(buf, '\0', 256);

    /*P IF RR SEQUENCE REVERSED */
    if(unit.RRSeqsHead.isReversed)
      RRrev[0] = 'R';
    else
      RRrev[0] = 'O';

    /*P IF DB REVERSED */
    if(unit.databaseSeqsHead.isReversed)
      DBrev[0] = 'R';
    else
      DBrev[0] = 'O';

    /*P WHICH SEQUENCE IS LONGER */
    if(unit.RRSeqsHead.len <= unit.databaseSeqsHead.len)
    {
      sprintf(buf,"     RR Seqs(%s)# %d[%d ... %d]\n     DB Seqs(%s)# %d[%d(offset:%d) ... %d]\n
Matched: %d\%\n",
              RRrev,unit.RRSeqsHead.id,unit.RRSeqsHead.start,
unit.RRSeqsHead.end,DBrev,unit.databaseSeqsHead.id,
              unit.databaseSeqsHead.start, unit.offset,unit.databaseSeqsHead.end, unit.maxMatch);
      seqsLen = unit.RRSeqsHead.len;
      convertChars(unit.RRSeqsList, 0, seqsLen, RRBuf);
      convertChars(unit.databaseSeqsList, unit.offset, seqsLen, DBBuf);

    } else
    {
      sprintf(buf,"     RR Seqs(%s)# %d[%d(offset:%d) ... %d]\n     DB Seqs(%s)# %d[%d ... %d]\n
Matched: %d\%\n",
              RRrev,unit.RRSeqsHead.id,unit.RRSeqsHead.start,unit.offset
,unit.RRSeqsHead.end,DBrev,unit.databaseSeqsHead.id,
              unit.databaseSeqsHead.start,unit.databaseSeqsHead.end, unit.maxMatch);
      seqsLen = unit.databaseSeqsHead.len;
      convertChars(unit.RRSeqsList, unit.offset, seqsLen, RRBuf);
      convertChars(unit.databaseSeqsList, 0, seqsLen, DBBuf);
    }/*P IF - ELSE */
    for(int i = 0; i<seqsLen; i++)
    {
      if(RRBuf[i] == DBBuf[i])
        ALBuf[i] = '|';
      else
        ALBuf[i] = ' ';
    }/*P END OF FOR LOOP */
```

```
  /*P PRINT ONE PAIR SEQUENCE RESULT TO OUTPUT FILE */
  fputs(buf, reportHandler);
  int printCount = 0, printS=0, printE=0;
  while(printCount < seqsLen)
  {
    /*P CALC # OF CHAR TO BE PRINTED*/
    printS = printCount;
    if((seqsLen-printCount)>=60)
    {
      printE = printCount+60;
    }
    else
    {
      printE = seqsLen;
    }

    /*P PRINT RR BUFFER */
    fputs("\n     ",reportHandler);
    for(int i=printS; i < printE; i++)
    {
      fputc(RRBuf[i], reportHandler);
    }

    /*P PRINT AL BUFFER*/
    fputs("\n     ",reportHandler);
    for(int i=printS; i < printE; i++)
    {
      fputc(ALBuf[i], reportHandler);
    }

    /*P PRINT DB BUFFER*/
    fputs("\n     ",reportHandler);
    for(int i=printS; i < printE; i++)
    {
      fputc(DBBuf[i], reportHandler);
    }
    fputs("\n",reportHandler);
    /*P RE-CLAC PRINTCOUNT */
    printCount = printE;
  }/*END IF INNER WHILE LOOP*/
  fputs("\n\n", reportHandler);
}/*END OF WHILE LOOP*/

/*P RELEASE MEMORY */
if(NULL != RRBuf)
{
  delete [] RRBuf;
  delete [] DBBuf;
  delete [] ALBuf;
}
```

```
  fputs("**********END OF ONE DATABASE**********\n",reportHandler);
}

void convertChars(vector<unsigned int>& v,unsigned int start,  unsigned int len, char* buf)
{
  /*P INIT TEMP INT HOLDER */
  unsigned int output = 0;
  int count = 0;
  int count1= start/16;

  output = v[count1];
  int j = (start%16)+1;

  while(1)
  {
    for(;j<=16;j++)
    {
      buf[count] = RETRIEVECHAR[RETRIEVE(output,j)];
      count++;
      if(count == len)
        return;
    }
    count1++;
    output = v[count1];
    j = 1;
  }/*P END OF WHILE LOOP */
}

int compareSequences(FILE* databaseHandler, FILE* RRHandler)
{
  static unsigned int count=0;
  /*P INITIAL RETURN VALUE */
  int rev = 0;
  /*P INITIAL TEMP COMPARE MAX TO ZERO */
  int tempMax = 0;
  /*P INITIAL DATABASE SEQUENCE'S LENGTH TO ZERO */
  seqsHead_t databaseSeqsHead;
  /*P INITIAL RR HANDLER SEQUENCE'S HEAD */
  seqsHead_t RRSeqsHead;
  /*P INITIAL LENGTH OF FILE FOR A SEQUENCE */
  int RRFileLen = 0;
  int databaseFileLen = 0;
  /*P INITIAL BUFFER FOR HOLD ONE SEQUENCE IN RR TO ZERO */
  unsigned int* RRSeqsBuf = NULL;
  /*P INITIAL BUFFER FOR HOLD ONE DATABASE SEQUENCE TO ZERO */
  unsigned int* databaseSeqsBuf = NULL;
  /*P INITIAL DRAW COUNT TO 0*/
  unsigned int drawCount = 0;
  /*P INIT FILE HANDLER'S POS TO START*/
  clearerr(RRHandler);
```

```
  fseek(RRHandler, 0, SEEK_SET);

  //FOR EACH SEQUENCE IN RR
  while(0 == feof(RRHandler))
  {
    RRFileLen = getNextSequenceSize(RRHandler, RRSeqsHead);
    /*P IF THERE ARE A VALID SEQUENCE */
    if(RRFileLen > 0)
    {
      /*P ALLOCATE MEMORY FOR RR SEQUENCE */
      RRSeqsBuf = new unsigned int[RRFileLen+2];
      fread((void*)RRSeqsBuf, RRFileLen*sizeof(unsigned int), 1, RRHandler);

      /*P FOR EACH SEQUENCE IN DATABASE*/
      /*P RESET START POS */
      clearerr(databaseHandler);
      fseek(databaseHandler, 0, SEEK_SET);
      while(0 == feof(RRHandler))
      {
        /*P seqsCompareResultUnit_t datastruct to hold temp max */
        seqsCompareResultUnit_t seqsComRst;

        databaseFileLen = getNextSequenceSize(databaseHandler, databaseSeqsHead);
        /*P IF THERE ARE A VALID SEQUENCE */
        if(databaseFileLen > 0)
        {
          /*P ALLOCATE MEMORY FOR DATABASE SEQUENCE */
          databaseSeqsBuf = new unsigned int[databaseFileLen+2];
          fread((void*)databaseSeqsBuf, databaseFileLen*sizeof(unsigned int), 1,
databaseHandler);
          //do some there
          //count++;
          tempMax = compareOnePairSequence(databaseSeqsBuf, databaseSeqsHead.len,  RRSeqsBuf,
RRSeqsHead.len, seqsComRst.offset);
          /*P UPDATA QUEUE */
          if(tempMax >= fthreshold)
          {
            seqsComRst.maxMatch = tempMax;
            seqsComRst.RRSeqsHead = RRSeqsHead;
            seqsComRst.databaseSeqsHead = databaseSeqsHead;
            for(int i=0; i < RRFileLen+2; i++)
              seqsComRst.RRSeqsList.push_back(RRSeqsBuf[i]);
            for(int i=0; i < databaseFileLen+2; i++)
              seqsComRst.databaseSeqsList.push_back(databaseSeqsBuf[i]);
            fOneToOneResult.push(seqsComRst);
          }

          if(rev < tempMax)
          {
            rev=tempMax;
```

```
        }
        /*P RELEASE MEMORY FOR DATABASE SEQUENCE */
        delete [] databaseSeqsBuf;
      }
      if(databaseFileLen < 0)
        break;
    }//FOR EACH SEQUENCE IN DATABASE
    /*P RELEASE MEMORY FOR RR SEQUENCE*/
    delete [] RRSeqsBuf;
  }
  drawCount++;
  cout<<'\b'<<drawEle[(drawCount%4)]<<flush;
  if(RRFileLen < 0)
    break;
 }//FOR EACH SEQUENCE IN RR
 return rev;
}

int compareOnePairSequence(unsigned int* databaseBuf, unsigned int databaseSeqsLen,
    unsigned int* RRBuf, unsigned int RRSeqsLen, unsigned int& offset)
{
  /*P INITIAL RETURN VALUE TO ZERO */
  int rev = 0;
  int temp = 0;
  /*P INITIAL LENGTH IN NUM INTEGER TO ZERO */
  int numOfInt = 0;

  /*P INITIAL SHIFT STEP TO ZERO */
  int shiftStep = 0;

  /*P INITIAL MAX BUFFER POINTER TO NULL */
  unsigned int* largeBuf = NULL;

  /*P INITIAL MIN BUFFER POINTER TO NULL */
  unsigned int* smallBuf =NULL;

  /*P INITIAL LARGE SEQUENCE NUM TO ZERO */
  unsigned int largeLen = 0;
  /*P INITIAL SMALL SEQUENCE NUM TO ZERO */
  unsigned int smallLen = 0;

  /*P INITIAL TEMP UNSIGNED INT BUF TO ZERO */
  unsigned int uintBuf = 0;

  /*P HELP TEMP VARIABLE */
  unsigned int bitStart = 0;
  unsigned int bitEnd = 0;
  unsigned int bitDiff = 0;
  unsigned int mod16 = 0;
  unsigned int op1 = 0, op2 = 2;
```

```
if(databaseSeqsLen > RRSeqsLen)
{
  largeBuf = databaseBuf;
  largeLen = databaseSeqsLen;
  smallBuf = RRBuf;
  smallLen = RRSeqsLen;
}else
{
  smallBuf = databaseBuf;
  smallLen = databaseSeqsLen;
  largeBuf = RRBuf;
  largeLen = RRSeqsLen;
}

if(smallLen%16 == 0)
  numOfInt = smallLen/16;
else
  numOfInt = 1 + (smallLen/16);

shiftStep = largeLen - smallLen;
for(int i=0; i <= shiftStep; i++)
{
  temp = 0;
  bitStart = i/16;
  bitEnd = bitStart + 1;
  bitDiff = (i - (bitStart * 16))*2;
  mod16 = i%16;
  /*P FOR EACH UNSIGNED INT HERE*/
  for(int j = 0; j < numOfInt;j++)
  {

    /*P COMBINE TWO INT TO ONE INT */
    if(mod16 != 0)
    {
      op1 = COMBINE(largeBuf[j + bitStart],largeBuf[j + bitEnd],bitDiff);
    }
    else
    {
      op1 = largeBuf[j + bitStart];
    }
    /*P GET RIDE OF GARBAGER */
    if(j==(numOfInt-1))
    {
      op1 = MASKRIGHT(op1,numOfInt*32-2*smallLen);
    }
    /*P GET SECOND OPERATOR */
    op2 = smallBuf[j];
    /*P TAKE OFF SAME BITS */
    uintBuf = op1 ^ op2;
```

```
      /*P GET NUM OF DIFF */
      temp += extractDiff(uintBuf);
    }//for
    temp =smallLen -temp;
    if(rev < temp)
    {
      offset = i;
      rev=temp;
    }
    if(rev == smallLen)
      break;
  }//for

  //
  rev = rev*100/RRSeqsLen;
  return rev;
}

int extractDiff(unsigned int uintBuf)
{
  int sumDiff = 0;
  unsigned int oneChar;
  oneChar = uintBuf & (0x000000ff);
  sumDiff +=(MAPBUF[oneChar]);
  oneChar = (uintBuf & (0x0000ff00)) >> 8;
  sumDiff += (MAPBUF[oneChar]);
  oneChar = (uintBuf & (0x00ff0000)) >> 16;
  sumDiff += (MAPBUF[oneChar]);
  oneChar = (uintBuf >> 24);
  sumDiff += (MAPBUF[oneChar]);

  return sumDiff;
}

void switchMap()
{
  for(int i=0; i<256; i++)
    MAPBUF[i]= 4 - MAPBUF[i];
}

int getNextSequenceSize(FILE* inFile, seqsHead_t& seqHead)
{
  /*P INITIAL rev to false */
  int rev = -10; /* return value */

  /*P CLEAR SEQHEAD */
  seqHead.len = 0;

  /*P read next sequence's head */
  fread((void*)&seqHead, sizeof(seqsHead_t), 1, inFile);
```

```
  /*P IF NOT REACH TO END OF FILE */
  if(0 == feof(inFile))
  {//IF
    rev = (int)(seqHead.len);

    /*P IF THE SIZE IS VALID NUMBER*/
    if(rev > 0)
    {
      if((rev % 16) == 0)
      {//IF
        rev = (rev/16);
      }else
      {
        rev = ((rev/16) +1);
      }//IF-ELSE
    }
  }//END IF
  return rev;
}
```

Appendix 3 Code for RegCompare II

```
##!/usr/local/bin/perl

#Program purpose: This program use for search a specificy pattern in DNA sequence database.
#Usage: perl RegCompare.pl [Detail level]
#For Rebecca Dan Zhu using only.
#

#function for check antisense
sub antisenseSingleCheck{
        #convert sequence sequence
        $firstPara=$_[0];
        $secondPara=$_[1];
        #sub char by a<->t g<->c
        $firstPara=~tr/atATcgCG/taTAgcGC/;
        #reverse
        $antisenseStr=reverse($firstPara);
        $maxLength=length($antisenseStr)-3;
        for($i=0;$i<$maxLength;$i++)
        {
                $temp = substr $antisenseStr, $i, 3;
                if($secondPara=~/$temp/og)
                {
                        return true;
                }
        }

        return false;

}
sub antisenseMatch{
        for($i=0;$i<4;$i++)
        {
                for($j=0;$j<4;$j++)
                {
                        next unless($i!=$j);
                        if(antisenseSingleCheck($_[$i],$_[$j]))
                        {
                                return true;
                        }
                }
        }
        return false;
}


#define search pattern here
#my $pattern = "(([gG]{2,6}?[atcgATCG]{2,15}?){4})";
```

```
$pattern="([gG]{2,6}?)([atcgATCG]{2,15}?)([gG]{2,6}?)([atcgATCG]{2,15}?)([gG]{2,6}?)([atcgATCG]{2
,15}?)([gG]{2,6}?)([atcgATCG]{2,15}?)";
my $DNAName = "";
my $DNAStart = "";
my $endPos = 0;
my $startPos = 0;
my $detailLevel= 0;
my $antisenseCheck=0;
if($#ARGV != 1)
{
  print "Usage: perl regCompare.pl [Detail level] [antisense]\n";
  print "Detail Level:\n";
  print "\t 0 -- Print number of match in sequence.\n";
  print "\t 1 -- Print number of match and matched position.\n";
  print "\t 2 -- Print detail info of matched sequence.\n";
  print "antisense:\n";
  print "\t 0 -- DO not do antisense check.\n";
  print "\t 1 -- DO antisense check.\n";
  die "Invalid argument list";
}

$dna_dir = ".";
$detailLevel = $ARGV[0];
$antisenseCheck=$ARGV[1];
#Open file for output search result.
$outFile = 'result.txt';
open(OUTINFO, ">$outFile");

#open Dir and read each file with .gbk
#open Dir
opendir(DIR, $dna_dir) || die "Cannot open dir $dna_dir: $!";
while (my $dna_sequence_file = readdir(DIR))
{
        #skip if entry is not a file
        next unless (-f "$dna_dir/$dna_sequence_file");
        #skip if file not end with .gbk
        next unless ($dna_sequence_file =~ m/\.gbk$/);
        #work with this dna sequence file
        #
        print "Searching DNA sequence file '$dna_sequence_file'\n";
        #open input file
        open(INPUTINFO, $dna_sequence_file);
         #read DNA name from file, it exist at second line of file
         $DNAName = <INPUTINFO>;
         #Got name here
         $DNAName = <INPUTINFO>;
         #Get DNA sequence start pos
         #Or may reach end of file
         $DNAStart = <INPUTINFO>;
         while((defined($DNAStart)) && ($DNAStart !~ /ORIGIN/))
```

```
{
  $DNAStart = <INPUTINFO>;
}

#reach end of file, print err msg
if(not defined($DNAStart))
{
  print OUTINFO "err on file: ",$dna_sequence_file, " --> ", $DNAName, "\n\n";
  print "err on file: ",$dna_sequence_file, " --> ", $DNAName, "\n\n";
  close(INPUTINFO);
}

#Go to next file if reach end of file
next unless(defined($DNAStart));

       #read rest of whole file to array
@seqFromFileArray = <INPUTINFO>;
#turn list to global variable string
$_ = "@seqFromFileArray";

#remove space char and control char from string
       s/\W//g;
#remove digit from string
       s/\d//g;

       #output result title to file
print OUTINFO "FILE: ", $dna_sequence_file,"\n", $DNAName;

       #output result title to screen
print "FILE: ", $dna_sequence_file,"\n", $DNAName;

       my $findCount = 0;
       my $matched=false;
#do a RE search
while(/$pattern/og)
{
  #RE exist in DNA seqs
  $endPos = pos($_) - 1;
         #Get the start pos of this RE in sequence
         $matchedStr=$1.$2.$3.$4.$5.$6.$7.$8;
         $startPos = $endPos - length($matchedStr);

         #if we need do more detail check for sequence match
         if(1==$antisenseCheck)
         {
                if(antisenseMatch($2,$4,$6,$8))
                {
                       $matched=true;
                }
                else
```

```
                        {
                                $matched=false;
                        }
                }
                else
                {
                        $matched=true;
                }
                if($matched)
                {
                        if($detailLevel > 0)
                        {
                #output resule to file
                    #output result to screen
                        print OUTINFO "\t-----> Matched  @ (", $startPos ,", ", $endPos, ")\n";
                                #print "\t-----> Matched @ (", $startPos ,", ", $endPos , ")\n";
                                if($detailLevel == 2)
                                {
                                        print OUTINFO "\t",$matchedStr, "\n";
                                        #print "\t",$matchedStr, "\n";
                                }
                        }
                        #increament count
                        $findCount++;
                }
        }
        if (0 == $findCount)
        {
          #RE not exist in DNA seqs
          #output result to file
          print OUTINFO "-----> Not Match.", "\n";
          #output result to screen
          print "-----> Not Match.", "\n";
        }
                else
                {
                                #print count of match
                                print OUTINFO "\t>>>>>>>>>>Total matched ", $findCount, "
times<<<<<<<<<<\n";
                                print "\t>>>>>>>>>>Total matched ", $findCount, "
times<<<<<<<<<<\n";
                }

                #print extra newline char to file and screen
                print OUTINFO "\n";
                print "\n";

        #close input file
        close(INPUTINFO);
}
```

```
closedir DIR;


#close opened file
#close output file
close(OUTINFO);
```

Appendix 4 List of Top 7 Candidates from Our Bioinformatics Analysis

Candidates	Adjacent Gene	Strand direction	Length	Position
UIG0242	ybhi / ybhj	→	84	802544-802627
Sequence: tattctcgtcatacttcaagttgcatgtgctgcgtctgcgttcgctcaccccagtcacttacttatgtaagctcctggggattc				
UIG0803	putative transposase	→	107	1529733-1529840
Sequence: caattcgcattttatgtttaaaaattgagatattccttattactaaagctgtttttttattgcttacacatgatcaaatactccttacataattaagg agaaaaaat				
UIG0985	ydhQ/ydhR	→	124	1744449- 1744572
Sequence: tacaacgttgcgttcatagctcagttggttagagcaccaccttgacatggtgggggtcgttggttcgagtccaattgaacgcaccat cctgcgtccgtagctcagttggttagagcaccacctt				
UIG1195	yecL/yecR	→	87	1986043-1986129
Sequence: catcacaaaaatcaatctttatgtgatacaaatcacataaataccccttaatgttataaaaatgataatcaaaaaacagcccccct				
UIG1259	yeeN/adhesin	→	118	2057869- 2057986
Sequence: gattcctctgtagttcagtcggtagaacggcggactgttaatccgtatgtcactggttcgagtccagtcagaggagccaaattcctgaa aagcccgcttttatagcgggatttttgct				
UIG1354	tRNA/tRNA	→	93	2192219-2192311
Sequence: agctgatagtttacctgaagaatatagagaagtacttacttaacattttcccatttggtactatctaaccccttttcactattaagaagtaat				
UIG1585	tRNA/DNA binding activator	→	117	2519021-2519137
Sequence: caccaactactttatgtagtctccgccgtgtagcaagaaattgagaagtgggtgattagctcagctgggagagcacctcccttacaagga gggggtcggcggttcgatcccgtcatc				

Note: Candidates' names are in the order when they are found in the analysis. Estimated start and end points of these candidates in the *E. coli* MG1655 genome were listed.

Appendix 5 Genomes that contain Our Top 7 Candidate Genes

Gene Name	Genomes	
242	Erwinia carotovora	
	Escherichia coli K12	Shigella flexneri
	Escherichia coli 536	Salmonella enterica
	Escherichia coli CFT073	Salmonella typhimurium LT2
	Pectobacterium carotovorum	Vibrio cholerae
	Pectobacterium chrysanthemi	Yersinia enterocolitica
	Photorhabdus luminescens	Yersinia pestis
	Serratia marcescens	Yersinia pseudotuberculosi
803	Escherichia coli K12	Escherichia coli CFT073
	Escherichia coli MG1655	Escherichia coli 536
	Escherichia coli W3110	Escherichia coli UTI89
	Escherichia coli O157	Shigella sonnei Ss046
	Escherichia coli H7	Shigella sonnei boydii serogroup 18 O
	Escherichia coli serotype O55:H7	Shigella sonnei dysenteriae Sd197
985	Erwinia carotovora; amylovora	Pseudomonas aeruginosa
	Escherichia coli K12	Rhodoferax ferrireducens Salmonella enterica
	Escherichia coli MG1655	Salmonella typhimurium LT
	Escherichia coli CFT073	Shigella flexneri 2a
	Escherichia coli UTI89	Shigella flexneri sonnei Ss046
	Escherichia coli O157:H7	Shigella flexneri dysenteriae Sd197
	Escherichia coli W3110	Shigella flexneri flexneri
	Escherichia coli 536	Shigella flexneri boydii Sb227
	Pseudomonas entomophila	Sodalis glossinidius;
	Pseudomonas putida	Yersinia pseudotuberculosis
	Pseudomonas fluorescens	Yersinia pestis biovar Medievalis
	Pseudomonas syringae	Yersinia pestis Antiqua; pestis
	Pseudomonas phaseolicola	
1195	Bacillus subtilis	Escherichia coli MG1655
	Bacillus licheniformis	Salmonella enterica
	Escherichia coli UTI89	Salmonella typhimurium

	Escherichia coli CFT073 Escherichia coli 536 Escherichia coli O157:H7 Escherichia coli W3110 Escherichia coli K12	Shigella boydii Shigella dysenteriae Shigella flexneri Shigella sonnei Ss046
1259	Erwinia carotovora Escherichia coli K12 Escherichia coli MG1655 Escherichia coli O157:H7 Escherichia coli UTI89 Escherichia coli CFT073 Klebsiella aerogenes Salmonella enterica Salmonella typhimurium	Shigella boydii Sb227 Shigella dysenteriae Shigella sonnei Shigella flexneri Sodalis glossinidius Yersinia enterocolitica Yersinia pestis Yersinia pseudotuberculosis
1354	Escherichia coli 536 Escherichia coli O157:H7 Escherichia coli W3110 Escherichia coli K12 Escherichia coli MG1655 Shigella sonnei Ss046 Shigella boydii Sb227l	Shigella dysenteriae Shigella flexneri Kluyveromyces lactis Theileria parva strain Medicago truncatula Crioceris duodecimpunctata Arabidopsis thaliana
1585	Bacillus halodurans; Bacillus clausii; Bacillus licheniformis Baumannia cicadellinicola Erwinia carotovora Escherichia coli K12 Escherichia coli MG1655 Escherichia coli CFT073 Escherichia coli 536 Escherichia coli O157:H7 Escherichia coli W3110 Escherichia coli EDL933 Geobacillus kaustophilus	Pseudomonas entomophila Pseudomonas putida Pseudomonas aeruginosa Pseudomonas aeruginosa Pseudomonas syringae Salmonella enterica Salmonella typhimurium Shigella boydii Shigella dysenteriae Shigella flexneri Shigella sonnei Ss046 Sodalis glossinidius

Appendix 6 pZE21-MCS-1 and pNYL9-MCS11 Sequences

A. Sequence of pZE21-MCS-1

Note:

(1) Sequences highlighted in yellow is the ribosomal binding site

(2) Sequences highlighted in red is the MCS site

(3) Sequences in italic form are were removed during construction of pNYL9-MCS11 vector

CTCGAGTCCC TATCAGTGAT AGAGATTGAC ATCCCTATCA GTGATAGAGA TACTGAGCAC ATCAGCAGGA
CGCACTGACC *GAATTCATTA* *AAGAGGAGAA* *AGGTA*CCGGG *CCCCCCCTCG* *AGGTCGACGG* *TATCGATAAG*
CTTGATATCG *AATTCCTGCA* GCCCGGGGGA TCCCATGGTA CGCGTGCTAG AGGCATCAAA TAAAACGAAA
GGCTCAGTCG AAAGACTGGG CCTTTCGTTT TATCTGTTGT TTGTCGGTGA ACGCTCTCCT GAGTAGGACA
AATCCGCCGC CCTAGACCTA
GGCGTTCGGC TGCGGCGAGC GGTATCAGCT CACTCAAAGG CGGTAATACG GTTATCCACA GAATCAGGGG
ATAACGCAGG AAAGAACATG TGAGCAAAAG GCCAGCAAAA GGCCAGGAAC CGTAAAAAGG CCGCGTTGCT
GGCGTTTTTC CATAGGCTCC GCCCCCCTGA CGAGCATCAC AAAAATCGAC GCTCAAGTCA GAGGTGGCGA
AACCCGACAG GACTATAAAG ATACCAGGCG TTTCCCCCTG GAAGCTCCCT CGTGCGCTCT CCTGTTCCGA
CCCTGCCGCT TACCGGATAC
CTGTCCGCCT TTCTCCCTTC GGGAAGCGTG GCGCTTTCTC AATGCTCACG CTGTAGGTAT CTCAGTTCGG
TGTAGGTCGT TCGCTCCAAG CTGGGCTGTG TGCACGAACC CCCCGTTCAG CCCGACCGCT GCGCCTTATC
CGGTAACTAT CGTCTTGAGT CCAACCCGGT AAGACACGAC TTATCGCCAC TGGCAGCAGC CACTGGTAAC
AGGATTAGCA GAGCGAGGTA TGTAGGCGGT GCTACAGAGT TCTTGAAGTG GTGGCCTAAC TACGGCTACA
CTAGAAGGAC AGTATTTGGT
ATCTGCGCTC TGCTGAAGCC AGTTACCTTC GGAAAAAGAG TTGGTAGCTC TTGATCCGGC AAACAAACCA
CCGCTGGTAG CGGTGGTTTT TTTGTTTGCA AGCAGCAGAT TACGCGCAGA AAAAAAGGAT CTCAAGAAGA
TCCTTTGATC TTTTCTACGG GGTCTGACGC TCAGTGGAAC GAAAACTCAC GTTAAGGGAT TTTGGTCATG
ACTAGTGCTT GGATTCTCAC CAATAAAAAA CGCCCGGCGG CAACCGAGCG TTCTGAACAA ATCCAGATGG
AGTTCTGAGG TCATTACTGG
ATCTATCAAC AGGAGTCCAA GCGAGCTCTC GAACCCCAGA GTCCCGCTCA GAAGAACTCG TCAAGAAGGC
GATAGAAGGC GATGCGCTGC GAATCGGGAG CGGCGATACC GTAAAGCACG AGGAAGCGGT CAGCCCATTC
GCCGCCAAGC TCTTCAGCAA TATCACGGGT AGCCAACGCT ATGTCCTGAT AGCGGTCCGC CACACCCAGC
CGGCCACAGT CGATGAATCC AGAAAAGCGG CCATTTTCCA CCATGATATT CGGCAAGCAG GCATCGCCAT
GGGTCACGAC GAGATCCTCG
CCGTCGGGCA TGCGCGCCTT GAGCCTGGCG AACAGTTCGG CTGGCGCGAG CCCCTGATGC TCTTCGTCCA
GATCATCCTG ATCGACAAGA CCGGCTTCCA TCCGAGTACG TGCTCGCTCG ATGCGATGTT TCGCTTGGTG
GTCGAATGGG CAGGTAGCCG GATCAAGCGT ATGCAGCCGC CGCATTGCAT CAGCCATGAT GGATACTTTC
TCGGCAGGAG CAAGGTGAGA TGACAGGAGA TCCTGCCCCG GCACTTCGCC CAATAGCAGC CAGTCCCTTC

CCGCTTCAGT GACAACGTCG
AGCACAGCTG CGCAAGGAAC GCCCGTCGTG GCCAGCCACG ATAGCCGCGC TGCCTCGTCC TGCAGTTCAT
TCAGGGCACC GGACAGGTCG GTCTTGACAA AAAGAACCGG GCGCCCCTGC GCTGACAGCC GGAACACGGC
GGCATCAGAG CAGCCGATTG TCTGTTGTGC CCAGTCATAG CCGAATAGCC TCTCCACCCA AGCGGCCGGA
GAACCTGCGT GCAATCCATC TTGTTCAATC ATGCGAAACG ATCCTCATCC TGTCTCTTGA TCAGATCTTG
ATCCCCTGCG CCATCAGATC
CTTGGCGGCA AGAAAGCCAT CCAGTTTACT TTGCAGGGCT TCCCAACCTT ACCAGAGGGC GCCCCAGCTG
GCAATTCCGA CGTCTAAGAA ACCATTATTA TCATGACATT AACCTATAAA AATAGGCGTA TCACGAGGCC
CTTTCGTCTT
CAC

B. Sequence of pNYL9-MCS11

Note: Sequences highlighted in yellow is the MCS11 site

CTCGAGTCCC TATCAGTGAT AGAGATTGAC ATCCCTATCA GTGATAGAGA TACTGAGCAC ATCAGCAGGA
CGCACTGACC CTCGAGTCGA CGGTATCGAT AAGCTTGATA TCGAATTCCT GCAGCCCGGG GGATCCCTG
CACGTGCTAG AGGCATCAAA TAAAACGAAA GGCTCAGTCG AAAGACTGGG CCTTTCGTTT TATCTGTTGT
TTGTCGGTGA ACGCTCTCCT GAGTAGGACA AATCCGCCGC CCTAGACCTA GGCGTTCGGC TGCGGCGAGC
GGTATCAGCT CACTCAAAGG CGGTAATACG GTTATCCACA GAATCAGGGG ATAACGCAGG AAAGAACATG
TGAGCAAAAG GCCAGCAAAA GGCCAGGAAC CGTAAAAAGG CCGCGTTGCT GGCGTTTTTC CATAGGCTCC
GCCCCCCTGA CGAGCATCAC AAAAATCGAC GCTCAAGTCA GAGGTGGCGA AACCCGACAG GACTATAAAG
ATACCAGGCG TTTCCCCCTG GAAGCTCCCT CGTGCGCTCT CCTGTTCCGA CCCTGCCGCT TACCGGATAC
CTGTCCGCCT TTCTCCCTTC GGGAAGCGTG GCGCTTTCTC AATGCTCACG CTGTAGGTAT CTCAGTTCGG
TGTAGGTCGT TCGCTCCAAG CTGGGCTGTG TGCACGAACC CCCCGTTCAG CCCGACCGCT GCGCCTTATC
CGGTAACTAT CGTCTTGAGT CCAACCCGGT AAGACACGAC TTATCGCCAC TGGCAGCAGC CACTGGTAAC
AGGATTAGCA GAGCGAGGTA TGTAGGCGGT GCTACAGAGT TCTTGAAGTG GTGGCCTAAC TACGGCTACA
CTAGAAGGAC AGTATTTGGT ATCTGCGCTC TGCTGAAGCC AGTTACCTTC GGAAAAAGAG TTGGTAGCTC
TTGATCCGGC AAACAAACCA CCGCTGGTAG CGGTGGTTTT TTTGTTTGCA AGCAGCAGAT TACGCGCAGA
AAAAAAGGAT CTCAAGAAGA TCCTTTGATC TTTTCTACGG GGTCTGACGC TCAGTGGAAC GAAAACTCAC
GTTAAGGGAT TTTGGTCATG ACTAGTGCTT GGATTCTCAC CAATAAAAAA CGCCCGGCGG CAACCGAGCG
TTCTGAACAA ATCCAGATGG AGTTCTGAGG TCATTACTGG ATCTATCAAC AGGAGTCCAA GCGAGCTCTC
GAACCCCAGA GTCCCGCTCA GAAGAACTCG TCAAGAAGGC GATAGAAGGC GATGCGCTGC GAATCGGGAG
CGGCGATACC GTAAAGCACG AGGAAGCGGT CAGCCCATTC GCCGCCAAGC TCTTCAGCAA TATCACGGGT
AGCCAACGCT ATGTCCTGAT AGCGGTCCGC CACACCCAGC CGGCCACAGT CGATGAATCC AGAAAAGCGG
CCATTTTCCA CCATGATATT CGGCAAGCAG GCATCGCCAT GGGTCACGAC GAGATCCTCG CCGTCGGGCA
TGCGCGCCTT GAGCCTGGCG AACAGTTCGG CTGGCGCGAG CCCCTGATGC TCTTCGTCCA GATCATCCTG
ATCGACAAGA CCGGCTTCCA TCCGAGTACG TGCTCGCTCG ATGCGATGTT TCGCTTGGTG GTCGAATGGG

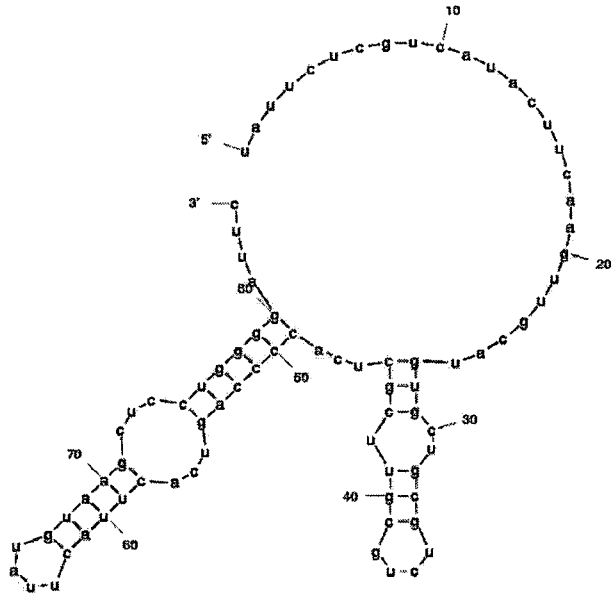
CAGGTAGCCG GATCAAGCGT ATGCAGCCGC CGCATTGCAT CAGCCATGAT GGATACTTTC TCGGCAGGAG
CAAGGTGAGA TGACAGGAGA TCCTGCCCCG GCACTTCGCC CAATAGCAGC CAGTCCCTTC CCGCTTCAGT
GACAACGTCG AGCACAGCTG CGCAAGGAAC GCCCGTCGTG GCCAGCCACG ATAGCCGCGC TGCCTCGTCC
TGCAGTTCAT TCAGGGCACC GGACAGGTCG GTCTTGACAA AAAGAACCGG GCGCCCCTGC GCTGACAGCC
GGAACACGGC GGCATCAGAG CAGCCGATTG TCTGTTGTGC CCAGTCATAG CCGAATAGCC TCTCCACCCA
AGCGGCCGGA GAACCTGCGT GCAATCCATC TTGTTCAATC ATGCGAAACG ATCCTCATCC TGTCTCTTGA
TCAGATCTTG ATCCCCTGCG CCATCAGATC CTTGGCGGCA AGAAAGCCAT CCAGTTTACT TTGCAGGGCT
TCCCAACCTT ACCAGAGGGC GCCCCAGCTG GCAATTCCGA CGTCTAAGAA ACCATTATTA TCATGACATT
AACCTATAAA AATAGGCGTA TCACGAGGCC CTTTCGTCTT CAC

Appendix 7 List of primers used in PCR

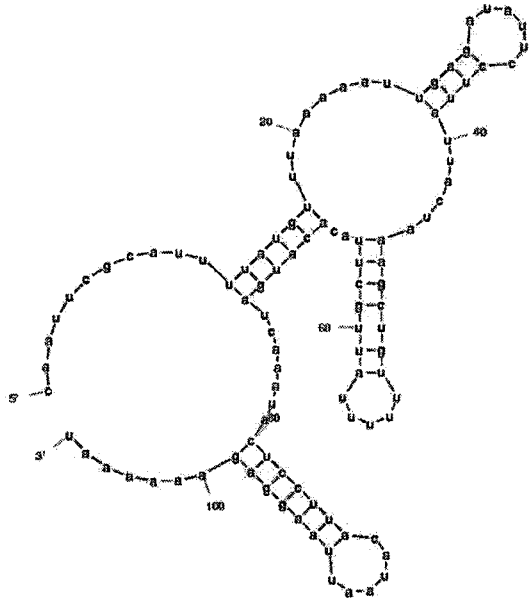
Gene Name	Forward Primer	Reverse Primer
UIG0242	tcgacgtcgacatcggatcctattctcgtcatacttcaagtt	cgatcggaccgattaagcttgaatccccaggagct
UIG0803	tcgacgtcgacatcggatccaattcgcattttatgtttaaaaattg	cgatcggaccgattaagcttattttgttctccttaattatgtaagg
UIG0985	tcgacgtcgacatcggatcctacaacgttgcgttcatagct	cgatcggaccgattaagcttaaggtggtgctctaaccaac
UIG1195	tcgacgtcgacatcggatccatcacaaaatcaatctttatgt	cgatcggaccgattaagcttaggggggctgtttttgatt
UIG1259	tcgacgtcgacatcggatccgattcctctgtagttcagtc	cgatcggaccgattaagcttagcaaaaatcccgctataaaagc
UIG1354	tcgacgtcgacatcggatccagctgatagtttacctgaagaat	cgatcggaccgattaagcttattacttcttaatagtgaaaaggg
UIG1585	tcgacgtcgacatcggatccaccaactactttatgtagt	cgatcggaccgattaagcttgatgacgggatcgaacc
tpke11	tcgacgtcgacatcggatcctcgccctataaacgggtaat	cgatcggaccgattaagctttcttttttaaattgccccta
C0293	tcgacgtcgacatcggatcctcacccgggactcgccaggg	cgatcggaccgattaagctttccgcatgtacctgaacgc
C0299	tcgacgtcgacatcggatccgccacgtgagcacaagataa	cgatcggaccgattaagcttcaatactgattcaggctatc
C0343	tcgacgtcgacatcggatcctgttccagtgcccgatccg	cgatcggaccgattaagcttatttaatcggttcttcatcg
sraD	tcgacgtcgacatcggatcctgaaagacgcgcatttgtta	cgatcggaccgattaagcttgaaaaaggccactcgtgagt
sraL	tcgacgtcgacatcggatccaaactaaagcgccacaaggg	cgatcggaccgattaagcttatcaacaccaaccggaacct
sraI	tcgacgtcgacatcggatccgcgatcaggaagaccctcgc	cgatcggaccgattaagcttaaaaaaaagccagcacccgg
ssrA	tcgacgtcgacatcggatccggggctgattctggattcga	cgatcggaccgattaagcttggtggagctggcgggagtt

Note: *SalI* site in each primer was highlighted in yellow, *BamHI* site was highlighted in green, *HindIII* site was highlighted in red.

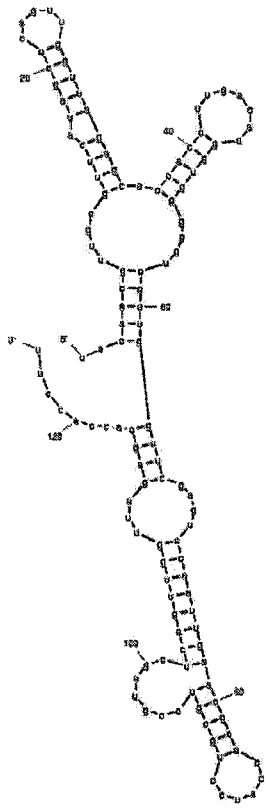
Appendix 10 Secondary Structures of all 7 candidates



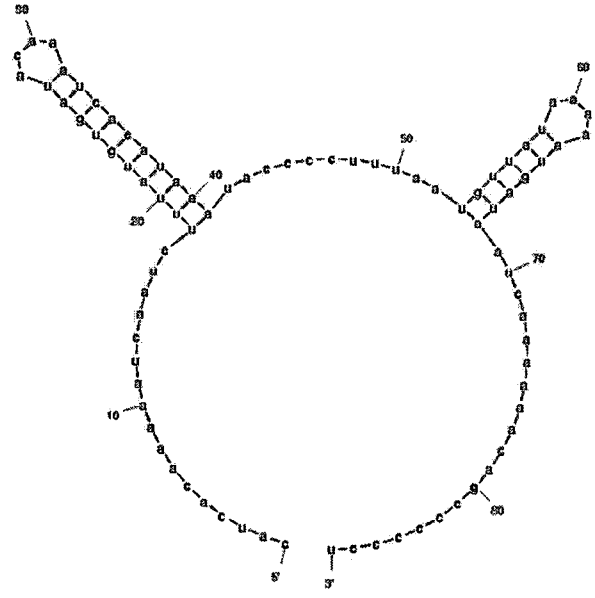
UIG0242



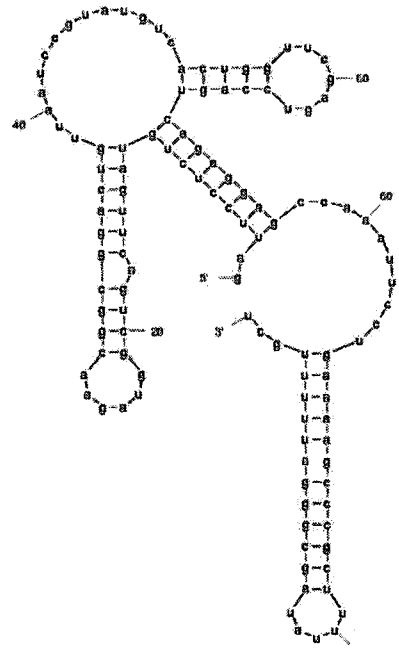
UIG0803



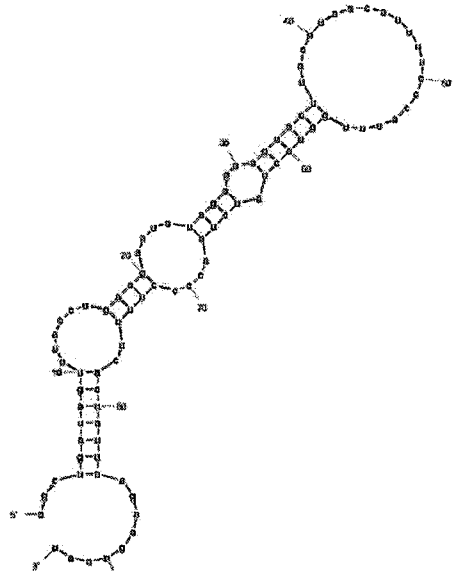
UIG0985



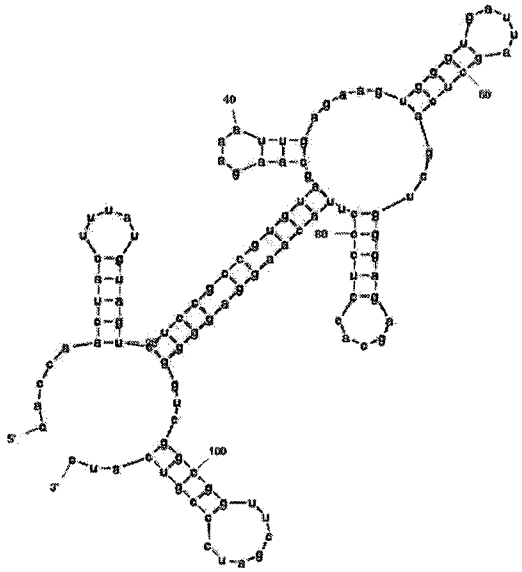
UIG1195



UIG1259



UIG1354



UIG1585

Appendix 11 List of sequences from Lambda library

Clone 140L

ATTTCTATACTCATCAAACTGTAGGGGTTGTAATAGTTTATCCGATTTCTCGCTGTAGGGGTACACGAGAACCACCGA
GCCTGATGTGGTTAAAAGACAGGCACAATCTTTACTACCGCAATCCACTATTTAAGGTGATATATGGAAGAAGAATTT
GAAGAGTTCGAAGAGCATCCTCAGGATGTGATGGAACAATACCAGGACTATCCGTATGACTACGACTATTGATAAAAA
TCAATGGTGTGGACAATTCAAGCGATGCAATGGATGCAAGCTGCAATCGGAATGCATGGTTAAGCCTGAAGAAATGTT
TCCTGTAATGGAAGATGGGAAATATGTCGATAAATGGGCAATACGAACGACGGCAATGATTGCCAGAGAACTTGGTAA
ACAGAACAACAAAGCTGCCTGATAGTGGCCTTTATTTTTGGCATAAATAACAGAATAAACACTGCACTGTGTATTCAT
TCCAACGAGTGAATACACGGAGCAATGTCGCTCGTAACTAAACAGGAGCCGACTTGTTCTGATTATTGGAAATCTTCT
TTGCCCTCCAGTGTGAGGGCGANTTTTTATCTGTGAGGATATGAACAGATGTCAAACATCAAAAAATACATCATTGAT
TACNACTGGNAAGCATCANTNGAAANTGAAATCGACCATGACGTA

Clone 141L

ATACGTTTGCCAGCGATGTGCAGGTTATGGTGATTAAGAAACAGGCGCTGGGCATCAGCGTGGTCTGAGTGTGTTACA
GAGGTTCGTCCGGGAACGGGCGTTTTATTATAAAACAGTGAGAGGTGAACGATGCGTAATGTGTGTATTGCCGTTGCT
GTCTTTGCCGCACTTGCGGTGACAGTCACTCCGGCCCGTGCGGAAGGTGGACATGGTACGTTTACGGTGGGCTATTTT
CAAGTGAAACCGGGTACATTGCCGTCGTTGTCGGGCGGGGATACCGGTGTGAGTCATCTGAAAGGGATTAACGTGAAG
TACCGTTATGAGCTGACGGACAGTGTGGGGGTGATGGCTTCCCTGGGGTTCGCCGCGTCGAAAAAGAGCAGCACAGTG
ATGACCGGGGAGGATACGTTTCACTATGAGAGCCTGCGTGGACGTTATGTGAGCGTGATGGCCGGACCGGTTTTACAA
ATCAGTAAGCAGGTCAGTGCGTACGCCATGGCCGGAGTGGCTCACAGTCGGTGGTCCGGCAGTACAATGGATTACCGT
AAGACGGAAATCACTCCCGGGTATATGAAAGAGACGACCACTGCCAGGGACGAAAGTGCAATGCGGCATACCTCAGTG
GCGTGGAGTGCAGGTATACAGATTAATCCGGCAGCGTCCGTCGTTGTTGATATTGCTTATGAAGGCTCCGGCAGTGGC
GACTGGCGTACTGACGGATTCATCGTTGGGGTCGGTTATAAATTCTGATTAGCCAGGTAACACAGTGTTATGACAGCC
CGNCGGAACCGGTGGGCTTTTTTGTGGGGTGAATATGGCAGTAAAGATTTCAGGAGTCCTGNAAGANGGCACAGGAAA
ACCGGTACAGAACTGCACCATTCAGCTGAAAGCCAGACGTACAGCACCACGTGGTGGTGAACACGGTGGGCTNNNANA
ATCCGGATGAAGCCGGGCGTTACA

Clone 152L

TATACGTTTGCCAGCGATGTGCAGGTTATGGTGATTAAGAAACAGGCGCTGGGCATCAGCGTGGT
CTGAGTGTGTTACAGAGGTTCGTCCGGGAACGGGCGTTTTATTATAAAACAGTGAGAGGTGAACG
ATGCGTAATGTGTGTATTGCCGTTGCTGTCTTTGCCGCACTTGCGGTGACAGTCACTCCGGCCCG
TGCGGAAGGTGGACATGGTACGTTTACGGTGGGCTATTTTCAAGTGAAACCGGGTACATTGCCGT
CGTTGTCGGGCGGGGATACCGGTGTGAGTCATCTGAAAGGGATTAACGTGAAGTACCGTTATGAG
CTGACGGACAGTGTGGGGGTGATGGCTTCCCTGGGGTTCGCCGCGTCGAAAAAGAGCAGCACAGT
GATGACCGGGGAGGATACGTTTCACTATGAGAGCCTGCGTGGACGTTATGTGAGCGTGATGGCCG
GACCGGTTTTACAAATCAGTAAGCAGGTCAGTGCGTACGCCATGGCCGGAGTGGCTCACAGTCGG
TGGTCCGGCAGTACAATGGATTACCGTAAGACGGAAATCACTCCCGGGTATATGAAAGAGACGAC
CACTGCCAGGGACGAAAGTGCAATGCGGCATACCTCAGTGGCGTGGAGTGCAGGTATACAGATTA
ATCCGGCAGCGTCCGTCGTTGTTGATATTGCTTATGAAGGCTCCGGCAGTGGCGACTGGCGTACT
GACGGATTCATCGTTGGGGTCGGTTATAAATTCTGATTAGCCAGGTAACACAGTGTTATGACAGC
CCGCCGGAACCGGTGGGCTTTTTTGTGGGGTGAATATGGCAGTAAAGATTTCNNGAGTCCTGAAA
GACGNACAGGAAAACCGGTACAGAACTGCACCATTCANCTGAAAGCCAGACGTAACAGCACCACG
GTGGTGGTGAACACGGTGGGCTCAGANATCCGGATGAAGCNGG