การระบุตำแหน่งอาร์เอ็นเอไม่แปลรหัสในจีโนมของไมโคแบคทีเรียมทูเบอร์คูโลซิส ด้วยวิธีทางคอมพิวเตอร์หลายวิธีร่วมกัน

นายณฐพล พรพุทธพงศ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาชีวเวชเคมี ภาควิชาชีวเคมี
คณะเภสัชศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2551

IDENTIFICATION OF NON-CODING RNAs IN *MYCOBACTERIUM TUBERCULOSIS* GENOME USING COMBINED COMPUTATIONAL APPROACH

Mr. Natapol Pornputtapong

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Sciences Program in Biomedicinal Chemistry
Department of Biochemistry
Faculty of Pharmaceutical Sciences
Chulalongkorn University
Academic Year 2008

Thesis Title	IDENTIFICATION OF NON-CODING RNAs IN *MYCOBACTERIUM TUBERCULOSIS* GENOME USING COMBINED COMPUTATIONAL APPROACH
By	Mr. Natapol Pornputtapong
Field of Study	Biomedicinal Chemistry
Thesis Principal Advisor	Associate Professor Duangdeun Meksuriyen, Ph.D.
Thesis Co-advisor	Chinae Thammarongtham, Ph.D.

Accepted by the Faculty of Pharmaceutical Sciences, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

……Dean of the Faculty of Pharmaceutical Sciences
(Associate Professor Pornpen Pramyothin, Ph.D.)

THESIS COMMITTEE

…….Chairman
(Assistant Professor Boonsri Ongpipattanakul, Ph.D.)

……..Thesis Principal Advisor
(Associate Professor Duangdeun Meksuriyen, Ph.D.)

……..…Thesis Co-advisor
(Chinae Thammarongtham, Ph.D.)

……..…….External Member
(Therdsak Prammananan, Dr. rer. biol. hum.)

……..….Member
(Associate Professor Thitima Pengsuparp, Ph.D.)

ณฐพล พรพุทธพงศ์ : การระบุตำแหน่งอาร์เอ็นเอไม่แปลรหัสในจีโนมของไมโคแบคทีเรียมทูเบอร์คูโลซิสด้วยวิธีทางคอมพิวเตอร์หลายวิธีร่วมกัน (IDENTIFICATION OF NON-CODING RNAs IN *MYCOBACTERIUM TUBERCULOSIS* GENOME USING COMBINED COMPUTATIONAL APPROACH) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : รศ.ดร. ดวงเดือน เมฆสุริเยนทร์, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : ดร. ชิเน ธำมรงค์ธรรม, 103 หน้า.

ปัจจุบันเป็นที่รู้กันดีว่าอาร์เอ็นเอไม่แปลรหัสมีหน้าที่สำคัญ ในการควบคุมการทำงานภายในเซลล์ ปกติการค้นหาตำแหน่งของยีนของอาร์เอ็นเอไม่แปลรหัสด้วยวิธีทางห้องปฏิบัติการทำได้ลำบาก จึงได้มีการนำวิธีทางคอมพิวเตอร์เข้ามาช่วยในการค้นหายีนของอาร์เอ็นเอไม่แปลรหัส แต่ว่าวิธีทางคอมพิวเตอร์ยังมีข้อจำกัดและผลที่ได้บางส่วนยังมีความคลาดเคลื่อน ในงานวิจัยนี้จึงใช้หลายวิธีรวมกันโดยสร้างเป็นข่ายงาน เพื่อใช้ในการค้นหายีนของอาร์เอ็นเอไม่แปลรหัส ซึ่งเป็นการนำข้อดีของวิธีการต่าง ๆ มารวมกัน แกนหลักของข่ายงานใช้วิธีที่เชื่อว่ามีความถูกต้องที่สุดคือการทำนายโครงสร้างในระดับทุติยภูมิของสายอาร์เอ็นเอ คำนวณโดยโปรแกรม RNAz ซึ่งผู้พัฒนาได้แนะนำให้ใช้คู่กับโปรแกรม TBA ที่ช่วยเปรียบเทียบจีโนม หลังจากที่ได้ทดสอบข่ายงานกับจีโนมของ *Escherichia coli* และจากการทบทวนวรรณกรรมพบว่าโปรแกรม TBA ทำให้เกิด false positive สูง และมีการเปรียบเทียบจีโนมที่ผิดพลาด ในงานวิจัยนี้จึงได้พัฒนาโปรแกรมที่ใช้ในการเปรียบเทียบจีโนมขึ้นใหม่ โดยใช้โปรแกรม BLAST และ MAFFT เป็นโปรแกรมหลัก ซึ่งผลที่ได้จากการทดสอบในเชื้อ *E. coli* ทำให้สามารถเพิ่มค่าความไวของการทำนายด้วย RNAz จาก 0.54 เป็น 0.84 และค่าความแม่นยำจาก 0.37 เป็น 0.56 ได้ หลังจากนั้นโปรแกรมได้ถูกนำไปใช้ในการทำนายหาตำแหน่งของยีนอาร์เอ็นเอไม่แปลรหัสใน *Mycobacterium tuberculosis* H37Rv ผลจากการทำนายพบบริเวณซึ่งน่าจะเป็นตำแหน่งของอาร์เอ็นเอไม่แปลรหัสทั้งหมด 61 บริเวณ เมื่อนำไปเปรียบเทียบกับยีนของอาร์เอ็นเอไม่แปลรหัสที่พบแล้วในจีโนมของ *M. tuberculosis* H37Rv พบว่าเป็นบริเวณซึ่งถูกรายงานแล้วในข้อมูลจีโนมของ *M. tuberculosis* H37Rv เป็นจำนวน 33 บริเวณ เมื่อนำไปประกอบกับผลที่ได้จากการทำนายตำแหน่งของโปรโมเตอร์ และเทอร์มิเนเตอร์พบบริเวณที่มีสัญญาณ 22 บริเวณ และเมื่อนำไปเปรียบเทียบกับฐานข้อมูลด้วยวิธี BLAST พบว่ามี 3 บริเวณที่ตรงกับลำดับเบสของอาร์เอ็นเอที่ทราบแล้วได้แก่อาร์เอ็นเอ *ykoK* ที่ควบคุมการแสดงออกของยีนที่เกี่ยวข้องกับโลหะที่มีประจุ +2 และอาร์เอ็นเอ SRP ซึ่งเกี่ยวข้องกับการนำส่งของโปรตีนภายในเซลล์ด้วย โดยบริเวณที่ทำนายได้ทั้งหมดมีความน่าสนใจในการนำไปศึกษาต่อในห้องปฏิบัติการ

ภาควิชา.....ชีวเคมี.....ลายมือชื่อนิสิต.....

สาขาวิชา.....ชีวเวชเคมี.....ลายมือชื่อ อ. ปรึกษาวิทยานิพนธ์หลัก.....

ปีการศึกษา.....2551.....ลายมือชื่อ อ.ที่ปรึกษาร่วม.....

4976567033: MAJOR BIOMEDICINAL CHEMISTRY
KEY WORD: NON-CODING RNAs / *MYCOBACTERIUM TUBERCULOSIS* / ncRNA COMPUTATIONAL IDENTIFICATION

NATAPOL PORNPUTTAPONG : IDENTIFICATION OF NON-CODING RNAs IN *MYCOBACTERIUM TUBERCULOSIS* GENOME USING COMBINED COMPUTATIONAL APPROACH. THESIS PRINCIPAL ADVISOR: ASSOC. PROF. DUANGDEUN MEKSURIYEN, Ph.D., THESIS CO-ADVISOR: CHINAE THAMMARONGTHAM, Ph.D., 103 pp.

Nowadays it has already been known that non-coding RNAs (ncRNAs), which are not translated to proteins, play important roles in cellular processes including regulatory functions. In order to identify putative ncRNAs of *Mycobacterium tuberculosis*, genome-wide screening by using computational approach is applied. Although the efficiency of currently available programs is limited, combined approach was the method of choice. New workflow development was required. The core program, RNAz, of the workflow was integrated with TBA. By testing the workflow with *Escherichia coli* genome, it was, however, observed that TBA generated a large number of false positives by generating missing alignment. This problem is challenging. In order to solve this, new genome wide alignment protocol was developed by combining BLAST search and MAFFT multiple sequence alignment. Evaluating this with *E. coli* ncRNA prediction, it can improve sensitivity of RNAz results from 0.54 to 0.84, precision from 0.37 to 0.56 and reduce time to calculate from over 6 hours to 70 minutes. Therefore, this protocol was used, instead of TBA, in *M. tuberculosis* ncRNA gene identification, resulting 61 predicted loci. Based on *M. tuberculosis* H37Rv ncRNA annotation, 33 predicted RNA loci were located in ncRNA gene region. Other loci were mapped with promoter and terminator prediction. There were 22 loci which had transcription signal and only a locus had double transcription signal. By sequence similarity search, there were 3 loci which matched with two known RNA sequences, *ykoK* and SRP, in database. The *ykoK* element is a regulatory element of divalent cation-related genes and SRP involves in protein translocation in cell. Resulting candidate putative loci were considered as putative ncRNAs for further experimental verification.

Department:……Biochemistry…… Student's signature:……..
Field of study:…Biomedicinal Chemistry… Principal advisor's signature:……..…
Academic year:……2008……..…Co-advisor's signature:……..……..

ACKNOWLEDGEMENTS

I would like to express sincere gratitude to my thesis advisor, Associate Professer Dr. Duangdeun Meksuriyen, Department of Biochemistry, Faculty of Pharmaceutical Sciences, Chulalongkorn University for her kindness, helpful, guidance, and non-technical advice throughout my thesis-research period. I thank her also for providing me an opportunity to grow as a good pharmacist with research career path.

I am also very grateful thank to Dr. Chinae Thammarongtham, Biochemical Engineering and Pilot Plant Research Development Unit, BIOTEC, for his great efforts to explain things clearly, valuable advice and encouragement. Throughout my thesis-research period, he provides encouragement, good teaching, and lots of good ideas.

I would like to thank Assistant Professor Dr. Boonsri Ongpipattanakul and Associate Professor Dr. Thitima Pengsuparp, Department of Biochemistry, Faculty of Pharmaceutical Sciences, Chulalongkorn University and Dr. Therdsak Prammananan, BIOTEC, for their helpful and valuable time to critical review of this thesis.

Thanks are also due to TGIST and NSTDA for scholarship support to fulfill this study.

I wish to thank all staff members and my best friends in Department of Biochemistry, Faculty of Pharmaceutical Sciences, Chulalongkorn University for their assistance and great encouragement.

Lastly, and most importantly, I wish to thank my family, on whose constant encouragement and love. Their unflinching courage and conviction will always inspire me. I thank them also for proving me to grow as a good human.

CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABBREVIATIONS

ΔG_{total}	total free energy
Δg_i	helix initial free energy
Δg_{sym}	free energy for self complementary sequence
BLAST	Basic Local Alignment Search Tool
BM	BLAST and MAFFT
bp	base pair
DNA	deoxy ribonucleic acid
ERPIN	Easy RNA Profile IdentificatioN
FFT	fast Fourier transform
FN	false negative
FP	false positive
HIV	human immunodeficiency virus
IUPAC	International Union of Pure and Applied Chemistry
MAFFT	Multiple Alignment using Fast Fourier Transform
MCC	Mathew Correlation Coefficient
MDR	multi-drug-resistant
MFE	minimal free energy
mRNA	messenger ribonucleic acid
ncRNA	non-coding ribonucleic acid
ORF	open reading frame
PPV	positive prediction value
PSSM	position specific scoring matrix
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
SCFG	stochastic context-free grammar
SCI	structure conservation index
Sig	sigma factor
siRNA	small interference ribonucleic acid
Sn	sensitivity
snRNA	small nuclear ribonucleic acid
snoRNA	small nucleolar ribonucleic acid
SRP	signal recognition particle

SVM	Support Vector Machine
TB	tuberculosis
TBA	Threaded Blockset Aligner
tmRNA	transfer-messenger ribonucleic acid
TN	true negative
TP	true positive
tRNA	transfer ribonucleic acid
WHO	World Health Organization
XDR	extensive-drug-resistant

CHAPTER I

INTRODUCTION

Traditionally, the role of RNA in the cell was considered mostly in the context of protein gene expression, limiting RNA to its function as mRNA, tRNA and rRNA. In the past few years, there have been many studies in various organisms particularly in *Escherichia coli*, on discovery and characterization of new sort of regulators for gene expression, small non-coding RNAs (ncRNAs). NcRNAs are transcripts that have function without being translated to protein. They have a number of roles in the cell including important regulatory roles (Storz, 2002), many of which remain to be discovered. This type of RNA possesses multifunction. Over the past, the gene expression has known mostly regulated by protein units. However, the RNA regulation units have been creeping discovery up. This new knowledge was a jigsaw of regulation map to understand about whole gene expression regulation. They have also been relevant to a number of regulatory effects which were unexplainable in the past (Gottesman, 2004). These untranslated RNA molecules are present in many organisms, ranging from mammals to bacteria (Erdmann *et al.*, 2001), including pathogens. In eukaryotic organisms, ncRNAs are involved in many molecular interactions, including defense against viruses and regulation of gene expression during development by acting as negative regulators, inhibiting protein synthesis in animal cells or promoting degradation of ncRNA target in plants (Carrington and Ambros, 2003). Consequently, efforts to identify the whole set of ncRNAs and then to elucidate their functions for better biological understanding and for discovering drug targets are more prominent.

Mycobacterium tuberculosis is a human pathogen that is a causative agent of tuberculosis (TB), one of the most serious infectious diseases worldwide. Although effective drugs exist, current therapy requires prolong treatment with three to four drugs, leading to compliance problem and the emergence of multi-drug resistance (Rattan, Kalia and Ahmad, 1998). In particular, the increasing prevalence of multi-drug-resistant (MDR)-TB has greatly contributed to the increased difficulties in the control of TB. Because of the global health problems of TB, the increasing rate of MDR-TB and the high rate of a co-infection with HIV, the development of potent new

anti-TB drugs without cross-resistance with known anti-microbacterial agents are urgently needed. Thus, new drugs that inhibit new targets and that are difficult to overcome by mutation are required. Greater understanding of biological aspects, including gene regulation, of this bacterium would open alternative way for identifying genes that code for new drug targets, besides metabolic drug targets. It is now becoming increasingly clear that besides regulatory proteins, bacteria also possess a significant number of regulatory ncRNAs. These are a heterogeneous group of functional RNA molecules showing up in all bacterial cells, including pathogenic species. It has been reported that RNAs are essential intracellular effectors of virulence traits and are parts of many regulatory networks in pathogenic bacteria (Johansson and Cossart, 2003; Romby, Vandenesch and Wagner, 2006). For *M. tuberculosis*, regulatory circuits involving ncRNAs can be expected as well. However, little had been known about ncRNAs in mycobacteria such as rRNA and tRNA. The ncRNA gene determination in *M. tuberculosis* was initially studied to understand about RNA gene expression regulation. The part of new regulatory pathway might also hold promise for more rapidly effective drug discovery against multi-drug resistant *M. tuberculosis*. Indeed, until year 2007 only 2 ncRNAs (10Sa RNA and ribonuclease P RNA), excluding tRNA and rRNA, were known in *M. tuberculosis* (Camus *et al.*, 2002). Studies of mycobacterial ncRNAs will gain insight of controls of many cellular processes that may be involved in pathogenesis. Therefore, this study is aimed to identify putative ncRNA genes from *M. tuberculosis* genome sequence.

However, it is relatively difficult to detect novel ncRNAs by conventional genetic and biological screening (Rivas and Eddy, 2000; Hersberg, Altuvia, and Margalit, 2003). Availability of complete genome data have made it possible to alternatively detect ncRNAs in sequenced genome using computational methods. Although experimental verification is necessary, there were several studies (Argaman *et al.*, 2001; Rivas *et al.*, 2001; McCutcheon and Eddy, 2003; Axmann *et al.*, 2005) demonstrated that computational identification may be an effective approach to first detect ncRNAs candidates, including novel ncRNA species, followed by biochemical assessment. Systematic identification of bacterial ncRNAs is mainly focused in enteric bacteria. Several experimental methods have been used to identify ncRNA genes but their efficiencies are varied since they were developed particularly for ncRNA analyses on specific genomes. Computational approach is adopted to predict

candidate ncRNA gene position in genomes. The protein coding gene finding tools cannot be used, since they predict gene position by finding open reading frame, stop codon, ribosomal binding sequence but ncRNAs seldom have these features. Many functional RNAs depend on defined secondary structures. In particular, evolutionary conservation of secondary structures serves as compelling evidence for biologically relevant RNA functions (Washietl, Hofacker and Stadler, 2005). In addition, particular genome sequences have various base-composition statistics. (G+C)%, (G-C)% difference, (A-T)% difference and dinucleotide frequency statistic were used to investigate the differences of ncRNA genes from another (Schattner, 2002). This hypothesis can be used to screen ncRNA gene by comparing with genome background. Generally, transcribable genes must have promoters. This approach was used to confirm transcribable ncRNA genes. In this research, comparative genomics approach was used based on published mycobacterial genome sequences combining with a measure for thermodynamic stability, based on Minimum Free Energy (MFE) of RNA folding, with a measure of structure conservation and converted to the normalize z-score (Washietl *et al.*, 2005) using RNAz. The protocol for genome wide ncRNA identification, which is recommended in RNAz manual, is to use TBA for aligning genomic sequences of closed species and then identifying ncRNA by RNAz. Although, the results, are generated from recommend protocol for RNAz, have many false positive. In this research, the new protocol is developed for improving quality of results, mainly reduced false positive by using combination of various methods. The *E. coli* genome was used as a model for evaluating new protocol. Finally, groups of high confident ncRNA genes were reported for further verification by molecular experimental techniques.

Rationale of the study

Tuberculosis is a pandemic and its causative agent - *M. tuberculosis* – is one of the most prolific infectious agents affecting humans. The World Health Organization (WHO) estimates that the largest number of new TB cases in 2005 occurred in the South-East Asia Region, which accounted for 34% of incident cases globally. It is estimated that 1.6 million deaths resulted from TB in 2005. Both the highest number of deaths and the highest mortality per capita are in the Africa region.

The TB epidemic in Africa grew rapidly during the 1990s, but this growth has been slowing each year and incidence rates now appear to have stabilized or begun to fall. The WHO/IUATLD Global Project estimated the magnitude of the multi-drug resistance problem worldwide with mathematical model in 2000. This model suggested an estimated 3.2% (273,000, 95% confidence intervals: 185,000 and 414,000) of all new estimated TB cases were multi-drug resistant tuberculosis (Dye *et al.*, 2002). That is the problem to control the case number of TB patient. For this reason, the novel knowledge about *M. tuberculosis* is urgently required. Relatively, tuberculosis is rather difficult to cure because there are a few antibiotics that are effective to this bacterium.

Recently many ncRNAs which regulated wide range of cell processes were found not only in eukaryotic organisms but also in prokaryotic cells. This is a very interesting and may open alternative way for drug development. In *M. tuberculosis*, ncRNAs were not completely identified. It has been reported that RNAs are essential intracellular effectors of virulence traits and are parts of many regulatory networks in pathogenic bacteria (Johansson and Cossart, 2003; Romby *et al.*, 2006). In addition it was reported that they can be applied in drug development, based on siRNA, for microbial infectious diseases. In this study, a computational tool, RNAz, was used to identify the ncRNAs genes by MFE calculation that is ideally suited for RNA secondary structure and sequence analysis which have been adopted from RNA folding energy calculation. This energy score would be converted to z-score using Support Vector Machine (SVM) regression. The whole *M. tuberculosis* H37Rv genome sequence was screened for structural RNAs by using RNAz. Among the candidate RNA genes, rRNA and tRNA were excluded. A number of ncRNA genes were found including known ncRNA genes identified previously. The novel ncRNAs were also gained. However, false positives were detected by this *in silico* prediction. Consequently, additional computational approaches were performed to reduce false positives and systematically identified promising ncRNAs.

Conceptual framework

There have been several computational identifications reported in some certain organisms. Although some programs for ncRNAs identification are already available,

some other characteristics of ncRNAs can be included in order to obtained integrated platform facilitating additional steps for identification. In the part of workflow development, three approaches of ncRNA identification are combined. The first approach is evolution of gene in closely related species that is defined based on similarity region of DNA sequences called conserved region. Conserved regions of closely related species are identified by genome comparison and generating multiple sequence alignments. The second approach is finding MFE of structure folding. Low MFE value infers ability of predicted RNA sequences, from the first approach, to fold forming RNA secondary structure. In this calculation, alignments only in intergenic regions are used to predict secondary structures and calculate MFE values. The third one is transcription feature approach. The transcription features are basic features of individual genes. The transcription units, promoter and terminator, will be identified using motif search tools with known patterns of sigma factor and rho-independent terminator. After finishing all identification, all results are combined to identify highly probable positions of ncRNA genes with some statistical approaches. The workflow is evaluated in the model organism, *E. coli*. After all, some parameters of workflow are adjusted in order to make it suitable for *M. tuberculosis* genome.

In the part of ncRNA gene identification in *M. tuberculosis* (right diagram in Figure 1), 5 genome sequences of mycobacteria are aligned, using Threaded Blockset Aligner (TBA), to find conserved regions and compared with GenBank annotation data in order to select non-protein coding sequences. Since genes in closely related species tend to be similar, sequences in conserved non protein-coding regions among 5 Mycobacterial genomes are likely to contain possible ncRNA genes. Therefore RNAz will then be used to identify ncRNAs, from multiple sequence alignment, by using secondary structure folding energy to calculate probability scores. The secondary structure folding energy is a thermodynamic property of RNA. Different sequences are able to fold into particular secondary structures with different levels of energy. Secondary structure folding makes RNA structures, which mainly contain hairpin and loop structures, become more stable with lower energy. However, only thermodynamic properties of sequences are not sufficient for detection of non-coding RNAs (Rivas and Eddy, 2000). Whole genome sequence of *M. tuberculosis* will be scanned to find promoter and terminator positions by using motif search tools such as sMotif, pftools and ERPIN.

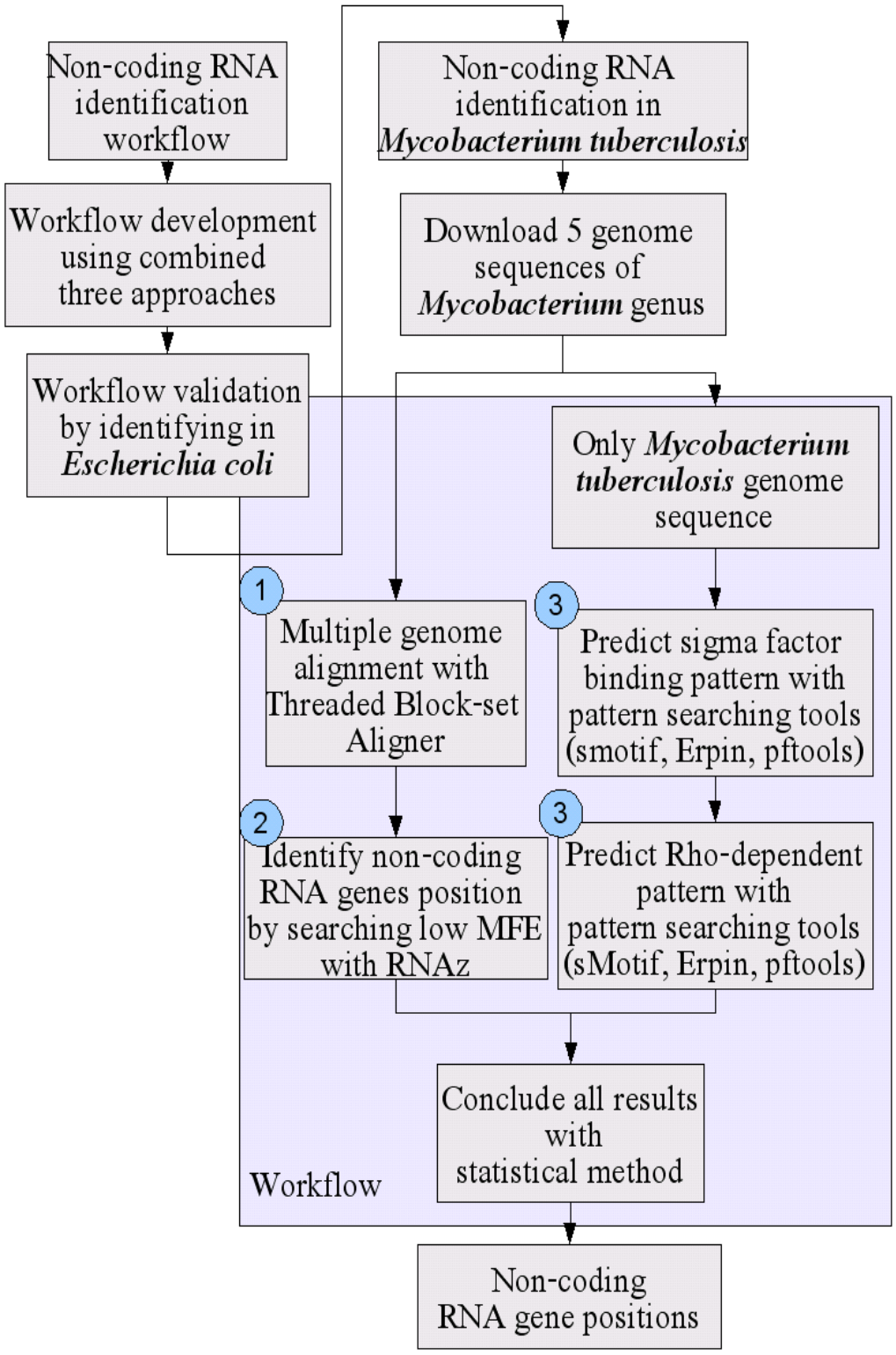

Figure 1 Hypothetical workflow of ncRNA identification using combined computational approaches.

Binding patterns, using for motif search tools, are comprised of 5 sigma factor of *M. tuberculosis*, sigma factor 70 of *E. coli*, sigma factor of *Bacillus subtilis* and rho-independent terminator. In the results obtained from RNAz and motif search tools, the position of ncRNA genes will be approximately located in *M. tuberculosis* genome. Predicted ncRNA genes will be compared with RNA gene annotation from GenBank and ncRNA databases to exclude known house keeping RNAs and other known ncRNAs in databases. In the final result of this work, novel ncRNA genes will hopefully be acquired which would be ready for further experimental verification and characterization.

Objectives

1. To develop ncRNA identification workflow using combined computational approaches.
2. To identify putative ncRNA gene positions on *M. tuberculosis* genome using computational method.

Impact of the study

The result would gain a list of putative ncRNAs from genome-wide screening which will facilitate experimental verification and further characterization. The information obtained would be beneficial to biology, gene regulation and pathogenesis of *M. tuberculosis*.

CHAPTER II

LITELATURE REVIEWS

Non-coding RNAs

The discovery of a diverse array of transcripts that are not translated to proteins but rather possessed other functions has changed the view of RNAs function only on protein expression. This is a typical function of RNAs. Besides mRNAs which are coding RNAs, certain types of ncRNAs have long been known namely rRNA, tRNA and RNase P. Expression of ncRNAs are found in many different organisms, spreading from bacteria to human. It was suggested that RNAs may be widely used as a signaling molecule in plants according to the identification and expression study of plant ncRNAs and the observation of RNA transport over long distant by phloem sieve tubes (Lucas, Yoo and Kragler, 2001). In human, several ncRNAs were found to act as gene expression regulators by stabilizing or degrading mRNAs using gene silencing pathway.

NcRNAs are transcripts that have several functions in many cellular processes without being translated to protein. However, most of ncRNAs have some common properties. Typically most ncRNAs exhibit size between approximately 20 and 500 nucleotides (Huttenhofer *et al.*, 2002). The bacterial ncRNAs usually found to act as translation regulators in size between one and two hundred nucleotides (Wassarman *et al.*, 1999). One of key features of ncRNAs is that they are able to form stem-loop structure, based on palindrome base composition in their sequences (Rivas and Eddy, 2001). This characteristic makes ncRNAs different from protein-coding RNAs and can be applied in RNA secondary structure determination which is a major part of ncRNA identification tools.

Generally, ncRNAs can be classified into two groups (Morey and Avner, 2004): housekeeping RNAs and regulatory RNAs. Housekeeping ncRNAs are constitutively expressed and essential for cell viability. Some of housekeeping ncRNAs are shown in Table 1.

Table 1. Functional classification of housekeeping ncRNAs (Toledo-Arana *et al.*, 2007)

rRNA	Translation of genetic information
tRNA	Translation of genetic information
snRNA	Pre-mRNA splicing; spliceosome components
snoRNA	RNA modifications,
	20-O-methylation and pseudouridylation
tmRNA	trans-translation
Telomerase RNA	Telomeric DNA synthesis
Ribonuclease RNA	RNA processing

Abbreviations: snRNA: small nuclear RNA, snoRNA: small nucleolar RNA, tmRNA, transfer-messenger RNA.

On the other hands, regulatory ncRNAs are expressed at certain stages along the cell life cycle. They can act as a response to external stimuli and can then affect level of gene expression by mainly controlling stability and degradation of mRNAs. Recent research has been revealed these functions of ncRNAs, which are essential for viability and adaptability of pathogens. In bacteria, several regulatory ncRNA were complementary binding with mRNA and provides a very specific and efficient regulation of protein expression in translation level. The effect can be either activation or inhibition depending on site and specificity of binding. One of well studied RNAs was *DsrA* that response for regulating translation of stress response sigma factor (*rpoS*). Binding of *DsrA* RNA to 5'-untranslated region of *rpoS* mRNA competes with their own mRNAs (Repoila, Majdalani and Gottesman, 2003).

Recent studies have been revealed that ncRNAs are key components in regulatory cascades of pathogenic bacteria as such controlling bacterial virulence and are important factors that response to environment changes. ncRNAs can adapt expression level of virulent genes against environment stresses and uncomfortable conditions (Toledo-Arana *et al.*, 2007). However, studies of ncRNAs controlling pathogenesis are at the beginning. Numerous of ncRNAs were discovered in many pathogenic bacteria. Several questions are waiting for answers which answers could open the new ways to generating drugs against bacteria signaling or metabolic pathway during infection.

Mycobacterium tuberculosis

M. tuberculosis; a rod-shaped, Gram-positive bacterium; is a very important pathogen, which causes TB in human. Genome sequence of *M. tuberculosis* H37Rv has been completed in 1998 (Cole *et al.*, 1998). It contains 4.4 Mbp s and 4,048 genes. Among theses genes, they code for 3,989 protein-coding genes and 50 ncRNAs. The main group of annotated ncRNAs is housekeeping ncRNAs. The rest, according to the *M. tuberculosis* H37Rv genome project, are rRNAs, RNase P and tmRNA. Therefore, very little has been known for other ncRNA species which have regulatory related functions. Generally, ncRNAs are located in intergenic regions which are mainly not overlapped with protein-coding sequences. The content of intergenic region is about 9.1%. Genome sequences of other species, including *M. avium*, *M. bovis*, *M. leprae*, have already been sequenced. *M. tuberculosis* was very hazardous air borne

pathogenesis bacteria for human and very slow growth in culture media. Therefore, any experiments of *M. tuberculosis* were complicated and must be investigated by specialist.

Identification of ncRNAs

Unlike regulatory protein identification, it is relatively difficult to detect novel ncRNAs by conventional genetic and biochemical screening (Rivas and Eddy, 2002; Hershberg, Altuvia and Margalit, 2003). The difficulty of ncRNAs detection by biochemical and genetic methods due to their small sizes which are complicates for mutagenesis. Inactivation of their functions is relatively difficult to be accomplished by single nucleotide changes, frame-shift or nonsense mutations. Therefore, nowadays, no exact methods for identifying only ncRNA were accomplished. The most identification of ncRNA was done by using general method for analyzed RNA and compared with ncRNA information to specify them (Vogel and Sharma, 2005). During the last decade, most of ncRNAs were discovered by chance. In 2000s, only 10 ncRNAs were recognized in *E. coli*, most of which were fortuitously discovered using polyacrylamide gel electrophoresis to analyze [^{32}P]-labeled total RNA and comparing with known mRNA sequences or hybridized with known ncRNA (Wassarman *et al.,* 1999). Lately, more than 50 novel ncRNAs were identified by systemic screening based on computational and microarray approaches, most of which had been overlooked by traditional biochemical methods.

Previous computational studies on ncRNAs

There are several approaches using computational methods to identify ncRNAs. They will cover on searching for ncRNAs by finding characteristics of ncRNAs. Most of them have been focused on ncRNAs mainly in *E. coli* and some other species of bacteria.

1. Transcription unit search

Positions of ncRNA genes were identified by searching promoters and Rho-independent positions only on intergenic regions of *E. coli* genome. The candidate regions between promoter and terminator were then collected and aligned with database to check conservation within closely-related species. They found 24

candidates and 14 of them were able to be detected by Northern blotting (Argaman *et al.*, 2001).

2. Structure prediction with pair stochastic context-free grammar

The second study; intergenic regions (IGRs) of *E. coli* were analyzed by statistical method called pair stochastic context-free grammar (SCFG) which detected base pairing probability to form secondary structures. Statistical model were then used to distinguish between ncRNAs and protein coding regions. They found 49 candidates from computational identification and 11 of them were detected by Northern blot analysis (Rivas and Eddy, 2001).

3. Sequence and structure conservation

According to a publication, ncRNA genes were identified from cyanobacteria especially from *Prochlorococcus* and *Synechococcus*. The computational method was based on sequence and structure conservation within closely-related species. The 7 candidates of ncRNAs were identified from this method and already verified by experimental method (Axmann *et al,* 2005).

Approaches for ncRNA identification

Recently, several approaches were established for located ncRNA genes in genome, but none of them was the most efficient approaches for using individually (Eddy, 2002). From previous computational study, some of the most successful methods were based on similarity sequence search between genomes of closely-related species. Other approaches were transcription unit searching or new efficient method for secondary structure stability prediction of ncRNAs.

1. Identification of ncRNA with comparative sequence analysis and secondary structure stability prediction

Based on evolution hypothesis, similarity of sequences between closely-related species infers functional motifs. This approach is a general method for identifying some functional motifs, such as protein coding genes, promoters and protein binding motifs; therefore, this approach is not specific for finding ncRNA genes. For example, in beginning, similarity search was used to identify conserved regions, which was likely to be ncRNAs, in *E. coli*. The 259 highest conservation groups from similarity search were comprised of conservation of regulatory region, translation leader and upstream of ORF (Wassarman *et al.*, 2001). From previous study, the similarity search was usually combined with other approach to improve

specificity. One of efficient combination was combined with secondary structure stability prediction (Washietl *et al.*, 2005).

2. Identification of ncRNA with base composition bias

Several reviewed approaches are mostly based on homology of sequences. This method is very simple and more efficient, but it is not suitable for identifying novel ncRNAs. The bias of base composition is very well known for searching the special motifs from raw genome sequence, such as promoter, origin of replication, transposons, genomic islands and protein coding gene. In particular, it is well known that the thermophile bacteria maintain stability of their ncRNAs by changing base content. Based on this idea, Rivas and Eddy proposed ncRNA genes finding based on CG content in thermophile bacteria. They also suggested using this approach with non-thermophile genomes (Rivas and Eddy, 2000). Afterwards, other base composition, such as G-C% Chargaff difference, A-T% Chargaff difference, AT content and di-nucleotide frequencies are possible to use in ncRNA genes identification. On the other hand, RNAs are deviation from Chargaff second law. For example, the non-zero of Chargaff difference is determined as protein coding messenger RNAs. Moreover, G/U base-pair is commonly found in general RNAs structure more than C/A base-pair; this exception might be lead to break Chargaff second law. From investigation in three test genomes, *Methanococcus jannasdchii*, *Plasmodium falciparum* and *Caenorhabditis elegans*, only GC content and CG di-nucleotide frequencies are used for identification ncRNA gene candidates (Schattner, 2002).

3. Prediction of ncRNA gene by searching transcribing unit

In generally, like mRNAs, ncRNAs must be transcribed from genes. There are many properties that share with the others genes in the processes of transcription. Some basic methods for annotating protein coding gene can be used for locating ncRNA gene in genome by searching for co-localize of several genetic features commonly associated with other encoding gene, including promoter, terminator and conservation of sequence in intergenic regions (Vogel and Sharma, 2005). The described methods were not specific for RNA genes identification but they can be used for utilized with RNA structural energy calculation for improving accuracy.

Computational tools

1. BLASTZ

BLASTZ is a whole genome alignment specified for finding orthologous regions among genomes. Aim of this program is used for eukaryotic genome but it is possible to apply in prokaryotic genomes. However, it cannot align highly dynamical region, *e.g.* tRNA genes, ncRNA genes. BLASTZ follows the three steps of gapped-BLAST, that is, find the short exact matches, extend the short matches without filling gaps and extend gap-free matches that exceed a certain threshold by using dynamic programming with filling gap. The difference between BLASTZ and gapped BLAST was the matching regions that must be reported in the same order and orientation in both sequences (Schwartz *et al.*, 2003).

2. TBA

The "threaded blockset" is a novel generalization of the classic notion of a multiple alignment. TBA program builds a threaded blockset under the assumption that all matching segments occur in the same order and orientation in the given sequences. This program is designed to be appropriate for aligning many megabase-sized regions mainly in multiple mammalian genomes. The output of TBA can be projected onto any genome chosen as a reference, thus guaranteeing that different projections present consistent predictions of which genomic positions are orthologous. Main algorithm in alignment process is performed by dynamic programming. BLASTZ is used in first step of TBA in pair-wise alignment. Sequentially, results from BLASTZ are processed with MULTIZ program to do multiple sequence alignments (Blanchette *et al.,* 2004).

3. MAFFT

The Multiple Alignment using Fast Fourier Transform (MAFFT) sequence aligner was originally developed to perform the rapid calculation of a multiple sequence alignment of the large number of sequences. A fast group-to-group alignment algorithm based on fast fourier transform (FFT) and an approximate distance calculation method (the 6mer method) facilitate the rapid calculation (Katoh and Toh, 2008). There were several methods of alignment available for specific alignment, such as L-INS-i and Q-INS-i. The L-INS-i method was specific to local alignment. and was included in the alignment protocol. In this study the MAFFT is used for doing multiple sequence alignmnet in new alignment protocol.

4. RNAz

There are two approaches using in RNAz, structural conservation and thermodynamic stability. These two parameters are calculated from multiple sequence alignment and trained in intelligence system called Support Vector Machine (SVM).

4.1. Thermodynamic stability

The MFE as a measure of thermodynamic stability for a sequence is calculated by using RNAfold (Hofacker and Stadler, 2006). However, the MFE depends on length and base composition of a sequence. Therefore, only the MFE energy score is difficult to interpret in absolute terms and hard to compare with another sequence. The RNAz calculates a normalized measure of thermodynamic stability by comparing the MFE of a given (native) sequence to the MFEs of a large number of random sequences of the same length and base composition. A standard score is calculated as $z = (m-\mu)/\sigma$, where μ and σ are the mean and standard deviations of the MFEs of the random sequences, respectively. Negative z-scores indicate that a given sequence is more stable than expected by chance in random sequences. RNAz does not actually sample random sequences to calculate z-scores but use some approximation, which is much faster but of the same accuracy.

4.2. Structural conservation

RNAz predicts a consensus secondary structure for an alignment by using the RNAalifold (Hofacker, 2007) approach. RNAalifold calculation is almost exactly as single sequence folding algorithms, with the main difference that the energy model is augmented by covariance information from alignment. In RNA sequence pairing, some mismatches called non-Watson and Crick pairing, such as consistent mutations (*e.g.* AU mutates to GU) and inconsistent mutations (*e.g.* CG mutates to CA), have some bonus scoring energy while energy scores are calculated in the term MFE. The MFE are calculated from individual sequences and from consensus of alignment. RNAz compares this consensus MFE to the average MFE of the individual sequences in alignments and calculates structure conservation. The SCI score shows the consensus of folding of individual sequence and consensus sequence. The SCI score will be high, if the consensus sequences fold together equally well as folded individually. On the other hand, SCI will be low, if no consensus fold can be found.

CHAPTER III

MATERIALS AND METHODS

Data sources

1. Genome sequences and annotation files

Based on the selected methods described below and using comparative genomic approach, genomes of closely related organisms were used for ncRNA gene identification. The genomic data files of 13 organisms, as shown in Table 2, were retrieved from NCBI Genbank genome database. For each genome, the files for whole genome sequence (fna file extension), annotation table (ptt file extension), nucleotide sequences of protein coding genes (ffn file extension), amino acid sequences of protein coding gene (faa file extension), RNA annotation table (rnt file extension) and RNA sequences (frn file extention) and Genbank annotation table were downloaded by file transfer protocol (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria).

2. Selecting genomes of closely related organisms

Initial alignment is an important step that has an effect on the results of ncRNA identification. For RNAz program, genome sequences of closely related species, in term of evolution, were recommended for generating multiple sequence alignment. In order to predict ncRNAs of *M. tuberculosis* H37Rv, genome sequences of *M. bovis, M. avium, M. avium* sub. *paratuberculosis*, *Mycobacterium* sp. MCS and *M. leprae* were chosen, based on phylogenetic analysis (Devulder *et al.,* 2005), for alignment with *M. tuberculosis* H37Rv. In workflow step, ncRNAs of *E. coli* were used as control. For *E. coli*, a model organism using in workflow evaluation, genome sequences of *Salmonella typhi, Salmonella typhymurium, Xanthomonas campestris, Xanthomonas citri and Yersinia pestis* were selected, according to phylogenetic trees of metabolic networks analysis (Oh *et al.,* 2006), to align with that of *E. coli*.

Table 2 Bacterial genomes for protocol development and ncRNA searching. Bacterial genomes in γ-proteobacteria were used in protocol development based on *E. coli* genome. Genome sequences of genus *Mycobacterium* were used for identification of ncRNA of *M. tuberculosis* H37Rv genome.

No	Species	Accession number (RefSeq.)	Group
1	*Escherichia coli* K12	NC_000913	Bacteria/γ-proteobacteria
2	*Salmonella typhi* str. CT18	NC_003198	Bacteria/γ-proteobacteria
3	*Salmonella typhimurium* LT2	NC_003197	Bacteria/γ-proteobacteria
4	*Xanthomonas citri* str. 306	NC_003919	Bacteria/γ-proteobacteria
5	*Xanthomonas campestris* str. 8004	NC_003902	Bacteria/γ-proteobacteria
6	*Yersinia pestis* CO92	NC_003143	Bacteria/γ-proteobacteria
7	*Mycobacterium tuberculosis* H37Rv	NC_000962	Bacteria/Actinobacteria
8	*Mycobacterium bovis*	NC_002945	Bacteria/Actinobacteria
9	*Mycobacterium leprae*	NC_002677	Bacteria/Actinobacteria
10	*Mycobacterium avium* sub sp. paratuberculosis	NC_002944	Bacteria/Actinobacteria
11	*Mycobacterium avium*	NC_008595	Bacteria/Actinobacteria
12	*Mycobacterium sp.* MCS	NC_008146	Bacteria/Actinobacteria
13	*Bacillus subtilis*	NC_000964	Bacteria/Firmicutes

3. Databases of ncRNAs

The sequences of known ncRNAs and ncRNA families were retrieved from three databases namely Rfam version 7 (Griffiths-Jones *et al.,* 2005), ncRNAdb (http://biobases.ibch.poznan.pl/ncR)NA/ (Szymanski *et al.,* 2007), and NONECODE (http://noncode.bioinfo.org.cn/) (Lui *et al.,* 2005). These databases were used in unknown ncRNA gene annotation. The data files are in fasta format.

Recommended protocol with RNAz

Since RNAz requires multiple sequence alignment as inputs, genome-wide alignment is a prior step before RNAz. While several genome scale alignment programs are available, MULTIZ using TBA protocol is recommended by RNAz developer. Therefore genome comparisons generating alignments by TBA were performed.

1. Genome-wide alignments

In order to obtain conserved regions among all genomes analyzed in this work, the regions that share sequence similarities were investigated. Genome-wide alignments of target genome and closely related species were started with sequence conservation search by local pair-wise alignment using BLASTZ (Schwartz *et al.*, 2003). The adjacent regions with distance of 20 base-pairs of less and also overlapping regions were joined together into single regions. Resulting pair-wise alignment were used to build multiple alignment by TBA program in MULTIZ package (Blanchette *et al.*, 2004) using default parameters.

In workflow evaluation step, genome sequences of *S. typhi, S. typhymurium, X. campestris, X.citri and Y. pestis* were aligned with that of *E. coli*. For *M. tuberculosis* ncRNA prediction, genome sequences of *M. bovis, M. avium, M. avium* sub. *paratuberculosis*, *M.* sp. MCS and *M. leprae* were aligned with that of *M. tuberculosis* H37Rv.

2. Pre-processing of alignments

One limitation of RNAz program is that it can not score alignments of 400 nucleotides in length or more. Practically, it is recommended to shorten the long alignments to approximately 200 nucleotides and in overlapping window pattern. Generally genome-wide scan for ncRNAs, sequences of 120-nucleotide window size

are recommended. Therefore, it is necessary to score long alignments in overlapping window manner. The alignments with poor quality are usually generated together with the reliable local alignments of genome wide comparison. The poor quality ones may involve in false positive result of ncRNA prediction. To reduce false positives, an additional pre-processing step before ncRNA prediction is required to eliminate gap rich regions and low complexity regions. The alignments from previous process are needed to process with rnazWindows.pl script included with RNAz package. By default, the following steps are performed:

2.1. Sliding windows of 120-nucleotide sequence were considered. The windows were slid in overlapping manner every 40 nucleotides.

2.2. Compare each pair-wise alignment with the first sequence as reference to all other sequences, removing common gaps and discard alignments with gap content higher than 25%.

2.3. Discard sequences which were shorter than 50 nucleotides and have CG base content higher than 75%.

2.4. Remove all sequence with 100% identical and left only pair-wise alignment.

3. RNAz scanning

The input alignments were screened for structural RNA genes using RNAz (version 1.0 local package). Alignments were screened in length varying between 50 to 1000 base pairs. This range is plausible length of ncRNA sizes from experimental data. Selected alignments were cut into windows of length 120 and slid with length 40. This windows size is small enough to detect local secondary structure of RNA within long ncRNAs. On In addition, this windows size is small enough to detect short ncRNAs without loosing signal in a much too big window (Washietl, Hofacker and Stadler, 2005). The resulting alignments were scored with RNAz using standard parameters. All alignments with classification score P-value > 0.5 was stored. Finally, overlapping hits (resulting from hits in overlapping windows and/or hits in both the forward and reverse strands) were combined into clusters. The corresponding regions in *M. tuberculosis* sequences were annotated as 'ncRNA' with maximum P-value of the single window hit in the cluster. Result from RNAz was a raw text file of calculation reports for each window as in Figure 2.

```
############################  RNAz 1.0  ##############################
 Sequences: 3
 Columns: 65
 Reading direction: forward
 Mean pairwise identity:  76.34
 Mean single sequence MFE: -10.27
 Consensus MFE: -10.68
 Energy contribution: -10.13
 Covariance contribution:  -0.55
 Combinations/Pair:   1.33
 Mean z-score:  -1.10
 Structure conservation index:   1.04
 SVM decision value:   4.62
 SVM RNA-class probability: 0.999930
 Prediction: RNA
########################################################################
>xca 1387699 65 + 5148708
ACUUUUAAUCUUUUGGUCGAUGGUUCGAAUCCAUCACGGCCCACCAUUCAAUUCAGUCAGCAAAA
..............(((((((((.......)))))..))))..................... (  -9.80)
>ecoli 780829 54 + 4639675
ACUUUUAAUCAAUUGGUCGCAGGUUCGAAUCCUGCACGACCCACCA---AU-------GUAAAA
..............(((((((((.......)))))..))))....----..-------..... ( -12.40)
>consensus
ACUUUUAAUCAAUUGGUCGCAGGUUCGAAUCCAGCACGACCCACCAU___AU_______GAAAA
..............(((((((((.......)))))..))))..................... (-10.68  =
-10.13 + -0.55)
```

Figure 2 A raw output file from RNAz.

In RNAz outputs, there are several values but only SVM RNA-class probability or P-value was meaningful for determining the results. In addition, other interested values in RNAz output are MFE, direction of windows and structural alignment of predicted ncRNAs.

4. Clustering result from RNAz

The output from RNAz holds all windows that have positive signal at P-value > 0.5. There is possibility that windows cover on the same position on genome. All overlapping windows are combined to loci. However, it is very important to note that the term "locus" is not related with the sense of genetic unit. Clustering process is performed with my own script. The script will observe position of windows based on reference genome. Windows lied in overlap region or not more than 20 base pairs were combined in the same locus. The score of locus were using the maximum score of windows that combined in the same locus excepted the direction of locus was using majority from direction that reported in windows.

5. Compare with protein coding genes

Results from RNAz with protocol for genomic search of ncRNAs were included protein coding genes. There were high possibility false positive loci. The RNAz hits, that were located in protein coding regions, were excluded from results by comparing positions of loci with annotated protein coding gene positions. The positions of protein coding genes were collected from Genbank annotation table. In addition, they will be classified into two groups; the putative ncRNA genes located in intergenic regions and the anti-sense putative RNA genes. Windows that were located in protein coding regions were excluded by comparing with position in annotation table. The excluding protein region processes were done by a Ruby language script. After this process, the results from RNAz were the loci that located only in intergenic regions.

BLAST and MAFFT alignment and alignment preparation for RNAz

Genome wide alignment protocol was proposed to improve ncRNA annotation protocol with RNAz. BLAST and MAFFT were combined in alignment protocol. The first program, BLAST, the fastest local pair search program, was used to find conserved regions among reference genome and others. Results were pair-wise alignment of conserved regions. After that, pair-wise alignments were joined together. MAFFT was used to create high accuracy of multiple sequence alignments (by using program that suitable for aligning RNA). High accuracy of alignment protocol can reduce false positive and increase evolution meaning of alignment sequences before identifying with RNAz. This protocol was developed and tested in *E. coli* genome with other genome in γ-proteobacteria and used instead of BLASTZ and TBA in MULTIZ package in suggested protocol. This protocol of alignment was called BM protocol which was illustrated in Figure 3.

1. Initial pair-wise alignments

In order to obtain conserved regions among all genome analyzed in this alignment, regions that share sequence similarities were investigated. Sequence conservation search were started with local pair-wise search of related genomes using only intergenic region sequences of reference genome. Genomic sequence of reference genome was split into separated intergenic regions sequence in fasta file format by using Ruby script based on position of protein coding gene in annotation table. The intergenic sequences are extended for 100 bases into protein coding regions both of head and tail of intergenic regions. Other genomic sequences, called target genome, were combined into a fasta format file and converted to BLAST database with formatdb program which included in BLAST package.

Local pair-wise alignments of potentially homologous region were determined by BLAST comparison against all genome in database. BLAST comparisons were conducted using BLASTN 2.2.17. E-value cutoff was set to e-10. All other search parameters were set to default values.

2. Combining pair-wise alignment

After local search, all alignments separated by only short distances lower than 20 bases and overlapping in reference genome were combined together with a Ruby script.

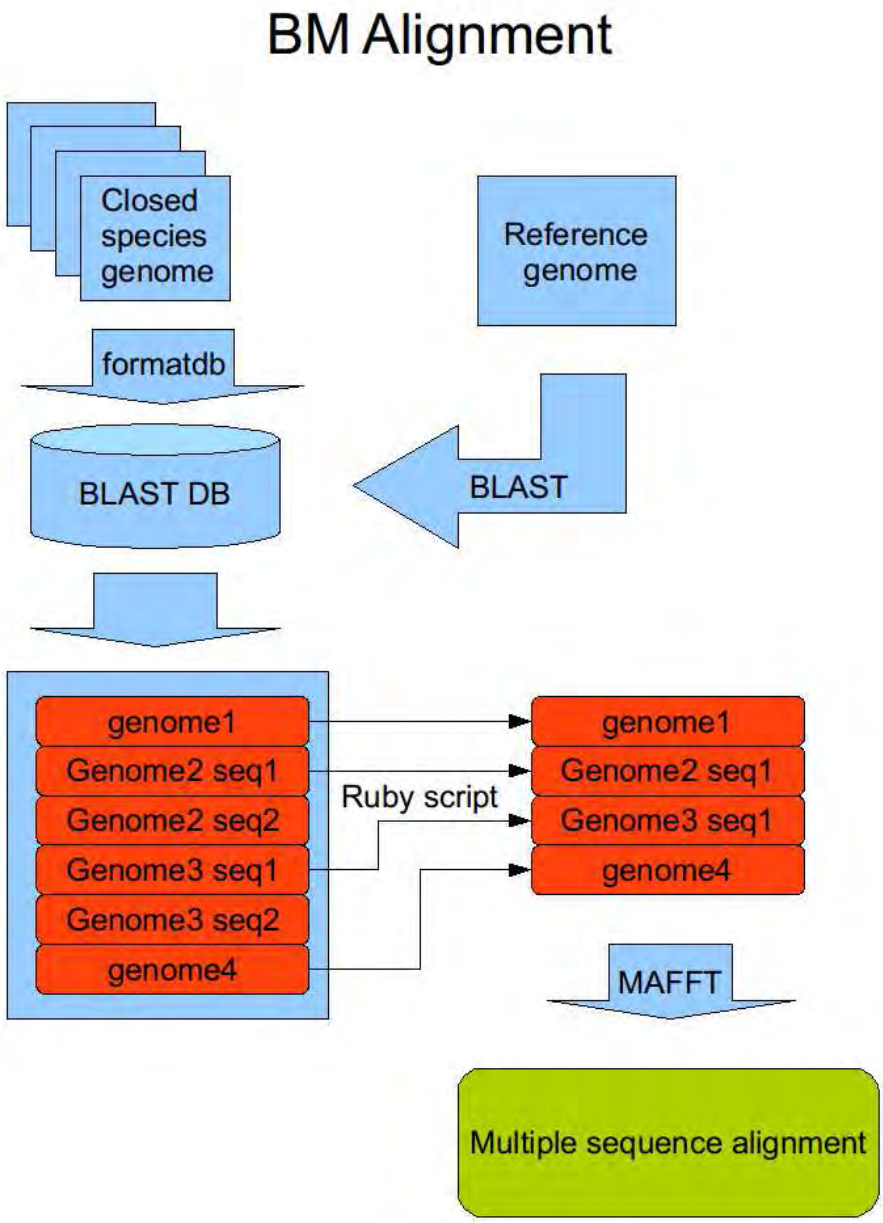

Figure 3 The BM alignment protocol start with pair-wise alignment with BLAST and then multiple align with MAFFT.

The multiple local alignments of the resulting regions were then computed by using program called MAFFT. MAFFT multiple sequence aligner was developed to perform the rapid and accurate calculation of the large number of sequences based on fast Fourier transform and iterative refinement method. All parameters of multiple alignments were set to default parameters. Alignments from MAFFT were default in CLUSTAL format, which less supported format as input for RNAz. It must be changed to maf format before identifying with RNAz by using Ruby scripts.

Promoter and rho-independent terminator finding

As ncRNA is coded by DNA sequence on its particular gene, appearance of promoter and terminator is one of sequence characteristics for ncRNA indentification computationally. Promoter prediction will be performed by the motif scanning program, sMotif (Zhang *et al.,* 2006) and pftools (http://www.isrec.isb-sib.ch/profile/profile.html) *E. coli* sigma factor 70 promoter Position Specific Scoring Matrix (PSSM) (Mulligan, *et al,* 1984) was used by Pftools to identify promoter position in *E. coli.* For another bacterial genome, in case of *M. tuberculosis* H37Rv, the specific sigma factor in prokaryote, from database of transcriptional regulation in *B. subtilis* (DBTBS) (Makita *et al.*, 2004) to located promoter. The PSSM tables from sigma factor of *B. subtilis* are shown in appendix A and pattern of sigma factor is shown in Table 3. These sigma factors were derived from PSSM (Makita *et al.*, 2004) in IUPAC code. In addition, some study that reported the specific sigma factor pattern in *M. tuberculosis* H37Rv is shown in Table 4. All sigma factor patterns were two specific motifs and combined with nonspecific sequence around 16 to 20 base pairs. The patterns were used to identify *M. tuberculosis* H37Rv promoter by using sMOTIF. The terminator scanning was performed by using ERPIN (Gautheret and Lambert, 2001) program with specific secondary structure profile of rho-independent factor. Training set of secondary structure profiles were obtained from the alignment of 1,201 rho-independent of *E. coli* and *B. subtilis* (Gautheret and Lambert, 2001).

Table 3 Pattern of sigma factor from *B. subtilis* (Makita *et al.*, 2004).

Sigma factor	Pattern
sigA	TTGACA(14)tgnTATAAT
sigB	AGGTTT(17)GGGTAT
sigD	TAAA(15)GCCGATAT
sigE	KMATATT(14)CATACAT
sigF	YGYWTA(...)GGMAWAMTA
sigG	GHATR(…)GGCATXHTA
sigH	AGGTATT(...)GAATT
sigK	MACM(16)CATATA
sigW	TGAAACN(16)CGTA
sigX	TGTAACN(17)CGAC

Genetic code in pattern is IUPAC alphabet code for nucleotide, H = A/C/T, K = G/T, M = A/C, R = A/G, S = G/C, W = A/T, Y = C/T and N = A/T/C/G.

Table 4 Pattern of sigma factor A, C, E, F, H and M of *M. tuberculosis*.

Sigma factor	Pattern	Reference
sigA	TTGACW(17)TATAMT	Manganelli *et al.,* 2004
sigC	SSSAAT(16-20)CGTSSS	Manganelli *et al.,* 2004
sigE	GGRMC(18)SGTTG	Manganelli *et al.,* 2004
sigF	GTTT(17)GGGTAT	Manganelli *et al.,* 2004
sigH	SGGAAC(17-22)SGTTS	Manganelli *et al.,* 2004
sigM	GGAAC(16-18)CGTCR	Agarwal 2006

Genetic code in pattern is IUPAC alphabet code for nucleotide, M = A/C, R = A/G, S = G/C, W = A/T and N = A/T/C/G.

Base composition calculation

For each sequence, the following statistics were computed:

(G+C)% = 100 (nG + nC) / (nA + nC + nG +nT)

(G-C)% Chargaff difference = 100 (nG - nC) / (nG + nC)

(A-T)% Chargaff difference = 100 (nA - nT) / (nA + nT)

$\rho(AB) = f(AB)/f(A)*f(B)$

Where nB and nAB are the number of occurrences of base 'B' or the dinucleotide AB (Schattner, 2002). Not only base composition are studied but also stability of Watson and Crick duplex structure were calculated in term of free enegy (ΔG).

$$\Delta G_{total} = -(\Delta g_i + \Delta g_{sym}) + \Sigma_x \Delta g_x$$

Where ΔG_{total} is free energy of DNA oligomer. Δg_i is a helix initial free energy which is assigned to 5 kcal for duplex containing with C and G. For the duplex formed from self –complementary sequence, in case of ncRNA prediction structure of RNA is formed duplex by its sequence, Δg_{sym} equals to 0.4 kcal. The last term, $\Sigma_x \Delta g_x$ is summation of number of duplex multiply by its free energy (Breslauer *et al.,* 1986). The free energy of each duplex is shown in Table 5.

In the first calculation, the comparison of positive group non-coding RNA genes and protein coding genes in genome was observed to confirm the difference of base-composition statistics and di-nucleotide free energy in these two functions of gene. For each group of sequence, median of each statistical value was calculated. Statistical significance of differences between these two groups was tested by using Wilcoxon Rank sum test comparison with 95% confidence level. The second calculation, small sequences of 100 bases were random extracted from whole genome sequence and calculated to compare with calculation from random sequences from RNA by using non-parametric Wilcoxon Rank sum test. The statistical values, which can exclude ncRNA gene signal from genomic background, are used in the follow step. For wide genome screening, genomic sequence of subject was cut into small fragments of DNA sequence with window size 100 base pairs, sliced 50 base pairs and calculated.

The comparison of positive group non-coding RNA genes and protein coding genes in genome was observe to confirm the difference of base-composition statistics and dinucleotide free energy in this two functions of gene.

Table 5 Standard free energy for each duplex of DNA (Breslauer *et al.,* 1986).

Interaction	**ΔG^0 (kcal/mol)**
AA/TT	1.9
AT/TA	1.5
TA/AT	0.9
CA/GT	1.9
GT/CA	1.3
CT/GA	1.6
GA/CT	1.6
CG/GC	3.6
GC/CG	3.1
GG/CC	3.1

The calculation was begun by cutting small sequences of 100 base pairs from whole genome sequence and calculated, pooled ncRNA sequences and pooled protein codings gene sequence randomly. The number of random sequences for calculation was 10,000 for genomic sequence and 500 for protein and ncRNA sequence. The small fragments of DNA sequence were used to compute base composition and duplex free energy. Wilcoxon Rank sum test at 99% confidential. The statistical values, which can exclude ncRNA gene signal from genomic background, are used in the follow step. For wide genome screening, genomic sequence of subject was cut into small fragments of DNA sequence with window size 100 base pairs and sliced every base and calculated score which calculated from absolute value of difference of value and median then divined by median absolute deviation. After that, plotting graphs showed the difference of ncRNA gene position with peak.

Annotation of ncRNA genes

The obtained result will be searched, using rnazBLAST script, against to available database of ncRNAs, as described in data source, in order to check if they looked like the known RNA species. BLAST database was generated with formatdb program comprising of Rfam, ncRNAdb and NONECODE. This comparing process is performed with NCBI-BLAST version 1.18 with default parameters. The results were generated and matched with the detected RNAz clustered locus of *M. tuberculosis*. Both matched and non-matched sequences will be analyzed.

Quality measurements

A model is evaluated on a positive and negative set of predicted loci; four values can be defined by counting number of bases. There are the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). This study uses three statistical values to define the efficiency of methods. There are precision or positive prediction value (PPV) (equation 1) sensitivity (Sn) (equation 2), and the Matthews correlation coefficient (MCC) (equation 3).

$$PPV = TP/(TP+FP) \quad \text{equation 1}$$

$$Sn = TP/(TP+FN) \quad \text{equation 2}$$

$$MCC = \frac{(TPxTN)-(FPxFN)}{\text{sqrt}((TP+FP)x(TP+FN)x(TN+FP)x(TN+FN))} \quad \text{equation 3}$$

CHAPTER IV

RESULTS AND DISCUSSION

There are several computational methods for identifying ncRNA genes. Although those available programs are developed for ncRNA prediction from genome sequences, but such programs is relatively not very high efficient comparing to protein coding gene prediction tools. Nevertheless several studies suggest using combination of several methods to improve accuracy of prediction (Gardner *and* Giegerich, 2004; Lindgreen, Gardner *and* Krogh, 2006). In this research, various approaches for identifying ncRNA genes were tried and used in combination protocol for searching ncRNA genes.

Workflow development for ncRNA prediction

1. TBA alignment and RNAz with *E. coli* K-12 genome

Since ncRNA prediction pipeline, starting from input genome sequences to the predicting results, is not available as standalone for local execution, workflow development must be conducted in the first step of this work. The selected tools namely BLASTZ, TBA and RNAz including some scripts, as mentioned in materials and methods, were integrated as a workflow. The developed workflow was then tested with *E. coli* genome sequence using information of *E. coli* ncRNA genes as control.

Prediction of *E. coli* ncRNA was performed to test the workflow. The *E. coli* conserved regions aligned with the other five γ-proteobacterium sequences were scanned by RNAz and 3,116 loci were predicted with P-value above 0.5. After excluded loci in protein coding regions, only 470 loci in intergenic regions were obtained. These loci were compared with position of ncRNA genes, which have been found and annotated in database, as positive set. Based on RNA annotation table, 172 ncRNA genes have been found in *E. coli* K-12 (NC_000913) genome. Annotated ncRNA positions of *E. coli* are shown in appendix B. positions of 470 predicted ncRNA loci compared with known RNA genes, along with protein coding genes are shown in Figure 4.

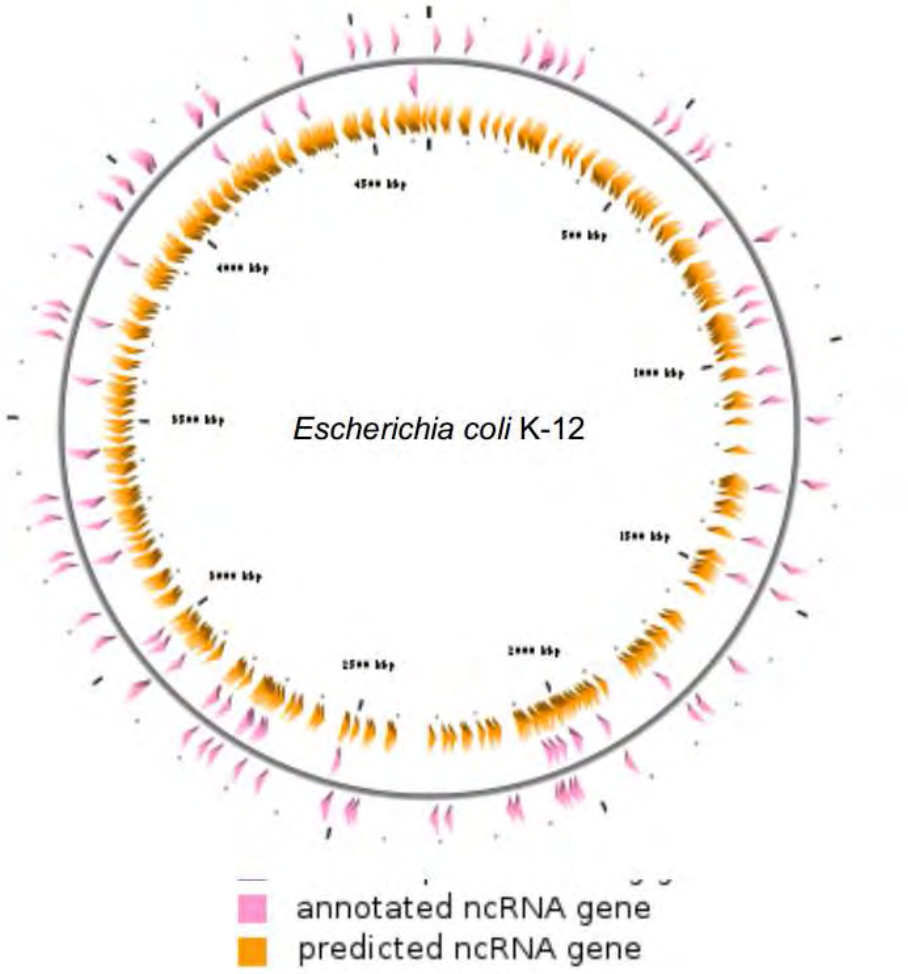

Figure 4 Positions of predicted loci of *E. coli* ncRNAs from developed workflow (using TBA method) and annotated ncRNA genes mapped on genome of *E. coli* K-12.

Results from prediction were compared with *E. coli* RNA positive set as described above. According to testing, the workflow could predict single RNase P, 14 of 22 ribosomal RNA (rRNA) genes, 59 of 88 transfer RNA genes and 29 of 54 other ncRNA genes. Table 6 showed the predicted "loci" positions which were not exact positions of genuine ncRNA genes annotated in genome sequence data. It was found that, for many authentic RNA genes, more than one predicted loci have their sequences matched with the regions of the same bona fide RNA genes (Table 7). The actual RNA genes consisting of more than one predicted loci especially are rRNA genes. This is because rRNA genes are relatively long ncRNA sequences which are typically consisted of more than one conserved stem-loop structures and linked with long non-conserved sequences between these stem-loop elements.

For measuring the efficiency gene annotation methods, the statistical values as described in chapter III were used. The sensitivity, precision and MCC of prediction using workflow consisting of TBA protocol were 0.54, 0.37 and 0.40 (table 8), respectively. The sensitivity of TBA protocol was affected by small number of TP that meant this protocol was less effective for detecting ncRNA loci. The precision was quite low that shown large number of FP in prediction results. The actual number of all ncRNA genes in genomes of any organisms, including that of *E. coli*, are not acknowledged so far. Presumably, the loci predicted as false positives can be novel ncRNA containing regions, if their actual transcripts are detected experimentally. It will, however, be laborious jobs and not practical for further experimental verification if very large numbers of FP are obtained. For this reason, alternative methods should be considered for reducing FP or prioritizing higher ranks of plausible regions for validating by laboratory techniques. Standard quality measurement of annotation protocol generally is the make used of MCC value to present the efficiency of gene allocated programs. The MCC value represents the correlation of the predicting results and the genuine items. The high value of MCC gene prediction tools means the protocols can predict positions of genes in agreement with actual positions in the genome. Generally MCC for effective protocols should be from 0.6 to 0.9 (Rogic, Mackworth *and* Ouellette, 2001). For TBA protocol, MCC was 0.40. It was rather low. This alignment protocol should be rectified for improving efficiency of ncRNA gene prediction.

Table 6 RNAz results from two alignment protocols compare with known ncRNA genes in *E. coli* K-12.

Type of RNAs	Known ncRNAs	BM	TBA
rRNA	22	22	14
tRNA	89	75	59
RNase P	1	1	1
Other ncRNA	60	31	29
total	172	129	113

Table 7 Predicted loci that matched with known *E. coli* K-12 RNA genes. The positions on *E. coli* K-12 genome of known RNA genes are shown.

Gene	Position	Predicted locus/loci
sgrS	77367..77593	62
tff	189712..189847	123
rrsH	223771..225312	153, 154, 155, 156
alaV	225500..225575	157
aspU	228928..229004	158
thrW	262095..262170	185
ffs	475672..475785	303
argU	563946..564022	368
metU	695887..695963	446
leuW	696186..696270	447
metT	696280..696356	447
valT	779988..780063	511
lysY	780370..780445	512
lysZ	780592..780667	514
lysQ	780800..780875	515
rybA	852175..852263	569
rybB	887199..887277	592
serT	1030848..1030935	709
serX	1096788..1096875	742
tyrV	1286467..1286551	851
tyrT	1286761..1286845	851, 852
rydC	1489467..1489530	960
valV	1744459..1744535	1117
rydB	1762737..1762804	1127
rprA	1768396..1768501	1131
ryeA	1921090..1921338	1239, 1240
ryeB	1921188..1921308	1239, 1240
leuZ	1989839..1989925	1286
cysT	1989938..1990011	1287

Table 7 Predicted loci that matched with known *E. coli* K-12 RNA genes. The positions on *E. coli* K-12 genome of known RNA genes are shown (cont).

Gene	Position	Predicted locus/loci
glyW	1990066..1990141	1287
rseX	2031673..2031763	1322
serU	2041492..2041581	1327
asnV	2060284..2060359	1332
ryeC	2151333..2151475	1389
ryeE	2165136..2165221	1399
micF	2311106..2311198	1502
argW	2464331..2464405	1601
alaX	2516063..2516138	1616
alaW	2516178..2516253	1617
valU	2518953..2519028	1620
valX	2519073..2519148	1621
valY	2519195..2519270	1621
lysV	2519275..2519350	1622
glmY	2689179..2689362	1731
rrlG	2724303..2727206	1756, 1757, 1758, 1759, 1760, 1761, 1762
gltW	2727391..2727466	1763
argQ	2815806..2815882	1807
argZ	2816081..2816157	1808
serV	2816575..2816667	1809
csrB	2922178..2922537	1890
gcvB	2940718..2940923	1911
metV	2945629..2945705	1914
omrA	2974124..2974211	1931
glyU	2997006..2997079	1945
ssrS	3054005..3054187	1964
rygC	3054871..3055010	1965
pheV	3108388..3108463	1999
rygE	3193121..3193262	2052

Table 7 Predicted loci that matched with known *E. coli* K-12 RNA genes. The positions on *E. coli* K-12 genome of known RNA genes are shown (cont).

Gene	Position	Predicted locus/loci
ileX	3213620..3213695	2062
psrN	3236396..3236583	2071
rnpB	3268238..3268614	2093, 2094
psrO	3309247..3309420	2124
leuU	3320094..3320180	2135
rrfF	3421445..3421564	2212
thrV	3421602..3421677	2213
rrlD	3421902..3424805	2214, 2215, 2216, 2217, 2218, 2219, 2220
alaU	3424980..3425055	2221
ileU	3425098..3425174	2221
rrsD	3425243..3426784	2222, 2223, 2224, 2225, 2226
ryhB	3578950..3579039	2327
proK	3706639..3706715	2404
sokA	3720099..3720128	2414
selC	3834245..3834339	2500
istR	3851141..3851280	2512
rrsC	3939831..3941372	2575, 2576, 2577, 2578
rrlC	3941727..3944630	2579, 2580, 2581, 2582, 2583, 2584
argX	3980398..3980474	2613
hisR	3980532..3980608	2613
leuT	3980629..3980715	2613
proM	3980758..3980834	2613
glmZ	3984455..3984626	2615, 2616
rrsA	4033554..4035095	2659, 2660, 2661, 2662
ileT	4035164..4035240	2663
alaT	4035283..4035358	2663
rrlA	4035542..4038446	2664, 2665, 2666, 2667,2668,2669,2670,2671
spf	4047922..4048030	2678
csrC	4049059..4049303	2679

Table 7 Predicted loci that matched with known *E. coli* K-12 RNA genes. The positions on *E. coli* K-12 genome of known RNA genes are shown (cont).

Gene	Position	Predicted locus/loci
rrsB	4164682..4166223	2769, 2770, 2771, 2 772, 2773
gltT	4166395..4166470	2774
rrlB	4166664..4169567	2775, 2776, 2777, 2778, 2779, 2780, 2781,2782
thrU	4173411..4173486	2785
tyrU	4173495..4173579	2786
glyT	4173696..4173770	2787
thrT	4173777..4173852	2787
rrsE	4206170..4207711	2821, 2822, 28 23, 2824
gltV	4207797..4207872	2825
rrlE	4208066..4210969	2825, 2826, 2827, 2828, 2829, 2830
rrfE	4211063..4211182	2831
pheU	4360574..4360649	2938
glyY	4390606..4390681	2968
leuX	4494428..4494512	3040
leuV	4604102..4604188	3099
leuQ	4604338..460442 4	3100

Table 8 Statistical evaluation of each method.

Methods	Sn	PPV	MCC
TBA	0.54	0.37	0.40
BM	0.84	0.56	0.66

2. Improved multiple alignment protocol (BM method)

The new alignment protocol was developed for improving efficiency of ncRNA gene prediction by RNAz. This protocol was used BLAST and MAFFT instead of BLASTZ and TBA to produce multiple genome wide alignment. Multiple sequence alignments were then pre-processed with rnazWindows.pl and subsequently scanned with RNAz, as same as suggested RNAz program. Using this protocol, 268 loci were predicted. They are distributed in intergenic regions of *E. coli* K-12 genome as in Figure 5. Similar to previous method, all loci were compared with annotated RNA genes of *E. coli* K-12. It was found that single RNase P RNA gene, all 22 rRNA genes, 75 of 89 tRNA genes and 31 of 60 other ncRNA genes were predicted by using workflow consisting of BM alignment protocol. The predicted loci that matched with known *E. coli* K-12 RNA genes are shown in Table 9.

According to previous studies, several approaches were used to predict ncRNAs in *E. coli*. According to sequence conservation study, the results of predictions gave 60 candidates predicted to be ncRNA elements other than rRNAs and tRNAs. After verifying by Northern blot analysis, only 17 (28%) RNA transcripts were detected (Wassarman *et al.*, 2001). In addition, resulting from combination approach of transcription unit search and sequence conservation, 24 candidates were predicted; but only 14 (58%) RNA transcripts were identified by Northern hybridization (Argaman *et al.*, 2001). This combination approach showed higher efficient than the other but its prediction result gave very low numbers of candidates. This implies its low sensitivity when compared with current numbers of non-house keeping ncRNAs. For comparative structure approach, 275 candidates were predicted by using SCFG structure comparison. Only 11 loci out of 49 tested candidates were positively observed on Northern analysis (Rivas *et al.*, 2001). The results from BM-RNAz, 31 RNA elements reported previously were predicted indicating fair and comparable efficiency comparing with those from other studies. However, all confirmed ncRNA genes were verified by Northern blot analysis only in the condition that cells were in stationary phase whereas some studies showed several ncRNA genes to be expressed only in specific condition (Eddy, 2002). Interestingly, the real number of ncRNA genes was possible higher than former reports.

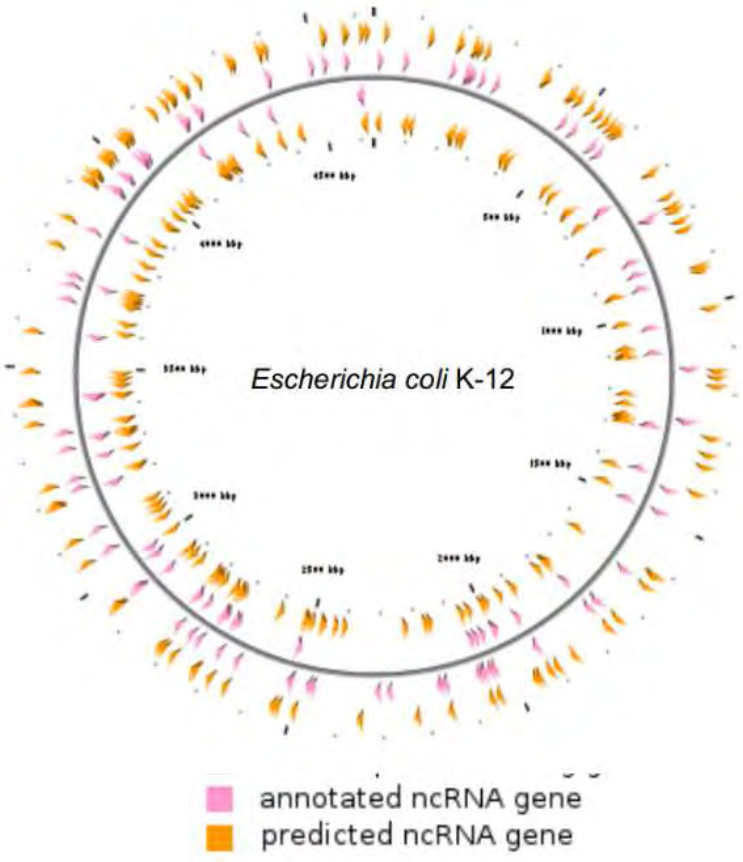

Figure 5 Positions of predicted loci of *E. coli* ncRNAs from workflow using improved alignment (BM) method and annotated ncRNA genes mapped on genome of *E. coli* K-12.

Table 9 Predicted loci from BM that matched with known *E. coli* K-12 RNA genes. The positions on *E. coli* K-12 genome of known RNA genes are shown.

Gene	Position	Predicted locus/loci
rrsH	223771..225312	9,10
alaV	225500..225575	11
rrlH	225759..228662	12, 13, 14, 15
rrfH	228756..228875	16
aspU	228928..229004	16
aspV	236931..237007	17
ffs	475672..475785	31
sroB	506428..506509	32
argU	563946..564022	36
pauD	585280..585324	39
glnX	695653..695727	42
glnV	695765..695839	43
metU	695887..695963	44
glnW	695979..696053	44
glnU	696088..696162	44
leuW	696186..696270	44
metT	696280..696356	44
lysT	779777..779852	46
valT	779988..780063	46
lysW	780066..780141	46
valZ	780291..780366	48
lysY	780370..780445	48
lysZ	780592..780667	48
lysQ	780800..780875	48
serW	925107..925194	53
serX	1096788..1096875	62
rdlA	1268546..1268612	69
rdlB	1269081..1269146	70
rdlC	1269616..1269683	71

Table 9 Predicted loci from BM that matched with known *E. coli* K-12 RNA genes. The positions on *E. coli* K-12 genome of known RNA genes are shown (cont).

Gene	Position	Predicted locus/loci
tyrV	1286467..1286551	73
tyrT	1286761..1286845	73
dicF	1647406..1647458	83
rydB	1762737..1762804	86
rprA	1768396..1768501	88
ryeA	1921090..1921338	93
ryeB	1921188..1921308	93
leuZ	1989839..1989925	98
cysT	1989938..1990011	98
glyW	1990066..1990141	98
asnT	2042573..2042648	100
asnW	2056051..2056126	101
asnV	2060284..2060359	102
ryeC	2151333..2151475	103
ryeD	2151668..2151803	104
ryeE	2165136..2165221	105
argW	2464331..2464405	111
alaX	2516063..2516138	116
alaW	2516178..2516253	116
valU	2518953..2519028	117
valX	2519073..2519148	117
valY	2519195..2519270	117
lysV	2519275..2519350	117
ryfA	2651877..2652180	121
glmY	2689179..2689362	122
rrfG	2724091..2724210	124
rrlG	2724303..2727206	125, 126, 127, 128, 129
gltW	2727391..2727466	130
rrsG	2727638..2729179	131

Table 9 Predicted loci from BM that matched with known *E. coli* K-12 RNA genes. The positions on *E. coli* K-12 genome of known RNA genes are shown (cont).

Gene	Position	Predicted locus/loci
ileY	2783784..2783859	135
argQ	2815806..2815882	139
argZ	2816081..2816157	139
argY	2816220..2816296	139
argV	2816495..2816571	140
serV	2816575..2816667	140
csrB	2922178..2922537	145
gcvB	2940718..2940923	147
metZ	2945409..2945485	148
metW	2945519..2945595	149
metV	2945629..2945705	150
glyU	2997006..2997079	151
ssrS	3054005..3054187	153
pheV	3108388..3108463	157
rygD	3192745..3192887	158
rygE	3193121..3193262	159
ileX	3213620..3213695	162
psrN	3236396..3236583	163
rnpB	3268238..3268614	164
psrO	3309247..3309420	165
metY	3316235..3316311	166
leuU	3320094..3320180	167
ryhA	3348599..3348706	169
rrfF	3421445..3421564	172
thrV	3421602..3421677	172
rrfD	3421690..3421809	172
rrlD	3421902..3424805	173, 174, 175, 176, 177
alaU	3424980..3425055	178
ileU	3425098..3425174	179

Table 9 Predicted loci from BM that matched with known *E. coli* K-12 RNA genes. The positions on *E. coli* K-12 genome of known RNA genes are shown (cont).

Gene	Position	Predicted locus/loci
rrsD	3425243..3426784	179
ryhB	3578950..3579039	184
rdlD	3698159..3698222	190
proK	3706639..3706715	191
sokA	3720099..3720128	192
selC	3834245..3834339	196
istR	3851141..3851280	197
rrsC	3939831..3941372	202
gltU	3941458..3941533	202
rrlC	3941727..3944630	203, 204, 205, 206
rrfC	3944723..3944842	207
aspT	3944895..3944971	207
trpT	3944980..3945055	207
argX	3980398..3980474	210
hisR	3980532..3980608	211
leuT	3980629..3980715	211
proM	3980758..3980834	211
rrsA	4033554..4035095	216
ileT	4035164..4035240	216
alaT	4035283..4035358	217
rrlA	4035542..4038446	217, 218, 219, 220
rrfA	4038540..4038659	221
spf	4047922..4048030	222
csrC	4049059..4049303	223
oxyS	4156308..4156417	227
rrsB	4164682..4166223	229
gltT	4166395..4166470	230
rrlB	4166664..4169567	231, 232, 233, 234
rrfB	4169660..4169779	235

Table 9 Predicted loci from BM that matched with known *E. coli* K-12 RNA genes. The positions on *E. coli* K-12 genome of known RNA genes are shown (cont).

Gene	Position	Predicted locus/loci
thrU	4173411..4173486	236
tyrU	4173495..4173579	236
glyT	4173696..4173770	236
thrT	4173777..4173852	236
rrsE	4206170..4207711	242
gltV	4207797..4207872	242
rrlE	4208066..4210969	243, 244, 245, 246
rrfE	4211063..4211182	247
pheU	4360574..4360649	254
glyV	4390383..4390458	255
glyX	4390495..4390570	256
glyY	4390606..4390681	257
symR	4577858..4577934	263

For measuring efficiency of gene annotation methods, the statistical values as described in chapter III were used (Table 8). Sensitivity of BM alignment protocol was 0.84 which was higher than that (0.54) of the protocol using TBA alignment. It was observed that, in the prediction result, the number of TP was higher than that of FP (39,509 bp of TP and 30,994 bp of FP). The prediction using BM alignment can predict larger number of loci that matched with known *E. coli* ncRNAs than the protocol using TBA can. Precision of the protocol using BM alignment was 0.56 since the number of FP was closed to that of TP. The result from BM alignment protocol gave the MCC of 0.66 which was in range of 0.6 to 0.9 for MCC values of efficient gene prediction programs (Rogic *et al.*, 2001).

3. Comparison of efficiency of alignment methods between TBA and BM protocol

To test the workflow developed for ncRNA prediction, ncRNA prediction from *E. coli* K-12 genome sequence was performed using known *E. coli* ncRNA genes as control. Two alignment protocols were also evaluated. The loci obtained from different alignment protocols were compared. For BM alignment, the predicted loci agreed with 129 known RNA genes while 103 known RNA genes matched up by loci predicted using TBA alignment. Although the number of predicted known-match loci (Table 6) by two alignments protocols were not much different, but the statistical evaluation values (Table 8) of BM protocol was improved to the acceptable range.

RNA secondary structure is conserved over evolutionary time-scales while the underlying sequences accumulate substitutions (Eddy and Durbin, 1994). These properties can be explored by computational methods such as RNAz (Washietl *et al*., 2005) to identify regions with stabilizing selection on RNA structure with in a sequence alignment. While genome-wide multiple sequence alignment is a necessary prerequisite for predicting ncRNAs from genome sequences using comparative genomic approach, quality of sequence alignment is a critical factor for sensitivity and specificity (equivalent to precision; positive predictive value: PPV) of ncRNA detection (Engelen and Tahi, 2007). The original protocol, using TBA, did not give reliable results according to workflow evaluation in this research. Recently, sporadic compelling evidence of TBA misalignment was reported (Wang, Ruzzo *and* Tompa,

2007). Besides, many conserved RNA structures in regions that TBA did not align at all were discovered (Torarinsson *et al*., 2006). New alignment protocol was therefore developed for using with RNAz. The goal of alignment protocol improvement is to reduce false positive from RNAz prediction. False positive value is represented in term of precision. The precision value of BM alignment (0.56) is higher than that (0.37) of TBA alignment. The precision value can be useful as guide to reduce number of experimental samples for verification of large number of positive ncRNAs predicted by computational tools. The sensitivity of each method is calculated to prove the efficiency of method. The sensitivity of ncRNA gene prediction using BM protocol was improved to 0.84 comparing with that of TBA protocol, 0.54.

House keeping ncRNA genes can be identified with BM alignment better than TBA alignments as showed in Table 6. In TBA protocol, though TBA involves in generating multiple alignments but BLASTZ (Schwartz *et al.,* 2003) performs sequence similarity searching. Algorithm of BLASTZ relies on order and orientation of sequence regions in given genome (Blachette *et al.*, 2004). It has been known that microbial genome sequences are diverse by evolutionary processes including recombination, such as DNA rearrangement and gene duplication, affecting in orders and orientations of sequence regions in the genome. On the other hand, BLASTN, universal BLAST search tool (Altschul *et al.,* 1990) was used in BM protocol. This regular BLAST only focuses on sequence similarity but not orders or orientations of sequences. In case of other types of ncRNAs only 31 genes were detected by BM alignment protocol. NcRNAs have many different functions such as regulatory roles. Those particular functions are rather specific in the cells. It may be that their primary sequences are varied for particular secondary structures reflecting specific functions (Eddy, 2001; Storz, 2002). Based on BLAST pair-wise alignment search, the sequences of interested were matched with those of other genomes according to their primary sequence similarity. It has been known that RNA sequence conservation may be very low but the RNA secondary conservation is relatively strong (Eddy and Durbin, 1994). This may be inferred that RNA structures involve in their functional properties. It might be possible that the BM alignment method is rather weak for finding the low-conserved or non-homologous ncRNA genes in the groups of genome in alignment. BM alignment usually matches only the highest conserved regions in the BLAST process. This problem is most disadvantage of the homology based

ncRNA gene finding methods. To avoid this problem, the selected genomes for alignment process must be from closely related species.

4. Results from base composition bias and free energy calculation

Median of average free energy of ncRNA genes was around 1.58 kcal/mol.base which significantly differ, by statistical test, from those of both protein-coding genes and of background genome, 1.62 kcal/mol.base. Mainly base composition biases were observed except the values for GA dinucleotide and AA dinucleotide. However, the interquartile ranges of some values of protein coding genes and genome background were very broad and overlapping with the median of ncRNA values. Figure 6 showed box plot analysis of 7 base composition parameters of ncRNA containing region which were significantly different from those of protein coding genes and genomic background. This will effect if they are used as parameters in ncRNA gene scanning in genome. According to this point only average free energy, (G-C) %, AT, CC, CG, GC and GT dinucleotide base compositions would be suitable for genome scanning in ncRNA screening. Technically, average free energy of DNA duplex was used to scan in *E. coli* genome. Using this technique 32 ncRNA genes could be identified at the cut off score 0.2. Relatively large number of false positives, however, was observed. Beside average free energy, other base composition values can be combined to increase specificity and sensitivity for ncRNA screening.

5. Identification of ncRNA genes by promotor and terminator prediction

A common transcription signal of any genes is promoter and terminator. They are important characteristics for not only protein coding genes but also ncRNA genes. Prediction of promoter and terminator in intergenic regions is therefore another computation approach for ncRNA detection. Integrating the result from this approach with those from the others may improve reliability of putative ncRNAs in prediction. Raw prediction of sigma factor-70 promoter and rho-independent terminator, there were 12,095 and 26,920 predictive promoter signals and terminator signals, respectively, located in intergenic regions on both strand of *E. coli* genome sequence. When compared with ncRNA gene position.

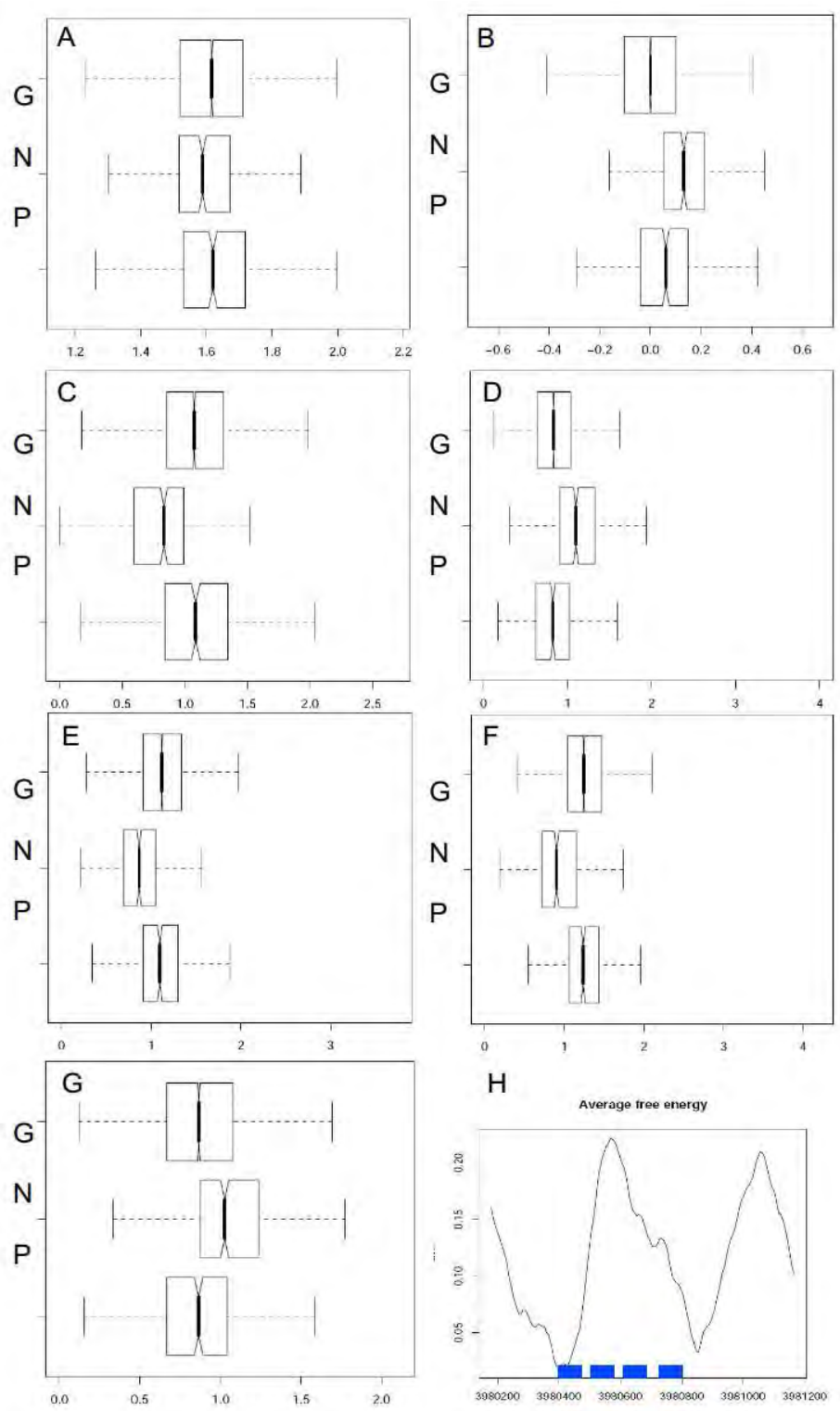

Figure 6 Box plot analysis, (A) average free energy, (B) (G-C) %, (C) AT, (D) CC, (E) CG, (F) GC and (G) GT**.** dinucleotide (H) genome wide search. G = genome background, N = ncRNA and P = protein coding

5.1. Transcription signal defination

In this study, only information for *E. coli* sigma factor 70 was exploited in prediction. Positions of predicted promoters and terminators were compared with the positions and directions of ncRNA genes. Only 45 ncRNA genes were found containing transcription signals (Table 10). Three types of transcription signals have been defined.

5.1.1. Promoter signal - a signal where the promoter position located in 100 bases upstream of the ncRNA it is compared to.

5.1.2. Terminator signal - a signal where the terminator position located in 100 bases upstream of the ncRNA it is compared to.

5.1.3. Double signal - a signal that is both a promoter hit and a terminator hit.

There were only small numbers of transcription signals found that could be mapped with ncRNA genes. This implies, according to promoter and terminator prediction method, that both known and suggested ncRNAs had weak transcriptional signals. For bacteria including *E. coli*, there are several types of sigma factors involving in transcription. There is no established information, so far, on a particular type of sigma factors involving in ncRNA gene transcription specifically. For the prediction in this research, information for *E. coli* sigma factors 32, 38, 54 and x may be useful for improving the prediction. Moreover, many bacterial genes are found to be organized in operon patterns. The genes in the same operon are transcriptionally drived by common promoter and terminator shared among these genes. This can imply that some ncRNA genes do not use promoter and terminator if they are in operons.

5.2. Locating promoters and terminators in predicted ncRNA loci

The transcription signals were mapped with predicted ncRNAs which did not overlap with known ncRNAs. The candidate loci with double transcription signal were only 18 loci and 105 loci with the single signal the results are shown in Table 11. All of loci that overlapped with known ncRNA had transcription signal. That meant this method had good sensitivity for detecting ncRNA genes but not specific. From comparing of result, this method alone can not be used to identify ncRNA gene but it can be combined with other methods to improve or prioritize results.

Table 10 The distribution of signal type compared with ncRNA genes.

Type of signal	Number of ncRNA gene
Promoter signal	9
Terminator signal	23
Double signal	13

Table 11 Transcription signal mapping with putative ncRNAs.

Locus	Start	Stop	strand	Promoter	Terminator	Signal type
locus1	11992	12163	-	0	1	terminator
locus2	29486	29651	+	1	0	promoter
locus3	89341	89387	-	2	2	double
locus4	107486	107544	-	1	0	promoter
locus5	127595	127735	+	0	2	terminator
locus6	127798	127839	+	3	1	double
locus7	190599	190857	+	0	0	-
locus8	192647	192764	+	0	2	terminator
locus9	223410	224718	-	0	0	-
locus10	224879	225361	+	1	2	double
locus11	225497	225576	+	0	2	terminator
locus12	225777	226156	-	0	3	terminator
locus13	226288	227343	-	1	2	double
locus14	227366	228387	-	0	3	terminator
locus15	228428	228706	+	0	3	terminator
locus16	228744	229014	-	1	1	double
locus17	236929	237007	+	1	3	double
locus18	249805	250043	-	1	2	double
locus19	253332	253467	+	0	1	terminator
locus20	262385	262490	-	0	2	terminator
locus21	389928	390244	+	1	2	double
locus22	392197	392556	+	0	1	terminator
locus23	392597	392716	-	1	2	double
locus24	392877	393036	+	0	2	terminator
locus25	393077	393196	-	0	2	terminator
locus26	393557	393676	+	0	0	-
locus27	410300	410339	-	0	0	-
locus28	410402	410438	-	0	2	terminator
locus29	454013	454357	+	0	0	-
locus30	460466	460675	+	0	0	-
locus31	475628	475796	+	0	2	terminator

Table 11 Transcription signal mapping with putative ncRNAs (cont).

Locus	Start	Stop	strand	Promoter	Terminator	Signal type
locus32	506466	506510	+	1	0	promoter
locus33	529216	529340	+	0	0	-
locus34	547731	547849	+	0	2	terminator
locus35	548130	548249	-	0	3	terminator
locus36	563946	564024	+	1	0	promoter
locus37	575015	575124	-	0	0	-
locus38	576448	576498	+	6	3	double
locus39	585270	585327	+	0	0	-
locus40	638731	638857	-	0	0	-
locus41	643236	643293	-	1	0	promoter
locus42	695609	695729	+	0	2	terminator
locus43	695764	695846	+	0	2	terminator
locus44	695887	696400	-	1	2	double
locus45	757671	757711	+	0	2	terminator
locus46	779744	780146	-	0	2	terminator
locus47	780169	780239	+	1	2	double
locus48	780275	781067	-	0	3	terminator
locus49	791278	791373	+	0	2	terminator
locus50	816004	816267	+	1	0	promoter
locus51	836666	836747	+	0	2	terminator
locus52	921813	922050	+	0	2	terminator
locus53	925080	925275	-	3	2	double
locus54	925666	925743	-	0	0	-
locus55	931580	931818	+	1	0	promoter
locus56	1006894	1007051	-	0	1	terminator
locus57	1014872	1014938	+	2	0	promoter
locus58	1078277	1078444	-	1	3	double
locus59	1096212	1096367	-	0	1	terminator
locus60	1096393	1096548	-	0	2	terminator
locus61	1096574	1096729	-	0	2	terminator
locus62	1096755	1096922	-	2	2	double

Table 11 Transcription signal mapping with putative ncRNAs (cont).

Locus	Start	Stop	strand	Promoter	Terminator	Signal type
locus63	1102444	1102511	-	7	0	promoter
locus64	1150698	1150838	+	0	0	-
locus65	1165137	1165308	+	0	0	-
locus66	1184817	1184951	+	0	2	terminator
locus67	1208631	1208846	-	0	3	terminator
locus68	1236472	1236536	-	2	0	promoter
locus69	1268498	1268682	-	2	0	promoter
locus70	1269033	1269217	-	2	0	promoter
locus71	1269568	1269751	-	2	0	promoter
locus72	1278577	1278698	-	0	0	-
locus73	1286466	1286980	-	0	1	terminator
locus74	1297508	1297609	+	2	2	double
locus75	1333735	1333855	+	1	0	promoter
locus76	1360537	1360664	+	0	2	terminator
locus77	1416466	1416625	-	0	2	terminator
locus78	1420873	1420992	+	0	3	terminator
locus79	1470961	1471298	-	0	2	terminator
locus80	1617047	1617139	+	0	0	-
locus81	1620863	1620908	+	2	1	double
locus82	1630030	1630149	-	2	2	double
locus83	1647418	1647617	+	0	0	-
locus84	1739238	1739437	+	2	0	promoter
locus85	1762440	1762497	-	2	0	promoter
locus86	1762699	1762867	-	0	3	terminator
locus87	1766812	1766921	+	0	2	terminator
locus88	1768315	1768493	-	2	1	double
locus89	1797006	1797133	-	0	0	-
locus90	1860600	1860795	+	1	0	promoter
locus91	1887775	1887957	-	0	0	-
locus92	1903364	1903566	+	1	2	double
locus93	1921126	1921271	+	3	2	double

Table 11 Transcription signal mapping with putative ncRNAs (cont).

Locus	Start	Stop	strand	Promoter	Terminator	Signal type
locus94	1932724	1932839	-	0	1	terminator
locus95	1948546	1948671	+	0	1	terminator
locus96	1956472	1956544	+	2	0	promoter
locus97	1976224	1976301	-	0	0	-
locus98	1989779	1990218	-	0	2	terminator
locus99	2033085	2033267	-	0	3	terminator
locus100	2042549	2042652	+	1	2	double
locus101	2056031	2056156	-	0	3	terminator
locus102	2060276	2060359	+	3	0	promoter
locus103	2151327	2151500	-	0	2	terminator
locus104	2151702	2151840	+	1	3	double
locus105	2165137	2165224	-	1	1	double
locus106	2225382	2225541	-	0	1	terminator
locus107	2225662	2225781	+	0	2	terminator
locus108	2350545	2350654	+	3	0	promoter
locus109	2403094	2403311	-	1	0	promoter
locus110	2428789	2428915	-	1	0	promoter
locus111	2464323	2464407	-	1	1	double
locus112	2468849	2469096	-	0	2	terminator
locus113	2494925	2495067	+	1	0	promoter
locus114	2496317	2496396	-	3	0	promoter
locus115	2510728	2510771	-	1	0	promoter
locus116	2516029	2516321	-	0	2	terminator
locus117	2518948	2519551	-	0	3	terminator
locus118	2531402	2531642	+	0	3	terminator
locus119	2541678	2541794	+	2	1	double
locus120	2595775	2595848	-	0	2	terminator
locus121	2651878	2652063	+	1	2	double
locus122	2689151	2689390	+	0	2	terminator
locus123	2714471	2714649	-	2	0	promoter
locus124	2724052	2724222	-	0	2	terminator

Table 11 Transcription signal mapping with putative ncRNAs (cont).

Locus	Start	Stop	strand	Promoter	Terminator	Signal type
locus125	2724244	2724482	-	0	2	terminator
locus126	2724579	2725599	+	0	3	terminator
locus127	2725620	2726743	+	1	2	double
locus128	2726808	2726927	-	0	1	terminator
locus129	2726968	2727167	+	0	2	terminator
locus130	2727235	2727474	+	2	2	double
locus131	2727573	2729490	+	2	3	double
locus132	2744207	2744313	-	0	0	-
locus133	2744347	2744446	-	0	2	terminator
locus134	2781498	2781553	+	0	2	terminator
locus135	2783781	2783859	-	1	2	double
locus136	2798630	2798745	-	0	2	terminator
locus137	2806375	2806494	-	0	3	terminator
locus138	2807398	2807517	+	0	2	terminator
locus139	2815664	2816329	-	0	2	terminator
locus140	2816360	2816768	+	1	2	double
locus141	2817168	2817355	-	0	0	-
locus142	2876457	2876513	-	1	2	double
locus143	2902401	2902435	-	5	2	double
locus144	2920122	2920169	-	2	0	promoter
locus145	2922188	2922386	-	0	0	-
locus146	2922683	2922737	-	0	2	terminator
locus147	2940611	2940915	+	0	0	-
locus148	2945383	2945489	+	1	2	double
locus149	2945517	2945597	+	1	2	double
locus150	2945627	2945728	+	0	0	-
locus151	2997005	2997115	-	0	2	terminator
locus152	3044147	3044186	-	0	2	terminator
locus153	3053963	3054229	+	0	0	-
locus154	3069291	3069418	-	0	0	-
locus155	3086196	3086265	-	0	1	terminator

Table 11 Transcription signal mapping with putative ncRNAs (cont).

Locus	Start	Stop	strand	Promoter	Terminator	Signal type
locus156	3107245	3107332	-	1	1	double
locus157	3108286	3108470	+	0	3	terminator
locus158	3192705	3192902	+	0	2	terminator
locus159	3193101	3193297	+	0	0	-
locus160	3199048	3199229	+	0	0	-
locus161	3208650	3208803	-	0	1	terminator
locus162	3213524	3213695	-	0	0	-
locus163	3236377	3236556	+	0	0	-
locus164	3268019	3268578	-	0	2	terminator
locus165	3309236	3309437	-	0	0	-
locus166	3316053	3316399	-	3	0	promoter
locus167	3320058	3320189	-	0	2	terminator
locus168	3320527	3320677	-	0	0	-
locus169	3348530	3348711	-	0	0	-
locus170	3352072	3352211	-	2	0	promoter
locus171	3376673	3376882	+	0	0	-
locus172	3421400	3421819	+	0	1	terminator
locus173	3421844	3422081	-	0	2	terminator
locus174	3422178	3423198	+	0	3	terminator
locus175	3423219	3424342	+	1	2	double
locus176	3424407	3424526	-	0	1	terminator
locus177	3424567	3424766	+	0	2	terminator
locus178	3424834	3425058	-	0	2	terminator
locus179	3425133	3427011	+	1	2	double
locus180	3451292	3451476	-	0	0	-
locus181	3468116	3468167	-	0	3	terminator
locus182	3483859	3484142	-	0	0	-
locus183	3491852	3492033	+	0	0	-
locus184	3578954	3579073	+	0	2	terminator
locus185	3598847	3599051	-	0	0	-
locus186	3628773	3628840	-	0	3	terminator

Table 11 Transcription signal mapping with putative ncRNAs (cont).

Locus	Start	Stop	strand	Promoter	Terminator	Signal type
locus187	3680030	3680088	-	0	1	terminator
locus188	3680121	3680177	-	0	2	terminator
locus189	3693301	3693555	-	0	0	-
locus190	3698114	3698281	-	1	0	promoter
locus191	3706462	3706699	-	2	2	double
locus192	3720109	3720189	-	2	0	promoter
locus193	3770078	3770304	-	0	2	terminator
locus194	3772375	3772447	+	1	0	promoter
locus195	3809697	3809906	-	0	0	-
locus196	3834181	3834343	+	5	2	double
locus197	3851177	3851268	-	2	3	double
locus198	3851665	3851712	-	1	0	promoter
locus199	3882190	3882276	-	1	2	double
locus200	3923656	3924028	+	0	0	-
locus201	3929207	3929319	+	1	0	promoter
locus202	3939382	3941538	-	1	0	promoter
locus203	3941754	3942124	+	1	2	double
locus204	3942256	3943311	-	1	2	double
locus205	3943334	3944355	-	0	3	terminator
locus206	3944396	3944673	+	0	3	terminator
locus207	3944711	3945057	-	1	1	double
locus208	3948446	3948565	+	0	2	terminator
locus209	3949006	3949125	+	0	2	terminator
locus210	3980396	3980515	+	1	2	double
locus211	3980596	3980879	-	1	3	double
locus212	3988954	3989043	+	2	3	double
locus213	3999172	3999449	+	0	0	-
locus214	4010952	4011008	-	0	2	terminator
locus215	4033181	4033307	-	0	0	-
locus216	4033328	4035205	-	1	2	double
locus217	4035280	4035939	-	0	2	terminator

Table 11 Transcription signal mapping with putative ncRNAs (cont).

Locus	Start	Stop	strand	Promoter	Terminator	Signal type
locus218	4036071	4037005	-	0	2	terminator
locus219	4037148	4038171	-	0	3	terminator
locus220	4038212	4038490	+	0	3	terminator
locus221	4038528	4038838	+	0	1	terminator
locus222	4047899	4048046	+	0	1	terminator
locus223	4049063	4049360	+	3	0	promoter
locus224	4054365	4054428	-	0	0	-
locus225	4056064	4056430	+	0	0	-
locus226	4124965	4125030	+	2	0	promoter
locus227	4156323	4156513	+	0	0	-
locus228	4164329	4164435	-	0	0	-
locus229	4164456	4166288	-	2	2	double
locus230	4166387	4166470	-	1	2	double
locus231	4166691	4167061	-	0	2	terminator
locus232	4167193	4168248	-	1	2	double
locus233	4168271	4169288	-	0	3	terminator
locus234	4169357	4169610	-	1	2	double
locus235	4169648	4169958	-	2	1	double
locus236	4173370	4173889	+	2	1	double
locus237	4175157	4175381	+	0	0	-
locus238	4177610	4178019	-	0	0	-
locus239	4178948	4179268	+	0	0	-
locus240	4194203	4194278	-	0	1	terminator
locus241	4205555	4205701	-	1	0	promoter
locus242	4205868	4207877	-	1	3	double
locus243	4208093	4208463	+	1	1	double
locus244	4208595	4209650	-	1	2	double
locus245	4209673	4210690	-	0	3	terminator
locus246	4210839	4211013	+	0	3	terminator
locus247	4211051	4211192	+	0	0	-
locus248	4213275	4213498	+	0	0	-

Table 11 Transcription signal mapping with putative ncRNAs (cont).

Locus	Start	Stop	strand	Promoter	Terminator	Signal type
locus249	4244673	4244807	+	0	0	-
locus250	4293855	4293894	-	0	0	-
locus251	4294417	4294456	-	0	3	terminator
locus252	4321268	4321303	+	0	1	terminator
locus253	4328298	4328426	+	0	2	terminator
locus254	4360574	4360673	-	0	2	terminator
locus255	4390381	4390468	+	1	2	double
locus256	4390490	4390572	+	0	2	terminator
locus257	4390601	4390686	+	0	3	terminator
locus258	4404010	4404203	+	0	0	-
locus259	4423035	4423141	-	0	2	terminator
locus260	4532076	4532249	+	0	3	terminator
locus261	4532762	4532827	+	0	2	terminator
locus262	4533305	4533505	+	0	2	terminator
locus263	4577868	4578087	+	0	0	-
locus264	4581138	4581267	+	0	0	-
locus265	4612324	4612366	-	0	0	-
locus266	4612627	4612668	-	0	2	terminator
locus267	4615207	4615316	+	0	0	-
locus268	4626762	4626847	+	0	0	-

Combining workflow

After testing efficiency of approaches with *E. coli* genome, BM alignment and RNAz were chosen as core of the workflow because of its high efficiency. Transcription unit search approach was generated large numbers of false positives; and efficiency of approach depended on specific features of individual motifs. For the genes of *M. tuberculosis*, promoter and terminator motifs are relatively diverse, not uniquely identical. Therefore, a certain transcription factor was able to bind not only to a defined motif sequence but also bind specifically to other diverse sequences. This point may reduce efficiency of prediction. Finally this approach was used only for prioritization the result from BM and RNAz. The last one, base composition and free energy bias was very low efficiency, since the normalized method and scanning method were low capability to differentiate positions of ncRNA genes. Then, the base composition and free energy bias were not combined in this workflow.

Identification of ncRNAs in *M. tuberculosis*

After workflow development, testing and alignment improvement, the obtained workflow with BM alignment protocol was used to predict *M. tuberculosis* ncRNAs. Promoter and terminator prediction was also utilized. The results were then combined.

There is no report on ncRNA identification of *M. tuberculosis* specifically. Only data of 50 ncRNA genes has been reported in genome information of *M. tuberculosis* H37Rv strain (Cole *et al.*, 1998; Camus *et al.*, 2002). They mainly are house-keeping ncRNA including rRNA and tRNA. Based on comparative genomic approach using the developed workflow, 61 loci were predicted containing putative structural RNAs. Predicted loci of *M. tuberculosis* H37Rv are lower than that of *E. coli*. Presumably it is because of difference of genome density of both organisms. According to data collected during statistical analysis, intergenic region portion (Table 12) of *M. tuberculosis* H37Rv genome (9.13% of whole genome sequence) were smaller than that of *E. coli* K-12 genome (14.9% of whole genome sequence).

Comparing with known *M. tuberculosis* H37Rv ncRNA annotated in genome project, 33 predicted RNA loci were located in ncRNA gene regions. For other 28 loci that were not matched with those known RNA genes were designated as ***putative loci and*** then mapped with promoter and terminator prediction result.

Table 12 Number of intergenic regions and whole genome in base pair. In parenthesis is percentage of intergenic region.

genome	Intergenic regions	Whole genome
E. coli K-12	693,499 (14.9%)	4,639,675
M. tuberculosis H37Rv	402,968 (9.13%)	4,411,532

There were 22 of 28 loci which had single transcription signal and only locus3 had double transcription signal as shown in Table 13. The graphical map of ncRNA gene compared with other feature was shown in Figure 7. The regulatory functions of ncRNAs are generally classified into two classes, *cis*- and *trans*- regulation, by the distances of ncRNAs and their targets. Therefore, neighbor genes were likely to be targets of ***putative loci***. However, the most of ***putative loci*** were located between hypothetical genes as shown in Table 14. Nevertheless, additional evidences from experimental techniques should be acquired for further functional characterizations of these putative ncRNAs.

For promoter and terminator prediction, most putative loci were mapped only with terminator, except just one locus that mapped with double transcription signal. Probably this is because of efficiency of promoter prediction. Most of promoter prediction tools usually generate potential candidates of promoter motifs together with false positives. Moreover, little has been known about promoters of mycobacterial genes. Therefore data for generating patterns used in prediction program was limited only *B. subtilis* sigma factor motifs and only few available sigma factor motifs of *M. tuberculosis* H37Rv. More efficient promoter prediction tools and additional defined information of mycobacterial sigma factors and promoters are required in order to improve this subject.

Although results of predicted promoter mapped onto ***putative loci*** could not make a clear signal for ncRNA identification, however, the RNAz result itself is rather interesting. The statistical value of each locus predicted from RNAz is P-value which is between 0 and 1. The default P-value for locus to be reported as RNA locus is 0.5 or higher. The higher P-value the more significant the locus is, inferring plausible secondary structure forming locus. The P-values of most of 28 ***putative loci*** are not only above the cut off value but also remarkable high (Table 13). Twenty of them are higher than 0.9. Since these ***putative loci***, which are in the intergenic regions, were predicted from alignments produced from genome sequences of closely relates species and predicted having abilities of secondary structure forming, they could be considered as putative ncRNA containing sequences and for further experimental verification.

Table 13 The putative ncRNAs of *M. tuberculosis* H37Rv.

locus	start	stop	strand	P value	Transcription signal type	BLAST search
locus3	293604	293656	+	0.99	double	
locus4	528387	528466	+	0.98	terminator	
locus5	800182	800411	+	0.99	terminator	
locus7	888914	888995	-	1	terminator	
locus10	965639	965719	-	0.86	terminator	
locus12	1057972	1058078	-	1	-	
locus15	1282075	1282216	-	0.68	terminator	
locus21	1473486	1473592	+	0.93	terminator	
locus30	1735490	1735679	-	1	terminator	ykoK
locus32	1852138	1852176	+	0.88	terminator	
locus34	2047593	2047687	-	0.99	terminator	ykoK
locus35	2096723	2096852	-	0.98	terminator	
locus39	2531900	2532210	-	0.96	-	
locus42	2744984	2745247	-	0.55	-	
locus44	2849541	2849703	+	0.87	terminator	
locus45	3155880	3156088	+	1	terminator	
locus46	3232650	3232865	+	1	-	
locus47	3239470	3239614	+	1	terminator	
locus49	3388874	3389001	-	0.89	-	
locus52	3551166	3551229	-	1	terminator	
locus53	3568770	3568814	+	0.98	terminator	
locus54	3820436	3820501	+	1	terminator	
locus55	3837332	3837515	-	1	-	
locus56	3862458	3862594	-	0.88	terminator	
locus57	4099383	4099515	-	0.93	terminator	
locus58	4156802	4156974	-	0.89	terminator	
locus59	4168193	4168298	-	0.91	-	SRP_bact
locus62	4273591	4273664	-	1	terminator	

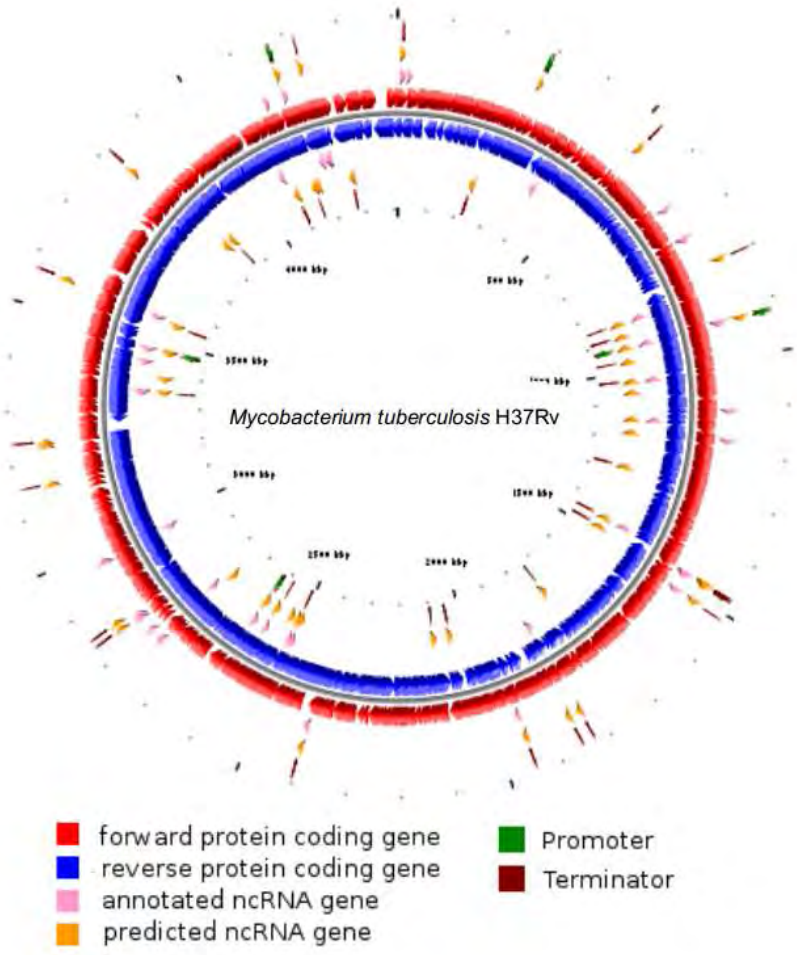

Figure 7 Position of predicted loci, protein coding genes, annotated ncRNA genes promoter and rho-independence terminator overlay on genome circle of *M. tuberculosis*.

Table 14 ***Putative loci*** and their neighbor.

	gene name		length	locus	length		gene name	
	fadA2	>	111	locus3	142	<	fadE5	
<	Rv0439c		73	locus4	142		groEL	>
	Rv0699	>	332	locus5	76		rpsJ	>
<	Rv0794c		278	locus7	77		Rv0795	>
<	rpfA		104	locus10	264	<	moaD2	
<	Rv0948c		9	locus12	182		uvrD1	>
	Rv1155	>	203	locus15	90		Rv1156	>
	murA	>	1909	locus21	3542	<	ogt	
	Rv1534	>	79	locus30	297		Rv1535	>
<	lysS		103	locus32	97		infC	>
<	Rv1804c		244	locus34	0	<	Rv1805c	
<	Rv1846c		124	locus35	25		Rv1847	>
<	Rv2258c		3	locus39	35		adhE2	>
<	rne		0	locus42	67	<	ndkA	
<	fas		209	locus44	149	<	Rv2525c	
<	cysG		9	locus45	60	<	cobB	
<	amt		144	locus46	6	<	ftsY	
<	fpg		0	locus47	215	<	rncS	
<	fixA		4	locus49	100		Rv3030	>
	Rv3183	>	122	locus52	52		Rv3184	>
<	whiB7		91	locus53	295	<	uvrD2	
	Rv3401	>	34	locus54	152	<	Rv3402c	
<	groES		44	locus55	40	<	gcp	
<	rplM		68	locus56	30	<	esxT	
<	Rv3660c		235	locus57	132		Rv3661	>
<	dnaQ		73	locus58	7		Rv3712	>
<	Rv3722c		65	locus59	238		Rv3723	>
<	glf		116	locus62	75		pirG	>

Length is number of base pair between ***putative loci*** and their neighbor. The signs, '>' and '<', refer to gene direction

By sequence similarity search and information on ncRNA database, 25 ***putative loci*** were not matched with known ncRNAs on non-coding database. Only 3 ***putative loci***, as shown in Table 13, were matched with already known ncRNA in databases. Two ***putative loci*** exhibited strong sequence similarity with known RNA element, *ykoK*, as shown in Table 15. The other ***putative locus*** had its sequence similar to signal recognition particle, SRP, sequence (Table 15).

The *ykoK* element is a divalent metal sensing RNA. Genes downstream of *ykoK* elements were reported to be similar to divalent transporter genes including those specific for Mg^{2+}, Mn^{2+}, Co^{2+}, Ni^{2+} and Fe^{2+} (Barrick *et al.*,2004). In *M. tuberculosis* H37Rv, several genes were annotated as putative metal transporters (Camus *et al.*, 2002). The comparison with annotation data, only downstream of locus34, that was located between nucleotides 2047593 and 2047687, had *mgtC* gene; possible Mg^{2+} transport; located from nucleotides 2053443 to 2054147 on the genome. The *B. subtilis ykoK* element was upstream, directly adjacent, of *mgtE* (Dann III *et al.*, 2007). Though, *M. tuberculosis* H37Rv, the locus 34 was upstream of *mgtC* gene but there were six genes, two hypothetical protein genes, one PE family gene and three PPE family genes, between them. The *mgtC* gene has been reported as an essential gene for growth in low concentration of Mg^{2+} such as in macrophage (Buchmeier *et al.,* 1999). The secondary structure of locus30 and 34 were predicted using UNAFOLD program (Markham and Zuker, 2008) and compared with known secondary structure of *ykoK* element of *B. subtilis* as shown in Figure 8. Overall secondary structure of locus 30 was quite similar to that of *B. subtilis ykoK* element with five stems. However, the prediction was not inferred any intramolecular interaction and binding properties between predicted structure of locus 30 and Mg^{2+}. Reasonably it is interesting to perform experimental verification of this locus. In the case of locus34, because locus was shorter than the *ykoK* element (94 to 175 bps), the predicted structure was substantial different from that of *B. subtilis*.

Signal recognition particle (SRP) RNA is an RNA element that is a component of signal recognition particle protein, forming RNA-protein complex involving in protein translocation and targeting. This RNA element is universally conserved and usually located in upstream region of the gene. By locating the position of locus59 on the genome, this locus was upstream of a hypothetical gene, Rv3722c (Camus *et al.*, 2002).

Table 15 Alignment from comparing between putative ncRNAs and ncRNA database. In the table was shown name of gene and its length

Alignment of locus 30 with ykoK gene; Length = 185

```
19  taagcacctcgttaggtgaggcggctacacgaacataggccactgaccccgaacgtcgag 78
     |||||||||||||||||||||| ||||||||| | ||||||||||||||||||||||
1   taagcacctcgttaggtgaggcgtctacacgaatacaggccactgaccccgaacgtcgaa 60

79  agacgccccgggtcaggacagctcttcccggcttaagggttgagcccaggt ggcttccgg 138
    |||||||||||||||||||||| ||||||||||||||||||||||||||||||||||||
61  agacgccccgggtcaggacagctcctcccggcttaagggttgagcccaggtggcttccgg 120

139 ct 140
    ||
121 ct 122
```

Alignment of locus34 with ykoK gene;Length = 175

```
2  caagcacctcgctaggtgaggcg tctgcgcggatataggccactgacctcgaacgtcgaa 61
   |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
1  caagcacctcgctaggtgaggcgtctgcgcggatataggccactgacctcgaacgtcgaa 60

62 agacgcccagggtcaggacagctcttcccggctt 95
   |||||||||||||||||||||||||||||||||
61 agacgcccagggtcaggacagctcttcccggctt 94
```

Alignment of locus59 with SRP gene;Length = 101

```
1  ggggaccccgcgcacccgacagagcccgttgacccttgctgccttccagccctgggggag 60
   |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
94 ggggaccccgcgcacccgacagagcccgttgac ccttgctgccttccagccctgggggag 35

61 ttcacaggatagacgccgcgcggggtccaccgtg 94
   |||||||||||||||||||||||||||||||||
34 ttcacaggatagacgccgcgcggggtccaccgtg 1
```

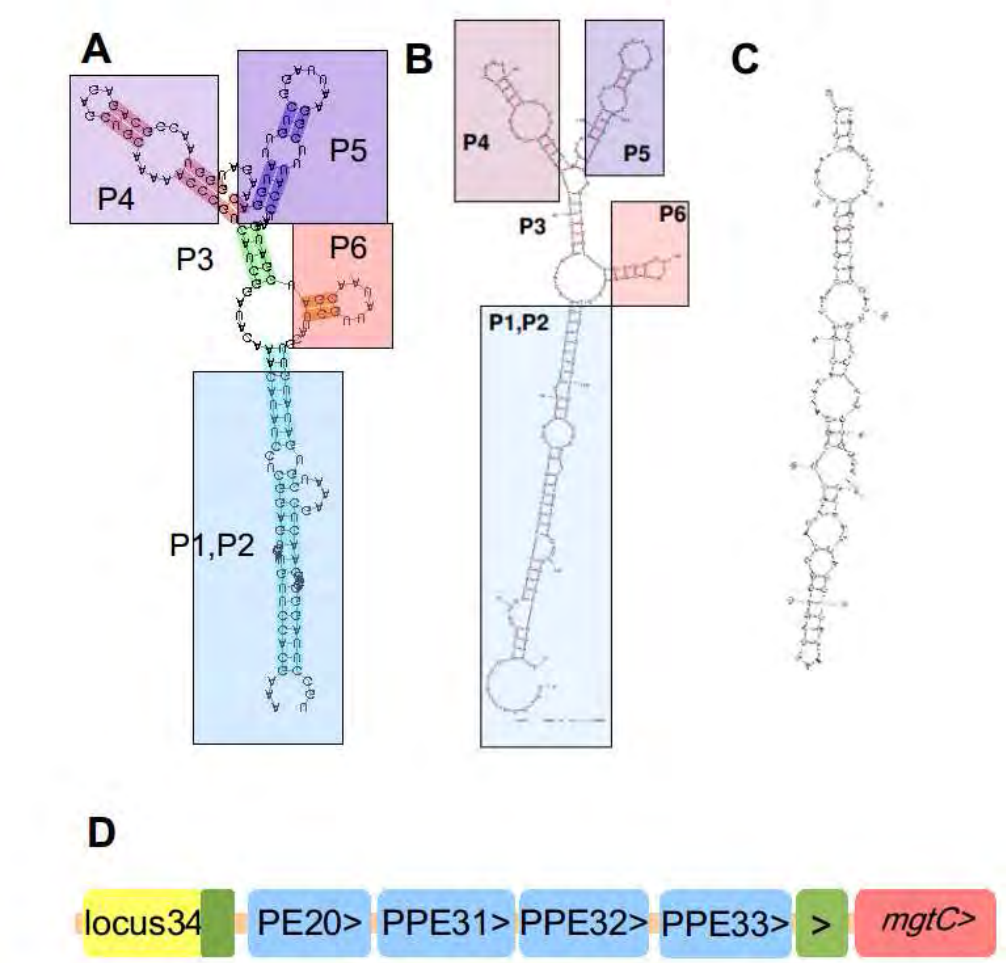

Figure 8 Secondary structure of locus30 and locus34 compared with *ykoK* of *B. subtilis*. (A) Secondary structure of *ykoK* in *B. subtilis* (Dann III *et al.*, 2007) with stem P1 to P6, (B) predicted secondary structure of locus30 compared to that of *B. subtilis ykoK*. Transparent boxes show regions that similar to *B. subtilis ykoK* regions with corresponding colors in (A). (C) Predicted secondary structure of locus 34. (D) Position of locus34 and their neighbor gene. The green box is hypothetical gene.

The SRP RNA element in eubacteria was classified into three types by using important domain called Alu, helix 6 and helix 8 (Regalia, Rosenblad and Samuelsson, 2002). From predicted secondary structure of locus59 with UNAFOLD (Markham and Zuker, 2008), structure had only helix 8 but lacked of Alu and helix 6 domain as shown in Figure 9.

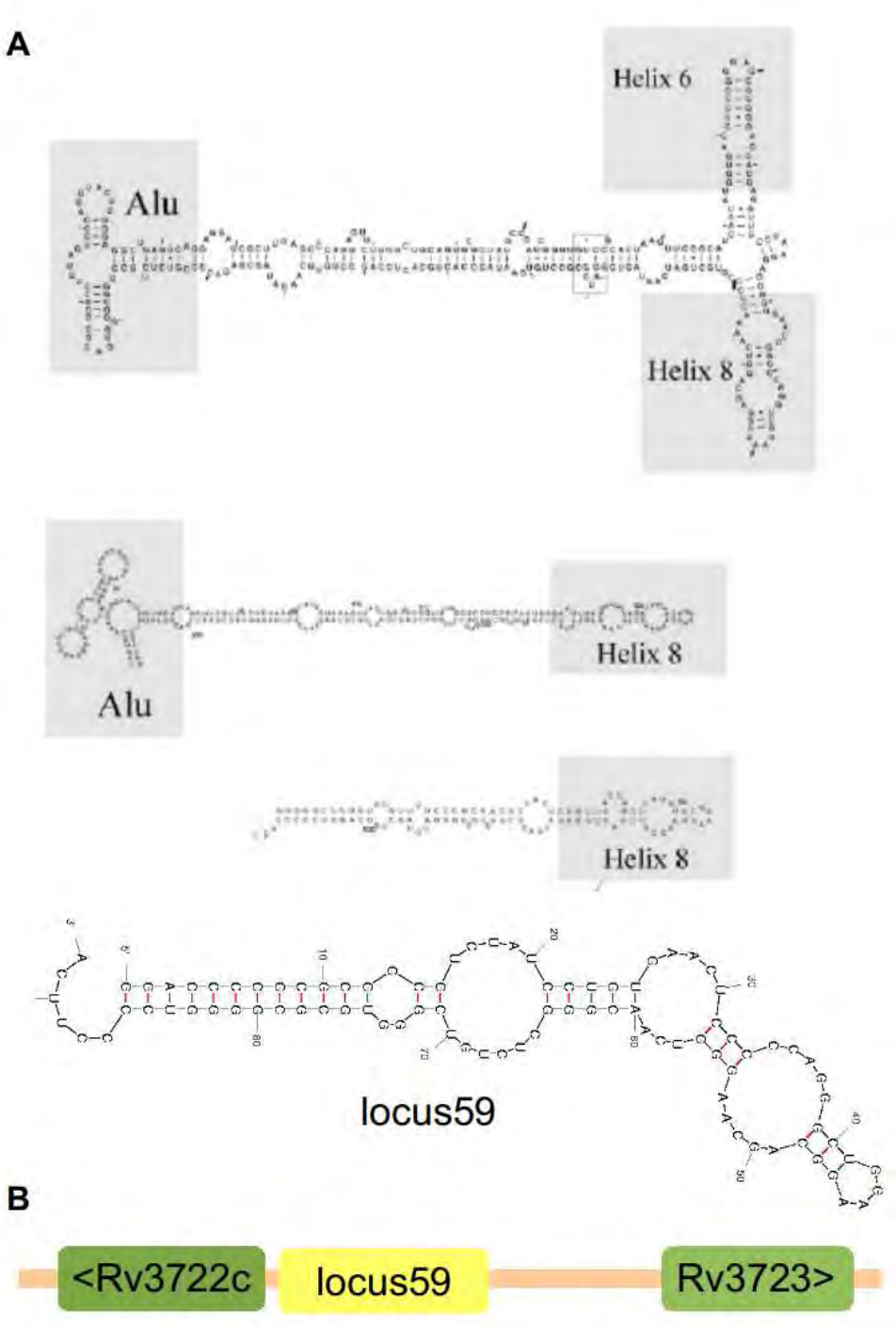

Figure 9 Predicted secondary structure of locus59 and gene position. (A) structure comparison between locus59 and three types of SRP in eubacteria (Regalia *et al.*, 2002). (B) Position of locus59 and their neighbor gene.

CHAPTER V

CONCLUSION

This study has created a novel protocol for ncRNA gene identification. This protocol is based upon secondary structure prediction of conserved DNA sequence. This identification protocol has been implemented with similarity search and secondary structure formation energy, and combined with other methods to improve quality of prediction. The main protocol of ncRNA identification was implemented by RNAz program. There were four methods planed to combine to RNAz. The result of ncRNA detection program suggested several ncRNA candidates and prioritized them. The number of candidates suggested is highly dependent upon the main protocol.

Firstly, the core programs in protocol were implemented by TBA alignment and RNAz. The advantages of these main protocols were fast and reliable with prediction of secondary forming energy. Generally, the main program was effective enough for identification of ncRNA but they generated too many false positives with low sensitivity and specificity. Results from RNAz depended on quality of input alignment. The first improvement in this protocol was the alignment program. In general, TBA alignment was suggested by developer of RNAz. Normally, TBA alignment is used in genome alignment upon the environment of aligned sequence. The new alignment protocol was developed by implementation of BLAST and MAFFT alignment. This new alignment protocol can improve the sensitivity and precision of prediction from RNAz, testing with *E. coli*. In this study, the new protocol was used in the prediction instead of TBA alignment.

The second improvement was promoter and terminator prediction. This method was usually used in gene annotating protocol but it generated many false positives in testing with *E. coli*. Notably, transcription process and transcription machines of ncRNA in bacteria were not clear. The difficulty of this method was the specification of promoter and terminator searching. In this study, the promoter and terminator prediction were used in prioritizing the loci in prediction.

The third method was base composition methods and energy of the duplex DNA. Several studies were suggested that this method was highly possible for

searching on single sequence of DNA, and it did not depend on similarity search. However, this method was generated untranslated signal and needed to be improved the algorithm and statistical model to differentiate the signal of ncRNA from background. This improved method was not completed for using in prediction.

The combined protocol was used in prediction of ncRNA in *M. tuberculosis* H37Rv. There were 62 predicted loci. In this number of candidates, there were 33 loci located in annotated ncRNA gene, 29 new ncRNA loci and 21 promising candidates with the transcription signals. This protocol is highly specific for identification of highly conserved ncRNA gene based on the first alignment. It cannot search for new species of ncRNA. The further study was the improvement the protocol and created new program to search ncRNA gene which does not depend on similarity search.

REFERENCES

Agarwal, N., Woolwine, S. C., Tyagi, S., and Bishai, W. R. Characterization of the *Mycobacterium tuberculosis* sigma factor SigM by assessment of virulence and identification of SigM-dependent genes. Infection Immunity 75 (2007): 452-461.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. Journal of Molecular Biology 215 (1990): 403-410.

Argaman, L., Hershberge, R., Vogel, J., Bejerano, G., Wagner, E. G., Margalit, H., and Altuvia, S. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. Current Biology 11 (2001): 941-950.

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

Barrick, J. E., Corbino, K. A., Winkler, W. C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I., Wickiser, J. K., and Breaker, R. R. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proceedings of the National Academy of Sciences USA 101 (2004): 6421-6426.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., and Miller, W. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Research 14 (2004): 708-715.

Breslauer, K. J., Frank, R., Blocker, H., and Marky, L. A. Predicting DNA Duplex Stability from the Base Sequence. Proceedings of the National Academy of Sciences USA 83 (1986): 3746-3750.

Camus, J. C., Pryor, M. J., Médigue, C., and Cole, S. T. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. Microbiology 148, (2002): 2967-2973.

Carrington, J. C., and Ambros, V. Role of microRNAs in plant and animal development. Science 301 (2003): 336-338.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., and Barrell, B. G. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393 (1998): 537-544.

Dann, C. E., Wakeman, C. A., Sieling, C. L., Baker, S. C., Irnov, I., and Winkler, W. C. Structure and mechanism of a metal-sensing regulatory RNA. Cell 130 (2007): 878-892

Devulder, G., de Montclos, M. P., and Flandrois, J. P. A multigene approach to phylogenetic analysis using the genus Mycobacterium as a model. International Journal of Systematic and Evolutionary Microbiology 55 (2005): 293-302.

Dye, C., Espinal, M. A., Watt, C. J., Mbiaga, C., and Williams, B. G. Worldwide incidence of multidrug-resistant tuberculosis. Journal of Infectious Diseases. (2002): 1197-1202.

Eddy, S. R. Non-coding RNA genes and the modern RNA world. Nature Reviews Genetics 2 (2001): 919-29.

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

Eddy, S. R., and Durbin, R. RNA sequence analysis using covariance models. Nucleic Acids Research 22 (1994): 2079-2088.

Engelen, S., and Tahi, F. Predicting RNA secondary structure by the comparative approach: how to select the homologous sequences. BMC Bioinformatics 8 (2007): 464.

Erdmann, V. A., Barciszewska, M. Z., Szymanski, M., Hochberg, A., Nathan de Groot, and Barciszewski, J. The non-coding RNAs as riboregulators. Nucleic Acids Research 29 (2001):189-193.

Gardner, P. P., and Giegerich, R. A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics 5 (2004): 140.

Gautheret, D., and Lambert, A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. Nucleic Acids Research 29 (2001): 4724-4735.

Gottesman, S. The small RNA regulators of *Escherichia coli*: role and mechanisms. Annual Review of Microbiology 58 (2004): 303-328.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Research 33 (2005): D121-D124.

Hershberg, R., Altuvia, S., and Margalit, H. A survey of small RNA-encoding genes in *Escherichia coli*. Nucleic Acids Research 31 (2003): 1813-1820.

Hofacker, I. L. RNA consensus structure prediction with RNAalifold. Methods in Molecular Biology 395 (2007): 527-544.

Hofacker, I. L., and Stadler, P. F. Memory efficient folding algorithms for circular RNA secondary structures. Bioinformatics 22 (2006): 1172-1176.

Huttenhofer, A., Brosius, J., and Bachellerie, J. P. RNomics: identification and function of small, non-messenger RNAs. Current Opinion in Chemical Biology 6 (2002): 835-43.

Johansson, J., and Cossart, P., RNA-mediated control of virulence gene expression in bacterial pathogens. Trends in Microbiology 11 (2003): 280-285.

Katoh, K., and Toh, H., Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. BMC Bioinformatics 9 (2008): 212.

Lindgreen, S., Gardner, P. P., and Krogh, A. Measuring covariation in RNA alignments: physical realism improves information measures. Bioinformatics 22 (2006): 2988-2995.

Liu, C., Bai, B., Skogerbø, G., Cai, L., Deng, W., Zhang, Y., Bu, D., Zhao, Y., and Chen, R. NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Research 33 (2005): D112-115.

Lucas, W. J., Yoo, B. C., and Kragler, F. RNA as a long-distance information macromolecule in plants. Nature Reviews Molecular Cell Biology 2 (2001): 849-857.

Manganelli, R., Provvedi, R., Rodrigue, S., Beaucher, J., Gaudreau, L., Smith, I., and Proveddi, R. Sigma factors and global gene regulation in Mycobacterium tuberculosis. Journal of Bacteriology 186 (2004): 895-902.

Markham, N. R., and Zuker, M. UNAFold: software for nucleic acid folding and hybridization. Methods Mololecular Biology 453 (2008): 3-31.

Makita, Y., Nakao, M., Ogasawara, N., and Nakai, K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. Nucleic Acids Research 32 (2004): D75-77.

McCutcheon, J. P., and Eddy, S. R. computational identification of noncoding RNAs in *Saccharomyces cerevisiae* by comparative genomics. Nucleic Acids Research 31 (2003): 4119-4128.

Morey, C., and Avner, P. Employment opportunities for non-coding RNAs. FEBS Letters 567 (2004): 27-34.

Mulligan, M. E., Hawley, D. K., Entriken, R., and McClure, W. R. *Escherichia coli* promoter sequences predict in vitro RNA polymerase selectivity. Nucleic Acids Research 12 (1984): 789-800.

Nawrocki, E. P., and Eddy, S. R., Query-dependent banding (QDB) for faster RNA similarity searches. PLoS Computational Biology 3 (2007): e56.

Oh, S. J., Joung, J., Chang, J., and Zhang, B. Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks. BMC Bioinformatics 7 (2006): 284.

Pittius, N. C. G. V., Gamieldien, J., Hide, W., Brown, G. D., Siezen, R. J., and Beyers, A. D. The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C Gram-positive bacteria. Genome Biology 2 (2001): RESEARCH0044.

Rattan, A., Kalia, A., and Ahmad, N. Multidrug-resistant *Mycobacterial tuberculosis*: molecular perspectives. Emerging Infectious Diseases 4 (1998): 1429-1449.

Repoila, F., Majdalani, N., and Gottesman, S. Small non-coding RNAs, co-ordinators of adaptation processes in *Escherichia coli*: the *RpoS* paradigm. Molecular Microbiology 48 (2003): 855-861.

Rivas, E., and Eddy, S. R., Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2 (2001): 8.

Rivas, E., and Eddy, S. R., Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics 16 (2000): 583-605.

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

Rogic, S., Mackworth, A. K., and Ouellette, F. B. Evaluation of gene-finding programs on mammalian sequences. Genome Research 11 (2001): 817-832.

Romby, P., Vandenesch, F. and Wagner, E. G. The role of RNAs in the regulation of virulence-gene expression. Current Opinion in Microbiology 9 (2006): 229-236.

Schattner, P. Searching for RNA genes using base-composition statistics. Nucleic Acids Research 30 (2002):2076–2082.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W. Human-mouse alignments with BLASTZ. Genome Research 13 (2003): 103-107.

Storz, G. An expending universe of noncoding RNAs. Science 296 (2002): 1260-1263.

Szymanski, M., Erdmann, V. A., and Barciszewski, J. Noncoding RNAs database (ncRNAdb). Nucleic Acids Research 35 (2007): D162-164.

Toledo-Arana, A., Repoila F., and Cossart P. Small noncoding RNAs controlling pathogenesis. Current Opinion in Microbiology 10 (2007): 182-188.

Torarinsson, E., Sawera, M., Havgaard, J. H., Fredholm, M., and Gorodkin, J. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. Genome Research 16 (2006): 885-889.

Vogel, J., and Sharma, C. M. How to find small non-coding RNAs in bacteria. Biological Chemistry 286 (2005):1219-1238.

Wang, A. X., Ruzzo, W. L., and Tompa, M. How Accurately Is ncRNA Aligned within Whole-Genome Multiple Alignments?. BMC Bioinformatics 8 (2007): 417

Washietl, S., Hofacker, I. L., Lukasser, M., Hüttenhofer, A., and Stadler, P. F. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nature Biotechnology 23 (2005): 1383-1390.

Washietl, S., Hofacker, I. L., and Stadler, P. F. Fast and reliable prediction of noncoding RNAs. Proceedings of the National Academy of Sciences, USA 102 (2005): 2454-2459.

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

World Health Organization (WHO). Tuberculosis. fact sheet no. 104. Geneva: WHO, 2007.

Zhang, Y., and Zaki, M. J. SMOTIF: efficient structured pattern and profile motif search. Algorithms for Molecular Biology 1 (2006): 22-24.

APPENDIX A

Annotation table of ncRNA genes in *E. coli*

Location	Strand	Length	Gene	Product
16952..17006	+	55	sokC	Antisense sRNA blocking mokC, and hence hokC, translation; IS186A interrupts hokC transcript downstream of hokC gene in K-12
77367..77593	+	227	sgrS	sRNA that destabilzes ptsG mRNA; regulated by sgrR
189712..189847	+	136	tff	Novel sRNA, function unknown
223771..225312	+	1542	rrsH	16S ribosomal RNA of rrnH operon
225381..225457	+	77	ileV	Ile tRNA
225500..225575	+	76	alaV	Ala tRNA
225759..228662	+	2904	rrlH	23S ribosomal RNA of rrnH operon
228756..228875	+	120	rrfH	5S ribosomal RNA of rrnH operon
228928..229004	+	77	aspU	Asp tRNA
236931..237007	+	77	aspV	Asp tRNA
262095..262170	+	76	thrW	Thr tRNA
296430..296478	+	49	ptwF	Xaa tRNA

Location	Strand	Length	Gene	Product
506428..506509	+	82	sroB	Novel sRNA, function unknown
563946..564022	+	77	argU	Arg tRNA
585280..585324	+	45	pauD	Xaa tRNA
695653..695727	-	75	glnX	Gln tRNA
695765..695839	-	75	glnV	Gln tRNA
695887..695963	-	77	metU	Met tRNA
695979..696053	-	75	glnW	Gln tRNA
696088..696162	-	75	glnU	Gln tRNA
696186..696270	-	85	leuW	Leu tRNA
696280..696356	-	77	metT	Met tRNA
779777..779852	+	76	lysT	Lys tRNA
779988..780063	+	76	valT	Val tRNA
780066..780141	+	76	lysW	Lys tRNA
780291..780366	+	76	valZ	Val tRNA
780370..780445	+	76	lysY	Lys tRNA
780592..780667	+	76	lysZ	Lys tRNA
780800..780875	+	76	lysQ	Lys tRNA
852175..852263	-	89	rybA	Novel sRNA, function unknown
887199..887277	-	79	rybB	sRNA effector of ompC and ompW mRNA instability; requires Hfq

Location	Strand	Length	Gene	Product
925107..925194	-	88	serW	Ser tRNA
1030848..1030935	-	88	serT	Ser tRNA
1096788..1096875	-	88	serX	Ser tRNA
1145812..1145980	+	169	psrD	Novel sRNA, function unknown
1268546..1268612	+	67	rdlA	Antisense sRNA RdlA affects LdrA translation; proposed addiction module in LDR-A repeat, with toxic peptide LdrA
1269081..1269146	+	66	rdlB	Antisense sRNA RdlB affects LdrB translation; proposed addiction module in LDR-B repeat, with toxic peptide LdrB
1269616..1269683	+	68	rdlC	Antisense sRNA RdlC affects LdrC translation; proposed addiction module in LDR-C repeat, with toxic peptide LdrC
1286289..1286459	-	171	rttR	rtT sRNA, processed from tyrT transcript; encodes putative Tpr protein; RNA itself may modulate the stringent response
1286467..1286551	-	85	tyrV	Tyr tRNA
1286761..1286845	-	85	tyrT	Tyr tRNA
1403676..1403833	-	158	isrA	Novel sRNA, function unknown
1435145..1435253	+	109	micC	MicC sRNA regulator of OmpC translation

Location	Strand	Length	Gene	Product
1489467..1489530	-	64	rydC	sRNA regulator of yejABEF
1490143..1490198	+	56	sokB	Antisense sRNA blocking mokB, and hence hokB, translation
1647406..1647458	+	53	dicF	DicF antisense sRNA, inhibits ftsZ, Qin prophage
1744459..1744535	+	77	valV	Val tRNA
1744540..1744616	+	77	valW	Val tRNA
1762737..1762804	-	68	rydB	Novel sRNA, function unknown
1768396..1768501	+	106	rprA	Positive regulatory sRNA for RpoS translation; non-essential gene
1921090..1921338	+	249	ryeA	Novel sRNA, function unknown
1921188..1921308	-	121	ryeB	Novel sRNA, function unknown
1985863..1986022	-	160	isrB	Novel sRNA, function unknown
1989839..1989925	-	87	leuZ	Leu tRNA
1989938..1990011	-	74	cysT	Cys tRNA
1990066..1990141	-	76	glyW	Gly tRNA
2023251..2023337	-	87	dsrA	Regulatory sRNA enhances translation of RpoS; component of acid resistance regulatory circuit; also antagonist of H-NS function by decreasing H-NS levels

Location	Strand	Length	Gene	Product
2031673..2031763	+	91	rseX	sRNA regulating ompA and ompC translation, with Hfq
2041492..2041581	-	90	serU	Ser tRNA
2042573..2042648	+	76	asnT	Asn tRNA
2056051..2056126	-	76	asnW	Asn tRNA
2057875..2057950	+	76	asnU	Asn tRNA
2060284..2060359	+	76	asnV	Asn tRNA
2069339..2069542	+	204	isrC	Novel sRNA, function unknown, CP4-44; putative prophage remnant
2151333..2151475	+	143	ryeC	Novel sRNA, function unknown; paralogous to the other QUAD sRNA genes unknown
2151668..2151803	+	136	ryeD	Novel sRNA, function unknown; paralogous to the other QUAD sRNA genes
2165136..2165221	+	86	ryeE	Novel sRNA, function unknown, prophage remnant PR-X hybrid
2284233..2284309	+	77	proL	Pro tRNA
2311106..2311198	+	93	micF	Regulatory antisense sRNA affecting ompF expression; member of soxRS regulon
2464331..2464405	+	75	argW	Arg tRNA

Location	Strand	Length	Gene	Product
2474606..2474620	+	15	pawZ	Xaa tRNA
2516063..2516138	-	76	alaX	Ala tRNA
2516178..2516253	-	76	alaW	Ala tRNA
2518953..2519028	+	76	valU	Val tRNA
2519073..2519148	+	76	valX	Val tRNA
2519195..2519270	+	76	valY	Val tRNA
2519275..2519350	+	76	lysV	Lys tRNA
2651877..2652180	+	304	ryfA	Novel sRNA, function unknown
2689179..2689362	-	184	glmY	sRNA activator of glmS mRNA
2698081..2698399	-	319	ryfB	Novel sRNA, function unknown
2698542..2698618	+	77	ryfC	Novel sRNA, function unknown
2724091..2724210	-	120	rrfG	5S ribosomal RNA of rrnG operon
2724303..2727206	-	2904	rrlG	23S ribosomal RNA of rrnG operon
2727391..2727466	-	76	gltW	Glu tRNA
2727638..2729179	-	1542	rrsG	16S ribosomal RNA of rrnG operon
2732175..2732317	-	143	ryfD	Novel sRNA, function unknown
2753615..2753977	+	363	ssrA	tmRNA
2775994..2776007	+	14	psaA	misc_RNA

Location	Strand	Length	Gene	Product
2783784..2783859	-	76	ileY	Ile tRNA
2812824..2812901	+	78	micA	sRNA effector of ompA mRNA instability in stationary phase; requires Hfq
2815806..2815882	-	77	argQ	Arg tRNA
2816081..2816157	-	77	argZ	Arg tRNA
2816220..2816296	-	77	argY	Arg tRNA
2816495..2816571	-	77	argV	Arg tRNA
2816575..2816667	-	93	serV	Ser tRNA
2922178..2922537	-	360	csrB	CsrA-binding sRNA, antagonizing CsrA regulation; blocks the CsrA binding of glgC mRNA
2940718..2940923	+	206	gcvB	GcvB sRNA gene divergent from gcvA
2945409..2945485	+	77	metZ	Met tRNA
2945519..2945595	+	77	metW	Met tRNA
2945629..2945705	+	77	metV	Met tRNA
2974124..2974211	-	88	omrA	sRNA down regulates OM proteins; positively regulated by OmpR/EnvZ; binds Hfq
2974332..2974407	-	76	omrB	sRNA down regulates OM proteins; positively regulated by OmpR/EnvZ; binds Hfq

Location	Strand	Length	Gene	Product
2997006..2997079	-	74	glyU	Gly tRNA
3054005..3054187	+	183	ssrS	6S sRNA inhibits RNA polymerase promoter binding; template for RNA-directed pRNA synthesis by RNAP; mimics an open promoter
3054871..3055010	+	140	rygC	Novel sRNA, function unknown; paralogous to the other QUAD sRNA genes
3108388..3108463	+	76	pheV	Phe tRNA
3192745..3192887	-	143	rygD	Putative sRNA, function unknown; paralogous to the other QUAD sRNA genes
3193121..3193262	-	142	rygE	Putative sRNA, function unknown; paralogous to the other QUAD sRNA genes
3213620..3213695	+	76	ileX	Ile tRNA
3236396..3236583	+	188	psrN	Novel sRNA, function unknown
3268238..3268614	-	377	rnpB	RNase P, M1 sRNA component; involved in transfer RNA and 4.5S RNA-processing
3309247..3309420	+	174	psrO	Novel sRNA, function unknown
3316235..3316311	-	77	metY	Met tRNA
3320094..3320180	-	87	leuU	Leu tRNA

Location	Strand	Length	Gene	Product
3348599..3348706	+	108	ryhA	Novel sRNA, function unknown
3421445..3421564	-	120	rrfF	5S ribosomal RNA of rrnD operon
3421602..3421677	-	76	thrV	Thr tRNA
3421690..3421809	-	120	rrfD	5S ribosomal RNA of rrnD operon
3421902..3424805	-	2904	rrlD	23S ribosomal RNA of rrnD operon
3424980..3425055	-	76	alaU	Ala tRNA
3425098..3425174	-	77	ileU	Ile tRNA
3425243..3426784	-	1542	rrsD	16S ribosomal RNA of rrnD operon
3578950..3579039	-	90	ryhB	Regulatory sRNA mediating positive Fur regulon response; requires Hfq for function; global iron regulator; degraded by RNase E when bound to target
3662887..3662991	+	105	gadY	sRNA regulator of gadAB transcriptional activator GadX mRNA
3698159..3698222	+	64	rdlD	Antisense sRNA RdlD affects LdrD translation; proposed addiction module in LDR-D repeat, with toxic peptide LdrD

Location	Strand	Length	Gene	Product
3706639..3706715	-	77	proK	Pro tRNA
3720099..3720128	+	30	sokA	misc_RNA
3834245..3834339	+	95	selC	Sec tRNA
3851141..3851280	-	140	istR	sRNAs IstR-1 and IstR-2, tisB regulators
3939831..3941372	+	1542	rrsC	16S ribosomal RNA of rrnC operon
3941458..3941533	+	76	gltU	Glu tRNA
3941727..3944630	+	2904	rrlC	23S ribosomal RNA of rrnC operon
3944723..3944842	+	120	rrfC	5S ribosomal RNA of rrnC operon
3944895..3944971	+	77	aspT	Asp tRNA
3944980..3945055	+	76	trpT	Trp tRNA
3980398..3980474	+	77	argX	Arg tRNA
3980532..3980608	+	77	hisR	His tRNA
3980629..3980715	+	87	leuT	Leu tRNA
3980758..3980834	+	77	proM	Pro tRNA
3984455..3984626	+	172	glmZ	sRNA activator of glmS mRNA, Hfq-dependent
4033554..4035095	+	1542	rrsA	16S ribosomal RNA of rrnA operon

Location	Strand	Length	Gene	Product
4035164..4035240	+	77	ileT	Ile tRNA
4035283..4035358	+	76	alaT	Ala tRNA
4035542..4038446	+	2905	rrlA	23S ribosomal RNA of rrnA operon
4038540..4038659	+	120	rrfA	5S ribosomal RNA of rrnA operon
4047922..4048030	+	109	spf	Spot 42 sRNA; antisense regulator of galK translation
4049059..4049303	+	245	csrC	CsrC sRNA sequesters CsrA, a carbon flux regulator; also affects biofilms and motilty
4156308..4156417	-	110	oxyS	OxyS sRNA activates genes that detoxify oxidative damage
4164682..4166223	+	1542	rrsB	16S ribosomal RNA of rrnB operon
4166395..4166470	+	76	gltT	Glu tRNA
4166664..4169567	+	2904	rrlB	23S ribosomal RNA of rrnB operon
4169660..4169779	+	120	rrfB	5S ribosomal RNA of rrnB operon
4173411..4173486	+	76	thrU	Thr tRNA
4173495..4173579	+	85	tyrU	Tyr tRNA
4173696..4173770	+	75	glyT	Gly tRNA

Location	Strand	Length	Gene	Product
4173777..4173852	+	76	thrT	Thr tRNA
4206170..4207711	+	1542	rrsE	16S ribosomal RNA of rrnE operon
4207797..4207872	+	76	gltV	Glu tRNA
4208066..4210969	+	2904	rrlE	23S ribosomal RNA of rrnE operon
4211063..4211182	+	120	rrfE	5S ribosomal RNA of rrnE operon
4275950..4276089	-	140	ryjA	Novel sRNA, function unknown
4360574..4360649	-	76	pheU	Phe tRNA
4390383..4390458	+	76	glyV	Gly tRNA
4390495..4390570	+	76	glyX	Gly tRNA
4390606..4390681	+	76	glyY	Gly tRNA
4494428..4494512	+	85	leuX	Leu tRNA
4526000..4526089	+	90	ryjB	Novel sRNA, function unknown
4577858..4577934	+	77	symR	sRNA destabilizing divergent and overlapping symE mRNA
4604102..4604188	-	87	leuV	Leu tRNA
4604223..4604309	-	87	leuP	Leu tRNA
4604338..4604424	-	87	leuQ	Leu tRNA

APPENDIX B

Position specific scoring matrixes of sigma factor binding motif in *B. subtilis*

Sigma factor A

position	A	C	G	T
1	-1.53	-1.29	-1.72	0.98
2	-1.18	-1.41	-0.95	0.91
3	-1.38	-0.64	1.30	-0.72
4	0.59	-0.28	-1.08	-0.34
5	-0.27	0.98	-0.70	-0.55
6	0.47	-0.64	-0.95	0.02
gap	13..22	basepairs		
1	-2.01	-1.13	-1.81	1.00
2	1.05	-2.47	-2.30	-1.67
3	-0.40	-0.04	-0.91	0.53
4	0.78	-1.29	-0.44	-0.84
5	0.73	-0.15	-0.73	-1.18
6	-1.83	-2.68	-2.30	1.06

Sigma factor B

position	A	C	G	T
1	-2.19	-1.66	1.56	-1.61
2	-0.51	-2.19	-0.42	0.74
3	-1.41	-1.32	-0.86	0.92
4	-1.24	-2.19	-2.19	1.00
5	0.27	-0.03	-0.42	-0.10
6	0.42	-0.55	-0.21	-0.20
gap	12..17	basepairs		
1	-1.41	-1.06	1.46	-1.41
2	-2.19	-1.66	1.58	-1.86
3	-1.61	-1.32	1.55	-2.19
4	0.51	-0.86	-2.19	0.16
5	0.97	-2.19	-1.32	-1.24
6	0.33	-0.69	-1.32	0.27

Sigma factor D

position	A	C	G	T
1	-0.87	-1.87	-1.17	0.90
2	0.70	-0.77	-1.87	-0.26
3	0.74	-0.25	-1.17	-0.87
4	0.78	-1.87	-1.17	-0.38
gap	12..16	basepairs		
1	-0.87	-0.77	1.23	-0.68
2	-1.10	1.40	-1.17	-1.10
3	-1.87	1.54	-1.17	-1.87
4	-1.10	-1.17	1.47	-1.87
5	1.00	-1.87	-1.87	-1.42
6	-1.42	-1.87	-1.87	1.00
7	1.00	-1.17	-1.87	-1.87
8	-0.38	-1.87	-1.17	0.78

Sigma factor E

position	A	C	G	T
1	-0.85	-1.73	0.54	0.49
2	-0.10	1.05	-1.73	-0.68
3	0.73	-0.30	-1.73	-0.60
4	-0.76	-0.78	-1.73	0.84
5	0.58	-0.51	0.09	-0.95
6	0.07	-0.40	-0.51	0.30
7	-0.35	0.15	-0.95	0.44
gap	13..15	basepairs		
1	-0.53	1.34	-2.24	-1.18
2	1.01	-1.40	-1.40	-2.24
3	-1.92	-2.24	-1.73	1.04
4	1.05	-2.24	-2.24	-1.92
5	-0.47	0.58	-0.30	0.07
6	0.65	-0.95	0.09	-0.95
7	0.10	0.15	-0.64	0.07
8	-0.60	-1.73	-1.40	0.85

Sigma factor F

position	A	C	G	T
1	-1.79	-1.79	1.35	-0.37
2	-1.30	0.37	-1.05	0.61
3	0.55	-0.33	-1.79	-0.11
4	-1.30	-1.79	-1.79	0.99
5	0.72	-1.79	0.08	-0.98
gap	13..16	basepairs		
1	-1.79	-0.63	1.39	-0.98
2	-0.37	-0.63	1.07	-0.73
3	-0.11	0.24	0.24	-0.23
4	0.55	0.24	-0.63	-0.98
5	0.86	-1.79	-1.05	-0.73
6	0.19	-0.10	-1.05	0.19
7	0.61	0.08	-0.33	-1.30
8	0.42	0.37	-1.79	-0.37
9	-1.79	-1.79	-1.05	0.99
10	0.55	-1.79	0.60	-1.30

Sigma factor G

position	A	C	G	T
1	-0.87	-0.58	1.22	-0.75
2	-0.26	0.59	-0.75	0.07
3	0.95	-2.13	-1.57	-0.99
4	-1.14	-2.13	-2.13	0.99
5	0.60	-1.57	0.37	-0.99
gap	16..18	basepairs		
1	-0.08	0.95	-0.75	-0.75
2	1.02	-1.57	-2.13	-1.78
3	0.16	-0.58	-0.96	0.36
4	0.55	-0.43	0.00	-0.75
5	0.36	0.59	-0.96	-0.75
6	-2.13	-1.57	-1.57	1.02
7	1.01	-1.57	-1.57	-1.78

Sigma factor H

position	A	C	G	T
1	1.02	-1.77	-1.77	-1.77
2	-0.95	-1.77	1.47	-1.77
3	-1.77	-1.77	1.56	-1.77
4	0.90	-1.77	-0.30	-1.77
5	0.23	-1.77	0.12	0.14
6	0.31	-1.77	-1.02	0.39
7	-0.34	-1.02	-0.60	0.65
gap	13..14	basepairs		
1	-0.07	0.90	-0.60	-0.70
2	-0.95	-1.77	1.38	-0.95
3	1.02	-1.77	-1.77	-1.77
4	1.02	-1.77	-1.77	-1.77
5	-0.20	-0.60	-1.02	0.59

Sigma factor K

position	A	C	G	T
1	0.07	0.77	-1.18	-0.43
2	0.76	-1.54	-0.39	-0.71
3	-2.11	1.58	-1.54	-2.11
4	0.44	0.64	-1.18	-1.10
gap	15..17	basepairs		
1	-0.61	1.32	-0.91	-1.48
2	0.92	-1.54	-0.71	-1.48
3	-1.74	-2.11	-2.11	1.04
4	0.92	-2.11	-1.54	-0.82
5	-0.21	0.05	-0.39	0.29
6	-0.03	-0.04	0.53	-0.43
7	0.21	0.21	-0.14	-0.35
8	-0.95	-1.18	-2.11	0.92
9	0.37	-2.11	0.35	-0.21

Sigma factor W

basepairs	A	C	G	T
1	-1.49	-0.86	-0.58	0.86
2	-0.96	-1.92	1.49	-1.92
3	0.95	-1.26	-1.26	-1.49
4	1.01	-1.92	-1.92	-1.49
5	1.04	-1.92	-1.92	-1.92
6	-1.92	1.58	-1.92	-1.92
7	-0.62	0.67	-0.86	0.22
8	-0.96	-1.26	-1.26	0.89
9	-1.19	-1.26	-0.36	0.82
10	-0.96	-1.26	-0.36	0.79
gap	12..13	basepairs		
1	-1.19	1.52	-1.92	-1.92
2	-1.92	-1.92	1.58	-1.92
3	-1.92	-1.92	-1.92	1.04
4	0.75	0.33	-1.92	-1.92
5	-0.96	-0.86	-0.86	0.82
6	0.59	-1.26	-1.26	0.01

Sigma factor X

basepairs	A	C	G	T
1	-0.99	-0.70	-0.24	0.71
2	-0.62	-1.58	1.39	-1.58
3	0.53	-0.70	-1.58	0.04
4	0.93	-1.58	-1.58	-0.99
5	0.99	-1.58	-1.58	-1.58
6	-1.58	1.52	-1.58	-1.58
gap	16..17	basepairs		
1	-0.99	1.46	-1.58	-1.58
2	-1.58	-1.58	1.52	-1.58
3	-0.35	-1.58	-1.58	0.79
4	-1.58	1.52	-1.58	-1.58
5	-0.14	-1.58	-1.58	0.71
6	0.71	-0.70	-1.58	-0.35

VITA

Mr. Natapol Pornputtapong was born on June 15, 1981 in Bangkok, Thailand. He was graduated Bachelor of Science in Pharmacy in 2004 from Faculty of Pharmaceutical Sciences, Chulalongkorn University.