UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

SEQUENCING AND COMPARATIVE GENOMICS OF *LEPTOSPIRA INTERROGANS* SEROVAR *POMONA* AND *LEPTOSPIRA KIRSCHNERI* SEROVAR *GRIPPOTYPHOSA*

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

JIAXIN QU
Norman, Oklahoma
2007

UMI Number: 3273881

UMI

UMI Microform 3273881

Copyright 2007 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

SEQUENCING AND COMPARATIVE GENOMICS OF
LEPTOSPIRA INTERROGANS SEROVAR *POMONA* AND
LEPTOSPIRA KIRSCHNERI SEROVAR *GRIPPOTYPHOSA*

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

By

____
Dr. Bruce A. Roe, Chair

____
Dr. Paul F. Cook

____
Dr. Han Wang

____
Dr. Ann H. West

____
Dr. Wai Tak Yip

©Copyright by JIAXIN QU 2007
All Rights Reserved.

Acknowledgements

I would like to thank my advisory committee, Drs. Bruce Roe, Paul Cook, Richard Taylor, Han Wang, and Ann West for their support, advice, and participation of my dissertation work. Especially, I would like to express my sincere gratitude to my major professor Dr. Roe for his teaching, inspiration, motivation, and patience to me through these years.

I am also thankful for all the members of Dr. Roe's laboratory at the Advanced Center for Genome Technology for their unselfish help. A special thanks goes to Dr. Fares Najar who first trained me when I started to work, taught me how to mine useful information from the genomic sequences of two *Leptospira* species, and developed various scripts to speedup large-scale data analysis. My thanks also goes to Hongshing Lai, Steve Kenton, and Dr. Axin Hua whose computer expertise was extremely instrumental for my data analysis. My thanks also goes to Jianfeng Li, Hung-Chun (James) Yu, Christopher Lau, and Shelly Oommen for the long term and beneficial scientific discussion.

I would like to thank my family for their love and support. My family has been separated for the past five years. There is no any doubt that this has been really hard on my wife, Chunxiao (Lily) Fu as she has to work and take care of our little son, Andy Fu, by herself at most time. She even carefully checked and edited all the references of my dissertation. Speaking of my family, I have to thank my major professor Dr. Roe again. During that hard time, I even tried to quit my study. It was Dr. Roe who creatively suggested me to take a flexible schedule, allowing me to work

on the weekends and stay at home on weekdays to take care of our little son. So Dr. Roe really saved my dream to earn a Ph. D.

I also am thankful for my nice friends in Tulsa, especially Liang Xi, Shunxin Wang, Chibing Tan, Qian Wang, and Ajula Vaid, for always being there when my family needed them.

Finally I would like to thank our collaborator Dr. Yung-Fu Chang, at Cornell University, for providing genomic DNA of *L. pomona* and *grippotyphosa* and advice throughout this research work.

I would like to dedicate this dissertation to my parents-in-law, Xing Fu and Xiuqin Li, my parents, Shiquan Qu and Lan Zhang, my wife, Chunxiao (Lily) Fu, and my little son, Andy Fu.

Table of Contents

List of Tables xi

List of Figures xiv

Abbreviations xvii

Abstract xviii

Chapter I Introduction 1

1.1 Leptospirosis and Leptospiries 1

1.1.1 Taxonomy and Classification 1

1.1.2 Biology of Leptospires 2

1.1.3 Molecular Biology 3

1.1.4 Pathogenesis and Virulence Factors 4

1.1.4.1 Motility 4

1.1.4.2 Attachment 4

1.1.4.3 Invasion 5

1.1.4.4 Iron Uptake 5

1.1.4.5 Immune Response 6

1.2 DNA, Gene, and Genome 6

1.2.1 DNA 6

1.2.2 Gene 8

1.2.3 Genome 8

1.3 Bacterial Genome Sequencing 9

1.3.1 Brief Introduction of Sequencing Methods 9

1.3.2 Strategies of Bacterial Genome Sequencing 13

1.3.2.1 Shotgun Phase 14

1.3.2.2 Closure Phase 14

Chapter II Materials and Methods 16

2.1 Construction of Shotgun Library and Sequencing 16

2.1.1 Small-insert Library 16

2.1.1.1 Fragmentation of Target Genomic DNA 16

2.1.1.2 End Repair and Size Selection 17

2.1.1.3 Ligation 18
2.1.1.3.1 Vector pUC18 18
2.1.1.3.2 Ligation of DNA Fragments with pUC18 19
2.1.1.4 Transformation 19
2.1.1.5 Automatic Isolation of Subclone DNA 21
2.1.1.6 Reaction and Clean Up 22
2.1.1.7 Sequencing 23
2.1.2 454/Roche GS-20 instrument Library 23
2.1.3 Large-insert Library 24
2.1.3.1 Vector pCC1 26
2.1.3.2 Methods to Insert Foreign DNA into Host Cells 26
2.1.3.3 DNA Isolation 27
2.2 Sequence Assembly 27
2.2.1 ABI Sequencing Reads 27
2.2.2 454/Roche GS-20 instrument Sequencing Reads 30
2.3 Gap Closure 31
2.3.1 Uniplex PCR-based Approach 31
2.3.2 Multiplex PCR-based Approach (MPCR) 34
2.3.2.1 Large-insert Clone Scaffolding 34
2.3.2.2 Genome Comparison 35
2.4 Sequencing Data Analysis 36
2.4.1 Annotation 36
2.4.1.1 ORFs Prediction 36
2.4.1.2 Assignment of Biological Function to the Putative ORFs 37
2.4.1.3 Assignment of Enzyme Commission (E.C.) Numbers 37
2.4.1.4 Identification of Secreted Proteins 38
2.4.1.5 Identification of Transmembrane Protein 38
2.4.1.6 Functional Domain Search 39
2.4.1.7 Identification of tRNA Genes 39
2.4.1.8 Identification of Ribosomal Frameshift 40
2.4.2 Reconstruction of Metabolic Pathways 41

2.4.2.1 Kyoto Encyclopedia of Genes and Genomes (KEGG) 41
2.4.2.2 Pathway Holes Filling 41
2.4.2.2.1 Domain Analysis 41
2.4.2.2.2 Isozymes Search 42
2.4.2.2.3 Literatures Mining 42
2.4.3 Virulence Genes 43
2.4.3.1 Definition of Virulence Genes 43
2.4.3.2 Identification of Virulence Genes 43
Chapter III Results and Discussion 44
3.1 Genome Overview 44
3.2 Reconstruction of Metabolic Pathways 52
3.2.1 Carbohydrate, Lipid Catabolism, and Oxidative Phosphorylation 52
3.2.1.1 Glycolysis 52
3.2.1.2 TCA Cycle 53
3.2.1.3 Pentose Phosphate Pathway (PPP) 56
3.2.1.4 Beta-oxidation of Long-chain Fatty Acids 58
3.2.1.5 Oxidative Phosphorylation 59
3.2.2 Lipid Biosynthesis 60
3.2.2.1 Fatty Acids Synthesis 60
3.2.2.2 Synthesis of Phospholipids 62
3.2.3 Biosynthesis of Amino Acids 64
3.2.3.1 Glutamate, Glutamine, Aspartate, Asparagine, and Alanine 64
3.2.3.2 Branched-chain Amino Acids 65
3.2.3.3 Methionine, Threonine, and Lysine 66
3.2.3.4 Serine, Glycine, and Cysteine 68
3.2.3.5 Aromatic Amino Acids 69
3.2.3.6 Arginine and Proline 70
3.2.3.7 Histine 72
3.2.4 Nucleotide Metabolism 72
3.2.4.1 Purine Nucleotides 73
3.2.4.2 Pyrimidine Nucleotides 74

3.2.5 Biosynthesis of Cofactors and Vitamins 75
3.2.5.1 Riboflavin Biosynthesis 75
3.2.5.2 Folate Biosynthesis 76
3.2.5.3 Thiamine Biosynthesis 77
3.2.5.4 Biotin Biosynthesis 78
3.2.5.5 Biosynthesis of Pantothenate and Coenzyme A 79
3.2.6 Macromolecular Metabolism 80
3.2.6.1 DNA Metabolism 81
3.2.6.2 RNA Metabolism 82
3.2.6.2.1 Transcription 82
3.2.6.2.2 Processing of Ribosomal and Transfer RNAs 83
3.2.6.2.3 Post Transcriptional Modification of Ribosomal and TransferRNA Precursors to Mature RNAs 86
3.2.6.3 Aminoacyl-tRNA Biosynthesis 87
3.2.6.4 Protein Biosynthesis 88
3.2.6.5 Protein Degradation 91
3.2.7 Cell Wall 92
3.2.7.1 Peptidoglycan Biosynthesis 92
3.2.7.2 Lipopolysaccharides (LPS) 94
3.2.8 Transport Proteins 96
3.2.8.1 Cation and Anion Transporters 97
3.2.8.2 Phosphotransferase System (PTS) 97
3.2.9 Protein Export and Secretion 99
3.3 Virulence Genes 104
3.3.1 Chemotaxis 104
3.3.2 Motility 106
3.3.3 Adherence 108
3.3.4 Invasion 109
3.3.5 Iron Acquisition and Utilization 110
3.3.6 Stress Response 113
3.3.7 Lipopolysacharides 114

3.3.8 Antibiotic Resistance 117

Chapter V Conclusion 119

References 128

Appendix A 454/Roche GS-20 instrument DNA Sequencing Protocol 144

Appendix B Fosmid DNA Isolation and End Sequencing Protocol 150

List of Tables

Table 3.1.1 General genome features of four *Leptospira* species 44
Table 3.1.2 Protein-coding genes comparison among four *Leptospira* species. 44
Table 3.1.3 Comparison of operon distribution among the small chromosomes of four *Leptospira* species. 51
Table 3.1.4 Comparison of operon distribution among the large chromosomes of four *Leptospira* species. 51
Table 3.2.1 Enzymes involved in the glycolic pathway, their EC numbers, a description, and an indication of the presence in four *Leptospira* species. 53
Table 3.2.2 Enzymes involved in the TCA, their EC numbers, a description, and an indicator of the presence in four *Leptospira* species. 54
Table 3.2.3 Enzymes involved in 3-hydroxypropionate pathway, their EC numbers, a description, and an indicator of the presence in four *Leptospira* species. 55
Table 3.2.4 Enzymes involved in the non-oxidative arm of Pentose Phosphate Pathway in four *Leptospira* species. 56
Table 3.2.5 Enzymes involved in NADPH biosynthesis, their EC numbers, a description, and an indication of the presence in four *Leptospira* species. 57
Table 3.2.6 Enzymes involved in the pathway of beta-oxidation of long-chain fatty acid, their EC numbers, a description, and an indicator of the presence in four *Leptospira* species. 59
Table 3.2.7 Enzymes involved in the predicted oxidative phosphorylation pathway in four *Leptospira* species. 60
Table 3.2.8 Enzymes involved in the fatty acid biosynthesis pathway present in four *Leptospira* species. 62
Table 3.2.9 Enzymes involved in the phosphoglyceride synthesis present in four *Leptospira* species. 64
Table 3.2.10 Enzymes involved in the biosynthesis of valine, leucine, and 66

isoleucine.
Table 3.2.11 Enzymes involved in the biosynthesis of lysine, threonine, and methionine. 68
Table 3.2.12 Enzymes involved in the synthesis of serine, glycine, and cysteine in four *Leptospira* species. 69
Table 3.2.13 Enzymes involved in aromatic amino acid biosynthesis. 70
Table 3.2.14 Enzymes involved in the biosynthesis of arginine and proline. 71
Table 3.2.15 Enzymes involved in the biosynthesis of histine. 72
Table 3.2.16 Enzymes involved in purine metabolism of four *Leptospira* species. 73
Table 3.2.17 Enzymes involved in pyrimidine metabolism of four *Leptospira* species. 74
Table 3.2.18 Enzymes involved in riboflavin biosynthesis, their EC numbers, a description, and an indication of the presence in four *Leptospira* species. 75
Table 3.2.19 Enzymes involved in folate biosynthesis, their EC numbers, a description, and an indication of the presence in four *Leptospira* species. 76
Table 3.2.20 Enzymes involved in thiamine biosynthesis, their EC numbers, a description, and an indication of the presence in four *Leptospira* species. 78
Table 3.2.21 Enzymes involved in biotin biosynthesis, their EC numbers, a description, and an indication of the presence in four *Leptospira* species. 79
Table 3.2.22 Enzymes involved in pantothenate and CoA synthesis in the four *Leptospira* species. 80
Table 3.2.23 Components of DNA polymerase III 81
Table 3.2.24 Complete list of all proteins involved in replication of *Leptospira* species. 82
Table 3.2.25 List of encoded proteins involved in transcription of *Leptospira* species. 83
Table 3.2.26 The distribution of ribosomal RNAs in four *Leptospira* species. 84

Table 3.2.27 Enzymes involved in the rRNA/tRNA processing of *Leptospira* species. 84
Table 3.2.28 RNA modification enzymes. 86
Table 3.2.29 Proteins involved in protein synthesis in four *Leptospira* species. 91
Table 3.2.30 Protease distribution in *Leptospira* species. 92
Table 3.2.31 Enzymes involved in peptidoglycan biosynthesis. 94
Table 3.2.32 Enzymes involved in Lipid A and KDO biosynthesis. 96
Table 3.2.33 Cation and anion transporters found in four *Leptospira* species. 97
Table 3.2.34 Proteins involved in the phosphotransferase system in four *Leptospira* species. 99
Table 3.2.35 Proteins involved in the Sec dependent pathway in four *Leptospira* species. 102
Table 3.2.36 Proteins involved in the Sec dependent pathway in four *Leptospira* species. 104
Table 3.3.1 Genes involved in the chemotaxis system of four *Leptospira* species. 105
Table 3.3.2 Genes involved in the motility of four *Leptospira* species. 107
Table 3.3.3 Genes involved in adherence of four *Leptospira* species. 109
Table 3.3.4 Genes involved in invasion of four *Leptospira* species. 110
Table 3.3.5 Enzymes involved in the heme synthesis pathway in four *Leptospira* species. 111
Table 3.3.6 Genes involved in the iron uptake systems of four *Leptospira* species. 112
Table 3.3.7 Stress response proteins identified from four *Leptospira* species. 114
Table 3.3.8 Genes involved in the *rfb* locus are compared among four *Leptospira* species. 117
Table 3.3.9 The putative antibiotic resistance genes identified in four *Leptospira* species. 118
Table 4.1 Strains of *Leptospira* species used in DNA sequence analysis studies. 121

List of Figures

Figure 1.1.1 Scanning electron micrograph of leptospiries 3

Figure 1.2.1 The structure of DNA 7

Figure 1.2.2 The evolution of Crick's Central Dogma from the 1950s to today 8

Figure 2.1.1 pUC18 plasmid 19

Figure 2.1.2 General overview of 454/Roche GS-20 instrument DNA sequencing protocol 23

Figure 2.1.3 Construction of a fosmid library using the EPICENTRE CopyControl™ BAC Cloning Kit 25

Figure 2.1.4 Map of vector pCC1 26

Figure 2.2.1 DNA sequence assembly pipeline 27

Figure 2.2.2 Assembly visualization by Consed 29

Figure 2.2.3 Exgap of *L. pomona* during final closure 30

Figure 2.2.4 Newbler™ Assembler: A whole genome shotgun assembler using flow signals. 31

Figure 2.2.5 A strategy of gap closure based on knowledge in gene organizations in operons or the existence of split genes between contigs 34

Figure 2.3.1 Scaffolding of fosmid DNA 35

Figure 2.4.1 ORF Prediction by Artemis 36

Figure 3.1.1 The small chromosome organization of one *Leptopsira* species compared with that of the other three. 47

Figure 3.1.2 The large chromosome organization of one *Leptopsira* species compared with that of the other three. 48

Figure 3.1.3 The operon distribution among the small chromosomes of four *Leptospira* species: (a) *pomona*, and (b) *grippotyphosa,* (c) *lai*, and (d) *copenhageni*. 49

Figure 3.1.4 The operon distribution among the large chromosomes of four *Leptospira* species. 50

Figure 3.2.1 The predicted glycolic pathway in the four *Leptospira* species 53

Figure 3.2.2 The predicted pathway of TCA cycle in the four *Leptospira* species. 54

Figure 3.2.3 Proposed 3-hydroxypropionate pathway in the four *Leptospira* 55

species
Figure 3.2.4 The non-oxidative arm of Pentose Phosphate Pathway in the four *Leptospira* species 56
Figure 3.2.5 NADPH biosynthesis of the four *Leptospira* species. 57
Figure 3.2.6 The predicted pathway of beta-oxidation of long-chain fatty acid in the four *Leptospira* species. 58
Figure 3.2.7 Possible pathways for aerobic respiration in the four *Leptospira* species. 59
Figure 3.2.8 Fatty acid biosynthesis in the four *Leptospira* species 62
Figure 3.2.9 Reconstructed pathway of phosphoglyceride synthesis in the four *Leptospira* species 63
Figure 3.2.10 Synthesis of glutamate, glutamine, aspartate, asparagine, and alanine. 65
Figure 3.2.11 Synthesis of branched-chain amino acids. 65
Figure 3.2.12 Synthesis of threonine, methionine, and lysine. 67
Figure 3.2.13 Synthesis pathways for serine, glycine, and cysteine. 69
Figure 3.2.14 Synthesis of aromatic amino acids. 70
Figure 3.2.15 Biosynthesis of arginie and proline. 71
Figure 3.2.16 Synthesis of histine. 72
Figure 3.2.17 The purine biosynthetic pathway 73
Figure 3.2.18 The pyrimdines biosynthetic pathway 74
Figure 3.2.19 Figure 3.2.19 Riboflavin biosynthetic pathway 75
Figure 3.2.20 Folate biosynthetic pathway 76
Figure 3.2.21 Thiamine biosynthetic pathway 78
Figure 3.2.22 Biotin biosynthetic pathway 79
Figure 3.2.23 Pantothenate and CoA biosynthesis of *L. pomona*. 79
Figure 3.2.24 Demonstration of DNA replication in the four *Leptospira* species with the proteins involved in the replication process. 82
Figure 3.2.25 Phylogenetic tree of 16s rRNA among four *Leptospira* species. 84
Figure 3.2.26 Schematic illustration of the processing steps of rRNA and tRNA. 85
Figure 3.2.27 Putative pathways for Gln-$tRNA^{Gln}$ and Asn-$tRNA^{Asn}$ synthesis in 88

four *Leptospira* species.
Figure 3.2.28 Schematic representation of the translation process as found in the four *Leptospira* species 90
Figure 3.2.29 Peptidoglycan biosynthesis 93
Figure 3.2.30 A schematic diagram of a lipopolysaccharide molecule 95
Figure 3.2.31 Lipid A biosynthetic pathway 96
Figure 3.2.32 The phosphotransferase system in *E. coli* 98
Figure 3.2.33 Sec dependent pathway in the four *Leptospira* species. 101
Figure 3.2.34 The Tat translocation pathway in bacteria. 103
Figure 3.3.1 The *E. coli* chemotaxis system 105
Figure 3.3.2 Representation of flagellar assembly in *E. coli*. 106
Figure 3.3.3 Heme synthesis pathway as deduced from the genomic sequence of four *Leptospira* species. 111
Figure 3.3.4 Genes involved in the first fragment of the *rfb* locus are compared among four *Leptospira* species: *grippotyphosa*, *pomona*, *copenhageni,* and *lai*. 116
Figure 4.1 The lifecycle of *Leptospira* species in human. 122
Figure 4.2a Overview of the metabolic scheme utilized by *Leptospira* species showing transporters and metabolic pathways. 124
Figure 4.2b Overview of the metabolic scheme utilized by *A. actinomycetemcomitans* showing transporters and metabolic pathways. 125

Abbreviations

aaRS	aminoacyl-tRNA Synthetase
CL	Cardiolipin
COG	Cluster of Orthologous Groups
ddNTP	Dideoxyribonucleoside Triphosphate
dITP	Deoxyinosine Triphosphate
DNA	Deoxyribonucleic Acid
FAD	Flavin Adenine Dinucleotide
FMN	Flavin Mononucleotide
FRET	Fluorescence Resonance Energy Transfer
IPTG	Isopropyl Thiogalactoside
KEGG	Kyoto Encyclopedia of Genes and Genomes
LPS	Lipopolysaccharide
MCP	Methyl-accepting Chemotaxis Protein
MPCR	Multiplex PCR
OMP	Orotate Monophosphate
ORF	Open Reading Frame
PA	Phosphatidic Acid
PE	Phosphatidylserine
PEP	Phosphoenol-pyruvate
PG	Phosphatidyl Glycerol
Phrap	Phragment Assembly Program
PPP	Pentose Phosphate Pathway
PRPP	5-phosphoribosyl-1-pyrophosphate
PTS	Phosphotransferase System
SOD	Superoxide Dismutase
Taq	*Thermo aquaticus*
Tat	Twin Arginine Translocation
X-gal	5-bromo-4-chloro-3-indolyl-D-galactoside

Abstract

The ~4.7Mb genomes of *Leptospira interrogans* serovar *pomona* and *Leptosrira kirschneri* serovar *grippotyphosa* respectively were sequenced to understand the molecular basis of leptospiral physiology, virulence, and pathogenesis in leptospirosis.

3,733 genes were predicted in *L. pomona* and 3,828 genes were predicted in *L. grippotyphosa* and compared with those of *L. lai* and *copenhageni*. The large chromosomes of *L. pomona*, *grippotyphosa*, *lai*, and *copenhageni* encode 133, 205, 854, and 98 species specific genes, respectively while 77, 99, and 66 genes are species specific in the small chromosomes of *L. pomona*, *grippotyphosa*, and *lai*, respectively, but none are specific in the *copenhageni* small chromosome.

Nearly 50 metabolic pathways, including glycolysis, pentose phosphate, TCA cycle, oxidative phosphorylation, and ATP synthesis, have been reconstructed for four *Leptospira* species using KEGG. Domain analysis, isozymes search, and literatures mining confirmed that all these pathways investigated were identical in *L. pomona*, *grippotyphosa*, *lai*, and *copenhageni*.

Virulence genes that include methyl-accepting chemotaxis protein, flagellar basal body-associated protein, flagellar motor switch protein, Lig A protein, thermolysin, multiple antibiotic resistance protein, acriflavine resistance protein, and rfb-related genes were identified and compared among the four *Leptospira* species: *pomona*, *grippotyphosa*, *lai*, and *copenhageni*. Based on the unique distribution of their virulence genes, *grippotyphosa* and *pomona* can be paired while *copenhageni*

and *lai* similarly can be grouped. This pairing correlates well with their respective host ranges.

Chapter I

Introduction

1.1 Leptospirosis and Leptospiries

Spirochaetes, consisting long and helical Gram-negative bacteria, distinguished from other bacterial phyla by the presence of flagella, or *axial filaments*, that run lengthwise between the cell membrane and outer membrane. There are three families of spirochaetes, *Treponema pallidum* (Fraser, C. M. et al., 1998), *Borrelia burgdorferi* (Fraser, C. M. et al., 1997), and *Leptospira* species. *T. pallidum*, is the causative agent of the sexually transmitted disease syphilis, *B. burgdorferi* is the causative agent of tick-transmitted Lyme disease, and *Leptospira* species that cause Leptospirosis are one of a group of emerging infectious diseases whose mammalian hosts include both livestock and humans (World Health Organization, 1999). The infection is transferred by exposure to water, damp soil, or vegetation contaminated with the urine of infected animals, because the infected animals pass the pathogens in their urine. Leptospirosis often results in fever, headache, and jaundice, that lead to renal failure, cardiopulmonary failure, and widespread hemorrhaging. In female cattle, pigs, horses, and dogs, leptospirosis often causes abortions, stillbirths and reproductive failure. There also is increasing prevalence of leptospirosis in humans who have come in contact with infected animals.

1.1.1 Taxonomy and Classification

There are two major methods to classify *Leptospira*, serologically and genotypicly. Serological classification is based on the agglutination after cross-absorption with lipopolysaccharide homologous antigen. Using this approach, the

genus *Leptospira* can be divided into two species, *L. interrogans*, comprising all pathogenic strains, and *L. biflexa*, comprising the saprophytic strains isolated from the environment (Faine and Stallman, 1982). Both *L. interrogans* and *L. biflexa* are further divided into many serovars with more than 200 serovars and 60 serovars, respectively. Genotypic classification is based on DNA hybridization, 16 genomospecies of *Leptospira* have been defined that interestingly do not discriminate between pathogenic and nonpathogenic serovars or phenotypic characteristics (Ramadass, et al., 1992).

In my dissertation research, the genome sequences of *Leptospira interrogans* serovar *pomona* and *Leptospira kirschneri* serovar *grippotyphosa* were determined. These serovars are most common in the United States. *L. pomona* infects pigs, dogs, cows, sheep, and horses, while *L. grippotyphosa* infects dogs, cows, and sheep, but rarely infects pigs or horses. The genomes of two additional serovars, *L. interrogans* serovar *lai* (Ren, et al., 2003) and *L. interrogans* serovar *copenhageni* (Nascimento, et al., 2004) were determined in 2003 and 2004, respectively. *L. lai* is the causative agent of rural leptospirosis in China, while *L. copenhageni* is responsible for the urban epidemics in Brazil. The hosts for serovar *lai* and *copenhageni* are the striped field mouse (*Apodemus agrarius*) and the domestic rat (*Rattus norvegicus*), respectively. All of the above four serovars are able to cause severe infection in humans.

1.1.2 Biology of Leptospires

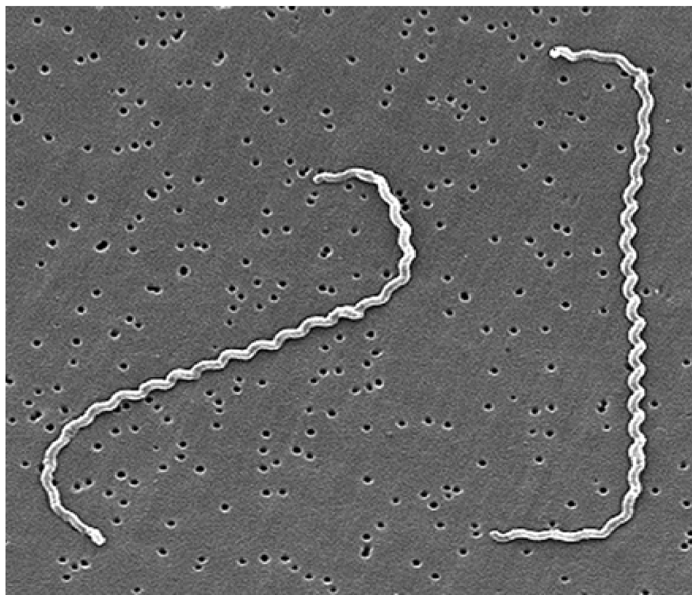


Figure 1.1.1 Scanning electron micrograph of leptospiries (Levett, 2001)

As shown in Figure 1.1.1, leptospires are tightly coiled spirochetes, usually 0.1 μm by 6-20 μm, with helical amplitude of 0.1-0.15 μm and a helical twist of approximately 0.5 μm. All leptospires are indistinguishable in morphology. The cells have a typical double membrane structure in common with other spirochetes, in which the cytoplasmic membrane and peptidoglycan cell wall are closely associated and surrounded by an outer membrane (Haake, 2000). The periplasmic space contains two flagella allowing leptospires to exhibit two distinct forms of movement, translational and nontranslational.

1.1.3 Molecular Biology

The genomic size of most *Leptospira* typically is approximately 5 Mbp (Smibert, 1977), and often is comprised of two chromosomes, a 4.4 Mbp chromosome and a smaller 350 Kbp chromosome, but no other plasmids are reported. Physical maps for serovars *pomona* subtype *kennewicki* and *icterohaemorrhagiae* have been constructed (Zuerner, 1991; Takahashi, et al., 1998). Most leptospires contain two

sets of 16S and 23S rRNA genes but only one set of 5S rRNA gene (Takahashi, et al., 1998), and the rRNA genes usually are widely separated on the large chromosome (Fukunaga and Mifuchi, 1989).

1.1.4 Pathogenesis and Virulence Factors

The mechanisms by which leptospires cause disease are not well understood. A number of putative virulence factors have been suggested, but the function of most of them in pathogenesis remains unexplored. Basically, virulence factors are involved in the following biological process: motility, adhesion to the host cells, penetration of the host cells, iron uptake, and the defense against the host immune response.

1.1.4.1 Motility

Prior to the infection, the pathogen has to move close to the host cell. *E. coli* cells depend on rotation of outside flagella filaments to move (Kim, et al., 2003). But motility of leptospires is dependent on the presence of two periplasmic flagella. Each flagellum is composed of a basal body, hook, and filament. The leptospires can swim through gel-like media, such as connective tissues, that inhibit the motility of most other bacteria (Li, et al., 2000), suggesting that mobility could play an important role as a virulence factor. A periplasmic flagellin gene, *flaB*, of *Leptospira* interrogans was expressed in *E. coli* for the production and antigenic characterization of the protein (Lin, et al., 1999). Two separate *flaB* genes of *Leptospira borgpetersenii* have been identified and characterized (Lin, et al., 2004).

1.1.4.2 Attachment

Usually, the first step in the infection process is the adherence of the pathogen. This process involves the interaction between surface proteins (adhesins) on the

bacteria and membrane components on the host cell surface. Virulent strains of leptospires can attach to the renal epithelial cells. A putative adhesin has been identified. It is a single 36-kDa fibronectin-binding protein specifically expressed by the virulent strains, but is not found in avirulent strains (Merien, et al., 2000). This protein interacts with the gelatine-binding domain of fibronectin and then the integrin cell-binding domain of fibronectin likely binds to the CR3 receptor on host cells (Cinco, et al., 2002).

1.1.4.3 Invasion

Once the pathogen has attached to the cell surface, it will release enzymes that degrade the host cell membrane. Hemolysins from several serovars have been characterized. The hemolysins of serovars *Leptospira pomona*, *hardjo*, *tarassovi*, and *ballum* are sphingomyelinase (Bernheimer and Bey, 1986; del Real, et al., 1989), while phospholipase A1 and lysophospholipase have been identified in *Leptospira biflexa* (Yanagihara, et al., 1984), a phospholipase C has been reported in serovar *canicola* (Yanagihara, et al., 1982), and a novel hemolysin (encoded by SphH) from serovar *lai* that is neither a sphingomyelinase nor phospholipase but rather a pore-forming protein (Lee, et al., 2000; Lee, et al., 2002).

1.1.4.4 Iron Uptake

A TonB-dependent outer membrane receptor, FecA, has been identified in *L. biflexa* (Louvel, et al., 2005) that is analogous to the bacterial ferric dicitrate iron transport system Fec (Ferguson, et al., 2002) that requires a direct physical interaction between TonB and FecA via a TonB box located at the N termini of all TonB-dependent receptors. Interestingly a TonB-independent receptor for iron acquisition,

called FeoB, recently was identified in *L. biflexa* and the inability of FeoB mutants to transport ferric dicitrate and iron sulfate suggests that FeoB is involved in the uptake of both $FeSO_4$ and Fe(3+) dicitrate (Louvel, et al., 2005).

1.1.4.5 Immune Response

Leptospiral lipopolysaccharide (LPS) is the basis of serovar identification and is the apparent target of naturally acquired immunity. Antibodies raised against LPS from different *Leptospira* strains during infections are related to polysaccharide structure in terms of its sugar composition, number, repetitiveness, and ramification (Faine, et al., 1999). Leptospiral LPS is the predominant component that signals to the innate immune system (Werts, et al., 2001). It activates macrophages through a Toll-like receptor 2 (TLR2) pathway rather than TLR4. The latter is the classical signaling receptor of the innate immune system that detects the lipid A moiety of most other Gram-negative LPSs (Poltorak, et al., 1998).

1.2 DNA, Gene, and Genome

1.2.1 DNA

Deoxyribonucleic acid (DNA) is a polymer with nucleotide monomer units that are composed of a 2'-deoxyribose sugar, one of four nitrogenous bases, a purine, adenine (A) or guanine (G), or a pyrimidine, thymine (T) or cytocine (C), attached to C-1 of 2-deoxyribose via a glycosidic bond from N-9 of a purine or N-1 of a pyrimidine, and a phosphate group esterified to the 3' OH group of one nucleotide to the 5' OH group of the next (Figure 1.2.1).

Prokaryotes and eukaryotes have double-stranded DNA genomes where the two anti-parallel chains interwine to form a double helix with the sugar-phosphate

backbone outside and nitrogenous bases inside. This structure of DNA was initially elucidated by Watson and Crick (1953) where they described that base pairing occurs between the two antiparallel DNA chains between adenine (A) and thymine (T) with two hydrogen bonds and cytosine (C) and guanine (G) with three hydrogen bonds. In addition, the double helical structure is stabilized by hydrophobic base stacking interactions between the aromatic rings of adjacent bases.

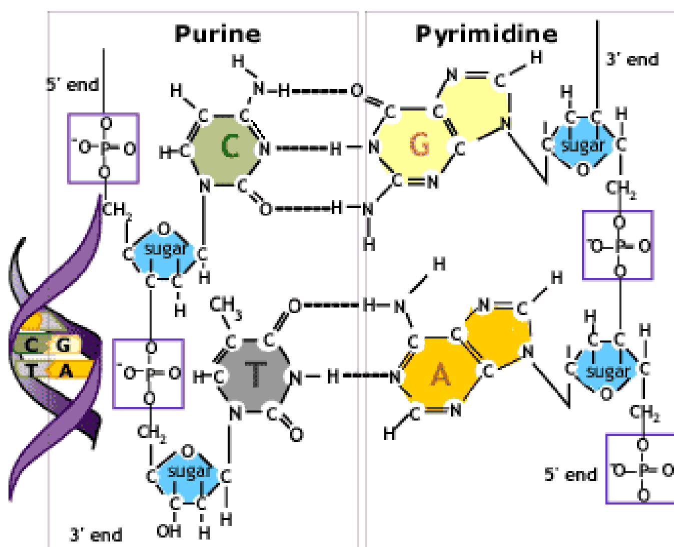


Figure 1.2.1 The structure of DNA
(http://nobelprize.org/chemistry/educational/dna/b/replication/dna_structure.html)

In 1970, Crick proposed the "Central Dogma of Molecular Biology" that predicted genetic information stored in DNA flows through RNA to protein (Crick, 1970). Subsequently this idea has evolved to include more details as illustrated below.

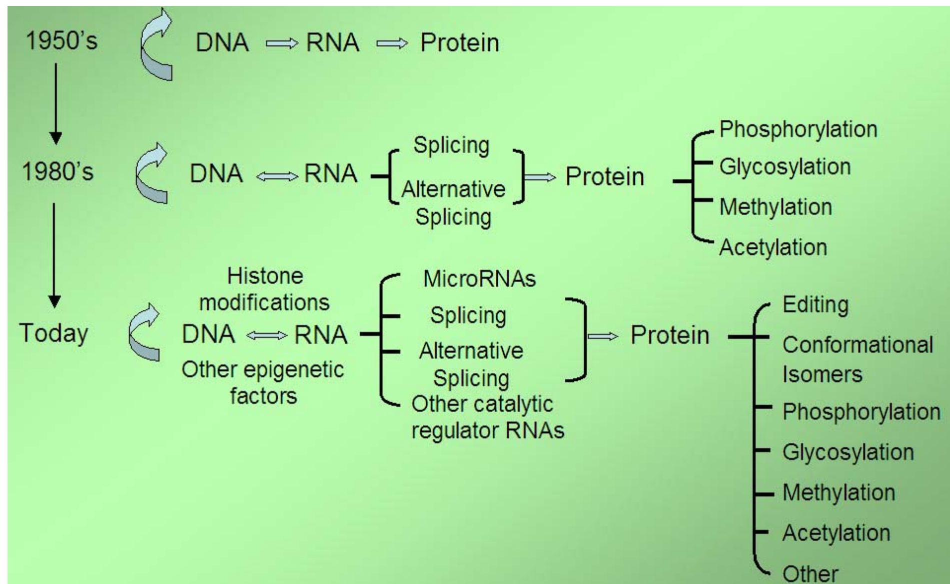


Figure 1.2.2 The evolution of Crick's Central Dogma from the 1950s to today (http://www.genome.ou.edu/5853/CentralDogma.ppt)

1.2.2 Gene

A gene is a sequence of DNA that is transcribed to RNA. Most genes encode proteins, such as enzymes, structural proteins, or regulatory proteins, but some encode stable RNA molecules. The molecular details of genes in prokaryotes and eukaryotes are different, even though they share some basic overall design. The first difference lies in that most prokaryotic genes are polycistronic (one transcription unit contains coding sequences for more than one type of protein or RNA), while eukaryotic genes are monocistronic (one transcription unit contains only a single coding sequence for a protein or stable RNA) with very few exceptions. The second difference is that most eukaryotic genes have exons separated by introns, i.e. they are discontinuous, while the prokaryotic genes are continuous.

1.2.3 Genome

Bacteria have a chromosome that either consists of double-stranded linear DNA (for example, *Borrelia burgdorferi*, the organism that causes Lyme disease

(Fraser, et al., 1997)) or double-stranded circular DNA (for example, *Vibrio cholerae*, the pathogenic bacterium that causes cholera (Clayton, et al., 2000)) and their genome sizes vary between 0.5 and 10 Mbp, but generally average between 2 to 5 Mbp. The smallest bacteria genome identified thus far is from *Mycoplasma genitalium*, an obligate intracellular pathogen with a genome size of 0.58 Mbp (Fraser, et al., 1995). In addition to their genomic DNA, many bacteria also contain smaller, double-stranded, autonomously replicating DNA plasmids that coexist with the larger genomic DNA in a bacterial cell. The plasmids can be transmitted from one cell to another by bacterial conjugation.

Bacterial genomes have compact genetic organizations with, for example, the non-coding DNA in the *E. coli* genome accounting for only 11% of the total. In addition, most bacterial genes are organized into operons where functionally related genes coding for a set of proteins that are involved in a single biochemical activity are expressed coordinatively. The lactose operon, the first to be discovered (Jacob and Monod, 1961), encodes those co-expressed genes involved in converting lactose to glucose. Bacteria that are related to each other often have a similar organization, although they have subtle genetic differences. For example, a comparison of *E. coli* K12 (Serres, et al., 2001) and *E. coli* O157:H7 (Perna, et al., 2001) reveals that O157:H7 has about 1,400 genes not present in K12, and K12 has about 500 genes not present in O157:H7, giving an overall difference of about 7.5×10^4 base pairs.

1.3 Bacterial Genome Sequencing

1.3.1 Brief Introduction of Sequencing Methods

Two basic methods of DNA sequencing were developed in the 1970's. One depends on specific chemical degradation (Maxam and Gilbert, 1977) and the other depends on enzymatic synthesis (Sanger, et al., 1977). Both methods produce nested DNA-fragment sets with one end sheared and the other differing in length by only one nucleotide that can be separated via polyacrylamide gel electrophoresis.

The Maxam-Gilbert sequencing method involves chemical degradation using base-specific chemical cleavage of end labeled DNAs. Each chemical reaction cleaves the DNA preferably at specific bases resulting in a mixture of DNA-strands of different lengths, with one end representing one of the bases and the other end containing a ^{32}P phosphate. By loading the A, C, G, and T specific cleavage reactions in different lanes on the same polyacrylamide gel, fragments in each mixture can be resolved by electrophoresis and the template DNA sequence can be deduced by comparing the four corresponding lanes in an autoradiogram.

The early Sanger dideoxy method required two steps, a labeling-reaction and a termination-reaction. During the labeling-reaction, DNA polymerase I and a short oligonucleotide primer were used to copy a complementary DNA template. The reactions include DNA, primer, four deoxyribonucleoside triphosphates (dNTP: dATP, dTTP, dCTP, and dGTP), trace amounts of ^{32}P-dATP, and one of the four dideoxynucleotide triphosphates (ddNTPs) in each of four tubes. Because ddNTP lacks the 3'-hydroxyl terminal needed to form the next phosphodiester bond, once it is incorporated at the 3' end, the chain can not be extended. In reaction mixtures containing a given ddNTP, chain-terminated fragments of various length are produced, with each fragment extending from the primer to one of the positions

represented by the ddNTP base. The fragments then are separated by electrophoresis and the template DNA sequence can be deduced by comparing the four corresponding lanes in an autoradiogram.

Because the Maxam-Gilbert method gives complex data, requires hazardous chemicals, and is difficult to scale-up, the Sanger's method is more widely used. Since 1977, a number of improvements have been introduced to the Sanger dideoxy method. The first improvement was to replace the radiolabled ^{32}P-dATP with fluorescence dyes. These dyes can be attached to either the four dideoxynucleotide terminators (Prober, et al., 1987) or the primers (Smith, et al., 1986). However the latter is used less frequently today since four separate reactions must be set up for each sequence. Originally, four single dyes (either fluorescein or rhodamine derivatives) (Swerdlow and Gesteland, 1990; Karger, et al., 1991; Metzker, et al., 1996) were utilized as fluorescent labels. The major disadvantage of single dye-labeled terminators is the presence of very small or very large peaks, which can result in errors in automated base-calling. Later, energy-transfer dyes were developed (Rosenblum, et al., 1997) as they have both improved spectral resolution and improved fluorescence yield compared to the single dyes, as they employ fluorescence resonance energy transfer (FRET).

The second important improvement was the development of improved enzymes for DNA sequencing. The enzyme originally used in the Sanger method was the Klenow fragment of *E. coli* DNA polymerase I (Klenoew, et al., 1971). Sequenase (T7) polymerase then was developed to replace the Klenow fragment for DNA sequencing with radioactive nucleotides because of its high processivity and low

error rate (Tabor and Richardson, 1990). The enzyme currently in general use is a variant of *Thermo aquaticus* (*Taq*) DNA polymerase. The advantage of using the thermostable *Taq* polymerase rather than Sequenase is that multiple rounds of sequencing can be performed without the need to add additional enzyme. Other amino acid modifications to the *Taq* DNA polymerase have enabled it to incorporate the fluorescent dye-labeled terminators more evenly and efficiently, resulting in very even peak heights over a sequence read. One of the typically used enzymes is Ampli*Taq* DNA polymerase, FS, a mutant form of *Taq* DNA polymerase that contains two point mutations, F667Y and G46D. The former mutation results in lower discrimination against dideoxynucelotides and leads to a more even peak intensity pattern (Tabor and Richardson, 1995), while the latter removes almost all of the 5'→3' exonuclease activity, thereby eliminating artifacts that arise from the activity.

However, the Sanger's method still faces limitations in both throughput and cost for most future applications. Many research groups have developed alternative approaches for DNA sequencing (Nyren, et al., 1993; Ronaghi, et al., 1996; Jett, et al., 1989). Recently, 454/Roche GS-20 instrument's “sequencing by synthesis” nanotechnology has been successfully used for both confirmatory and *de novo* sequencing (Margulies, et al., 2005). Compared with the conventional Sanger technology, this new method does not require any cloning but instead uses PCR for DNA fragment amplification. Here an emulsion-based PCR (emPCR) is used to amplify DNA fragment immobilized on a bead in water-in-oil micelles that contain the DNA template capture beads and PCR reagents (Dressman, et al., 2003). During the emulsification, millions of micelles are created, and millions of individual PCR

reactions, instead of only one as in the conventional sequencing method, are carried out simultaneously in a single tube. The amplified DNA then is sequenced by a pyrosequencing rather than dideoxy chain termination approach (Ronaghi, et al., 1998). Pyrosequencing is based on detecting the release of pyrophosphate (PP_i) when a nucleotide is incorporated into an extended DNA chain by DNA polymerase. This released PP_i subsequently is converted to ATP by ATP sulfurylase, and the hydrolysis of this ATP provides the energy for luciferase to oxidize luciferin and generate light. Since the added nucleotide is known, the sequence of the template can be determined when a light flash occurs. In addition, it should be emphasized that the sequencing reactions are run in picoliter-sized wells on a slide, containing approximately 1.6 million wells (Leamon, et al., 2003). With massive parallelization, more than 20 million bases can be sequenced per 4.5-hour instrument run, making this technique approximately 30 times faster and cheaper than current Sanger sequencing. However, this new method still has some disadvantages. For example, the average reads initially were only ~ 100 bp but now have been extended by members of our laboratory to ~ 300 bp. Initially all of the pyrosequencing reads were unpaired rather than paired as in conventional double stranded DNA sequencing, but our laboratory now has begun implementing a novel paired-end approach. Although at present the 454/Roche GS-20 instrument platform is better used as a complement to, rather than a replacement of, the existing Sanger sequencing method, this may change in the near future as improvements in read length and read accuracy are introduced.

1.3.2 Strategies of Bacterial Genome Sequencing

Sequencing of a bacterial genome generally begins with a shotgun approach (Anderson, et al., 1982) followed by a more directed closure phase. In the shotgun phase, the genomic DNA is randomly sheared into short fragments. After the ends of the fragments are sequenced, these sequences are assembled into hundreds of contigs to obtain the genomic draft sequence. In the closure phase, various methods, such as sequencing off PCR products and construction of a large-insert library, are used to join the contigs into a single contig representing the whole genome sequence. Gap closure currently is both a time-consuming and often difficult phase in large-scale sequencing whether its based on a Sanger, a pyrosequencing or a combined approach.

1.3.2.1 Shotgun Phase

A hybrid sequencing approach was adopted to carry out this phase in the late study of this dissertation. Two shotgun libraries were constructed. The first was a small-insert 2~4 Kbp library. Both ends of fragments were sequenced on ABI sequencers. The second was a 454/Roche GS-20 instrument library in which the end sequences of DNA fragments of 300-800 bp long are obtained on the 454/Roche GS-20 instrument GS 20 sequencer. The end sequences from the different sequencers were assembled separately and integrated by Phrap (Green, pers. comm.). The details about shotgun library construction, sequencing, and assembly will be discussed in section **2.1 Construction of Shotgun Library and Sequencing**.

1.3.2.2 Closure Phase

A variety of methods typically are employed to join the contigs and reduce the overall error to less than 1 uncertain base per 10 kb. The first method is based on the small-insert library. If the terminal sequences of the single insert DNA fragments

belong to different contigs, it is highly probable that these two contigs are neighbors. The second method is based on a large-insert shotgun library typically a fosmid library, in which the insert DNA fragments are ~40 kb long. This library is used to obtain a "scaffold" of the genome during the closure phase. In addition, the partial sequence of the genome being sequenced also may be conserved in the sequenced genome of a related organism. In this way, a control genome may be of help in the completion of the genome under study.

To insure the sequence is assembled correctly, the predictions of all the potential contig neighbors need to be verified by standard PCR (Barnes, 1994). For the contigs without identified neighbors, multiplex PCR (Claustres, et al., 1989) also can be helpful as it involves PCR amplification using a mixture of primers that are located in the vicinity of contig ends. After sequencing the PCR products, the contigs can be ordered and the gaps filled. The details of the closure phase are discussed later in section **2.3 Gap Closure**.

Chapter II

Materials and Methods

2.1 Construction of Shotgun Library and Sequencing

Genomic DNA of *L. pomona* strain RM211 and *L. grippotyphosa* strain RM52 were supplied by Dr. Yung-Fu Chang at Cornell University. The strategy of whole genome shotgun was employed to obtain the draft sequence of these genomes followed by several directed closure methods.

2.1.1 Small-insert Library

A small-insert library was generated and randomly end-sequenced to obtain the initial genomic draft sequence. At first, the genomic DNA was sheared to 2-4 kb fragments. After end-repair, DNA fragments then were size-selected by purification on a low-melt agarose gel, inserted into pUC18 vectors, and transformed into *E. coli* XL1-Blue MRF'. After the subcloned DNA molecule was isolated, cycle sequencing was performed with forward and reverse universal primers and fluorescent-labeled *Taq* ddNTP terminators. End sequences of the shotgun DNA sub-clones were collected on ABI3700 sequencers, converted to base calls with Phred, and assembled with Phrap (Green, pers. comm.).

2.1.1.1 Fragmentation of Target Genomic DNA

The point of this step is to break the DNA into random small clonable pieces (2-4 kb). The major methods include sonication (Deininger, 1983), partial restriction enzyme digestion (Fitzgerald, et al., 1992), nebulization (Bodenteich, et al, 1994), transposon insertion (Phadnis, et al., 1989), or the HydroShear (Oefner, et al., 1996).

The Hydroshear shearing method was chosen in this dissertation because of the limited amount of DNA available. Genomic DNA was sheared according to the manufacturer recommended procedure, but using a solution cooled to 4°C to give more random library.

After 100 µl of DNA (100-200 µg) is randomly sheared into the fragments, the sheared DNA is precipitated with 2.5× ethanol-acetate (95% ethanol and 0.12 M sodium acetate), washed with 2.5× 75% ethanol, dried in a vacuum, and then dissolved in 27 µl of ddH_2O.

2.1.1.2 End Repair and Size Selection

The mechanical method of DNA fragmentation produces a collection of DNA fragments with heterogeneous ends, 5' and 3' overhangs, with or without phosphate groups at the ends. Therefore, it is necessary to repair the ends of the resulting fragments prior to cloning.

DNA fragments were end-repaired and phosphorylated by Klenow fragment of DNA polymerase and T4 polynucleotide kinase. Klenow fragment is a large fragment of DNA Polymerase I, *E. coli* (Ollis, et al., 1985), that has the 5'→3' polymerase activity and the 3'→5' exonuclease activity, but lacks 5'→3' exonuclease activity. Therefore, if the 5' end of one strand of some certain fragment is overhang, the complementary strand will be elongated. If the 3' end of one strand is overhang, this strand will be cleaved to the blunt end. In this way, all the fragments are made blunt-ended. T4 polynucleotide kinase catalyzes the transfer of phosphate group from the gamma position of ATP to the 5'-hydroxyl terminus of DNA fragments (Richardson, 1981).

The end repair solution contains 27 µl of the sheared DNA, 5 µl of 10× Kinase buffer, 5 µl of 10 mM rATP, 7 µl of 0.25 mM dNTPs, 1 µl of T4 Polynucleotide Kinase (3U/µl), and 2 µl of Klenow DNA polymerase (5U/µl). The reaction was incubated for 30 minutes at 37°C and loaded on to a 1% low-melt agarose gel with molecular size markers (Hind III-digested λ-DNA and HaeI-digested φX174-DNA). Electrophoresis was performed at 120 mA for 1.50 hours. Fragments with the size between 2-4 kb were excised from the gel into a 1.5 ml snap-cap tube and then frozen at –80°C.

After thawing the tube at the room temperature for 5 minutes and centrifugation in a table-top microcentrifuge at 13,000 rpm for 15 minutes, the supernatant containing the DNA fragments was transferred into a new tube. The step of centrifugation was repeated once more and all the supernatants were pooled. Then the DNA fragments were precipitated with 2.5× ethanol-acetate (95% ethanol and 0.12 M sodium acetate), washed with 2.5× 75% ethanol, dried in a vacuum, and then dissolved in 15 µl of ddH_2O.

2.1.1.3 Ligation

2.1.1.3.1 Vector pUC18

pUC18 vector, 2,686 bp, is a small *E. coli* plasmid with the high copy number of ~200 (Takeshita, et al., 1987). pUC18 plasmid contains (1) the replicon *ori* responsible for the replication of plasmid, (2) *bla* gene, coding for beta-lactamase that confers resistance to ampicillin, and (3) region of *E. coli* operon *lac* containing CAP protein binding site, promoter P_{lac}, *lac* repressor binding site, and 5'-terminal part of the *lacZ* gene encoding the N-terminal fragment of beta-galactosidase (Figure 2.1.1).

At first, pUC18 is digested with the endonuclease *Sma*I to generate the correct ends required for cloning. Then the plasmid is dephosphorylated by calf intestine alkaline phosphatase to decrease the possibility of self ligation.

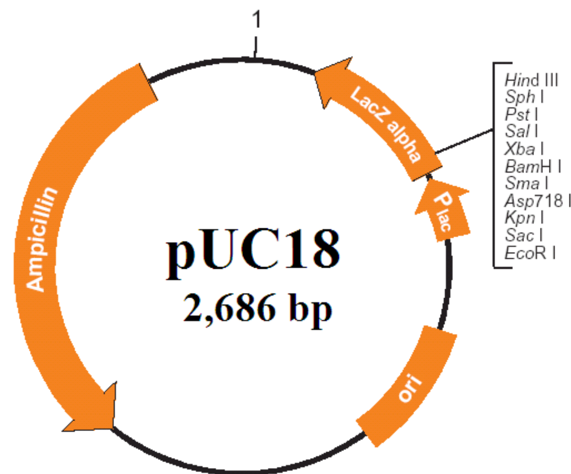


Figure 2.1.1 pUC18 plasmid

(http://www.genomex.com/vector_maps/puc18_map.pdf)

2.1.1.3.2 Ligation of DNA Fragments with pUC18

T4 DNA ligase is used to join the sheared DNA fragments with the vector pUC18 to obtain the subclones. DNA ligase catalyzes formation of a phosphodiester bond between the 5' phosphate of DNA fragment and the 3' hydroxyl of the vector (Pheiffer and Zimmerman, 1983). The ligation reaction solution contained 2 µl (~ 20 ng) of pUC18 vector, 1 µl of 10× ligase buffer, and 1 µl of T4 DNA ligase (400U/µl). A set of different volume of sheared DNA (such as 0.5 µl, 1 µl, and 2 µl) was added to the ligation reactions, respectively. Finally, ddH_2O was added to the final volume of 10 µl. The reactions were incubated at 4°C overnight.

2.1.1.4 Transformation

Following the ligation, the recombinant pU18 was transformed into the electrocompetent cell *E. coli* XL1-Blue MRF'. Here 2 µl of ligation solution was mixed with 40 µl of electrocompetent cell in the cold room. Then the cell-DNA mixture was transferred to an electroporation chamber. After an electrical pulse of 2.5 kV was applied for 5 microseconds at 4°C, 1 ml of cold YENB medium was immediately added to the chamber. After mixing, the cells were transferred to a new Falcon tube and incubated at 37°C for 30 minutes with shaking at 250 rpm. The transformed cells then were harvested by centrifugation at 2,000 rpm for 5 minutes. After decanting the supernatant, 30 µl of 25 mg/ml 5-bromo-4-chloro-3-indolyl-D-galactoside (X-gal) and 30 µl of 25 mg/ml isopropyl thiogalactoside (IPTG) were added to the tube. The resuspended cells were poured on the surface of an LB agar plate with 100 µg/ml ampicillin that then were incubated at 37°C for 18 hours.

After the incubation, two types of colonies, white and blue, are obtained. Without the foreign DNA, the *lacZ* gene region on the vector of pUC18 is intact. The presence of IPTG, an inducer of the *lac* operon, allows transcription of the beta-galactosidase by strongly binding and inhibiting the *lac* repressor. The beta-galactosidase cleaves X-Gal and results in a blue colored metabolite. Insertion of DNA within the *lacZ* gene on the vector of pUC18 inactivates the N-terminal fragment of beta-galactosidase and therefore bacteria carrying recombinant plasmids give rise to white colonies on Petri dishes containing both IPTG and X-gal.

Using Flexys colony picker, white colonies were picked from 20×20 cm Petri dishes into 384-well flat bottom microtiter plates, containing 72 µl TB broth and 8 µl of 10× TB salt supplemented with 100 µg/ml ampicilin. The plates were incubated in

a HiGro incubator at 37°C with shaking at 350 rpm. After 3.5 hours' shaking, an oxygen flow begins with the setting at 0.5 second on and 0.5 minute off. 18 hours later, cell pellets in the plate were harvested at 3,000 rpm for 10 minutes, and frozen overnight at –80°C.

2.1.1.5 Automatic Isolation of Subclone DNA

The subclone DNA was isolated by single acetate cleared lysis method (Birnboim and Doly, 1979). In this procedure, SDS was used to dissolve the phospholipid and protein components of the cell membrane of *E. coli*. Once the cell contents were released, NaOH and RNase A/T1 are added to destroy RNA. Subsequent treatment with KOAc and HOAc forms an insoluble precipitate of SDS/lipid/protein and neutralizes NaOH from the previous step. At neutral pH, the chromosomal DNA is trapped in the SDS/lipid/protein precipitate, while the plasmid DNA remains in solution. After centrifugation, the supernatant was collected, and the plasmid DNA was precipitated by isopropanol. Finally, the pellet was washed with ethanol and resuspended in sterile-distilled deionized water (Micklos and Freyer, 1999).

Briefly, the procedure for the automatic isolation of subclone DNA initially requires using a ZyMark robot to transfer the cell pellets containing 384-well flat-bottom microtiter plates to the bed of a SiClone robot where the cells were suspended in 23 µl of TE-RNase solution (50 mM Tris-HCl, pH 7.6, 0.5 M EDTA, 40 µg/ml RNase A, and 0.04 U/µl RNase T1). After 10 minutes of shaking at 1,000 rpm, 23 µl of lysis buffer (1% SDS and 0.2 M NaOH) was added and the plates were shaken for another 10 minutes at 1000 rpm. Then 23 µl of 3 M KOAc (pH 4.5) was added and

the plates shaken for another 10 minutes at 1,000 rpm and frozen at –80°C overnight. Next, the plates were thawed and centrifuged at 3,000 rpm for 45 minutes in Beckman C56R centrifuge. Using the Vprep, 50 µl of the resulting supernatant was transferred to a new 384-well plate, and the DNA was precipitated by adding 50 µl of 100% isopropanol. Then 50 µl of air bubbles were added to mix the isopropanol. After centrifugation at 3,000 rpm for 30 minutes in Beckman C56R centrifuge, the obtained DNA pellet was washed with 50 µl of 70% ethanol. After centrifuging the plates at 3,000 rpm for 10 minutes and decanting the supernatant, the DNA templates were dried in a vacuum for 10 minutes and then dissolved in 20 µl of sdd-water. An aliquot then was evaluated by electrophoresis on a 1% agarose gel.

2.1.1.6 Reaction and Clean Up

The cycle sequencing method was used to sequence the DNA templates (Mardis and Roe, 1989; Chissoe, et al., 1991). In cycle sequencing, the sequencing reaction is incubated for several cycles consisting of three different temperatures: one for denaturation of double-stranded DNA, one for primer annealing, and one for chain elongation.

Approximately 150-200 ng of subclone DNA was used for the sequencing reaction. The other reagents include 2 µl of 6.5 µM universal forward or universal reverse primer and 2 µl of the 20× diluted ET reaction kit containing Ampli*Taq* FS, dATP, dCTP, dTTP (100 µM each), dITP (500 µM), ddATP, ddCTP, ddTTP, and ddGTP (~0.11 µM each). The reaction mix was thermocycled for 60 cycles of 95°C for 30 seconds, 50°C for 20 seconds, and 60°C for 4 minutes.

Once the cycling reaction was complete, the unincorporated terminators were removed from the sequencing reactions by ethanol-acetate (95% ethanol and 0.12 M sodium acetate) precipitation, followed by a 70% ethanol rinse. Then the plates were dried for 10 minutes at room temperature and stored at –20°C until ready for loading onto the sequencer.

2.1.1.7 Sequencing

The products of sequencing reaction initially were dissolved in 20 µl of ddH_2O, but more recently in 0.1mM EDTA, then loaded on the ABI 3700 or 3730 DNA sequencer. After 2.5 hours of electrophoresis at 6.5 kV, the trace files of DNA sequencing were collected automatically and analyzed using the ABI base caller on the attached computer. These trace files then were transferred to a Unix-based SUN work station for further analysis by Phred and assembled by Phrap that could be viewed by Consed.

2.1.2 454/Roche GS-20 instrument Library

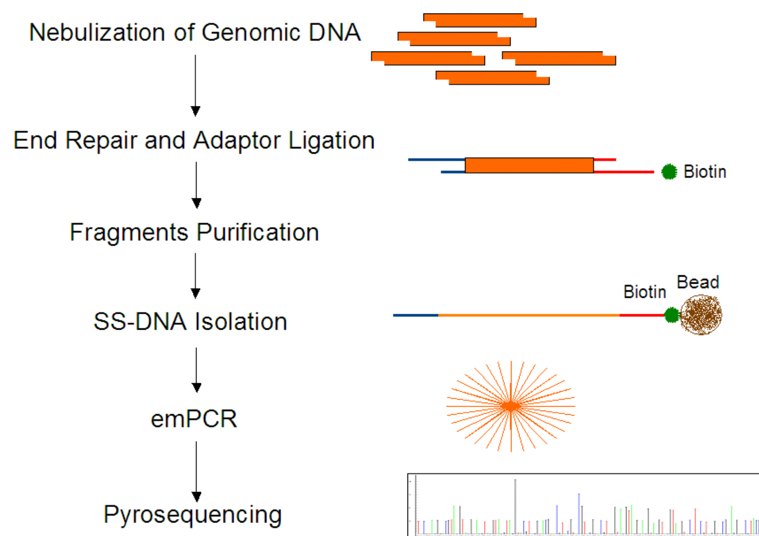


Figure 2.1.2 General overview of 454/Roche GS-20 instrument DNA sequencing protocol.

The flowchart of 454/Roche GS-20 instrument DNA sequencing is described in Figure 2.1.2. Briefly, the DNA sample is nebulized into 300-800 bp fragments, compared with 2-4 kb fragments in the Small-insert Library. Each fragment then is ligated to DNA adaptors that carry the required sequence key, as well as the amplification and sequencing primer sequences. The ligation products are bound to a solid support that permits the isolation of a library of random single-stranded template DNA fragments.

After purification and quantitation of the fragments, the single-stranded DNA library is immobilized onto beads with each bead carrying no more than one amplifiable single-stranded DNA molecule. The entire bead-bound library then is emulsified with the amplification reagents in a water-in-oil mixture, such that bead is captured within its own microreactor (micelle) for the amplification of a single single-stranded DNA fragment.

After amplification, the DNA-carrying beads are spun into the wells of a PicoTiterPlate device such that each well contains only single bead. The loaded PicoTiterPlate device then is inserted into the 454/Roche GS-20 instrument GS 20 sequencer, and sequencing reagents are sequentially flowed over the plate. Information from all the wells of the PicoTiterPlate device is captured simultaneously by the camera, and processed by the associated computer. The detailed 454/Roche GS-20 instrument DNA sequencing protocol is described in **Appendix A 454/Roche GS-20 instrument DNA Sequencing Protocol**.

2.1.3 Large-insert Library

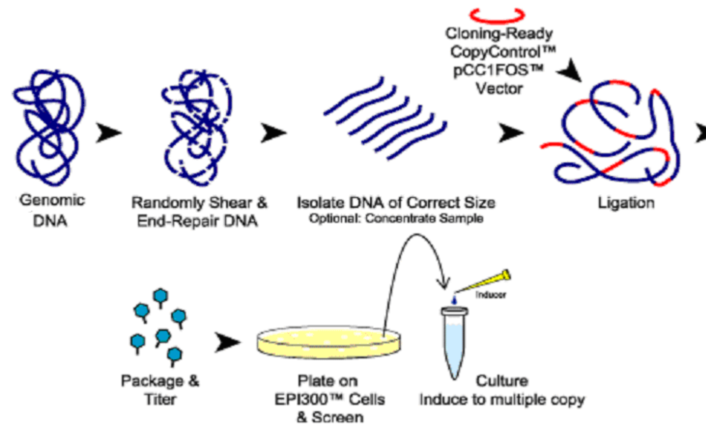


Figure 2.1.3 Construction of a fosmid library using the EPICENTRE CopyControl™ BAC Cloning Kit (http://www.epibio.com/item.asp?ID=385#figure1)

The large-insert library can be used to aid in assembly verification and determination of the future gap sizes and provide a minimal scaffold for order and orientation across assembly gaps. To construct the large-insert library, *L. pomona* genomic DNA first was sheared into ~40 kb fragments by passing it through a 200 µl syringe 50-100 times (Figure 2.1.3). The sheared DNA was end-repaired to generate 5'-phosphorylated blunt ends and size-selected using a low-melt agarose gel. The sheared DNA was ligated into the linearized and dephosphorylated CopyControl pCC1 Cloning-Ready vector, and packaged using ultra-high efficiency MaxPlax™ Lambda Packaging Extracts (>10^9 pfu/µg for phage lambda), and plated on phage T1-resistant EPI100™-T1R *E. coli* plating cells. Individual CopyControl clones were picked from the plate and grown in culture. Then the induction solution (L-arabinose) was added to amplify the clones to high copy number. Finally, both ends of the fosmid DNA were sequenced after purification via the protocol described in **Appendix B Fosmid DNA Isolation and End Sequencing Protocol**. The major differences between small-insert and large insert library construction are discussed below.

2.1.3.1 Vector pCC1

The vector pCC1 (Figure 2.1.4) used for the construction of the large-insert library is a fertility factor (f-factor) fosmid that is capable of containing much larger pieces of DNA, up to 50 kb compared to about 10 kb in a plasmid, and like plasmids, fosmids are double-stranded circular DNA. However, unlike multicopy plasmids, fosmids have much lower copy number.

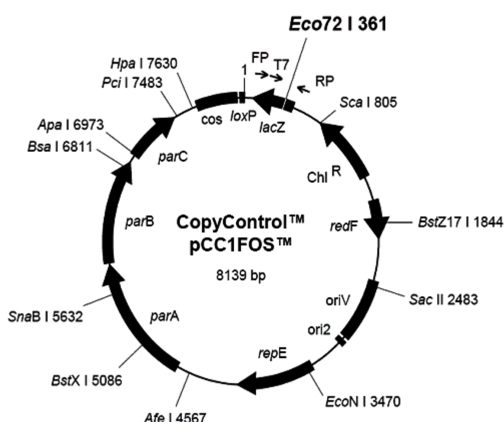


Figure 2.1.4 Map of vector pCC1

Note:
Features of the pCC1 Vector include:
- Chloramphenicol-resistance as an antibiotic selectable marker.
- *E. coli* F factor-based partitioning and single copy origin of replication.
- *ori*V high copy origin of replication.
- Bacteriophage lambda *cos* site for lambda packaging or lambda-terminase cleavage.
- Bacteriophage P1 *lox*P site for Cre-recombinase cleavage.
- Bacteriophage T7 RNA polymerase promoter flanking the cloning site.

2.1.3.2 Methods to Insert Foreign DNA into Host Cells

During the construction of small-insert library, the competent cells obtained the foreign plasmid by the way of transformation (see **2.1.1.4 Transformation**), while transduction was adopted during the construction of large-insert library (see Figure 2.1.3). In this process, a shuttle organism (phage) was used. Here the phage packages

the ligated fosmid DNA and then infects bacteria as a parasite to mediate the transfer and replication of fosmid DNA.

2.1.3.3 DNA Isolation

As discussed in **2.1.1.5 Automatic Isolation of Subclone DNA**, plasmid DNA is isolated automatically by single acetate cleared lysis method. However fosmid DNA initially was isolated manually using the detailed protocol described in **Appendix B Fosmid DNA Isolation and End Sequencing Protocol**. At later stage of this research, fosmid DNA isolation was automated using the Zymark robot.

2.2 Sequence Assembly

2.2.1 ABI Sequencing Reads

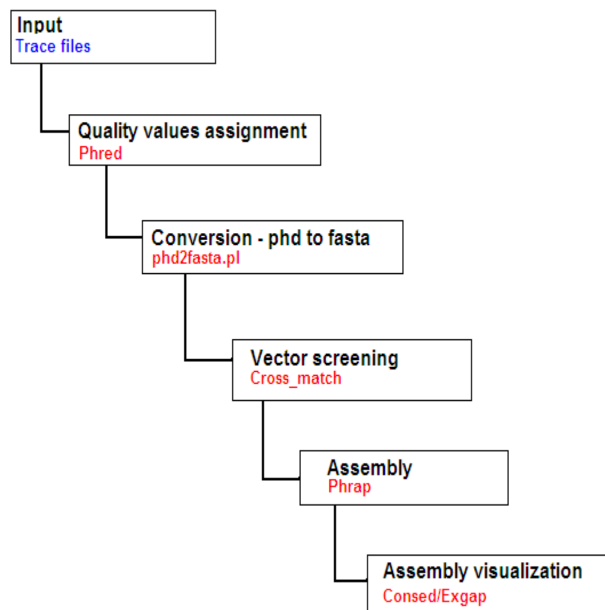


Figure 2.2.1 DNA sequence assembly pipeline

Once the end sequencing trace files of small- or large-insert DNA were transferred to the Unix-based SUN work station, they were re-analyzed by Phred

(Ewing, et al., 1998; Ewing and Green, 1998) and assembled by Phrap (Green, pers. comm.) as shown in Figure 2.2.1.

Phred

The Phred software performs several tasks: reads DNA sequencing trace files from the sequencers, calls bases, assigns a quality value to each called base, and writes these to output files.

The quality value is a log-transformed error probability, specifically

$Q = -10 \log10(P_e)$

where Q and P_e are respectively the quality value and estimated probability error for a base call.

Phrap

Phrap ("phragment" assembly program) then assembles the individual sequence reads into contiguous sequences using the Phred quality values. The main features of Phrap are that it allows use of the entire read and not just the trimmed high quality portion. Phrap uses a combination of user-supplied and internally computed data quality information to improve assembly accuracy in the presence of repeats to construct the contig sequence as a mosaic of the highest quality read segments rather than a consensus. It provides extensive assembly information to assist in trouble-shooting assembly problems, and it handles large datasets while generating output files (such as *.ace) that can be visualized by Consed.

Consed

Consed (Gordon, et al., 1998; Baxevanis and Devison, 2004) is a tool for viewing, editing, and finishing sequence assemblies created with Phrap (Figure 2.2.2). It provides extensive information about the assembled sequence, including the original trace files, the quality of the data, the error rate of the contiguous sequence, and the presence of repetitive sequences. Consed also can view and compare traces, align and compare two regions, look for homologous sequence regions, and pick primers.

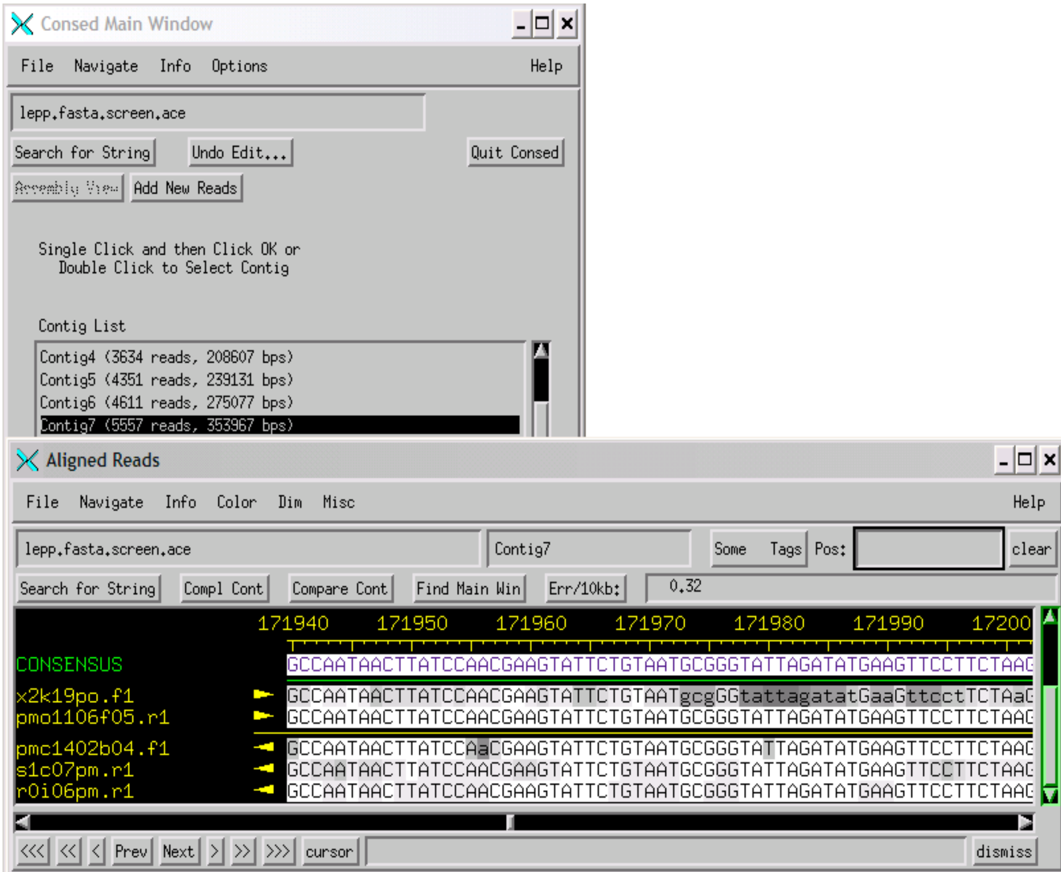


Figure 2.2.2 Assembly visualization by Consed

Exgap

Exgap (http://www.genome.ou.edu/informatics), developed by Dr. Axin Hua in our laboratory, can view paired end relationships from assembled shotgun sequencing data (Figure 2.2.3). After reading the new ace file generated from the

Phrap/Consed assembly suite, Exgap automatically orders contigs based on their forward-reverse mate pairs, lists the subclones covering gaps for selecting gap-closing PCR primers, and locates potential miss-joints for further examination.

Project: lepp
Total #contigs = 13
Total length = 4770.6 kb

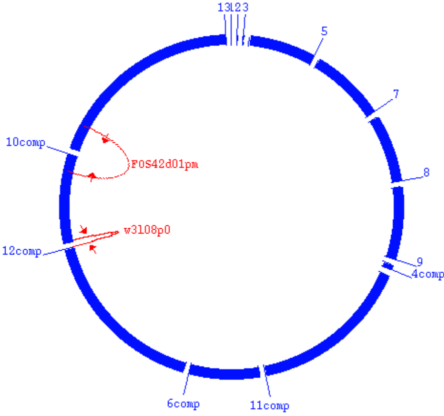


Figure 2.2.3 Exgap of *L. pomona* during final closure
Note: The thick arches represent the contigs. The longer arch means that the corresponding contig has more base pairs. And the thin loop refers to a pair of end sequences of a certain insert, either small or large.

2.2.2 454/Roche GS-20 instrument Sequencing Reads

454/Roche GS-20 instrument sequencing reads are assembled by Newbler Assembler, the new *de novo* assembly software developed by 454 Life Sciences. It operates in flowgram signal space, as opposed to the standard nucleotide space. The advantage lies in that Newbler Assembler utilizes the abundant information stored in the flowgram signals that is lost after base calling. It has three main components, overlap generation, contig layout, and consensus generation (Figure 2.2.4). The overlap generator aligns raw reads in flowgram signal space using a proprietary algorithm. Then all aligned flowgram signals at each position are averaged. Based on the averaged signal, the consensus is generated. Finally, the contigs from the

consensus are converted into chromatogram files with the phred quality values and reassembled with ABI sequencing reads by Phrap.

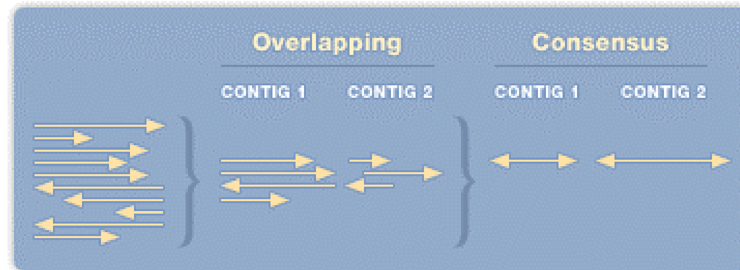


Figure 2.2.4 Newbler™ Assembler: A whole genome shotgun assembler using flow signals (http://www.454.com/enabling-technology/the-software.asp)

2.3 Gap Closure

As mentioned above, the data collected from the sequencers were assembled into a draft of sequence using Phred/Phrap. Once the coverage of the genome reached 8×, alternative strategies were needed to close the remaining gaps and reduce the overall error rate to less than one per 10,000 nucleotides. Viewing the assembly with exgap reveals two types of contigs (see Figure 2.2.3), those with their orientation identified using linking information from forward and reverse reads, and contigs with unknown order and orientation. For example as shown in Figure 2.2.3, the small-insert subclone w3l08po spans the reverse end of Contig 12 and the forward end of Contig 10, while the large-insert subclone FOS42d01pm bridges the reverse end of Contig 13 and the reverse end of Contig 10. Contig 5, Contig 7, and Contig 8 are examples of unordered and unoriented contigs. Different approaches were utilized to join the different type of contigs.

2.3.1 Uniplex PCR-based Approach

For the contigs with the known orientation, uniplex PCR (Mullis and Falooan, 1987) was performed to amplify the gap region. The primer pairs for uniplex PCR were selected at both ends of the contigs with Consed using the following criteria:

(1) Within an acceptable range of 50-65°C, a pair of primers should have similar melting temperatures with 40-60% GC content.

(2) Every primer base should have a corresponding consensus quality value at or above the threshold (default 30), because the data with a quality of 20 is not completely accurate.

(3) Every primer picked should be unique by pre-screening against either a single contig or the entire sequencing project target contig set.

(4) No primer should contain more than 4 contiguous base pairs of homology to itself or counterpart, in order to decrease the possibility of the primer dimers.

Once primers were picked and synthesized, 50 μl of PCR reaction typically containing 2 μl of each 6.5 μM primer, 5 μl of 10× PCR buffer (containing 500 mM KCl, 100 mM Tris-HCl, and 10 mM $MgCl_2$, pH 7.6), about 10 ng of genomic DNA, 5 μl of 2 mM dNTPs, 2 μl of *Taq* DNA Polymerase (2 units) was used to amplify the gap. The reaction mix first was denatured at 95°C for 5 minutes, and then thermocycled for ~35 cycles of 95°C for 1 minute, 55°C for 1 minute, and 72°C for 2 minutes. The resulting PCR products were evaluated by electrophoresis on a 1% agarose gel prior to removing the excess primers by treatment with 5 U of Exonuclease I (ExoI) and 1 U of Shrimp Alkaline Phosphatase (SAP). ExoI digests single stranded DNA into free nucleotides and SAP dephosphorylates free nucleotides making them unavailable for polymerisation. The clean up reaction was incubated at

37°C for 30 minutes, 80°C for 10 minutes, and then held at 4°C. For sequencing, 4-6 μl of PCR product, 2 μl of primer used in the PCR reaction, and 2 μl of 20× diluted ET reaction kit were mixed and sequencing was performed by incubating as described in **2.1.1.6 Reaction and Clean Up**.

However, there are instances in which a particular DNA region proves difficult to amplify by PCR. Typically these target gaps were GC rich or they formed secondary structures that resulted in little or no yield of expected product. Furthermore, amplification may result in products derived from regions other than the target DNA region, as indicated by multiple bands on a stained agarose gel. To overcome these problems, two PCR enhancing agents, betaine and DMSO, are commonly used. Both of these agents facilitate strand separation as DMSO disrupts base pairing whereas betaine equalizes the contribution of GC- and AT-base pairing to the stability of the DNA duplex (Baskaran, et al., 1996). In addition, two nucleotide analogs, deoxyinosine triphosphate (dITP) and 7-deaza-dGTP (Dierick, et al., 1993), can be used in PCR to release secondary structure of the PCR product that will be used as a template for DNA sequencing. Either dITP or 7-deaza-dGTP forms fewer H-bonds with dCTP than dGTP does, and makes the secondary structure more easily disrupted. Elimination of spurious GC hydrogen bonding and release of the secondary structure result in more efficient and specific synthesis of PCR products.

Since often the gap region may result in PCR products longer than 2 kb, multiple rounds of end sequencing often are needed to sequence across the entire gap. In this case, primer walking typically is used. Here a primer that is 100 base pair

upstream or downstream of sequence end is used to "walk" on the template of PCR product. This process is repeated until the entire gap sequence is obtained.

2.3.2 Multiplex PCR-based Approach (MPCR)

For the contigs with the unknown orientation, Multiplex PCR (MPCR) is utilized (Claustres, et al., 1989). As a variant of uniplex PCR, MPCR enables simultaneous amplification of many targets by including more than one pair of primers in a single reaction. The primers for MPCR were selected at both ends of the orphan contigs using Consed. As many as 96 primers were pooled, together with the other PCR reaction components. The PCR products then were distributed into individual cycle reaction wells, and the individual primer was added to each well for sequencing. The detail protocol for MPCR, cleaning up of PCR product, and sequencing reaction of PCR products is available at http://www.genome.ou.edu/ MultiplexPCRbasedSeq. html.

In some cases, MPCR failed because of the presence of multiple repeat regions that were misassembled into the incorrect contigs by Phrap. These misassembled contigs cannot be used properly unless they are corrected. In this dissertation, two ways are utilized to locate and correct the misjoined regions.

2.3.2.1 Large-insert Clone Scaffolding

The read pairing information from the large insert (fosmid DNA) was used to locate a misjoined region. For example, the forward end sequence of FOS42d11 was in Contig 9 of *L. pomona*, with the distance from the end of Contig 9 of 157 kb, while the reverse end sequence was in Contig 11, with the distance from the end of Contig 11 of 441 kb (Figure 2.3.1).

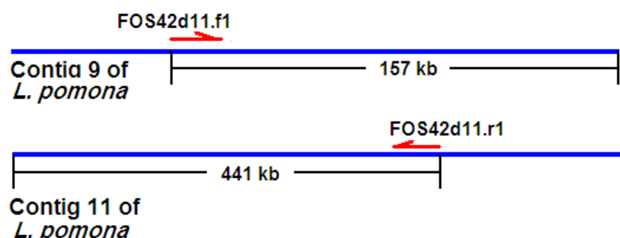


Figure 2.3.1 Scaffolding of fosmid DNA

Since the average size of fosmid DNA is approximately 40 kb, Contigs 9 and 11 must have been misjoined by Phrap. In this case, the fosmid DNA of FOS42d11 was shared into small fragments using Hydroshear and subcloned into the pUC18 vector as described in **2.1.1 Small-insert Library**. End sequences of these library clones were assembled separately to obtain a consensus sequence, and the consensus sequence subsequently was copied into the shotgun library database to cover the repeat region to correct the misjoined region.

2.3.2.2 Genome Comparison

Since *L. lai* and *L. pomona* are close related species, it is assumed that both of them share the similar sequence in most regions. The contigs of *pomona* could be aligned with the complete genome sequence of *lai* by using cross_match program. A portion of homology search result is shown as follows:

L. lai	Start	End	*L. pomona*	Start	End
lai.fa	211	2511	Contig 7	151594	149309
lai.fa	2508	4460	Contig 7	149208	147268
Possible Misjoined Region					
lai.fa	4459	48026	Contig 10	526507	570077
lai.fa	48026	63754	Contig 10	570465	586193
lai.fa	63747	67711	Contig 10	586227	590207

Based on the above result, we can see that the 147268 to 151594 region of *pomona* Contig 7 is homologous to the region of 211 to 4460 of *lai*, while the 526507 to 590207 region of *pomona* Contig 10 is homologous to the region of 4459 to 67711

of *lai*. In this way, the misjoined regions in Contig 7 and 10 were recognized and a uniplex PCR could be used to correct the misjoined region.

2.4 Sequencing Data Analysis

The genomic sequences of *L. pomona* and *grippotyphosa* were analyzed to determine the features present and extract scientifically interesting knowledge from the raw genomic sequence data. This annotation provided information on the predicted open reading frames (ORFs), rRNA and tRNA genes, repeat elements, and transposable elements. After biological functions were assigned to all ORFs, metabolic pathways of *L. pomona* and *grippotyphosa* were reconstructed, and then potential virulence genes involved in leptospirosis were identified.

2.4.1 Annotation

2.4.1.1 ORFs Prediction

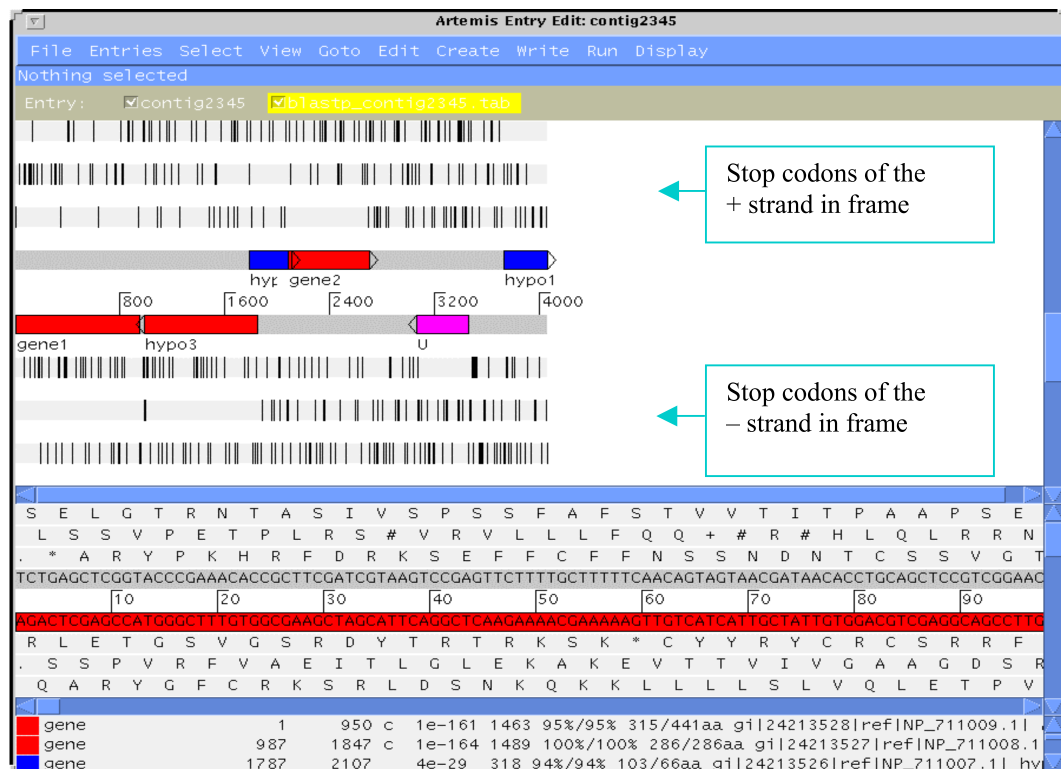


Figure 2.4.1 ORF Prediction by Artemis

Artemis (Rutherford, et al., 2000) is one of the widely used methods for analysis of microbial genomes. The complete microbial genomic DNA sequence was input to the Artemis program to display six-frame translation of the sequence. Then all the six translated amino acid sequences were scanned for stop codons and the region between two consecutive stop codons (> 100 bp long) was predicted as an open reading frame (ORF) (Figure 2.4.1).

2.4.1.2 Assignment of Biological Function to the Putative ORFs

The derived amino acid sequences of all the ORFs were searched by Blastp (Altshul, et al., 1990) against the NCBI non-redundant protein databases to determine the relationship to previously sequenced genes. The ORFs could be classified into three classes based on this homology search: homology to a gene of known function (with E value as $1.0e^{-10}$), homology to a gene of unknown function (usually referred to as a conserved hypothetical protein, also with E value as $1.0e^{-10}$), or no database homology. Functional protein coding genes then were grouped into physiological and metabolism subclasses, while those genes with unknown function or no database homology were analyzed further to determine if they harbor known functional or structural motifs.

2.4.1.3 Assignment of Enzyme Commission (E.C.) Numbers

Before assignment of E.C. numbers to the ORFs of a microbial genome, a genome-specific database was set up that included all known enzymes involved in the physiology of *E. coli* and other studied prokaryotes. Each enzyme then was attached to its EC number, when available, enzyme name, gene name, and amino acid sequence. All the amino acid sequences of the bacterial genome then were analyzed

by BlastP against the above enzyme database. A Perl program (Extract_KEGG.pl), written by Dr. Najar Fares in our laboratory, was utilized to search the output files generated by BlastP to assign the E.C. number to some certain ORFs.

2.4.1.4 Identification of Secreted Proteins

Secreted proteins possess an N-terminal signal peptide that typically is 15-40 amino acids long (Rapoport, 1992). The common structure of signal peptides is described as a positively charged *n*-region, followed by a hydrophobic *h*-region and a neutral but polar *c*-region (von Heijne, 1985). Based on neural network and hidden Markov model algorithms, the program of SignalP can analyze the amino acid sequences of all predicted ORFs and identify secreted proteins with the signal peptides (Bendtsen, et al., 2004).

2.4.1.5 Identification of Transmembrane Protein

Transmembrane proteins constitute an important subset of the proteins encoded by a genome, typically making up ~25% of the proteome (Krogh, et al., 2001). These proteins are crucial for many cellular processes including signaling and transport processes. The major features of transmembrane proteins include helices that are on average more hydrophobic than the loop regions of transmembrane proteins and positively charged amino acids are more common in the cytoplasmic than in the external loop regions (Claros and von Heijne, 1994). The program of TMHMM was used to analyze the amino acid sequences of all microbial ORFs and predict the transmembrane proteins (Krogh, et al., 2001).

2.4.1.6 Functional Domain Search

ORFs with no significant homology in the GenBank database or with homology to other proteins with no known functions then were analyzed for known motifs by cluster of orthologous groups (COG) analysis (Tatusov, et al., 2003). A COG is defined if a group of three or more proteins from different lineages that have homology to each other. It is therefore thought to correspond to evolutionary ancient conserved domain. The basic idea is that if the function is known for at least one member of the COG it can be assumed that the other members of the COG have an identical or a very similar function. This relation automatically yields a number of functional predictions for poorly characterized ORFs. The current COG set consists of 4852 clusters of orthologs, which include 59,838 proteins from 63 sequenced prokaryotic genomes and three genomes of unicellular eukaryotes (Tatusov, et al., 2003).

2.4.1.7 Identification of tRNA Genes

All tRNAs have similar sequences of 73-93 nucleotides, terminated with the sequence CCA at 3' end. They have cloverleaf secondary structure due to four base-paired stems. The cloverleaf also contains three non-base-paired loops: D, anticodon, and TpsiC loop (Voet and Voet, 1995). tRNAscan-SE (Lowe and Eddy, 1997) was the most widely used program to identify tRNA genes. This program uses probabilistic models that flexibly describe the primary sequence consensus and secondary structure (such as the conserved bases, the position of four stems, and the length of three loops) of a RNA sequence family.

2.4.1.8 Identification of Ribosomal Frameshifts

Maintenance of correct reading frame is fundamental to the integrity of the translation process. However, the translating ribosome sometimes is intentionally directed to shift reading frame at a specific site in response to special signals in the messenger RNA. This frameshift plays a significant role in morphogenesis, autogenous control, and in producing alternative enzymatic activities (Farabaugh, 1996).

The ribosomal frameshift mainly includes two categories, –1 frameshift and +1 frameshift. In the first category, the ribosome slips a single nucleotide in the upstream direction. It includes three components: a slippery site, where the ribosome changes reading frames, a 5-9 nt of spacer, and a stem-loop structure (Baranov, et al., 2002). The slippery site often consists of a heptameric sequence, X XXY YYZ. –1 frameshift typically produces fusion proteins in which the N- and C-terminal domains are encoded by two distinct, overlapping ORFs (Hammell, et al., 1999). In the second category, +1 frameshift, the ribosome slides back a single nucleotide. It also includes three components: a Shine–Dalgarno (SD) sequence, a 5-9 nt of spacer, and a slippery sequence, normally CUU UGA C, and in a single known case CUU UAA C (Baranov, et al., 2002). –1 frameshift is much more common than +1 frameshift. Frameshift Signal Finder is a useful tool to discover unknown genes that use either –1 or +1 frameshift (Moon, et al., 2004).

All the above annotation was integrated into the Artemis file where an overall representation of the genome can be viewed.

2.4.2 Reconstruction of Metabolic Pathways

2.4.2.1 Kyoto Encyclopedia of Genes and Genomes (KEGG)

Once the EC numbers were assigned to the ORFs (see **2.4.1.3 Assignment of Enzyme Commission (E.C.) Numbers**), the results were put into KEGG (Kanehisa and Goto, 2000) to reconstruct metabolic pathways computationally by correlating genes in the genome with gene products (enzymes) in the reference pathways according to the matching E.C. numbers (see **3.2 Reconstruction of Metabolic Pathways**).

2.4.2.2 Pathway Holes Filling

A pathway hole is a reaction in a pathway for which no corresponding gene has been identified in the genome. Pathway holes may exist for a number of possible reasons. They may represent true enzymatic functions in the organism for which the gene has not yet been found, or they could represent false positive pathway predictions or cases in which the pathway in this organism differs slightly from the reference pathway in KEGG. Three ways are mainly used to fill the pathway holes in this dissertation study.

2.4.2.2.1 Domain Analysis

The principle behind the domain analysis lies in the divergent evolution, which occurs when two or more genes have a common evolutionary origin but have diverged over evolutionary time such that their functional domains still are conserved. Therefore, if an enzyme is not found for a given pathway but one ORF is identified that includes the same functional domain conserved in the missing enzyme, this ORF

may function as the missing enzyme. Several examples were discussed in **3.2 Reconstruction of Metabolic Pathways**.

2.4.2.2.2 Isozymes Search

Isozymes are different variants of the same enzyme that have identical functions and occur as genecopies in an individual species (Hunter and Merkert, 1957). Often isozymes result from convergent evolution that occures when two or more genes evolved toward a similar biological function. Since their structure and function have arisen independently, the amino acid sequences of isozymes are not necessarily homologous. Thus to perform an isozymes search, the amino acid sequences of isozymes of the enzyme absent in a pathway are retrieved from NCBI and then compared, using BLASTP, with the database of all *Leptospira* ORFs. The candidate isozyme will have an E value of less than $1.0e^{-10}$. If this is observed, it is highly likely that this gene candidate is an isozyme that functions as the missing enzyme (Green and Karp, 2004).

2.4.2.2.3 Literatures Mining

The research progress from the relevant literatures can be used to fill the pathway holes reliably. For example, hemD gene involved in the heme biosynthesis is absent in the genome of *Leptospira* species. Guégan and colleagues discovered that the leptospiral hemC gene encodes a bifunctional enzyme, one function from the enzyme encoded by hemC and another function from the enzyme encoded by hemD (Guégan, et al., 2003). In this way, this pathway hole is filled (see **3.3.5 Iron Acquisition and Utilization**).

2.4.3 Virulence Genes

2.4.3.1 Definition of Virulence Genes

Most bacterial pathogens use the common strategies to cause the disease, such as adhesion to the host cells, penetration of the membranes, and the defense against the host immune response. So the genes involved in these biological processes are considered as the pathogenic genes (Wilson, et al., 2002).

2.4.3.2 Identification of Virulence Genes

Although the mechanisms by which leptospires cause disease are not well understood, virulence genes still can be predicted. At first, the query sequences were set up. They include all ORFs of *L. pomona* or *grippotyphosa*. Then a database of virulence genes was constructed. It includes about 20 bacterial pathogens, such as *E. coli*, two sequenced *Leptospira* species, *lai* and *copenhageni*. Then the homolog search was performed by the query sequences against the database. In this way, the potential virulence genes of *pomona* or *grippotyphosa* were identified (see **3.3 Virulence Genes**).

Chapter III

Results and Discussion

3.1 Genome Overview

The small chromosome of *L. pomona* has 353,967 bp with a G+C content of 35.1% and a ~4.3 Mbp large chromosome with a G+C content of 35.0%. The *L. pomona* genome encodes 3,733 predicted genes (Table 3.1.1) with 380 on the small chromosome and 3,311 on the large chromosome as well as 37 tRNAs, and 5 rRNAs (one 5S, two 16S, and two 23s). The rRNA genes are not organized in operons, as in most other bacteria, but are scattered through the large chromosome. The genome features of the other three *Leptospira* species including *L. grippotyphosa* sequenced as part of this dissertation research also were analyzed and are shown in Table 3.1.1.

Table 3.1.1 General genome features of four *Leptospira* species

	pomona		*grippotyphosa*		*lai*[1]		*copenhageni*[2]	
Features	lg. chr.[3]	sm. chr.[4]	lg. chr.	sm. chr.	lg. chr.	sm. chr.	lg. chr.	sm. chr.
Chromosome size	~4.3Mb	353,967bp	~4.3Mb	352,307bp	4,332,241bp	358,943bp	4,277,185bp	350,181bp
G+C content	35.0%	35.1%	35.9%	35.0%	36.0%	36.1%	35.1%	35.0%
Total	3,353	380	3,436	392	4,361	367	3,393	276
Protein-coding genes	3,311	380	3,395	392	4,319	367	3,351	276
tRNA genes	37	0	37	0	37	0	37	0
5s rRNA	1	0	1	0	1	0	1	0
16s rRNA	2	0	1	0	2	0	2	0
23s rRNA	2	0	2	0	1	0	2	0

Note: [1]Ren, S. X., et al., 2003; [2]Nascimento, A.L.T.O. et al., 2004
[3]lg. chr.: large chromosome; [4]sm. chr.: small chromosome

Table 3.1.2 Comparison of protein-coding genes among four *Leptospira* species.

Leptospira	large chromosome		small chromosome	
	shared genes	species specific genes*	shared genes	species specific genes
pomona	2,741	133	261	77
grippotyphosa	2,825	205	261	99
lai	2,788	854	263	66
copenhageni	2,724	98	256	0

Note: * A species specific gene is one that has a E-value of less than 1.0e-5 at the amino acid level.

In addition, all the protein-coding genes were compared among the four *Leptospira* species shown in Table 3.1.2. On the small chromosomes, 261, 261, 263,

and 256 genes from *pomona*, *grippotyphosa*, *lai*, and *copenhageni*, respectively, are shared by the other three *Leptospira* species. 77, 99, and 66 genes are species specific to *pomona*, *grippotyphosa*, and *lai*, respectively. However, *copenhageni* has no species specificgenes in its small chromosome. Among 77 *pomona* specific genes, all of them encode hypothetical proteins. Among 99 *grippotyphosa* specific genes, 92 encode hypothetical proteins and 8 encode proteins with known function, including three transposases, two lipoproteins, one integrase, and one chromate transport protein. Among 66 *lai* specific genes, 65 encode hypothetical proteins and one encodes a transposase. In addition, the gene organization also was compared among four *Leptospira* species. There are 69, 78, 85, 73 operons in the small chromosome of *L. pomona*, *grippotyphosa*, *lai*, and *copenhageni*, respectively. The operon distributions were compared in Figure 3.1.3 and in Table 3.1.3.

On the large chromosomes, 2,741, 2,825, 2,788, and 2,724 genes from *pomona*, *grippotyphosa*, *lai*, and *copenhageni*, respectively, are shared by the other three *Leptospira* species, while 133, 205, 854, and 98 genes are species specificto *pomona*, *grippotyphosa*, *lai*, and *copenhageni*, respectively. Among 133 *pomona* specific genes, 123 encode hypothetical proteins and 10 encode proteins with known function, including four integrases, two lipoproteins, two transcriptional regulators, one glucosamine-6-phosphate deaminase, one Cobalamin B12-binding protein. Among 205 *grippotyphosa* specificgenes, 183 encode hypothetical proteins and 22 proteins with known function. It is worth mentioning that one 16S rRNA gene from *Corynebacterium* was identified from *grippotyphosa* genomic sequence. Therefore, the *grippotyphosa* genomic DNA is likely to be contaminated. Among 854 *lai*

specificgenes, all of them encode hypothetical proteins. Among 98 *copenhageni* specificgenes, 93 encode hypothetical proteins and 5 encode proteins with known function, including one transcriptional repressor, one transcriptional regulator, one cytoplasmic membrane protein, one tautomerase, and one lipoprotein. In addition, the gene organization was also compared among four *Leptospira* species. There are 787, 729, 754, 723 operons in the large chromosome of *L. pomona*, *grippotyphosa*, *lai*, and *copenhageni*, respectively. The operon distributions were compared in Figure 3.1.4 and in Table 3.1.4.

Besides, the genome organization of four *Leptospira* species is compared with each other. For the small chromosome, *lai*, *copenhageni*, and *pomona* are quite similar, while *grippotyphosa* has one large rearrangement and one large flip (inversion) (Figure 3.1.1a-d). For the large chromosome, *lai* and *copenhageni* are quite similar except one large inversion. Interestingly there are more regions inverted between *pomona* and *lai* than *pomona* and *copenhageni*, while *grippotyphosa* has more small rearrangements and inversions making it far different from the other three *Leptospira* species (Figure 3.1.2a-d). However, it is very interesting to notice that the origin (upleft in Figure 3.1.2) and terminus (downright in Figure 3.1.2) of replication in all four *Leptospira* species are conserved and encode several replication-related genes, including the chromosomal replication initiator protein (dnaA), DNA polymerase III beta subunit (dnaN), DNA gyrase subunit (gyrB1), DNA gyrase subunit A (gyrA1), integrase, and DNA polymerase III subunits gamma and tau (dnaX2).

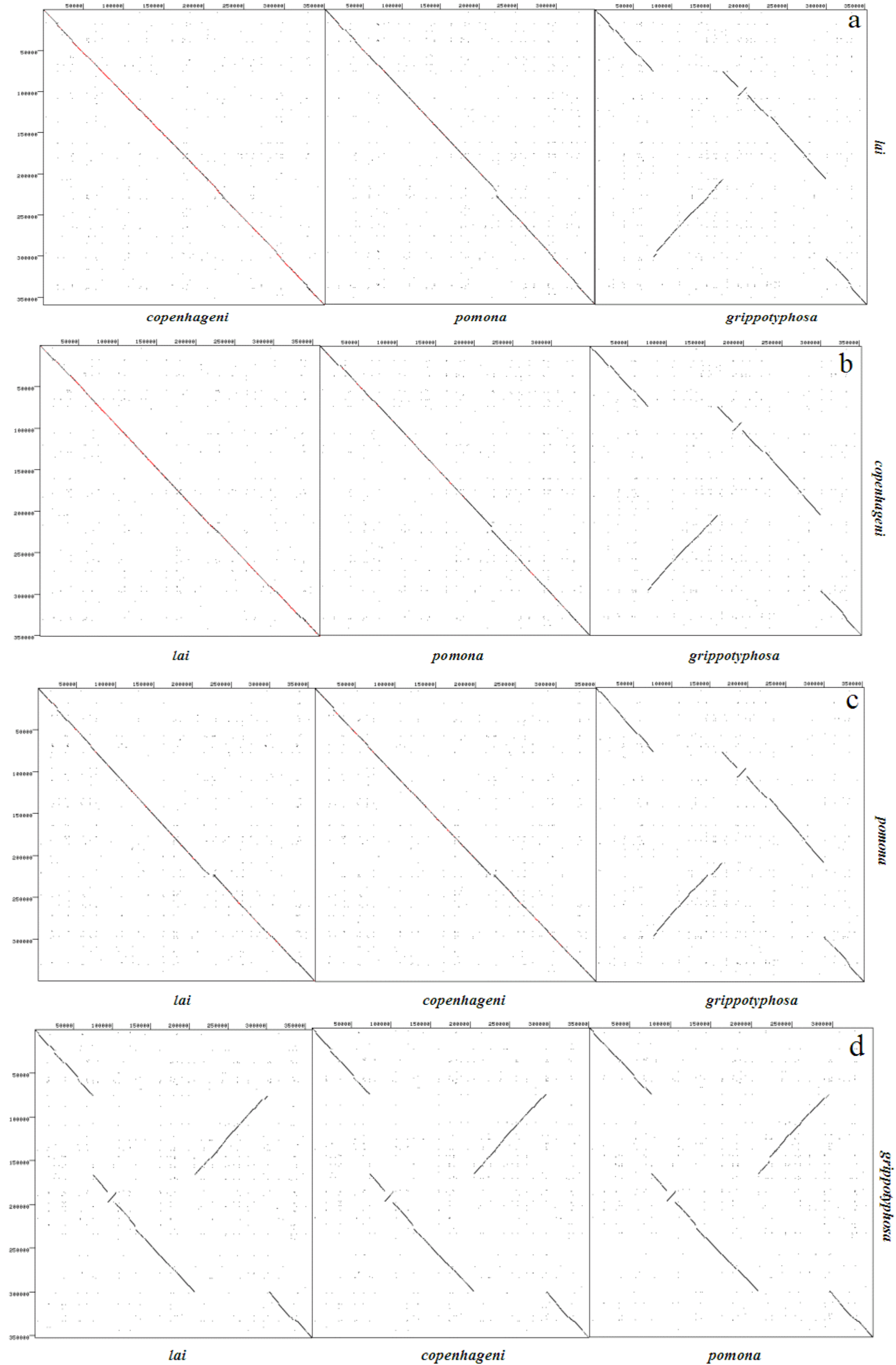


Figure 3.1.1 The small chromosome organization of one *Leptopsira* species compared with that of the other three: (a) *lai*, (b) *copenhageni*, (c) *pomona*, and (d) *grippotyphosa*.

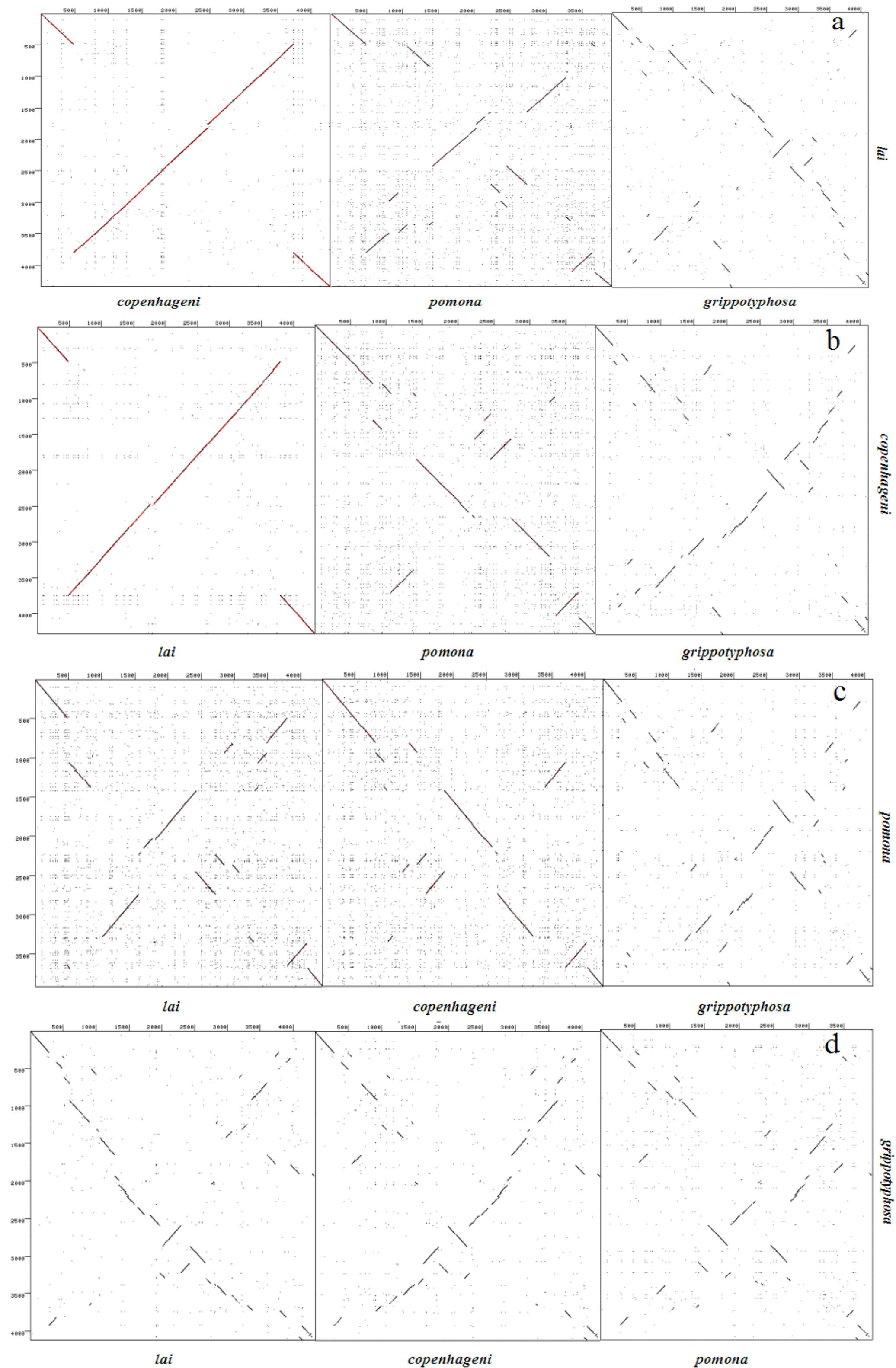


Figure 3.1.2 The large chromosome organization of one *Leptopsira* species compared with that of the other three: (a) *lai*, (b) *copenhageni*, (c) *pomona*, and (d) *grippotyphosa*.

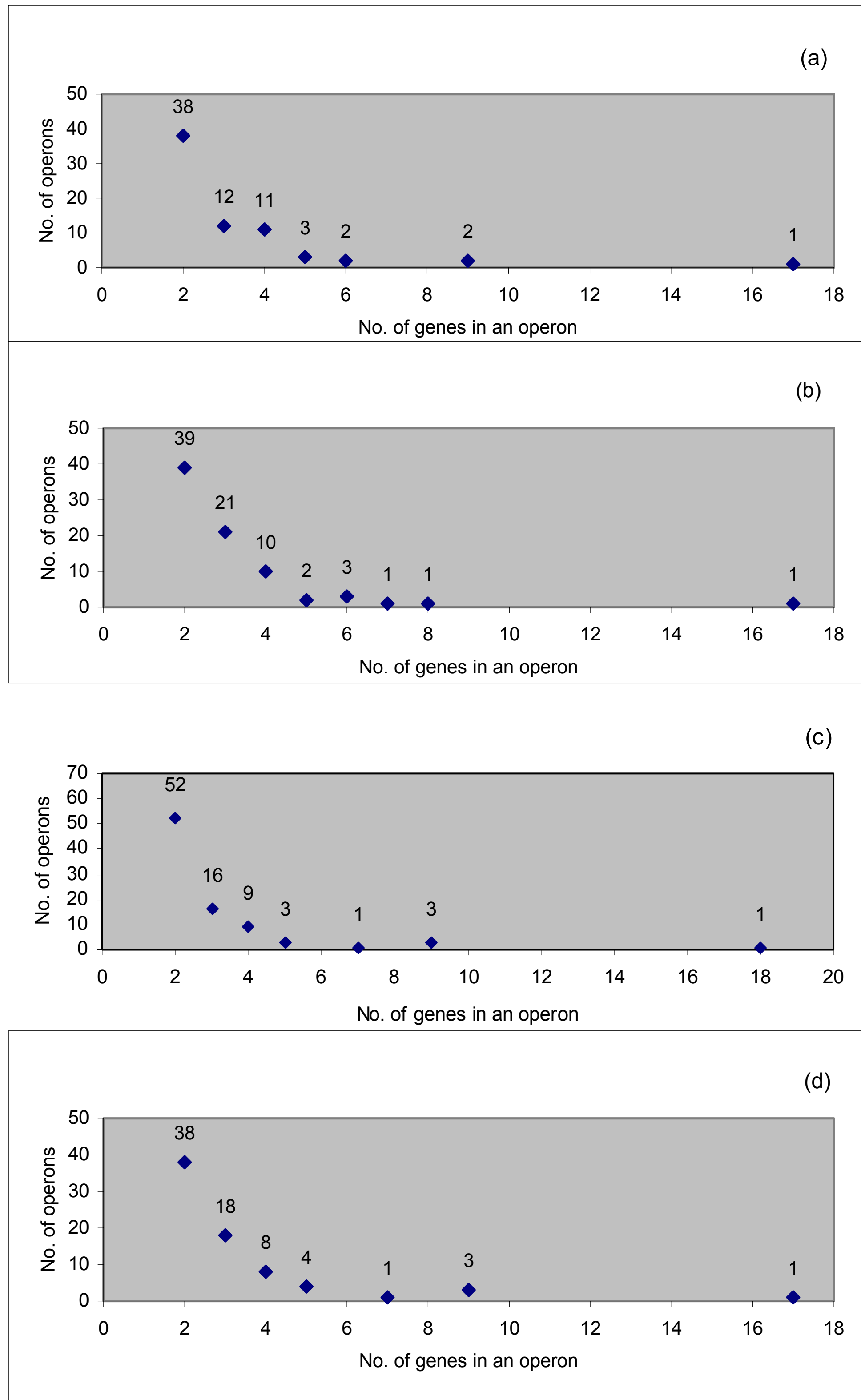


Figure 3.1.3 The operon distribution among the small chromosomes of four *Leptospira* species: (a) *pomona*, and (b) *grippotyphosa,* (c) *lai*, and (d) *copenhageni*.

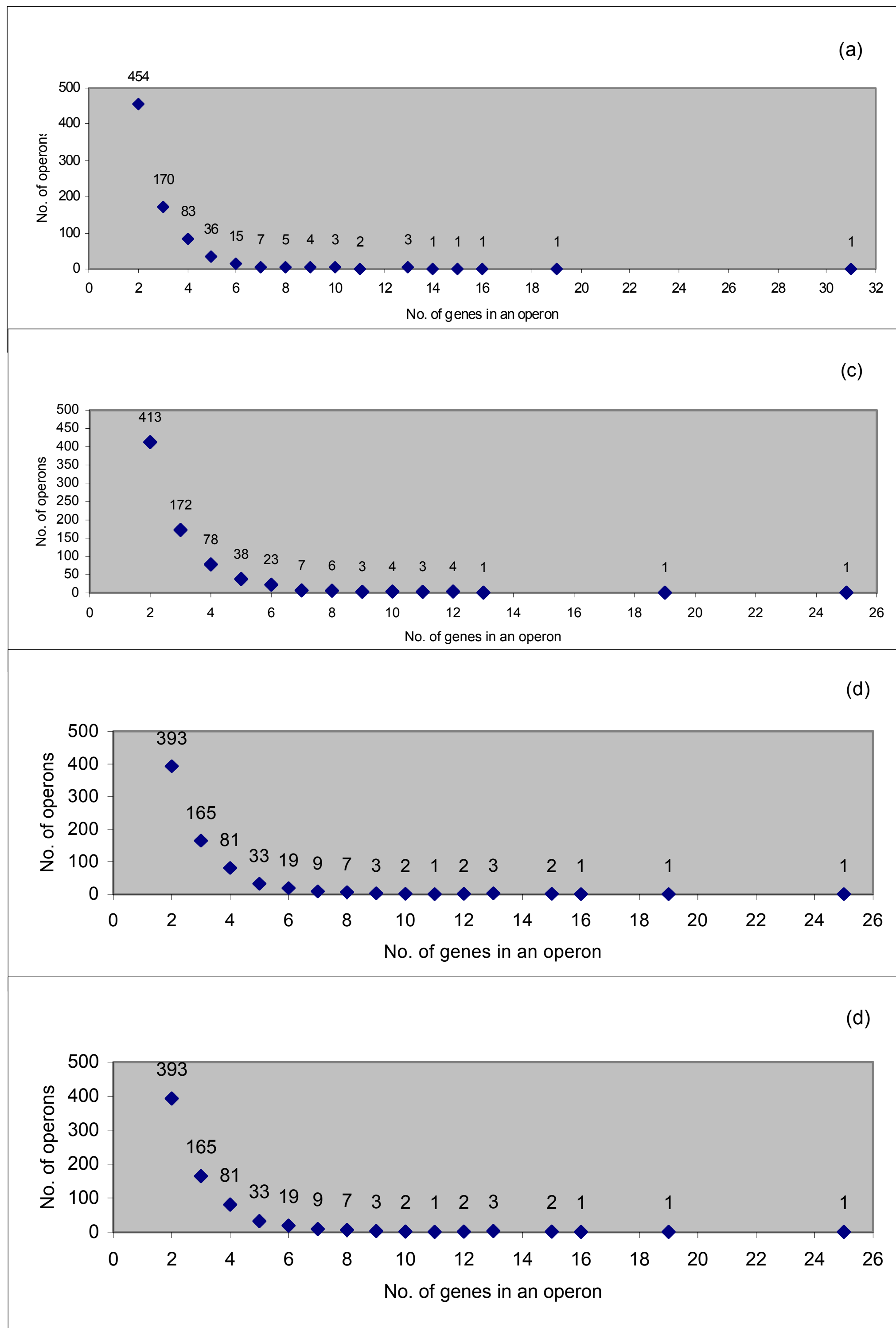


Figure 3.1.4 The operon distribution among the large chromosomes of four *Leptospira* species.

Table 3.1.3 Comparison of operon distribution among the small chromosomes of four *Leptospira* species.

	No. of operons			
No. of genes	*pomona*	*grippotyphosa*	*lai*	*copenhageni*
2	38	39	52	38
3	12	21	16	18
4	11	10	9	8
5	3	2	3	4
6	2	3	0	0
7	0	1	1	1
8	0	1	0	0
9	2	0	3	3
17	1	1	0	1
18	0	0	1	0

Table 3.1.4 Comparison of operon distribution among the large chromosomes of four *Leptospira* species.

	No. of operons			
No. of genes	*pomona*	*grippotyphosa*	*lai*	*copenhageni*
2	454	405	413	393
3	170	145	172	165
4	83	78	78	81
5	36	43	38	33
6	15	13	23	19
7	7	12	7	9
8	5	8	6	7
9	4	3	3	3
10	3	2	4	2
11	2	3	3	1
12	0	3	4	2
13	3	2	1	3
14	1	1	0	0
15	1	0	0	2
16	1	0	0	1
19	1	0	1	1
25	0	0	1	1
26	0	1	0	0
31	1	0	0	0

3.2 Reconstruction of Metabolic Pathways

Analysis of the genome sequences of four *Leptospira* species revealed as shown in Figure 3.2.1 and 3.2.2 that both aerobic and anaerobic glycolysis pathways are complete, even though it has been reported that beta-oxidation of long-chain fatty acids is utilized as the major energy and carbon source (Nascimento, et al., 2004). Interestingly, *Leptospira* species seem to lack the enzymes needed for the oxidative arm of the pentose phosphate pathway, but encode the NADH dehydrogenase and NAD(P) transhydrogenase needed to produce NADPH from NAD. Most genes encoding the enzymes needed for the citric acid cycle, the synthesis of amino acids, fatty acids and phospholipids also were observed. In addition, the lipoprotein and peptidoglycan biosynthetic pathways seem to be complete, as are pathways for vitamin and cofactor synthesis including thiamin, biotin, pantothenate, molybdenum cofactor, and riboflavin. Four *Leptospira* species also encode the replication, transcription, and translation machineries similar to those in *E. coli*. From these studies, we can conclude that the central metabolic pathways identically are conserved in four *Leptospira* species: *pomona*, *grippotyphosa*, *lai* and *copenhageni*.

3.2.1 Carbohydrate, Lipid Catabolism, and Oxidative Phosphorylation

3.2.1.1 Glycolysis

Glycolysis is the sequence of reactions that converts glucose into pyruvate with the concomitant production of a relatively small amount of ATP. Glycolysis also serves as a source of raw materials for the synthesis of other compounds needed for many other pathways. Under normal laboratory conditions, *L. interrogans* cannot utilize glucose as a carbon and energy source for the unknown reasons, however my

comparative genomic analysis revealed that the glucose utilization pathway is complete in all four *Leptospira* species (Figure 3.2.1).

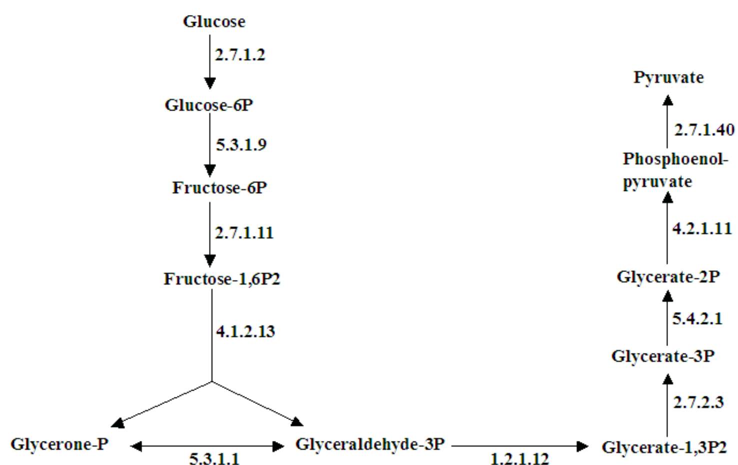


Figure 3.2.1 The predicted glycolic pathway in the four *Leptospira* species

Table 3.2.1 Enzymes involved in the glycolic pathway, their EC numbers, a description, and an indication of the presence in four *Leptospira* species.

EC#	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
2.7.1.2	glucokinase	Yes	Yes	Yes	Yes
5.3.1.9	glucose-6-phosphate isomerase	Yes	Yes	Yes	Yes
2.7.1.11	6-phosphofructokinase	Yes	Yes	Yes	Yes
4.1.2.13	fructose-bisphosphate aldolase	Yes	Yes	Yes	Yes
5.3.1.1	Triosephosphate isomerase	Yes	Yes	Yes	Yes
1.2.1.12	glyceraldehyde 3-phosphate dehydrogenase	Yes	Yes	Yes	Yes
2.7.2.3	phosphoglycerate kinase	Yes	Yes	Yes	Yes
5.4.2.1	2,3-bisphosphoglycerate-independent phosphoglycerate mutase	Yes	Yes	Yes	Yes
4.2.1.11	Enolase	Yes	Yes	Yes	Yes
2.7.1.40	pyruvate kinase	Yes	Yes	Yes	Yes

3.2.1.2 TCA Cycle

TCA cycle produces reducing equivalents (NADH and $FADH_2$) for the electron transport chain and provides anabolic precursors to different amino acid synthetic pathways. The genes coding all the required enzymes for TCA cycle are observed in *L. pomona*, *grippotyphosa, lai*, and *copenhageni*. Figure 3.2.2

summarizes the pathway of TCA cycle with the enzymes involved listed in Table 3.2.2.

Acetyl_CoA
CoA
Oxaloacetate
2.3.3.1
Citrate
4.2.1.3
Isocitrate
1.1.1.42
NADH
CO:
1.1.1.37
NADH
Malate
α -Ketoglutarate
1.2.4.2
CO:
4.2.1.2
NADH
FADH:
CoA
1.3.99.1
6.2.1.5
Fumarate
Succinate
Succinal-CoA

Figure 3.2.2 The predicted pathway of TCA cycle in the four *Leptospira* species.

Table 3.2.2 Enzymes involved in the TCA, their EC numbers, a description, and an indicator of the presence in four *Leptospira* species.

EC#	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
2.3.3.1	citrate synthase	Yes	Yes	Yes	Yes
4.2.1.3	aconitate hydrase	Yes	Yes	Yes	Yes
1.1.1.42	isocitrate dehydrogenase	Yes	Yes	Yes	Yes
1.2.4.2	α-ketoglutarate decarboxylase	Yes	Yes	Yes	Yes
6.2.1.5	succinyl-CoA synthetase	Yes	Yes	Yes	Yes
1.3.99.1	succinate dehydrogenase	Yes	Yes	Yes	Yes
4.2.1.2	fumarase	Yes	Yes	Yes	Yes
1.1.1.37	malate dehydrogenase	Yes	Yes	Yes	Yes

Although several microorganisms bypass the decarboxylation steps of the citric acid cycle by employing the glyoxylate cycle, that requires isocitrate lyase and malate synthase. However, neither of these two key enzymes has been observed in *Leptospira* species, even after an extensive analysis for the two conserved domains, COG2224 in isocitrate lyase and COG2225 in malate synthase. Thus, the glyoxylate pathway likely does not occur in at least these four *Leptospira* species.

Figure 3.2.3 Proposed 3-hydroxypropionate pathway in the four *Leptospira* species (KEGG Propanoate Metabolism)

Table 3.2.3 Enzymes involved in 3-hydroxypropionate pathway, their EC numbers, a description, and an indicator of the presence in four *Leptospira* species.

EC#	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
6.4.1.2	acetyl-CoA carboxylase	Yes	Yes	Yes	Yes
4.2.1.17	enoyl-CoA hydratase	Yes	Yes	Yes	Yes
1.3.99.3	acyl-CoA dehydrogenase	Yes	Yes	Yes	Yes
6.4.1.3	propionyl-CoA carboxylase beta chain	Yes	Yes	Yes	Yes
5.1.99.1	methylmalonyl CoA epimerase	No	No	No	No
5.4.99.2	methylmalonyl-CoA mutase	Yes	Yes	Yes	Yes

In the initial analysis, all 3-hydroxypropionate pathway enzymes, except methylmalonyl CoA epimerase were observed in the four *Leptospira* genomes analyzed. However, since the 3-hydroxypropionate pathway also is involved in the oxidative carboxylation of propionyl-CoA to succinyl-CoA in *Chloroflexus aurantiacus* (Sylvia, et al., 2001; Figure 3.2.3), further domain analysis was performed in an effort to find a possible *Leptospira* glyoxylase. This analysis revealed a candidate gene containing the same functional domain (COG0346: lactoylglutathione lyase) that occurs in methylmalonyl CoA epimerase. Therefore this additional enzyme likely is the missing methylmalonyl CoA epimerase, and if so, *Leptospira* species contain all the genes needed for the complete 3-hydroxypropionate pathway.

3.2.1.3 Pentose Phosphate Pathway (PPP)

Pentose phosphate pathway has both an oxidative and a non-oxidative arm. The oxidation arm is used to generate reducing power, NADPH. The non-oxidative arm is designed to synthesize ribose 5-phosphate, the precursor for the synthesis of the nucleotides and nucleic acids. All enzymes required for the PPP non-oxidative arm were observed encoded in the four analyzed *Leptospira* genomes and are shown in Figure 3.2.4 and Table 3.2.4.

Figure 3.2.4 The non-oxidative arm of Pentose Phosphate Pathway in the four *Leptospira* species.

Table 3.2.4 Enzymes involved in the non-oxidative arm of Pentose Phosphate Pathway in four *Leptospira* species.

EC#	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
5.3.1.9	glucose-6-phosphate isomerase	Yes	Yes	Yes	Yes
2.7.1.11	pyrophosphate--fructose 6-phosphate 1-phosphotransferase	Yes	Yes	Yes	Yes
4.1.2.13	fructose-bisphosphate aldolase	Yes	Yes	Yes	Yes
2.2.1.1	transketolase C-terminal section	Yes	Yes	Yes	Yes
2.2.1.2	transketolase N-terminal section	Yes	Yes	Yes	Yes
5.1.3.1	ribulose-phosphate 3-epimerase	Yes	Yes	Yes	Yes
5.3.1.6	Ribose 5-phosphate isomerase B	Yes	Yes	Yes	Yes
2.7.1.15	ribokinase	Yes	Yes	Yes	Yes
2.7.6.1	Ribose-phosphate pyrophosphokinase	Yes	Yes	Yes	Yes

However, most enzymes involved in the oxidative arm are absent in *Leptospira* species and without the oxidative arm they cannot produce NADPH. Further analysis of the genomic sequences of four *Leptospira* species revealed the complete pathway for the synthesis of NAD as shown in Figure 3.2.5 and Table 3.2.5. In addition, the genes for three other enzymes, NADH dehydrogenase (EC 1.6.5.3), NAD(P) transhydrogenase (EC 1.6.1.2), and malate dehydrogenase (EC 1.1.1.40), were identified in the *Leptospira* genomes. NADH dehydrogenase catalyzes NAD to produce NADH, that will be converted to NADPH by NAD(P) transhydrogenase or malate dehydrogenase. Therefore, *Leptospira* species may use the pathway shown in Figure 3.2.5 to produce NADPH.

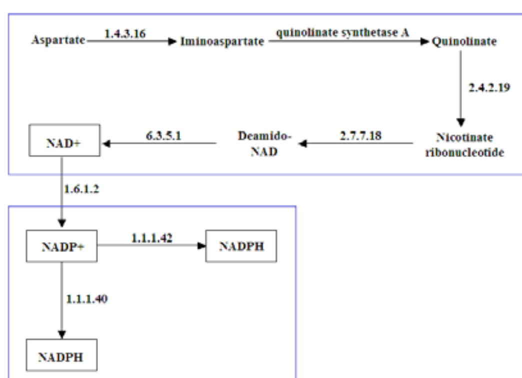


Figure 3.2.5 NADPH biosynthesis of the four *Leptospira* species.

Table 3.2.5 Enzymes involved in NADPH biosynthesis, their EC numbers, a description, and an indication of the presence in four *Leptospira* species.

EC#/gene name	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
1.4.3.16	aspartate oxidase	Yes	Yes	Yes	Yes
nadA	quinolinate synthetase A	Yes	Yes	Yes	Yes
2.4.2.19	nicotinate-nucleotide pyrophosphorylase	Yes	Yes	Yes	Yes
2.7.7.18	nicotinate-nucleotide adenylyltransferase	Yes	Yes	Yes	Yes
6.3.5.1	glutamine-dependent NAD(+) synthetase	Yes	Yes	Yes	Yes
1.6.5.3	NADH dehydrogenase	Yes	Yes	Yes	Yes
1.6.1.2	NAD(P) transhydrogenase	Yes	Yes	Yes	Yes
1.1.1.40	malate dehydrogenase	Yes	Yes	Yes	Yes

3.2.1.4 Beta-oxidation of Long-chain Fatty Acids

Leptospira species utilize beta-oxidation of long-chain fatty acids as the major energy and carbon source (Henneberry and Cox, 1970) instead of the more common sugar oxidative pathway. Beta-oxidation of fatty acids is initiated by the ATP driven coupling of Coenzyme A (CoA) to the fatty acid by fatty acid-CoA ligase. Next, the acyl-CoA is oxidized with oxidation of the β-carbon and a series of steps that each releases a two-carbon fragment, in the form of acetyl-CoA, from the fatty acid undergoing oxidation. The degradation of the fatty acid yields the other high-energy compounds, NADH and $FADH_2$. Acetyl-CoA is catabolized via the citric acid cycle, and $FADH_2$ and NADH transfer electrons to the respiratory chain. A complete set of genes for the long-chain fatty-acid utilization was identified in all four *Leptospira* species. Figure 3.2.6 summarizes the pathway of fatty acid oxidization with the enzymes involved listed in Table 3.2.6.

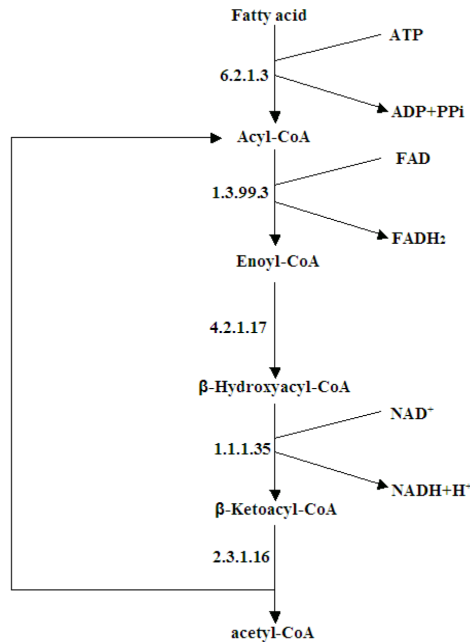


Figure 3.2.6 The predicted pathway of beta-oxidation of long-chain fatty acid in the four *Leptospira* species.

Table 3.2.6 Enzymes involved in the pathway of beta-oxidation of long-chain fatty acid, their EC numbers, a description, and an indicator of the presence in four *Leptospira* species.

EC#	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
6.2.1.3	long-chain fatty acid-CoA ligase	Yes	Yes	Yes	Yes
1.3.99.3	acyl-CoA dehydrogenase	Yes	Yes	Yes	Yes
4.2.1.17	enoyl-CoA hydratase	Yes	Yes	Yes	Yes
1.1.1.35	3-hydroxyacyl-CoA dehydrogenase	Yes	Yes	Yes	Yes
2.3.1.16	3-ketoacyl-CoA thiolase	Yes	Yes	Yes	Yes

3.2.1.5 Oxidative Phosphorylation

Photosynthetic bacteria often adopt an oxidative phosphorylation pathway similar to that in eukaryotic cells (Bernard, 1990). However, bacteria such as *E. coli* and *Leptospira* species use a slightly modified oxidative phosphorylation pathway (Figure 3.2.7). Besides Complexes I, II, and IV, *Leptospira* species have additional substrate-specific electron carriers, that include lactate dehydrogenase, glycerol 3-phosphate dehydrogenase, and pyruvate dehydrogenase. However, since complex III is absent in the genomes of *Leptospira* species, they generate ATP via an F_0F_1-type ATPase that is encoded in a single operon with a conserved gene organization in *L. pomona*, *grippotyphosa*, *lai*, and *copenhageni*.

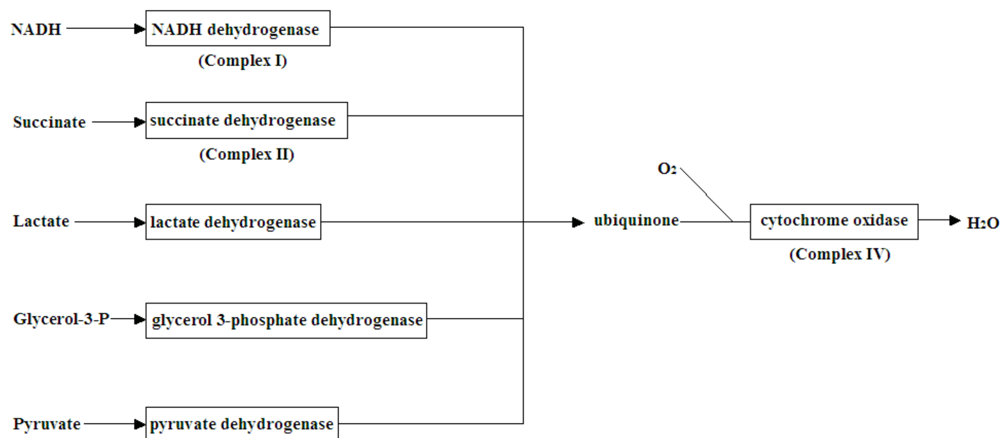


Figure 3.2.7 Possible pathways for aerobic respiration in the four *Leptospira* species.

Table 3.2.7 Enzymes involved in the predicted oxidative phosphorylation pathway in four *Leptospira* species.

Description	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
NADH dehydrogenase, subunit from A to N	Yes	Yes	Yes	Yes
succinate dehydrogenase, subunit A, B, and C	Yes	Yes	Yes	Yes
lactate dehydrogenase	Yes	Yes	Yes	Yes
glycerol-3-phosphate dehydrogenase	Yes	Yes	Yes	Yes
pyruvate dehydrogenase	Yes	Yes	Yes	Yes
cytochrome C oxidase assembly factor	Yes	Yes	Yes	Yes
cytochrome C oxidase, subunit I, II, III, and XV	Yes	Yes	Yes	Yes
ATP synthase F1, subunit alpha, beta, gamma, delta, and epsilon	Yes	Yes	Yes	Yes
ATP synthase F0, subunit A, B, and C	Yes	Yes	Yes	Yes

3.2.2 Lipid Biosynthesis

3.2.2.1 Fatty Acids Synthesis

Fatty acids are the essential building blocks for membrane phospholipid formation. Fatty acid metabolism is a fundamental component of the cellular metabolic network (Rock and Cronan, 1996). Fatty acid biosynthesis can be separated into two stages, stage I (initiation) and stage II (cyclic elongation) (Cronan and Rock, 1996). In stage I, the acetate moiety is transferred from acetyl-CoA to acetyl-ACP, and then the acetyl-ACP is condensed with malonyl-ACP to form acetoacetyl-ACP. In stage II, acetoacetyl-ACP is reduced and dehydrated. After the second reduction, acyl-ACP is formed, and served as a substrate for another round of elongation (Figure 3.2.8).

Three essential enzymes involved in the pathway for fatty acid synthesis are absent in the four *Leptospira* genomes: ACP S-malonyltransferase (FabD, EC 2.3.1.39), enoyl-ACP reductase (FabI, EC 1.3.1.9), and 3-hydroxydecanoyl-ACP dehydratase (FabA, EC 4.2.1.60). Through the domain analysis, one putative fatty acid synthase from four *Leptospira* species was identified that included the same

functional domain (COG0331: ACP S-malonyltransferase) conserved in FabD. In addition, two conserved domains (COG0623: enoyl-ACP reductase and COG0764: 3-hydroxydecanoyl-ACP dehydratase) were identified from *E. coli* FabI and FabA, respectively. However, neither of the two above domains was identified from any ORF of *Leptospira* species. Heath and the colleagues discovered that expression of *B. subtilis* FabL complemented the FabI defect in *E. coli* (Heath, et al., 2000). Therefore, *B. subtilis* FabL may function as FabI. A possible short-chain dehydrogenase in all four *Leptospira* species was identified as it included the same functional domain (COG1028: dehydrogenases) conserved in *B. subtilis* FabL. In order to find the third missing gene, FabA, isozymes search was employed (Green and Karp, 2004). 902 amino acid sequences of FabA isozymes were downloaded from NCBI and each sequence was compared with each *L. pomona* ORF by blastp. When the isozyme from *Saccharophagus degradans* was used as the query sequence, one ORF of *L. pomona* (ORF2849108) could be identified as the best substitute with E value of $4.0e^{-49}$ for the missing FabA. In the same way, the best substitutes were also identified from the other three *Leptospira* species. Therefore, since all three missing genes that likely encoded the enzymes involved in this pathway were identified, *Leptospira* species may be capable of synthesizing fatty acids *de novo*.

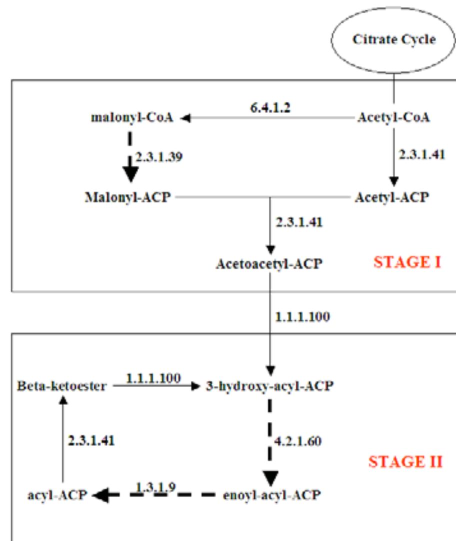


Figure 3.2.8 Fatty acid biosynthesis in the four *Leptospira* species where the enzymes missing are indicated in dotted arrows and those present by the solid arrow.

Table 3.2.8 Enzymes involved in the fatty acid biosynthesis pathway present in four *Leptospira* species.

EC#	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
6.4.1.2	acetyl-CoA carboxylase	Yes	Yes	Yes	Yes
2.3.1.41	3-oxoacyl-ACP synthase (FabH)	Yes	Yes	Yes	Yes
2.3.1.39	ACP S-malonyltransferase	Yes	Yes	Yes	Yes
1.1.1.100	3-oxoacyl-ACP reductase	Yes	Yes	Yes	Yes
4.2.1.60	3-hydroxydecanoyl-ACP dehydratase (FabA)	Yes	Yes	Yes	Yes
1.3.1.9	enoyl-ACP reductase (NADH)	Yes	Yes	Yes	Yes

3.2.2.2 Synthesis of Phospholipids

Phospholipids are amphipathic molecules that are major structural components of cellular membranes (Cronan and Rock, 1996). In addition, phospholipids provide precursors for the synthesis of macromolecules, serve as molecular chaperons, serve in protein modification for membrane association, and are reservoirs of second messengers. Thus, phospholipids are essential for vital cellular processes (Iwanyshyn, et al., 2004).

Figure 3.2.9 shows the pathway responsible for the biosynthesis of the three major phospholipids: phosphatidylserine (PE), phosphatidyl glycerol (PG), and

cardiolipin (CL). After phosphatidic acid (PA) is synthesized, it is converted to CDP-diacyl-glycerol, which serves as an intermediate in the biosynthesis of all membrane phospholipids. PE is formed by the condensation of CDP-diacyl-glycerol with serine followed by decarboxylation. PG is formed by the condensation of CDP-diacyl-glycerol with glycerol-3 phosphate followed by removal of the phosphate. CL is formed by the condensation of two molecules of PG. Analysis of the four *Leptospira* genomic sequences revealed that all the other genes except cardiolipin synthase (2.7.8.-) are present that could encode the enzymes associated with phospholipid biosynthesis as shown in Figure 3.2.9 and listed in Table 3.2.9. The domain analysis revealed that one putative phospholipase D family protein includes the same functional domain (COG1502: cardiolipin synthase) conserved in cardiolipin synthase. Therefore, *Leptospira* species likely encode for the complete pathway for the phospholipid biosynthesis.

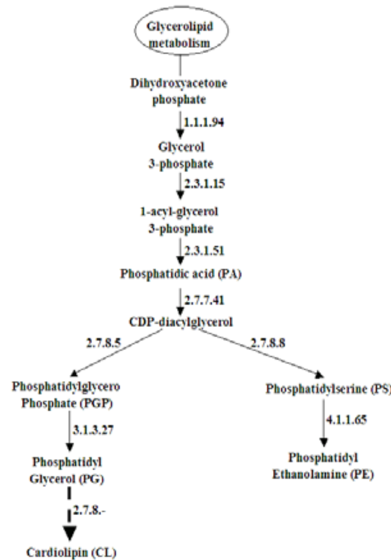


Figure 3.2.9 Reconstructed pathway of phosphoglyceride synthesis in the four *Leptospira* species where the enzymes missing are indicated in dotted arrows and those present by the solid arrow.

Table 3.2.9 Enzymes involved in the phosphoglyceride synthesis present in four *Leptospira* species.

EC#	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
1.1.1.94	Glycerol-3-phosphate dehydrogenase	Yes	Yes	Yes	Yes
2.3.1.15	Glycerol-3-phosphate acyltransferase	Yes	Yes	Yes	Yes
2.3.1.51	1-acyl-sn-glycerol-3-phosphate acyltransferase	Yes	Yes	Yes	Yes
2.7.7.41	Phosphatidate cytidylyltransferase	Yes	Yes	Yes	Yes
2.7.8.8	Phosphatidylserine synthase	Yes	Yes	Yes	Yes
4.1.1.65	Phosphatidylserine decarboxylase	Yes	Yes	Yes	Yes
2.7.8.5	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase	Yes	Yes	Yes	Yes
3.1.3.27	Phosphatidylglycerophosphatase	Yes	Yes	Yes	Yes
2.7.8.-	Cardiolipin synthase	Yes	Yes	Yes	Yes

3.2.3 Biosynthesis of Amino Acids

The central pathways provide the precursors needed to synthesize 20 amino acids. All the four *Leptospira* species seem capable of synthesizing all 20 amino acids except asparagine.

3.2.3.1 Glutamate, Glutamine, Aspartate, Asparagine, and Alanine

As illustrated below in Figure 3.2.10, all of the enzymes needed for the formation of glutamate, glutamine, aspartate, and alanine were found encoded in the four *Leptospira* genomes. However, asparagine synthetase (EC 6.3.1.1) involved in the synthesis of asparagine was not observed. Domain analysis revealed that no *Leptospira* ORF could be found that corresponded to the functional domain (COG2502: asparagine synthetase A) conserved in asparagine synthetase. How is protein synthesis affected if asparagine cannot be synthesized? It will be discussed in **3.2.6.3 Aminoacyl-tRNA Synthesis.**

Figure 3.2.10 Synthesis of glutamate, glutamine, aspartate, asparagine, and alanine. All the enzymes except asapariginr synthetase (dotted arrow) were present in the four *Leptospira* genomes.

3.2.3.2 Branched-chain Amino Acids

Analysis of four *Leptospira* genomic sequences reveals that they all contain the enzymes necessary for the synthesis of the branched-chain amino acids.

Figure 3.2.11 Synthesis of branched-chain amino acids. All enzymes except threonine dehydratase (EC 4.3.1.19), represented by the dotted arrow, were observed encoded in the four *Leptospira* genomes.

Table 3.2.10 Enzymes involved in the biosynthesis of valine, leucine, and isoleucine.

EC#	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
2.2.1.6	acetolactate synthase	Yes	Yes	Yes	Yes
1.1.1.86	isomeroreductase	Yes	Yes	Yes	Yes
4.2.1.9	dihydroxy-acid dehydratase	Yes	Yes	Yes	Yes
2.6.1.42	branched-chain-amino-acid transaminase	Yes	Yes	Yes	Yes
2.3.3.13	2-isopropylmalate synthase	Yes	Yes	Yes	Yes
4.2.1.33	isopropylmalate isomerase	Yes	Yes	Yes	Yes
1.1.1.85	3-isopropylmalate dehydrogenase	Yes	Yes	Yes	Yes
4.1.3.-	citramalate synthase	Yes	Yes	Yes	Yes

In most microorganisms, isoleucine is synthesized from aspartate via threonine deaminase pathway (Umbarger, 1978) with threonine dehydratase (EC 4.3.1.19) being the key enzyme in this pathway. Because this enzyme could not be identified by either homolog search or domain analysis, *Leptospira* species might use the alternative pathway from pyruvate (Xu, et al., 2004) for isoleucine biosynthesis. All the essential enzymes involved in this alternative pathway are present (Figure 3.2.11).

3.2.3.3 Methionine, Threonine, and Lysine

All the genes encoding the enzymes required for the biosynthesis of methionine and threonine are present in the four *Leptospira* genomes (Figure 3.2.12). However, two essential enzymes, succinyl-diaminopimelate desuccinylase (EC 3.5.1.18) and 2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase (EC 2.3.1.117), involved in lysine biosynthesis, were not identified by the homolog search. Domain analysis revealed that one possible enzyme from four *Leptospira* species includes the same functional domain (COG0624: succinyl-diaminopimelate desuccinylase) conserved in succinyl-diaminopimelate desuccinylase. However, no ORF was observed in any of the four *Leptospira* species that includes the functional domain (COG2171: DapD) conserved in 2,3,4,5-tetrahydropyridine-2-carboxylate N-

succinyltransferase. To investigate further, an isozyme analysis (Green and Karp, 2004) was performed to search for the missing succinyltransferase. 384 isozyme amino acid sequences were downloaded from NCBI and each sequence was compared with each *L. pomona* ORF by blastp. When the isozyme from *Streptococcus pneumoniae* was used as the query sequence, one ORF of *L. pomona* (ORF642477) could be identified as the most likely substitute with E value of $6.0e^{-12}$ for the missing succinyltransferase. Similarly, the corresponding gene was also identified from *lai*, *copenhageni*, and *grippotyphosa*. Taken together, all the four *Leptospira* species likely synthesize lysine *de novo* via the pathway shown in Figure 3.2.12.

Figure 3.2.12 Synthesis of threonine, methionine, and lysine. The reaction catalyzed by succinyl-diaminopimelate desuccinylase (EC 3.5.1.18) and 2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase (EC 2.3.1.117) are represented by dotted arrows.

Table 3.2.11 Enzymes involved in the biosynthesis of lysine, threonine, and methionine.

EC#	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
2.7.2.4	aspartate kinase	Yes	Yes	Yes	Yes
1.2.1.11	aspartate semialdehyde dehydrogenase	Yes	Yes	Yes	Yes
4.2.1.52	dihydrodipicolinate synthase	Yes	Yes	Yes	Yes
1.3.1.26	dihydrodipicolinate reductase	Yes	Yes	Yes	Yes
2.3.1.117	2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase	Yes	Yes	Yes	Yes
2.6.1.7	N-succinyldiaminopimelate aminotransferase	Yes	Yes	Yes	Yes
3.5.1.18	succinyl-diaminopimelate desuccinylase	Yes	Yes	Yes	Yes
5.1.1.7	diaminopimelate epimerase	Yes	Yes	Yes	Yes
4.1.1.20	diaminopimelate decarboxylase	Yes	Yes	Yes	Yes
1.1.1.3	homoserine dehydrogenase	Yes	Yes	Yes	Yes
2.3.1.31	homoserine acetyltransferase	Yes	Yes	Yes	Yes
2.5.1.49	methionine synthase	Yes	Yes	Yes	Yes
2.1.1.13	methionine synthase	Yes	Yes	Yes	Yes
2.7.1.39	homoserine kinase	Yes	Yes	Yes	Yes
4.2.3.1	threonine synthase	Yes	Yes	Yes	Yes

3.2.3.4 Serine, Glycine, and Cysteine

As illustrated in Figure 3.2.13 and Table 3.2.12, all the four *Leptospira* species appear to encode all enzymes required for the complete biosynthetic pathways for serine, glycine and cysteine. It is interesting that the last reaction for cysteine biosynthesis pathway (catalyzed by cysteine synthase, EC 2.5.1.47) utilizes hydrogen sulfide. Even though the complete sulfate transport system was observed from *Leptospira* species (see **3.2.8.1 Cation and Anion Transporters**), it is still unclear how sulfate is reduced to hydrogen sulfide. It is possible that a different pathway might exist to utilize sulfate or that *Leptospira* species rely on external cysteine through the different amino acid and peptide transport systems that the analysis of the genome revealed (Schiemann, 1973).

Figure 3.2.13 Synthesis pathways for serine, glycine, and cysteine.

Table 3.2.12 Enzymes involved in the synthesis of serine, glycine, and cysteine in four *Leptospira* species.

EC#	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
1.1.1.95	phosphoglycerate dehydrogenase	Yes	Yes	Yes	Yes
2.6.1.52	phosphoserine aminotransferase	Yes	Yes	Yes	Yes
3.1.3.3	phosphoserine phosphatase	Yes	Yes	Yes	Yes
2.1.2.1	serine hydroxymethyltransferase	Yes	Yes	Yes	Yes
2.3.1.30	serine acetyltransferase	Yes	Yes	Yes	Yes
2.5.1.47	Cysteine synthase	Yes	Yes	Yes	Yes

3.2.3.5 Aromatic Amino Acids

All the four *Leptospira* species are capable of synthesizing all three aromatic amino acids as all of the genes needed for the enzymes involved in the aromatic amino acid biosynthetic pathways are present as shown in Figure 3.2.14 and Table 3.2.13.

Figure 3.2.14 Synthesis of aromatic amino acids.

Table 3.2.13 Enzymes involved in aromatic amino acid biosynthesis.

EC#	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
2.5.1.54	Phospho-2-dehydro-3-deoxyheptonate aldolase	Yes	Yes	Yes	Yes
4.2.3.4	dehydroquinate synthase	Yes	Yes	Yes	Yes
4.2.1.10	dehydroquinate dehydratase	Yes	Yes	Yes	Yes
4.2.1.11	phosphopyruvate hydratase	Yes	Yes	Yes	Yes
1.1.1.25	shikimate dehydrogenase	Yes	Yes	Yes	Yes
2.7.1.71	shikimate kinase	Yes	Yes	Yes	Yes
2.5.1.19	enolpyruvylshikimate phosphate synthase	Yes	Yes	Yes	Yes
4.2.3.5	chorismate synthase	Yes	Yes	Yes	Yes
4.1.3.27	chorismate lyase	Yes	Yes	Yes	Yes
2.4.2.18	phosphoribosyltransferase	Yes	Yes	Yes	Yes
5.3.1.24	phosphoribosylanthranilate isomerase	Yes	Yes	Yes	Yes
4.1.1.48	indoleglycerol phosphate synthetase	Yes	Yes	Yes	Yes
4.2.1.20	tryptophan synthase	Yes	Yes	Yes	Yes
5.4.99.5	chorismate mutase	Yes	Yes	Yes	Yes
4.2.1.51	prephenate dehydratase	Yes	Yes	Yes	Yes
2.6.1.1	aspartate transaminase	Yes	Yes	Yes	Yes
1.3.1.12	prephenate dehydrogenase	Yes	Yes	Yes	Yes

3.2.3.6 Arginine and Proline

Four *Leptospira* genomes contain all the genes needed to encode the enzymes which synthesize arginine and proline. A summary of the arginine and proline

biosynthetic pathway is shown in Figure 3.2.15 and the genes present for the biosynthetic enzymes in *Leptospira* species are indicated in Table 3.2.14.

Figure 3.2.15 Biosynthesis of arginie and proline.

Table 3.2.14 Enzymes involved in the biosynthesis of arginine and proline.

EC#	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
2.7.2.11	glutamate kinase	Yes	Yes	Yes	Yes
1.2.1.41	glutamate-5-semialdehyde dehydrogenase	Yes	Yes	Yes	Yes
1.5.1.2	proline oxidase	Yes	Yes	Yes	Yes
2.3.1.1	acetylglutamic synthetase	Yes	Yes	Yes	Yes
2.7.2.8	acetylglutamate kinase	Yes	Yes	Yes	Yes
1.2.1.38	N-acetyl-gamma-glutamyl-phosphate reductase	Yes	Yes	Yes	Yes
2.6.1.11	acetylornithine transaminase	Yes	Yes	Yes	Yes
3.5.1.14	aminoacylase	Yes	Yes	Yes	Yes
2.1.3.3	ornithine carbamoyltransferase	Yes	Yes	Yes	Yes
6.3.4.5	argininosuccinate synthase	Yes	Yes	Yes	Yes
4.3.2.1	arginosuccinase	Yes	Yes	Yes	Yes

3.2.3.7 Histidine

All the genes that are needed to encode the histidine biosynthesis pathway enzymes were observed in the four *Leptospira* species as shown in Figure 3.2.16 and Table 3.2.15.

Figure 3.2.16 Synthesis of histidine.

Table 3.2.15 Enzymes involved in the biosynthesis of histidine.

EC#	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
2.4.2.17	ATP phosphoribosyltransferase	Yes	Yes	Yes	Yes
3.6.1.31	phosphoribosyl-ATP pyrophosphohydrolase	Yes	Yes	Yes	Yes
3.5.4.19	phosphoribosyl-AMP cyclohydrolase	Yes	Yes	Yes	Yes
5.3.1.16	phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase	Yes	Yes	Yes	Yes
2.4.2.-	Amidotransferase hisH	Yes	Yes	Yes	Yes
4.2.1.19	Imidazoleglycerol-phosphate dehydratase	Yes	Yes	Yes	Yes
2.6.1.9	histidinol-phosphate transaminase	Yes	Yes	Yes	Yes
3.1.3.15	histidinol-phosphatase	Yes	Yes	Yes	Yes
1.1.1.23	histidinol dehydrogenase	Yes	Yes	Yes	Yes

3.2.4 Nucleotide Metabolism

Analysis of the four *Leptospira* genomic sequences revealed the presence of the genes for all the enzymes needed for *de novo* synthesis of purine (Figure 3.2.17)

and pyrimidine (Figure 3.2.18). These pathways lead to production of nucleoside-5-phosphates through the utilization of an activated sugar intermediate, 5-phosphoribosyl-1-pyrophosphate (PRPP), and a class of enzymes, phosphoribosyltransferases.

3.2.4.1 Purine Nucleotides

Figure 3.2.17 The purine biosynthetic pathway

Table 3.2.16 Enzymes involved in purine metabolism of four *Leptospira* species.

EC#	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
2.4.2.14	amidophosphoribosyltransferase	Yes	Yes	Yes	Yes
6.3.4.13	phosphoribosylamine-glycine ligase	Yes	Yes	Yes	Yes
2.1.2.2	phosphoribosylglycinamide formyltransferase	Yes	Yes	Yes	Yes
6.3.5.3	phosphoribosylformylglycinamidine synthase	Yes	Yes	Yes	Yes
6.3.3.1	phosphoribosylformylglycinamidine cyclo-ligase	Yes	Yes	Yes	Yes
4.1.1.21	phosphoribosylaminoimidazole carboxylase	Yes	Yes	Yes	Yes
6.3.2.6	phosphoribosylaminoimidazolesucci nocarboxamide synthase	Yes	Yes	Yes	Yes
4.3.2.2	adenylosuccinate lyase	Yes	Yes	Yes	Yes
2.1.2.3	imidazolecarboxamide formyltransferase	Yes	Yes	Yes	Yes
3.5.4.10	IMP cyclohydrolase	Yes	Yes	Yes	Yes
6.3.4.4	adenylosuccinate synthetase	Yes	Yes	Yes	Yes
2.7.4.3	adenylate kinase	Yes	Yes	Yes	Yes
2.7.4.6	nucleoside diphosphate kinase	Yes	Yes	Yes	Yes
1.17.4.1	ribonucleotide reductase	Yes	Yes	Yes	Yes
1.1.1.205	IMP dehydrogenase	Yes	Yes	Yes	Yes
6.3.5.2	GMP synthase	Yes	Yes	Yes	Yes
2.7.4.8	guanylate kinase	Yes	Yes	Yes	Yes

As shown in Figure 3.2.16 and Table 3.2.16, purine biosynthesis begins with PRPP and the purine base is built upon the ribose by several amidotransferase and transformylation reactions. The first fully formed nucleotide, inosine 5-monophosphate (IMP), represents a branch point for purine biosynthesis that results in the synthesis of either AMP or GMP through two distinct reaction pathways.

3.2.4.2. Pyrimidine Nucleotides

Figure 3.2.18 The pyrimidines biosynthetic pathway

Table 3.2.17 Enzymes involved in pyridine metabolism of four *Leptospira* species.

EC#	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
6.3.5.5	carbamoyl-phosphate synthase	Yes	Yes	Yes	Yes
2.1.3.2	aspartate carbamoyltransferase	Yes	Yes	Yes	Yes
3.5.2.3	dihydroorotase	Yes	Yes	Yes	Yes
1.3.3.1	dihydroorotate oxidase	Yes	Yes	Yes	Yes
2.4.2.10	orotate phosphoribosyltransferase	Yes	Yes	Yes	Yes
4.1.1.23	orotodylate decarboxylase	Yes	Yes	Yes	Yes
2.7.4.14	cytidylate kinase	Yes	Yes	Yes	Yes
2.7.4.6	nucleoside-diphosphate kinase	Yes	Yes	Yes	Yes
6.3.4.2	CTP synthase	Yes	Yes	Yes	Yes
1.17.4.1	ribonucleotide reductase	Yes	Yes	Yes	Yes
3.5.4.13	dCTP deaminase	Yes	Yes	Yes	Yes
2.7.4.9	thymidylate kinase	Yes	Yes	Yes	Yes
2.1.1.45	thymidylate synthase	Yes	Yes	Yes	Yes

All four *Leptospira* species contain the complete set of genes needed for pyrimidine biosynthesis (Table 3.2.17). It is well established that the synthesis of pyrimidine differs in two significant ways from that of purine as shown in Figure 3.2.18. First, the pyrimidine ring structure is assembled as a free base, rather than being built upon PRPP. Then, PRPP is added to the first fully formed pyrimidine base (orotic acid), forming orotate monophosphate (OMP), which subsequently is decarboxylated to UMP. Second, there is no branch in the pyrimidine synthesis pathway as there is no purine biosynthetics as UMP is phosphorylated twice to yield UTP, and then UTP is aminated by the action of CTP synthase, generating CTP.

3.2.5 Biosynthesis of Cofactors and Vitamins

3.2.5.1 Riboflavin Biosynthesis

Figure 3.2.19 Riboflavin biosynthetic pathway

Table 3.2.18 Enzymes involved in riboflavin biosynthesis, their EC numbers, a description, and an indication of the presence in four *Leptospira* species.

EC#	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
3.5.4.25	GTP-8-formylhydrolase	Yes	Yes	Yes	Yes
3.5.4.26	2,5-diamino-6-hydroxy-4-(5-phosphoribosylamino)pyrimidine 2-aminohydrolase	Yes	Yes	Yes	Yes
1.1.1.193	5-amino-6-(5-phosphoribosylamino)uracil reductase	Yes	Yes	Yes	Yes
2.5.1.9	riboflavin synthase	Yes	Yes	Yes	Yes
2.7.1.26	riboflavin kinase	Yes	Yes	Yes	Yes
2.7.7.2	FMN adenylyltransferase	Yes	Yes	Yes	Yes

Riboflavin, also known as vitamin B2 or vitamin G, is a water-soluable vitamin. It supports metabolism of fats, carbohydrates, and proteins and biosynthesis of a number of compounds through its coenzyme forms, flavin adenine dinucleotide (FAD) and flavin adenine mononucleotide (FMN). As shown in Figure 3.2.19 and Table 3.2.18, all of the genes that encode the enzymes required for the synthesis of riboflavin are present in all four *Leptospira* genomes.

3.2.5.2 Folate Biosynthesis

Purine metabolism
GTP
3.5.4.16
2-Amino-4-hydroxy-6-(erythro-1,2,3-trihydroxypropyl)dihydropteridine triphosphate
3.1.3.1
dihydropteridine
4.1.2.25
2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine
Phenylalanine Biosynthesis
Chorismate
6.3.5.8
2.7.6.3
4-Amino-4-deoxychorismate
4.1.3.38
4-Aminobenzoate
7,8-Dihydropterin pyrophosphate
2.5.1.15
Dihydropteroate
6.3.2.12
Dihydrofolate
1.5.1.3
Tetrahydrofolate (THF)

Figure 3.2.20 Folate biosynthetic pathway

Table 3.2.19 Enzymes involved in folate biosynthesis, their EC numbers, a description, and an indication of the presence in four *Leptospira* species.

EC#	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
3.5.4.16	GTP cyclohydrolase	Yes	Yes	Yes	Yes
3.1.3.1	glycerophosphatase	Yes	Yes	Yes	Yes
4.1.2.25	dihydroneopterin aldolase	Yes	Yes	Yes	Yes
2.7.6.3	7,8-dihydro-6-hydroxymethylpterin pyrophosphokinase	Yes	Yes	Yes	Yes
6.3.5.8	aminodeoxychorismate synthase	Yes	Yes	Yes	Yes
4.1.3.38	4-amino-4-deoxychorismate lyase	Yes	Yes	Yes	Yes
2.5.1.15	dihydropteroate synthase	Yes	Yes	Yes	Yes
6.3.2.12	dihydrofolate synthase	Yes	Yes	Yes	Yes
1.5.1.3	dihydrofolate reductase	Yes	Yes	Yes	Yes

Folate, a water-soluble B vitamin, is the precursor of coenzyme tetrahydrofolate. Folate plays a role in protein metabolism, the formation of genetic material, cell growth and division, and pyridine dimer repair. All the genes for all the complete folate biosynthetic pathway enzymes are present in all four *Leptospira* species as shown in Figure 3.2.20 and Table 3.2.19.

3.2.5.3 Thiamine Biosynthesis

Thiamine, also known as vitamin B1, is a water-soluble vitamin. It is the precursor of coenzyme thiamine pyrophosphate. Thiamine is needed to process carbohydrates, fat, and protein. It is also necessary for the formation of ATP. It seems that all but one of the genes for the thiamine biosynthetic pathway enzymes are present in four *Leptospira* genomes (Figure 3.2.21 and Table 3.2.20). The last enzyme in thiamine biosynthesis, bifunctional isocitrate dehydrogenase kinase/phosphatase (AceK), was not observed in any of the four *Leptospira* genomes, and it has been reported that thiamine is required for the *Leptospira* culture medium base (Johnson and Gary, 1962). However, both the bifunctional isocitrate dehydrogenase kinase/phosphatase (AceK) with its conserved domain (COG4579: isocitrate dehydrogenase kinase/phosphatase) and phosphohistidine phosphatase (SixA) with its conserved domain (COG2062: phosphohistidine phosphatase) are annotated as EC 3.1.3.- (Phosphoric monoester hydrolases) by KEGG. Since all those *Leptospira* species do encode SixA, but do not encode AceK, it is possible that the leptospira species cannot synthesize thiamine *de novo*, or that SixA can replace the AceK that most other microorganisms use to convert thiamine phosphate to thiamine, but at a much slower rate that required for rapid cell growth in culture media.

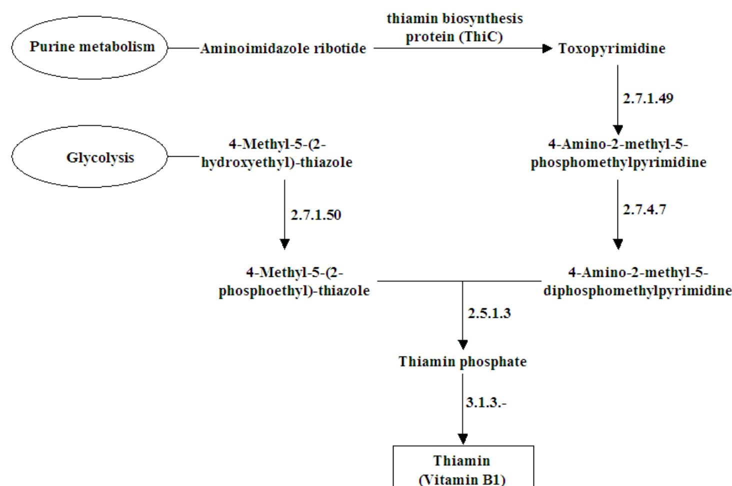


Figure 3.2.21 Thiamine biosynthetic pathway

Table 3.2.20 Enzymes involved in thiamine biosynthesis, their EC numbers, a description, and an indication of the presence in four *Leptospira* species.

EC#	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
2.7.1.49	hydroxymethylpyrimidine kinase	Yes	Yes	Yes	Yes
2.7.4.7	phosphomethylpyrimidine kinase	Yes	Yes	Yes	Yes
2.7.1.50	4-methyl-5-(beta-hydroxyethyl)thiazole kinase	Yes	Yes	Yes	Yes
2.5.1.3	thiamine-phosphate diphosphorylase	Yes	Yes	Yes	Yes
2.7.4.16	thiamine-phosphate kinase	Yes	Yes	Yes	Yes
3.1.3.-	phosphohistidine phosphatase (SixA)	Yes	Yes	Yes	Yes
3.1.3.-	bifunctional isocitrate dehydrogenase kinase/phosphatase (AceK)	No	No	No	No

3.2.5.4 Biotin Biosynthesis

Biotin, commonly referred to as vitamin H, is a water-soluble B vitamin. It is involved in the biosynthesis of fatty acids, gluconeogenesis, energy production, the metabolism of the branched-chain amino acids, and the *de novo* synthesis of purine nucleotides (Zempleni and Mock, 1999). As shown in Figure 3.2.22 and Table 3.2.21, all of the genes that encode the enzymes required for the synthesis of biotin are present in the all four *Leptospira* genomes. In addition, it is very interesting to notice

that all four genes involved in Biotin biosynthesis pathway form an operon in the large chromosome and conserved in all four *Leptospira* species.

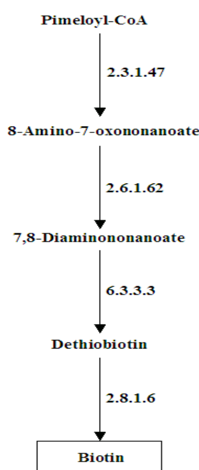


Figure 3.2.22 Biotin biosynthetic pathway

Table 3.2.21 Enzymes involved in biotin biosynthesis, their EC numbers, a description, and an indication of the presence in four *Leptospira* species.

EC#	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
2.3.1.47	8-amino-7-oxononanoate synthase	Yes	Yes	Yes	Yes
2.6.1.62	adenosylmethionine-8-amino-7-oxononanoate aminotransferase	Yes	Yes	Yes	Yes
6.3.3.3	dethiobiotin synthase	Yes	Yes	Yes	Yes
2.8.1.6	biotin synthase	Yes	Yes	Yes	Yes

3.2.5.5 Biosynthesis of Pantothenate and Coenzyme A

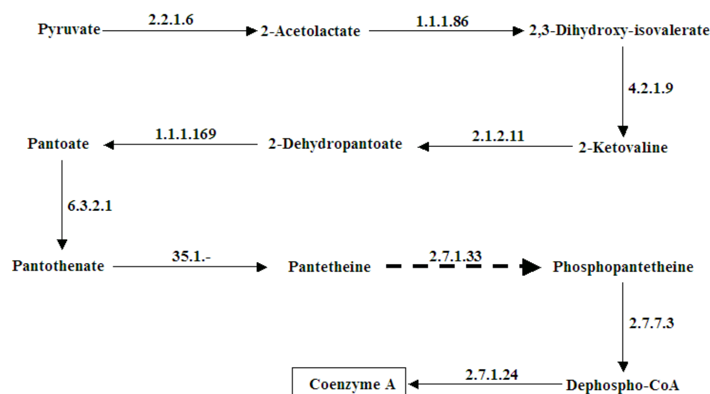


Figure 3.2.23 Pantothenate and CoA biosynthesis of four *Leptospira* species. The dotted arrow represents reactions carried out by pantothenate kinase, the candidate *Leptospira* gene found by isozyme analysis.

Pantothenate, a component of the vitamin B complex, serves as a precursor for coenzyme A and acyl carrier protein. A thorough examination of the four *Leptospira* genomic sequences revealed the presence of all the genes that encode the enzymes involved in the biosynthetic pathway for pantothenate and CoA except pantothenate kinase (EC 2.7.1.33) (Figure 3.2.23 and Table 3.2.22). A domain analysis revealed that *E. coli* pantothenate kinase includes the conserved domain, COG1072: panthothenate kinase, but there was no ORF from any of four *Leptospira* species that included this conserved domain. However, a novel pantothenate kinase encoded by *coaX* was recently identified from *Helicobacter pylori* (Brand and Strauss, 2005). It includes the conserved domain, COG1521: putative transcriptional regulator, distinct from the previously characterized *E. coli* pantothenate kinase. The domain analysis revealed that one possible gene from *L. pomona*, *lai*, *copenhageni*, and *grippotyphosa*, respectively, could encode the missing pantothenate kinase as it includes the same functional domain (COG1521). Thus all the four *Leptospira* species may be capable of synthesizing CoA *de novo* (Figure 3.2.23).

Table 3.2.22 Enzymes involved in pantothenate and CoA synthesis in the four *Leptospira* species.

EC#	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
2.2.1.6	acetolactate synthase	Yes	Yes	Yes	Yes
1.1.1.86	acetolactate reductoisomerase	Yes	Yes	Yes	Yes
4.2.1.9	dihydroxy-acid dehydratase	Yes	Yes	Yes	Yes
2.1.2.11	ketopantoate hydroxymethyltransferase	Yes	Yes	Yes	Yes
1.1.1.169	ketopantoate reductase	Yes	Yes	Yes	Yes
6.3.2.1	pantothenate synthetase	Yes	Yes	Yes	Yes
3.5.1.-	hydrolase	Yes	Yes	Yes	Yes
2.7.1.33	pantothenate kinase	Yes	Yes	Yes	Yes
2.7.7.3	dephospho-CoA pyrophosphorylase	Yes	Yes	Yes	Yes
2.7.1.24	dephospho-CoA kinase	Yes	Yes	Yes	Yes

3.2.6 Macromolecular Metabolism

Macromolecular metabolism broadly includes the metabolic processes involving DNA, RNA, and protein synthesis.

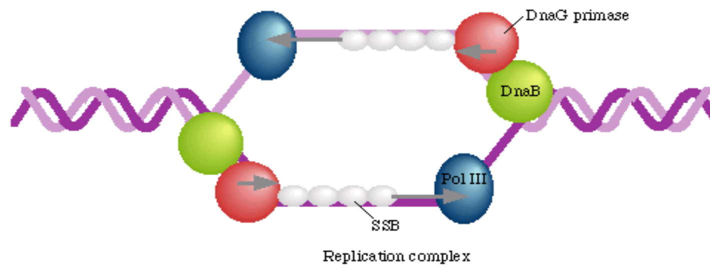
3.2.6.1 DNA Metabolism

Generally, replication in *Leptospira* species is similar to that of *E. coli* as suggested by the genomic analysis, although the gene for DNA polymerase II used in *E. coli* for DNA repair was absent in all four *Leptospira* genomes studied. Thus it is likely that DNA repair is accomplished using DNA polymerase I and/or III. In *E. coli*, the DNA polymerase III holoenzyme is assembled from subunits tau, beta, core, and clamp loader. The core consists of subunits alpha, epsilon, and theta, while the clamp loader consists of subunits gamma, delta, delta, prime, chi, and psi (Table 3.2.23). Among all the subunits, only alpha, tau, gamma, delta, and beta are conserved in all the sequenced bacteria, while subunits theta, delta prime, chi, psi, and epsilon of DNA polymerase III are absent from *Leptospira* species.

Table 3.2.23 Components of DNA polymerase III

<table>
<tr><th colspan="3">DNA polymerase III subunits and subassemblies</th></tr>
<tr><th>Subunit</th><th>Function</th><th>Subassembly (complex)</th></tr>
<tr><td>alpha[1,2]</td><td>DNA polymerase</td><td rowspan="3">core</td></tr>
<tr><td>epsilon[1]</td><td>3'→5' exonuclease</td></tr>
<tr><td>theta[1]</td><td>stimulates 3'→5' exonuclease</td></tr>
<tr><td>tau[1,2]</td><td>dimerizes cores, activates DnaB helicase activity</td><td></td></tr>
<tr><td>gamma[1,2]</td><td>binds ATP</td><td rowspan="5">clamp loader</td></tr>
<tr><td>delta[1,2]</td><td>unknown</td></tr>
<tr><td>delta prime[1]</td><td>stimulates clamp loading</td></tr>
<tr><td>chi[1]</td><td>removal of primase</td></tr>
<tr><td>psi[1]</td><td>unknown</td></tr>
<tr><td>beta[1,2]</td><td>sliding clamp.</td><td></td></tr>
</table>

Note: 1 present in *E. coli*, 2 present in *Leptospira* species.



Proteins Required for DNA Replication

SSB	DnaA	DnaB	DnaC
DnaG	PolA	LigA	GyrA/B

DNA polymerase

III		I
alpha	gamma/tau	PolA
delta	beta	

Figure 3.2.24 Demonstration of DNA replication in the four *Leptospira* species with the proteins involved in the replication process.

Table 3.2.24 Complete list of all proteins involved in replication of four *Leptospira* species.

Gene name	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
SSB	single-stranded DNA binding protein	Yes	Yes	Yes	Yes
DnaA	chromosomal replication initiator protein	Yes	Yes	Yes	Yes
DnaB	replicative DNA helicase B	Yes	Yes	Yes	Yes
DnaC	DNA replication protein	Yes	Yes	Yes	Yes
DnaG	DNA primase	Yes	Yes	Yes	Yes
LigA	DNA ligase	Yes	Yes	Yes	Yes
GyrA	DNA gyrase subunit A	Yes	Yes	Yes	Yes
GyrB	DNA gyrase subunit B	Yes	Yes	Yes	Yes
PolA	DNA polymerase I	Yes	Yes	Yes	Yes
DnaE	DNA polymerase III, subunit alpha	Yes	Yes	Yes	Yes
DnaX1	DNA polymerase III, subunit gamma and tau	Yes	Yes	Yes	Yes
HolA	DNA polymerase III, subunit delta	Yes	Yes	Yes	Yes
DnaN	DNA polymerase III, subunit beta	Yes	Yes	Yes	Yes

3.2.6.2 RNA Metabolism

3.2.6.2.1 Transcription

Analysis of the four *Leptospira* genomes reveals that they contain the transcription machinery similar to that of *E. coli* since homologs of all genes for the proteins required for the transcription in *E. coli* are present as shown in Table 3.2.25. In addition, three other sigma factors (RpoN, RpoD, and RpoE) needed for the transcription of most cellular genes also were identified in the four *Leptospira* species. Each species has one sigma 54 factor and two sigma 70 factors. But it is very interesting to notice that *L. copenhageni* has one more copy of RNA polymerase ECF-type sigma factor (11) than the other three *Leptospira* species (10).

Table 3.2.25 List of encoded proteins involved in transcription of four *Leptospira* species.

Gene name	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
RpoN	transcription initiation factor	Yes	Yes	Yes	Yes
NusA	transcription elongation factor	Yes	Yes	Yes	Yes
Rho	transcription termination factor	Yes	Yes	Yes	Yes
NusB	transcription antitermination protein	Yes	Yes	Yes	Yes
RpoA	DNA-directed RNA polymerase alpha chain	Yes	Yes	Yes	Yes
RpoB	DNA-directed RNA polymerase beta chain	Yes	Yes	Yes	Yes
RpoC	DNA-directed RNA polymerase beta prime chain	Yes	Yes	Yes	Yes
SrmB	ATP-dependent RNA helicase	Yes	Yes	Yes	Yes
RpoN	RNA polymerase sigma 54 factor	Yes	Yes	Yes	Yes
RpoD	RNA polymerase sigma 70 factor	Yes	Yes	Yes	Yes
RpoE	RNA polymerase ECF-type sigma factor	Yes	Yes	Yes	Yes

3.2.6.2.2 Processing of Ribosomal and Transfer RNAs

As in *E. coli*, the 70S ribosomes are composed of three rRNAs, 5S, 16S, and 23S, and more than 50 ribosomal proteins. 21 ribosomal proteins (S1-S21) makeup the 30S ribosomal subunit, while 36 ribosomal proteins (L1-L36) are components of the 50S ribosomal subunit. It is interesting to note that all four *Leptospira* species lack the gene of L33 (rpmG). A mutagenesis study of *E. coli* revealed that protein

L33 has no significant effect on ribosome synthesis or function (Maguire and Wild, 1997) and if the same is true in *Leptospira*, the 70S *Leptospira* ribosomes can be assembled normally even without rpmG gene.

Generally, several ribosomal RNA operons are encoded in a bacterial genome, but they are scattered throughout the large chromosomes of four *Leptospira* species. It needs to be mentioned that each species has different number of rRNAs as shown in Table 3.2.26.

Table 3.2.26 The distribution of ribosomal RNAs in four *Leptospira* species.

	pomona	***grippotyphosa***	***lai***	***copenhageni***
5S rRNA	1	1	1	1
16S rRNA	2	1	2	2
23S rRNA	2	2	1	2

Phylogenetic analysis was performed to 16s rRNA genes from all four *Leptospira* species. It showed clearly that these four leptospiries can be grouped: *grippotyphosa* with *pomon,* and *copenhageni* with *lai*, two groups that correlates well with their observed host ranges.

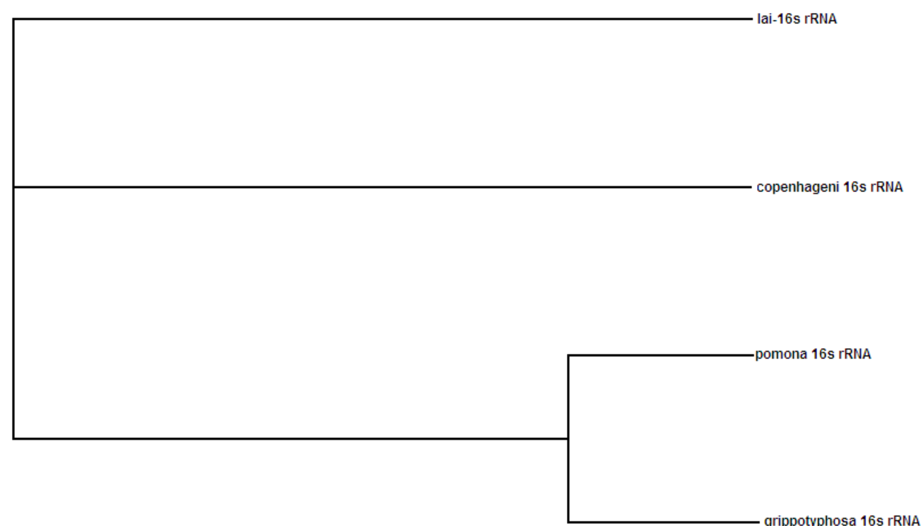


Figure 3.2.25 Phylogenetic tree of 16s rRNA among four *Leptospira* species.

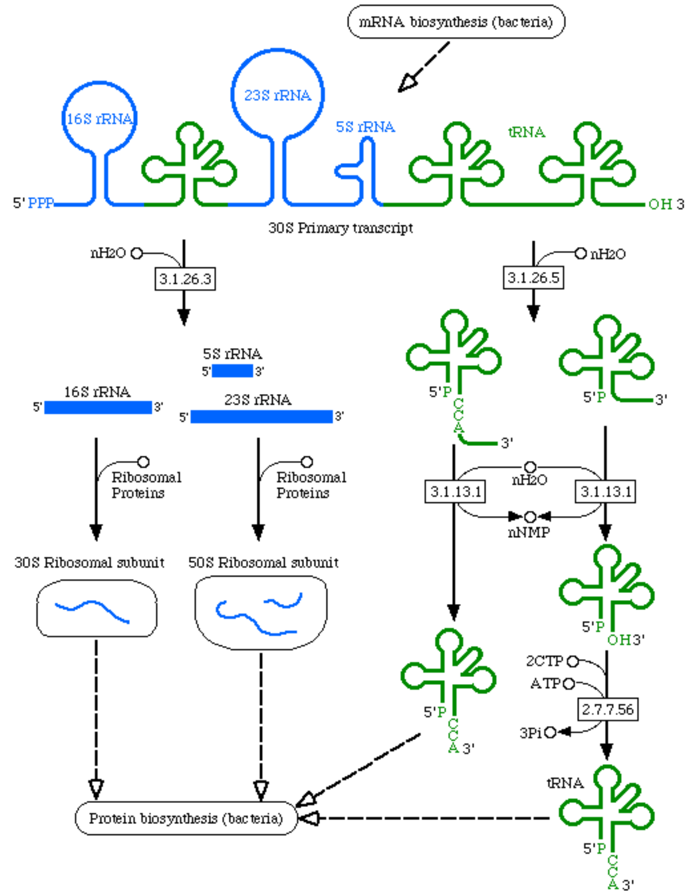


Figure 3.2.26 Schematic illustration of the processing steps of rRNA and tRNA (adapted from KEGG).

Table 3.2.27 Enzymes involved in the rRNA/tRNA processing of four *Leptospira* species.

EC#	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
3.1.26.3	Ribonuclease III	Yes	Yes	Yes	Yes
3.1.26.5	Ribonuclease P	Yes	Yes	Yes	Yes
3.1.13.1	Ribonuclease II	Yes	Yes	Yes	Yes
2.7.7.56	Ribonuclease PH	Yes	Yes	Yes	Yes

Several enzymes are needed to process the rRNA/tRNA transcription (Figure 3.2.26). Ribonuclease III (EC 3.1.26.3) cleaves multimeric tRNA precursor at the spacer region, and also is involved in processing precursor rRNA (Rech, et al., 1980). Ribonuclease P (EC 3.1.26.5) cleaves sequences from the 5' ends of precursors of tRNAs to produce the mature 5' termini of the tRNAs (Pace and Smith, 1990).

Ribonuclease II (EC 3.1.13.1) processes 3'-terminal extra-nucleotides of monomeric tRNA precursors, following the action of ribonuclease P (Shimura, et al., 1978). Ribonuclease PH (EC 2.7.7.56) trims the 3'-terminus of tRNA precursors to produce a mature 3'-terminus on tRNA (Deutscher, et al., 1988). Of the four above genes, ribonuclease PH could not be observed in the four *Leptospira* species. The domain analysis revealed that one putative polynucleotide phosphorylase from *Leptospira* species includes the functional domain (COG0689: RNase PH) conserved in ribonuclease PH, indicating that this gene may replace ribonuclease PH.

3.2.6.2.3 Post Transcriptional Modification of Ribosomal and Transfer RNA Precursors to Mature RNAs

Table 3.2.28 RNA modification enzymes.

EC#/Gene name	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
yjfH	rRNA methylase	Yes	Yes	Yes	Yes
rluC	ribosomal large subunit pseudouridine synthase C	Yes	Yes	Yes	Yes
rluD	ribosomal large subunit pseudouridine synthase D	Yes	Yes	Yes	Yes
ksgA	dimethyladenosine transferase	Yes	Yes	Yes	Yes
trmD	tRNA (Guanine-N1)-methyltransferase	Yes	Yes	Yes	Yes
trmU	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase	Yes	Yes	Yes	Yes
2.1.1.-	RNA methyltransferase	Yes	Yes	Yes	Yes
2.4.2.29	tRNA-guanine transglycosylase	Yes	Yes	Yes	Yes
miaA	tRNA delta(2)-isopentenylpyrophosphate transferase	Yes	Yes	Yes	Yes
truA	tRNA pseudouridine synthase A	Yes	Yes	Yes	Yes
truB	tRNA pseudouridine synthase B	Yes	Yes	Yes	Yes
queA	S-adenosylmethionine--tRNA ribosyltransferase-isomerase	Yes	Yes	Yes	Yes

The post transcriptional modification of RNA in bacteria requires a large number of enzymes (Björk, 1996). As in *E. coli*, several genes encoding rRNA modification enzymes were identified in the four *Leptospira* genomes (Table 3.2.28). They include rRNA methylase, required for 16S rRNA processing, and pseudouridine

synthase, required for ribosomal large subunit processing. In addition, at least 8 genes encoding tRNA modification enzymes are present in all four *Leptospira* species (Table 3.2.28).

3.2.6.3 Aminoacyl-tRNA Biosynthesis

The aminoacyl-tRNA synthetases catalyse the attachment of an amino acid to its cognate transfer RNA molecule. These proteins differ widely in size and oligomeric state, and have limited sequence homology (Eriani, et al., 1990). Based on their secondary structure, the 20 aminoacyl-tRNA synthetases are divided into two classes, class I and class II (Sugiura, et al., 2000; Perona, et al., 1993). In reactions catalyzed by the class I aminoacyl-tRNA synthetases, the aminoacyl group is coupled to the 2'-hydroxyl of the tRNA, while, in class II reactions, the 3'-hydroxyl site is preferred. Class I synthetases are specific for arginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine, tryptophan, and valine, while class II synthetases are specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine. Some organisms, such as *E. coli*, use a distinct aminoacyl-tRNA synthetase (aaRS) for each amino acid species to be charged on the cognate tRNA, while others lack one or several aaRSs (Tumbula, et al., 2000). Specifically, all aaRSs except glutaminyl-tRNA synthetase (EC 6.1.1.18) were observed in the four *Leptospira* genomes. This indicates that an alternative pathway is required to aminoacylate the corresponding glutaminyl-tRNA(s) correctly (Figure 3.2.27). The first step of this pathway is the misacylation of $tRNA^{Gln}$ with glutamate, catalyzed by glutamyl-tRNA synthetase (EC 6.1.1.17) (Lapointe, et al., 1986), followed by the transamidation of $Glu\text{-}tRNA^{Gln}$ into Gln-

tRNA^{Gln}, catalyzed by a tRNA-dependent amidotransferase (EC 6.3.5.-) (Curnow, et al., 1997). All three subunits of tRNA-dependent amidotransferase (gatA, gatB, and gatC) were identified in four *Leptospira* genomes.

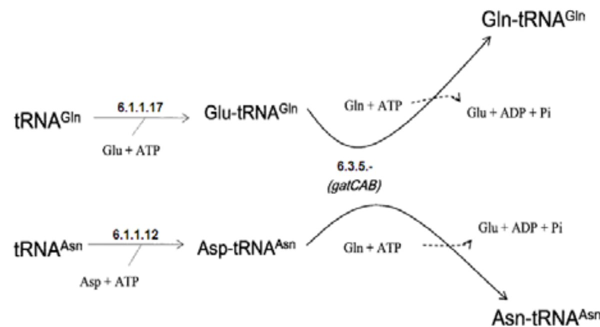


Figure 3.2.27 Putative pathways for Gln-tRNA^{Gln} and Asn-tRNA^{Asn} synthesis in four *Leptospira* species.

As mentioned early, the four *Leptospira* species cannot synthesize asparagine as it lacks the gene for aspartate aminotransferase (EC 6.3.1.1). However, Asn-tRNA^{Asn} synthesis still is possible because tRNA-dependent amidotransferase (EC 6.3.5.-) can transamidate both Glu-tRNA^{Gln} and Asp-tRNA^{Asn} (Akochy, et al., 2004).

In addition, initiation of protein synthesis in prokaryotes employs fmet-tRNA^{fmet}, where the esterified methionine is formylated. At first, the methionine is esterified to tRNA^{fmet} by the methionyl-tRNA synthetase (EC 6.1.1.10) and then the esterified methionine is formylated via the anzyme methionyl-tRNA formylatransferase (EC 2.1.2.9). The genes for both of these enzymes were observed in the genomes of all four *Leptospira* species.

3.2.6.4 Protein Biosynthesis

Proteins are synthesized from mRNA templates by a highly conserved process as shown in Figure 3.2.28 and the genes coding for all enzymes necessary for translation are listed in Table 3.2.29. The translation process can be divided into three stages: initiation, elongation, and termination. In the first stage, two initiation factors

(IF-1 and IF-3) are bound to the 30S ribosomal subunit. The mRNA and initiator N-formylmethionyl tRNA then join the complex, with IF-2 specifically recognizing the initiator tRNA. IF-3 then is released, allowing a 50S ribosomal subunit to associate with the complex. This association leads to the release of IF-1 and IF-2. Finally, a 70S initiation complex is formed.

In the second stage, the polypeptide chain is elongated. The ribosome has three sites for tRNA binding, designated the “P” (peptidyl), “A” (aminoacyl), and “E” (exit) sites. The initiator methionyl tRNA is bound at the “P” site. The aminoacyl tRNA is escorted to into the “A” site by an elongation factor EF-Tu, which is complexed to GTP. Then, elongation factor EF-Tu is released with the hydrolysis of GTP to GDP and elongation factor EF-Ts regenerates EF-Tu/GDP to EF-Tu/GTP. Once EF-Tu has left the ribosome, a peptide bond can be formed between the initiator methionyl tRNA at the “P” site and the second aminoacyl tRNA at the “A” site by peptidyl transferase. Another elongation factor P (EF-P) enhances the synthesis of the first peptide bond initiated by N-formylmethionine (Aoki, et al., 1997). Finally, the ribosome translocates along the mRNA to the next codon accompanied by the hydrolysis of the GDP associated with elongation factor EF-G. This frees the “A” site for the next aminoacyl-tRNA to bind while the growing peptide chain is residing on the “P” site.

In the third stage, the peptide chain is terminated when a stop codon (UAA, UAG, or UGA) is translocated into the “A” site of the ribosome. Cells contain release factors that recognize the signals and terminate protein synthesis. The first release

factor RF-1 recognizes UAA or UAG, and the second release factor RF-2 recognizes UAA or UGA. The third release factor RF-3 does not recognize specific termination codons but acts together with RF-1 and RF-2. The release factors (RF-1, RF2, and RF-3) bind to a termination codon at the "A" site and stimulate hydrolysis of the bond between the tRNA and the polypeptide chain at the "P" site, resulting in release of the completed polypeptide from the ribosome. The fourth release factor RF-4 enhances the dissociation of ribosomes from mRNA template after termination of translation (Heurgué-Hamard, et al., 1998). As stated above, the genes encoding the enzymes required for protein translation identified in all four *Leptospira* species are listed in Table 3.2.29.

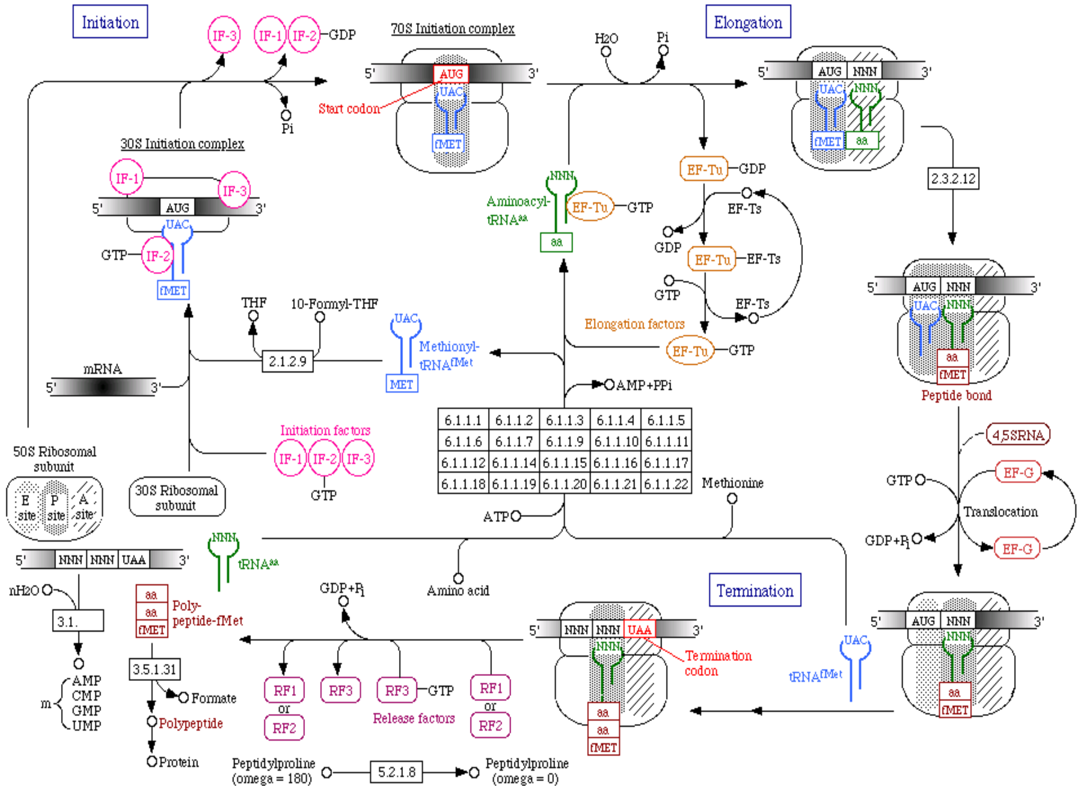


Figure 3.2.28 Schematic representation of the translation process as found in the four *Leptospira* species (adapted from KEGG)

Table 3.2.29 Proteins involved in protein synthesis in four *Leptospira* species.

EC#/Gene Name	Desription	Gene in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
2.1.2.9	methionyl-tRNA formyltransferase	Yes	Yes	Yes	Yes
2.3.2.12	peptidyl transferase (50S ribosomal protein L16)	Yes	Yes	Yes	Yes
3.4.11.18	methionine aminopeptidase	Yes	Yes	Yes	Yes
3.5.1.31	formylmethionine deformylase	Yes	Yes	Yes	Yes
InfA	translation initiation factor IF-1	Yes	Yes	Yes	Yes
InfB	translation initiation factor IF-2	Yes	Yes	Yes	Yes
InfC	translation initiation factor IF-3	Yes	Yes	Yes	Yes
Tsf	elongation factor EF-Ts	Yes	Yes	Yes	Yes
Tuf	elongation factor EF-Tu	Yes	Yes	Yes	Yes
FusA	elongation factor EF-G	Yes	Yes	Yes	Yes
Efp	elongation factor EF-P	Yes	Yes	Yes	Yes
PrfA	peptide chain release factor RF-1	Yes	Yes	Yes	Yes
PrfB	peptide chain release factor RF-2	Yes	Yes	Yes	Yes
PrfC	peptide chain release factor RF-3	Yes	Yes	Yes	Yes
Frr	ribosome recycling factor	Yes	Yes	Yes	Yes

3.2.6.5 Protein Degradation

Hydrolysis of peptides by the protease (also called peptidase) to free amino acids is one of the central activities within a cell. The degradation of proteins is an important mechanism for regulating many pathways, degrading mis-folded proteins, and degrading proteins during starvation to provide amino acids for energy. Usually, proteases are classified into the different families, based on the evolutionary relationship between each other. Each family is identified by an upper-case letter representing the catalytic type (S for serine-type, T for threonine-type; C for cysteine-type, A for aspartic-type, M for metallo-type, and U for unknown type) (Puente, et al., 2003). As shown in Tbale 3.2.30, analysis of the genomes of the four *Leptospira* species revealed a slightly different distribution of the three protease families, C, M, and S.

Table 3.2.30 Protease distribution in *Leptospira* species.

Protease Family	members in each family			
	pomona	*grippotyphosa*	*lai*	*copenhageni*
A	1	1	1	1
C	2	2	3	4
M	34	38	35	36
T	0	0	0	0
S	7	8	7	8
U	1	1	1	1

3.2.7 Cell Wall

3.2.7.1 Peptidoglycan

In order to maintain shape and withstand intracellular pressure, most bacteria are surrounded by a cell wall consisting mainly of the peptidoglycan, a polymer of a repeating disaccharide-peptide unit, where the pentapeptide chains attached to adjacent sugar molecules are cross-linked. The synthesis of peptidoglycans can be divided into three stages (van Heijenoort, 1998; Mirelman, et al., 1976).

The first stage involves synthesis of two amino sugars precursors UDP-NAG and UDP-NAM in the cytoplasm. After D-glutamate and D-alanyl-D-alanine are synthesized respectively, UDP–NAM is linked to pentapeptide, forming the basic subunit of peptidoglycan.

In the second stage, the lipid carrier (lipid P) transfers UDP–NAM-pentapeptide through the inner membrane to the periplasm, where UDP–NAM-pentapeptide is linked to the UDP-NAG sugar to form the disaccharide precursor. Then, lipid PP is hydrolyzed to lipid P that can reenter the cycle.

In the third stage, after the newly synthesized peptidoglycan subunit is transferred to the growing point of the cell wall's peptidoglycan, the sugars are polymerized, and the peptide chains are cross-linked.

As shown in Table 3.2.31, all the enzymes involved in peptidoglycan synthesis except UDP-N-acetylmuramoylpentapeptide lysine N6-alanyltransferase (EC 2.3.2.10) were found in four *Leptospira* species. The domain search revealed that one possible ORF could encode the missing transferase as it includes the same functional domain (COG2348: uncharacterized protein involved in methicillin resistance) conserved in UDP-N-acetylmuramoylpentapeptide lysine N6-alanyltransferase. Figure 3.2.29 illustrates peptidoglycan synthesis as reconstructed in the four *Leptospira* species.

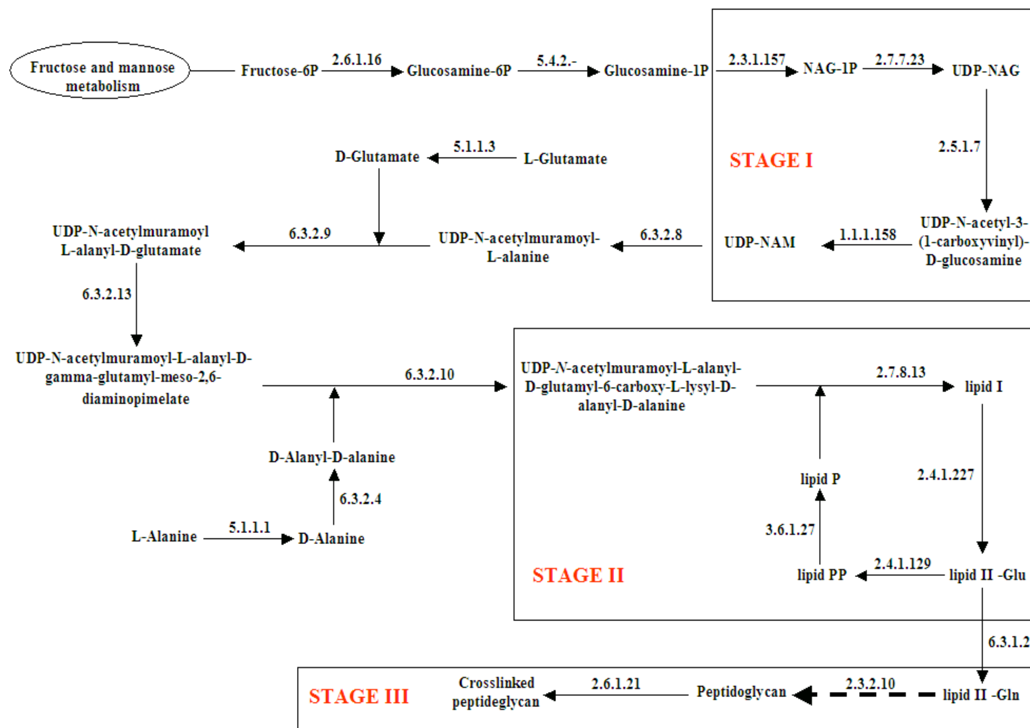


Figure 3.2.29 Peptidoglycan biosynthesis

Table 3.2.31 Enzymes involved in peptidoglycan biosynthesis

EC#	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
2.6.1.16	glucosamine--fructose-6-phosphate aminotransferase	Yes	Yes	Yes	Yes
5.4.2.-	Phosphoglucosamine mutase	Yes	Yes	Yes	Yes
2.3.1.157	glucosamine-1-phosphate N-acetyltransferase	Yes	Yes	Yes	Yes
2.7.7.23	UDP-N-acetylglucosamine pyrophosphorylase	Yes	Yes	Yes	Yes
2.5.1.7	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	Yes	Yes	Yes	Yes
1.1.1.158	UDP-N-acetylmuramate dehydrogenase	Yes	Yes	Yes	Yes
6.3.2.8	UDP-N-acetylmuramate--alanine ligase	Yes	Yes	Yes	Yes
5.1.1.3	glutamate racemase	Yes	Yes	Yes	Yes
6.3.2.9	UDP-N-acetylmuramoylalanine--D-glutamate ligase	Yes	Yes	Yes	Yes
6.3.2.13	UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase	Yes	Yes	Yes	Yes
5.1.1.1	alanine racemase	Yes	Yes	Yes	Yes
6.3.2.4	D-alanylalanine synthetase	Yes	Yes	Yes	Yes
6.3.2.10	UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl-D-alanine ligase	Yes	Yes	Yes	Yes
2.7.8.13	phospho-N-acetylmuramoyl-pentapeptide-transferase	Yes	Yes	Yes	Yes
2.4.1.227	UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase	Yes	Yes	Yes	Yes
2.4.1.129	peptidoglycan glycosyltransferase	Yes	Yes	Yes	Yes
3.6.1.27	undecaprenyl-diphosphatase	Yes	Yes	Yes	Yes
6.3.1.2	glutamine synthetase	Yes	Yes	Yes	Yes
2.3.2.10	UDP-N-acetylmuramoylpentapeptide lysine N6-alanyltransferase	Yes	Yes	Yes	Yes
2.6.1.21	penicillin-binding protein	Yes	Yes	Yes	Yes

3.2.7.2 Lipopolysaccharides (LPS)

Lipopolysaccharide (LPS) contributes greatly to the structural integrity of the bacteria and protects them from the host immune defenses. An LPS contain three parts: lipid A, core polysaccharide, and O-specific chain (Figure 3.2.30).

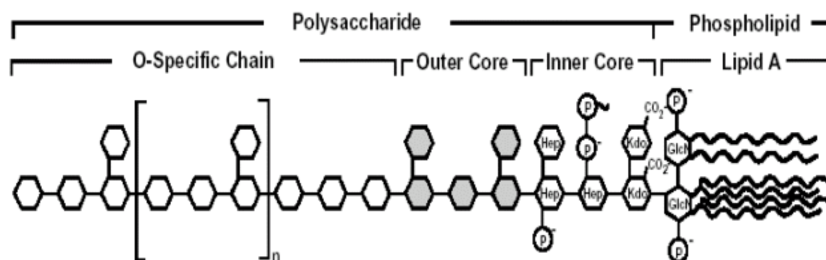


Figure 3.2.30 A schematic diagram of a lipopolysaccharide molecule (Bulach, 2000)

The first part, lipid A, consists of six fatty acyl chains linked to two glucosamine residues. The second part, core polysaccharide, is attached to lipid A through 3-hydroxy-D-manno-octulosonate (KDO). This core is further divided into two regions, an inner core and an outer core. The inner core consists of KDO, heptose, and phosphate, and the outer core consists of hexoses. The third part, O-specific chain, determines the antigenic specificity of the organism. The O-specific chain consists of four to six sugars that may be repeated up to 50 times, making it the most variable region, while lipid A is the most conserved.

The core and O-specific chain are synthesized in a manner similar to peptidoglycans, where the sugar residues are synthesized and assembled on a lipid P carrier. Figure 3.2.30 described the biosynthesis pathway of lipid A. All the genes in this pathway except KDO 8-P phosphatase (EC 3.1.3.45) were identified from *Leptospira* species. A further domain analysis revealed that no any ORF from *Leptospira* species includes the functional domain (COG1778: low specificity phosphatase) conserved in KDO 8-P phosphatase. However, since an amino acid sequence analysis indicates that KDO 8-P phosphatase is a member of the haloacid dehalogenase hydrolase superfamily (Wu and Woodard, 2003) and one haloacid dehalogenase-like hydrolase could be identified in all four *Leptospira* species, this enzyme may function as the missing KDO 8-P phosphatase (Table 3.2.32).

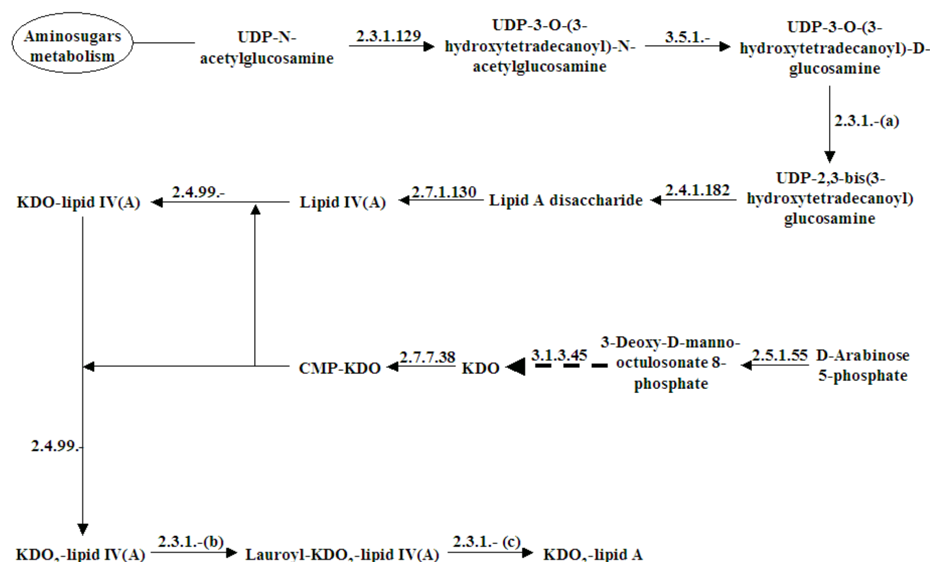


Figure 3.2.31 Lipid A biosynthetic pathway

Table 3.2.32 Enzymes involved in Lipid A and KDO biosynthesis.

EC#	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
2.3.1.129	UDP-N-acetylglucosamine acyltransferase (lpxA)	Yes	Yes	Yes	Yes
3.5.1.-	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase (lpxC)	Yes	Yes	Yes	Yes
2.3.1.-(a)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (lpxD)	Yes	Yes	Yes	Yes
2.4.1.182	Lipid-A-disaccharide synthase (lpxB)	Yes	Yes	Yes	Yes
2.7.1.130	Tetraacyldisaccharide 4'-kinase	Yes	Yes	Yes	Yes
2.4.99.-	3-deoxy-D-manno-octulosonic-acid transferase (KDO transferase) (kdtA)	Yes	Yes	Yes	Yes
2.3.1.-(b)	Apolipoprotein N-acyltransferase (lnt)	Yes	Yes	Yes	Yes
2.3.1.-(c)	Lipid A biosynthesis lauroyl acyltransferase;lauroylacyltransferase (htrB)	Yes	Yes	Yes	Yes
2.5.1.55	KDO 8-P synthase	Yes	Yes	Yes	Yes
3.1.3.45	3-deoxy-manno-octulosonate-8-phosphatase	Yes	Yes	Yes	Yes
2.7.7.38	3-deoxy-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase) (kdsB)	Yes	Yes	Yes	Yes

3.2.8 Transport Proteins

Transport proteins are responsible for the passage of substances, from small ions to large molecules, across the cell membrane. Therefore, transport proteins can

compensate for the incomplete metabolic pathways by providing the cell with needed precursors for these reactions.

3.2.8.1 Cation and Anion Transporters

Analysis of the four *Leptospira* genomes revealed the presence of similar ion transporter genes as listed in Table 3.2.33.

Table 3.2.33 Cation and anion transporters found in four *Leptospira* species.

Gene name	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
cysT/cysW	sulfate ABC transport system	Yes	Yes	Yes	Yes
eriC	chloride ion channel	Yes	Yes	Yes	Yes
---	phosphate transporter	Yes	Yes	Yes	Yes
trkH	potassium uptake protein	Yes	Yes	Yes	Yes
kefA	potassium efflux system	Yes	Yes	Yes	Yes
kefB	potassium efflux system	Yes	Yes	Yes	Yes
kefC	potassium efflux system	Yes	Yes	Yes	Yes
kdpA	potassium-transporting ATPase, A chain	Yes	Yes	Yes	Yes
kdpB	potassium-transporting ATPase, B chain	Yes	Yes	Yes	Yes
kdpC	potassium-transporting ATPase, C chain	Yes	Yes	Yes	Yes
nctP1	sodium transporter	Yes	Yes	Yes	Yes
nctP2	sodium transporter	Yes	Yes	Yes	Yes
amtB	ammonium transporter	Yes	Yes	Yes	Yes
corA	magnesium and cobalt transporter	Yes	Yes	Yes	Yes
atc	copper-transporting ATPase	Yes	Yes	Yes	Yes
mgtA	magnesium-transporting ATPase	Yes	Yes	Yes	Yes
mgtE	magnesium transporter	Yes	Yes	Yes	Yes
---	mercuric ion permease	Yes	Yes	Yes	Yes
feoB	ferrous iron transporter	Yes	Yes	Yes	Yes

3.2.8.2 Phosphotransferase System (PTS)

PTS is used to import carbohydrates into the cell, and it has been described previously in several bacteria (Reizer, et al., 1993; 1999). PTS consists of two cytoplasmic energy-coupling proteins (enzyme I (EI) and HPr) and a permease enzyme II (EII) (Figure 3.2.32). At first, the EI transfers the phosphate group (***P***)

from phosphoenol-pyruvate (PEP) to the phosphocarrier protein HPr. Then, the phosphoryl group is delivered to the carbohydrate via EII with three functional domains IIA, IIB, and IIC. The IIA domain becomes phophorylated by HPr and further passes the phosphate to the IIB domain. Finally, IIB domain phosphorylates the carbohydrate as IIC domain translocates it into the cell. It is noteworthy that EI and HPr are general PTS components, while EIIs are specific for one or a few carbohydrates (Saier, 2000).

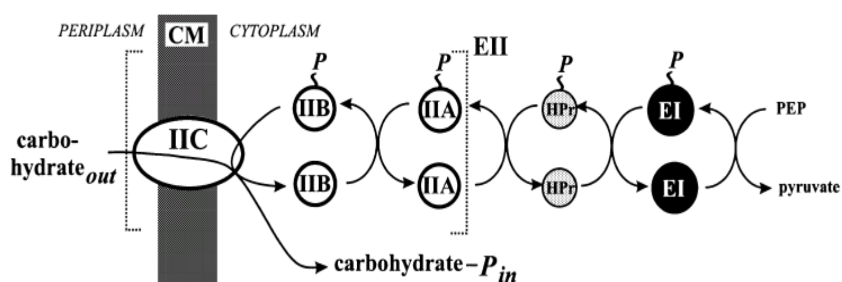


Figure 3.2.32 The phosphotransferase system in *E. coli* (Kotrba, et al., 2001)
Note: EI-enzyme I, EII-enzyme II, PEP-phosphoenol-pyruvate, CM-cytomembrane.

Enzyme I (ptsI, EC 2.7.3.9), phosphocarrier protein HPr (ptsH), and fructose-specific IIA component (ptsN, EC 2.7.1.69) were identified from the four *Leptospira* genomes. However, IIB and IIC were not present. To investigate further, an isozyme analysis (Green and Karp, 2004) was performed to search for the missing IIB. To accomplish this, 1599 isozyme amino acid sequences were downloaded from NCBI and each sequence was compared with each *L. pomona* ORF by blastp. When the isozyme from *E. coli UTI89* is used as the query sequence, the best candidate identified from *L. pomona* was ORF163092 with E value of 0.25 for the missing IIB. Then an isozyme analysis (Green and Karp, 2004) was performed to search for the missing IIC. Each of 640 isozyme amino acid sequences was compared with each *L.*

pomona ORF by blastp. When the isozyme from *Salmonella enterica* is used as the query sequence, the best candidate identified from *L. pomona* was ORF1380300 with E value of 0.007 for the missing IIC. Actually, *L. pomona* ORF163092 is assigned as Glycerol-3-phosphate dehydrogenase, and *L. pomona* ORF1380300 is assigned as cytochrome C oxidase polypeptide I. Therefore, ORF163092 and ORF1380300 could not function as IIB and IIC, respectively. Similarly, the isozyme analysis was performed to the other three *Leptospira* species. No gene was identified that could substitute for the function of IIB or IIC. Taken together, neither of the four *Leptospira* species includes the complete phosphotransferase system.

Table 3.2.34 Proteins involved in the phosphotransferase system in four Leptospira species.

Gene name	**Desription**	**Gene in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
EI	PTS system Enzyme I	Yes	Yes	Yes	Yes
HPr	phosphocarrier protein	Yes	Yes	Yes	Yes
EII A	PTS system Enzyme II A component	Yes	Yes	Yes	Yes
EII B	PTS system Enzyme II B component	No	No	No	No
EII C	PTS system Enzyme II C component	No	No	No	No

It is interesting to notice that *Leptospira* species seem to have one glucose uptake system, a glucose-sodium symporter that is dependent on a sodium gradient across the bacterial membrane. *pomona* encodes two sugar transporters, *grippotyphosa* two, and *lai* also two, but *copenhageni* encodes only one. Therefore, the *Leptospira* species may uptake sugar only through the transporter, as opposed to other bacteria that import sugar via the phosphotransferase system.

3.2.9 Protein Export and Secretion

Prokaryotes possess two parallel and complementary pathways for the export of proteins across the cytoplasmic membrane and into different cellular compartments. The first pathway is known as the general secretory (Sec) pathway (Fekkes and Driessen, 1999) and has been studied most extensively in *E. coli* (Stephenson, 2005). Multiple proteins are involved in this process (Figure 3.2.33). SecB, a chaperone, binds to the mature part of the preprotein and prevent folding. SecA, an ATPase, binds the signal sequence of the preprotein. SecY, SecE, and Sec G form a heterotrimeric complex. SecD, SecF, YajC, and YidC form a heterotetrameric complex. These two complexes then form the translocation channel. Once SecA approaches to the channel, SecB is released with a concomitant binding of ATP by SecA. Then ATP binding inserts the signal sequence into the membrane. SecA pushes the preprotein through the transmembrane pore upon hydrolysis of ATP. Once translocated into the membrane, the exported protein is fastened to the membrane by the signal peptide. Signal peptidase cleaves off the signal peptide thereby allows the protein to travel to it final destination.

The core of the Sec pathways of *Leptospira* species is generally equivalent to that of *E. coli*. However, the striking difference lies in that the chaperone SecB is absent in *Leptospira* species (Table 3.2.35). Domain analysis revealed that no *Leptospira* ORF could be found that corresponded the functional domain (COG1952: preprotein translocase subunit SecB) conserved in *E. coli* SecB. To investigate further, an isozyme analysis (Green and Karp, 2004) was performed to search for the missing SecB. To accomplish this, 779 isozyme amino acid sequences were downloaded from NCBI and each sequence was compared with each *L. pomona* ORF

by blastp. When the isozyme from *Rickettsia conorii* was used as the query sequence, the best candidate identified from *L. pomona* was ORF399057 with a very high E value of 0.039 for the missing SecB. Therefore, ORF399057 might not function as SecB. The similar results were obtained from the other three *Leptospira* species. It is very interesting to also note that no homologous gene for *E. coli* SecB was identified in *Bacillus subtilis* as the *Leptospira* species (Kunst, et al., 1997). Yamane and colleges discovered that gene Ffh from *Bacillus subtilis* may function as SecB (Yamane, et al., 2004). However, further domain analysis did reveal that one possible flagellar GTP-binding protein from *Leptospira* species includes the same functional domain (COG0541: signal recognition particle GTPase) conserved in FfH. Therefore, *Leptospira* species may use the modified Sec dependent pathway for the protein secretion in case of the absence of SecB (Figure 3.2.32).

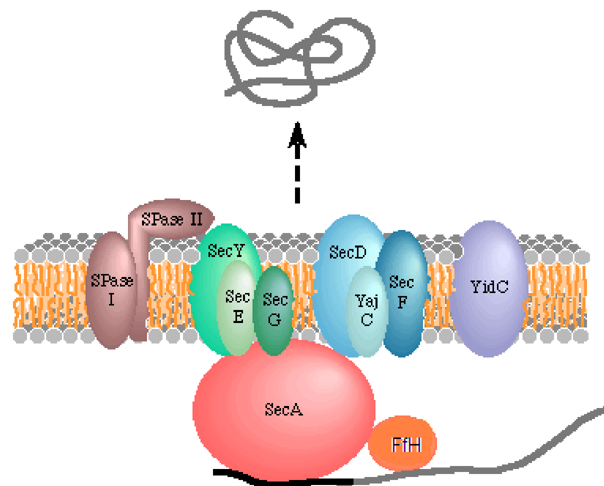


Figure 3.2.33 Sec dependent pathway in *L. pomona*.

Table 3.2.35 Proteins involved in the Sec dependent pathway in four *Leptospira* species.

Gene name	**Desription**	**Gene in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
SecA	preprotein translocase SecA subunit	Yes	Yes	Yes	Yes
SecB	preprotein translocase SecB subunit	No	No	No	No
FfH	signal recognition particle GTPase	Yes	Yes	Yes	Yes
SecD	preprotein translocase SecD subunit	Yes	Yes	Yes	Yes
SecE	preprotein translocase SecE subunit	Yes	Yes	Yes	Yes
SecF	preprotein translocase SecF subunit	Yes	Yes	Yes	Yes
SecG	preprotein translocase SecG subunit	Yes	Yes	Yes	Yes
SecY	preprotein translocase SecY subunit	Yes	Yes	Yes	Yes
YajC	preprotein translocase YajC subunit	Yes	Yes	Yes	Yes
YidC	preprotein translocase YidC subunit	Yes	Yes	Yes	Yes
Spase I	signal peptidase I	Yes	Yes	Yes	Yes
Spase II	signal peptidase II	Yes	Yes	Yes	Yes

The second pathway for the export of proteins is the twin arginine translocation (Tat) pathway (Berks, et al., 2005). It is distinct from the Sec pathway because Tat substrates are secreted in a folded conformation (Thomas, et al., 2001), Tat signal peptides contain a highly conserved twin-arginine motif (Niviere, et al., 1992), and the energy driving translocation is provided solely by the proton motive force (Santini, et al., 1998).

Four genes (TatA, TatB, TatC, and TatE) have been identified that encode the components of the *E. coli* Tat translocation apparatus (Muller and Klosgen, 2005). TatA, TatB, and TatE are sequence-related proteins. TatE encodes a protein with high sequence similarity to TatA, and the two proteins can functionally substitute for each other (Jack, et al., 2001; Blaudeck, et al., 2005). TatA and TatB show the weaker

sequence identity. However, TatA may replace TatB (Blaudeck, et al., 2005). TatA and TatBC form separate highly oligomeric complexes (Figure 3.2.34). The signal peptide of the preprotein binds to a site in TatC. This TatBC-preprotein complex induces a protein-conducting channel in TatA open, allowing the mature domain of the preprotein across the membrane. After protein transport has been completed, the signal peptide is cleaved by the signal peptidase. Then the TatA and TatBC components dissociate and the system returns to the initial state.

Analysis of the four *Leptospira* genomes revealed that genes for TatA and TatC were present, while TatE and TatB were not (Table 3.2.36). As mentioned above, since TatA may function as TatE and it also may replace TatB, *Leptospira* species may be capable of employing the Tat system for the protein excretion.

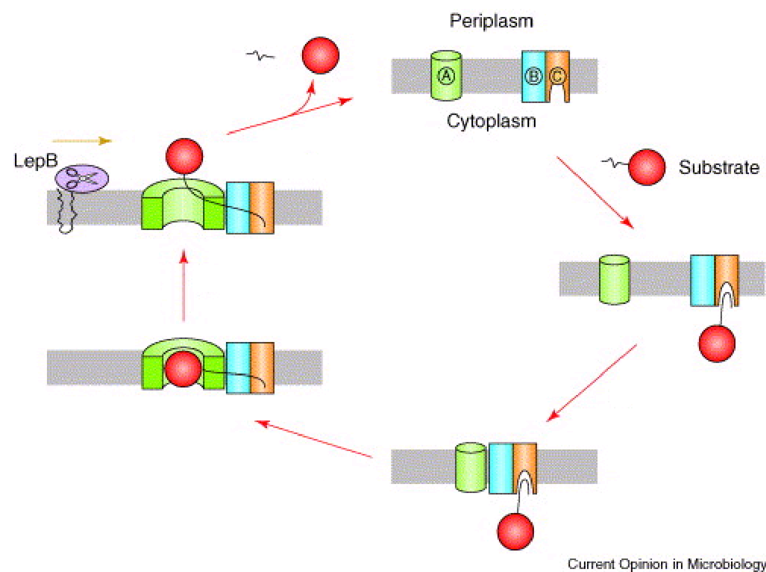


Figure 3.2.34 The Tat translocation pathway in bacteria (Berks, et al., 2005). This figure was refined manually to show the systems as present in the four *Leptospira* species.

Table 3.2.36 Proteins involved in the Sec dependent pathway in four *Leptospira* species.

Gene name	**Desription**	**Gene in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
TatA	Sec-independent protein translocase protein	Yes	Yes	Yes	Yes
TatB	Sec-independent protein translocase protein	Yes	Yes	Yes	Yes
TatC	Sec-independent protein translocase protein	Yes	Yes	Yes	Yes
TatE	Sec-independent protein translocase protein	Yes	Yes	Yes	Yes
LepB	signal peptidase I	Yes	Yes	Yes	Yes

3.3 Virulence Genes

Leptospirosis is the most widespread zoonosis in the world and has emerged as an important public health problem (Levett, 2001). Its severe disease form, known as Weil's syndrome, is an acute febrile illness associated with multiorgan system complications, including jaundice, renal failure, meningitis, and pulmonary hemorrhage, with a mortality rate of more than 15% (Marotto, et al., 1999). Leptospirosis is caused by the pathogens-leptospiries. Although the mechanisms underlying leptospirosis are not well understood (Vinetz, 2001), potential virulence genes have been inferred from this dissertation research and earlier genomic sequencing studies (Ren, et al., 2003; Nascimento, et al., 2004). Most bacterial pathogens use common virulence strategies that include, for example, adhesion to the host cells, penetration of the membranes, and the defense against the host immune response (Wilson, et al., 2002) as described below.

3.3.1 Chemotaxis

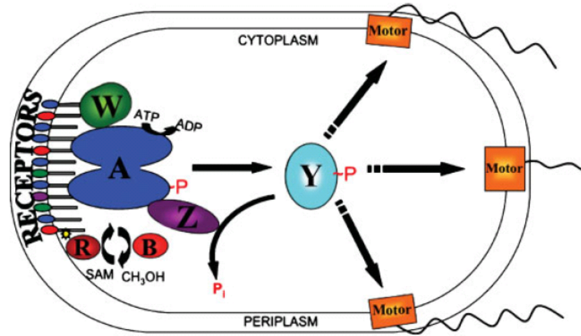


Figure 3.3.1 The *E. coli* chemotaxis system (Baker, et al., 2005)

Note: RECEPTORS-methyl-accepting chemotaxis protein (MCP); A-CheA; B-CheB; R-CheR; W-CheW; Y-CheY; Z-CheZ; P-phosphoryl group.

Chemotaxis and motility, critical for the virulence of pathogenic leptospires, enable the bacteria to move towards attractants and avoid repellents, and thus respond quickly to different environments as well as allow them to penetrate host tissue barriers during infection (Charon and Goldstein, 2002).

Table 3.3.1 Genes involved in the chemotaxis system of four *Leptospira* species.

Gene name	**Function**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
MCP	methyl-accepting chemotaxis protein	11	11	12	11
CheA	histidine protein kinase	2	2	2	2
CheB	chemotaxis response regulator protein-glutamate methylesterase	3	3	3	3
CheR	chemotaxis protein methyltransferase	2	2	2	2
CheW	coupling protein	3	3	3	3
CheY	response regulator	1	1	1	1
CheZ	CheY-P phosphatase	0	0	0	0
CheX	CheY-P phosphatase	1	1	1	1
FliY	flagellar motor switch protein	1	1	1	1

In *E. coli* chemotaxis (Baker, et al., 2005), see Figure 3.3.1, the transmembrane receptors first recognize the signal (the change of the concentration of a chemical) in the surrounding environment. This signal then is transduced to the histidine kinase, CheA, and the coupling protein, CheW, and the phosphorylation of CheA induces phosphorylation of the response regulator, CheY. Finally, phosphorylated CheY (CheY-P) interacts with the flagellar motors to direct the

motility of cell. In addition, the activity of CheA can be altered by the adaptional proteins CheB and CheR, and the phosphatase CheZ can enhance the dephosphorylation of CheY-P by interacting with CheA.

The *Leptospira* species seem to have adopted a similar chemotaxis system as *E. coli*, but they lack an obvious CheZ gene. It was reported that FliY and CheX genes from *Bacillus subtilis* (Szurmant, et al., 2003) and *Borrelia burgdorferi* (Motaleb, et al., 2005), respectively, encode proteins that catalyze the dephosphorylation of CheY-P. Therefore they may have functional role analogous to that of CheZ in *E. coli*. Both FliY and CheX genes were identified from the genomes of all four *Leptospira* species as shown in Table 3.3.1. Furthermore, it is very interesting to notice that there are different numbers of methyl-accepting chemotaxis protein in the different species: 11 in *pomona*, 11 in *grippotyphosa*, 12 in *lai*, and 11 in *copenhageni*.

3.3.2 Motility

Figure 3.3.2 Representation of flagella assembly in *E. coli*.

Table 3.3.2 Genes involved in the motility of four *Leptospira* species.

Gene name	**Function**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
FliD	flagellar hook-associated protein	1	1	1	1
FliC/ FlaB	periplasmic flagellin	4	4	4	4
FlgL	flagellin and related hook-associated proteins	1	1	1	1
FlgK	flagellar hook-associated protein	1	1	1	1
FlgE	flagellar hook protein	1	1	1	1
FliK	flagellar hook-length control protein	1	1	1	1
FlgD	flagellar basal-body rod modification protein	1	1	1	1
FlgF	flagellar hook protein	0	0	0	0
FlgG	flagellar hook protein	2	2	2	2
FlgH	flagellar L-ring protein precursor	1	1	1	1
FlgI	flagellar basal-body P-ring protein	1	1	1	1
MotA	chemotaxis motA protein	2	2	2	2
MotB	chemotaxis motB protein	4	4	4	4
FliE	flagellar hook-basal body protein	1	1	1	1
FlgB	flagellar basal body protein	1	1	1	1
FlgC	flagellar basal body rod protein	1	1	1	1
FliF	flagellar MS-ring protein	1	1	1	1
FliG	flagellar motor switch protein	3	3	3	3
FliL	flagellar basal body-associated protein	1	2	2	1
FliM	flagellar motor switch protein	1	1	1	1
FliN	flagellar motor switch protein	2	3	2	2
FlhA	flagellar biosynthesis protein	1	1	1	1
FlhB	flagellar biosynthetic protein	1	1	1	1
FliH	flagellar assembly protein	1	1	1	1
FliI	flagellum-specific ATP synthase	1	1	1	1
FliO	flagellar protein required for flagellar formation	1	1	1	1
FliP	flagellar biosynthesis protein	1	1	1	1
FliQ	flagellar biosynthetic protein	1	1	1	1
FliR	flagellar biosynthetic protein	1	1	1	1

The flagellum, used by *E. coli* for motility, has three parts (Figure 3.3.2). The first part is the basal body associated with the cell wall and the cytoplasmic membrane that functions as a rotary motor to generate torque and consists of both rings and rod like structures (Ikeda, et al., 1996). The second part is the hook that acts as a universal joint to transmit the motor torque to the long helical propeller in its different orientations. The third part is the filament with a thin helical structure and rapid rotation that propels the cell locomotion in viscous environments.

All genes except FlgF involved in the flagellar assembly were identified in four *Leptospira* species as shown in Table 3.3.2. A domain analysis revealed that FlgG and FlgF shared the common functional domain, COG1749: flagellar basal body and hook proteins. Therefore, FlgG, identified from *Leptospira* species, may function as FlgF. In addition, it is noteworthy that different *Leptospira* species have different copy numbers of FliL and FliN as shown in Table 3.3.2.

3.3.3 Adherence

Adherence often is an essential step in bacterial pathogenesis or infection and in an early event in the establishment of the infection. To effectively adhere to mammalian host surfaces, *Leptospira* pathogens produce multiple families of adhesions as shown in Table 3.3.3 (Ren, et al., 2003; Nascimento, et al., 2004).

The first family of adhesins consists of three integrin alpha-like proteins (Springer, 1997). The N-terminal region of the alpha subunit is composed of seven FG-GAP (phenyl-alanyl-glycyl and glycyl-alanyl-prolyl) repeats that fold into a beta-propeller domain. These adhesins recognize the diverse ligands on the host cell surface and in the extracellular matrix. Three common genes encoding integrin alpha-like proteins were identified in *L. pomona*, *grippotyphosa*, *lai*, and *copenhageni*.

The second family consists of the Lig (leptospira immunoglobulin-like) proteins that include about 10 bacterial immunoglobulin-like (Big) repeat domains (Palaniappan, et al., 2002). These adhesins mediate host cell attachment and entry. It is notable that Lig genes have been reported to be present only in pathogenic but not saprophytic *Leptospira* species (Matsunaga, et al., 2003). This family comprises three genes: ligA, ligB, and ligC. ligA and ligB encode large lipoproteins (128 and 212

kDa, respectively). But ligC contains mutations that disrupt the reading frames. All the four *Leptospira* species harbor single copy of each Lig gene except *L. lai* as it lacks LigA gene (Nascimento, et al., 2004).

The third family consists of fibronectin-binding protein (cadF) (Merien, et al., 2000). These adhesions are 36kDa outer-membrane protein with a conserved ompA functional domain (Pautsch and Schulz, 1998). Each *Leptospira* species (*pomona*, *grippotyphosa*, *lai*, or *copenhageni*) harbors three genes encoding fironectin-binding proteins that likely mediate the binding of the *Leptospira* pathogens to the extracellular matrix component fibronectin of the host.

Table 3.3.3 Genes involved in adherence of four *Leptospira* species.

Gene name	Description	Genes in *pomona*	Genes in *grippotyphosa*	Genes in *lai*	Genes in *copenhageni*
itgA	integrin alpha-like proteins	3	3	3	3
LigA	Ig-like repeat domain protein 1	1	1	0	1
LigB	Ig-like repeat domain protein 3	1	1	1	1
LigC	---	1	1	1	1
cadF	fibronectin-binding protein	3	3	3	3

3.3.4 Invasion

In addition to chemotaxis and motility, leptospiral invasion may be mediated by secretion of enzymes capable of degrading host cell membranes (Ren, et al., 2003). Basically, there are three families of protein involved in invasion as shown in Table 3.3.4.

The first family, related to the hemolysin, consists of proteins that likely bind to the outer membrane of the susceptible host cell to form a transmembrane channel that allows water, ions, and small organic molecules to pass through the transmembrane channel that can result in host cell death by irreversible osmotic swelling, a major mechanism by which protein toxins can damage cells (Gentschev, et

al., 2002). Eighteen common genes encoding hemolysins were identified from *pomona*, *grippotyphosa*, *lai*, and *copenhageni*: 5 sphingomyelinase C-type hemolysins, 3 orthologs of the *Serpulina hyodysenteriae* tlyABC hemolysins, and 10 other hemolysins.

The second family is a phospholipase that converts phospholipids into fatty acids and other lipophilic substances. There are four major classes, termed A, B, C and D, that act by cleaving the triglyceride molecule into various fragments (Birch, et al., 1996). Only one copy of phospholipase C and one copy of phospholipase D were identified from each of these four *Leptospira* species.

The third family is a protease that degrades extracellular matrix proteins to facilitate invasion of host tissues. One gene encoding collagenase, and four genes encoding metalloprotease were identified in all four *Leptospira* species. A particularly interesting finding from our analyses was that there are four predicted thermolysins in *pomona*, five in *grippotyphosa*, three in *lai*, and four in *copenhageni*.

Table 3.3.4 Genes involved in invasion of four *Leptospira* species.

Family	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
hemolysin	sphingomyelinase C-type hemolysins	5	5	5	5
	tlyA tlyB tlyC	1 1 1	1 1 1	1 1 1	1 1 1
	other hemolysins	10	10	10	10
phospholipase	phospholipase C	1	1	1	1
	phospholipase D	1	1	1	1
protease	collagenase	1	1	1	1
	metalloprotease	4	4	4	4
	thermolysin	4	5	3	4

3.3.5 Iron acquisition and utilization

Iron is an essential nutrient to support the growth of most organisms. However, despite the relative abundance of iron in nature, oxidation creates insoluble

iron complexes that are unavailable to bacteria. To satisfy their iron requirements, bacteria have evolved numerous strategies for iron uptake and the ability to acquire iron and iron complexes also has long been recognized as an important determinant of bacterial virulence (Braun, 2001).

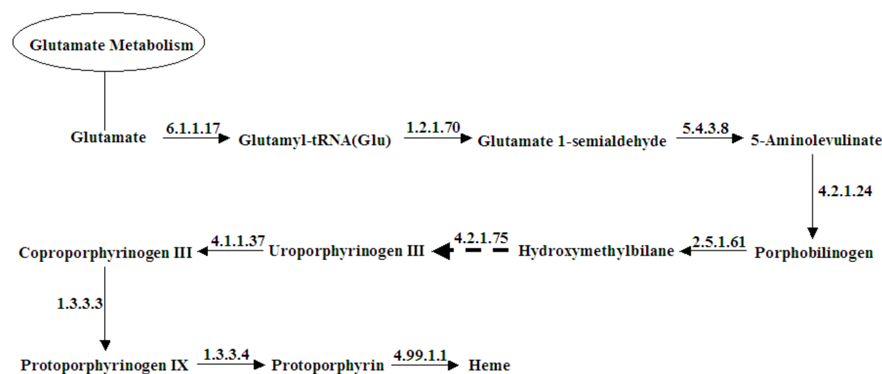


Figure 3.3.3 Heme synthesis pathway as deduced from the genomic sequences of four *Leptospira* species.

Table 3.3.5 Enzymes involved in the heme synthesis pathway in four *Leptospira* species.

EC#	**Description**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
6.1.1.17	glutamyl-tRNA synthetase (gltX)	Yes	Yes	Yes	Yes
1.2.1.70	glutamyl-tRNA reductase (hemA)	Yes	Yes	Yes	Yes
5.4.3.8	glutamate-1-semialdehyde 2,1-aminomutase (hemL)	Yes	Yes	Yes	Yes
4.2.1.24	porphobilinogen synthase (hemB)	Yes	Yes	Yes	Yes
2.5.1.61	hydroxymethylbilane synthase (hemC)	Yes	Yes	Yes	Yes
4.2.1.75	uroporphyrinogen-III synthase (hemD)	No	No	No	No
4.1.1.37	uroporphyrinogen decarboxylase (UPD) (hemE)	Yes	Yes	Yes	Yes
1.3.3.3	coproporphyrinogen III oxidase (hemF)	Yes	Yes	Yes	Yes
1.3.3.4	protoporphyrinogen oxidase (hemG)	Yes	Yes	Yes	Yes
4.99.1.1	ferrochelatase (hemH)	Yes	Yes	Yes	Yes

Analysis of the *Leptospira* species genomes shows that all heme biosynthetic genes (hemA, hemL, hemB, hemC, hemE, hemN, hemY, and hemH) except hemD were identified. A domain analysis revealed that all four *Leptospira* species lack the conserved hemD uroporphyrinogen III synthase functional domain (COG1587: uroporphyrinogen III synthase). Guégan and colleagues reported that the leptospiral hemC gene (encoding porphobilinogen deaminase) could restore the activity of uroporphyrinogen III synthase in an *E. coli* ΔhemD mutant (Guégan, et al., 2003). Their discovery indicated that the leptospiral hemC gene encodes a bifunctional enzyme, allowing *Leptospira* species to synthesize heme *de novo* even without the hemD gene. In addition, it is very interesting to notice that all these nine genes form an operon in the small chromosome and conserved in all four *Leptospira* species.

Table 3.3.6 Genes involved in the iron uptake systems of four *Leptospira* species.

Gene name	**Description**	**Genes in *pomona***	**Genes in grippotyphosa**	**Genes in *lai***	**Genes in *copenhageni***
FhuA	the receptor for deferoxamine	Yes	Yes	Yes	Yes
HmuR	the receptor of heme	Yes	Yes	Yes	Yes
TonB	periplasmic protein, linking inner and outer membranes	Yes	Yes	Yes	Yes
ExbB	biopolymer transport protein	Yes	Yes	Yes	Yes
ExbD	biopolymer transport protein	Yes	Yes	Yes	Yes
FeoB	ferrous iron transport protein	Yes	Yes	Yes	Yes

Although as shown in Figure 3.3.3 and Table 3.3.5, all four *Leptospira* species possesses the genes for all the enzymes necessary for synthesizing heme, they still require an iron uptake system. Previous studies have showed that iron uptake systems can be classified into two categories, TonB-dependent and TonB-independent. The first category includes two distinct systems based on the form of iron entering the cell. One system takes up iron in the form of deferoxamine, a hydroxamate siderophore, with FhuA serving as the deferoxamine receptor (Pawelek, et al., 2006). The other

system imports heme using the HmuR receptor (Simpson, et al., 2000). These two systems share the same TonB energy transducing system that consists of three proteins: TonB, ExbB, and ExbD. The second category consists of only FeoB (Louvel, et al., 2005), a protein that can transport either ferric dicitrate or iron sulphate. All the genes involved in both iron uptake systems were identified from each of the four *Leptospira* species and listed in Table 3.3.6.

3.3.6 Stress Response

Depending upon the environment stress, a large number of stress response proteins are induced to protect the bacteria against further damage. This response is crucial for bacterial survival. All four *Leptospira* species respond to two major stress conditions, oxidation and heat.

Under the oxidative stress, the toxic superoxide radicals and H_2O_2 are produced and enzymes such as superoxide dismutase (SOD), catalase, and peroxidase function as the scavengers. No SOD orthologs have been observed in the four *Leptospira* genomes, consistent with the observation that *Leptospira* interrogans serovars lacked significant SOD activity (Austin, et al., 1981). However, catalase, thiol peroxidase, and glutathione peroxidase have been identified and they function to replace the missing SOD in *Leptospira* species as shown in Table 3.3.7.

Under the heat stress, a group of heat-shock proteins are induced that protect cells against the adverse effects of hypothermia. Basically, heat-shock proteins can be classified into two categories, molecular chaperons and ATP-dependent proteases. A sudden heat shock results in some denatured and mis-folded proteins immediately. These denatured proteins are recognized and bound by chaperons, and then degraded

by ATP-dependent proteases (Schumann, 2003). The two categories of heat-shock proteins were identified from four *Leptospira* species and listed in Table 3.3.7.

Table 3.3.7 Stress response proteins identified from four *Leptospira* species.

Gene name	**Function**	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
SOD	superoxide dismutase	0	0	0	0
CatA	catalase	1	1	1	1
Tpx	thiol peroxidase	1	1	1	1
Gpo	glutathione peroxidase	2	2	2	2
Hsp20	chaperone	2	2	2	2
Hsp33	chaperone	1	1	1	1
Hsp70 (groEL)	chaperone	1	1	1	1
Hsp90 (htpG)	chaperone	2	2	2	2
DnaJ	chaperone	5	5	5	5
DnaK	chaperone	1	1	1	1
GroES	chaperone	1	1	1	1
GrpE	chaperone	1	1	1	1
HtpX	protease	1	1	1	1
Lon	protease	1	1	1	1
LonA	protease	1	1	1	1
hslU (clpY)	protease	1	1	1	1
hslV (clpQ)	protease	1	1	1	1
ClpA	protease	1	1	1	1
ClpB	protease	1	1	1	1
ClpC	protease	1	1	1	1
ClpP	protease	2	2	2	2
ClpS	protease	1	1	1	1
ClpX	protease	1	1	1	1

3.3.7 Lipopolysacharides (LPS)

As discussed in **3.2.7.2 Biosynthesis of Lipopolysaccharides**, LPS is an outer membrane chemical moiety consisting of three sections: a toxic lipid (Lipid A) anchored in the outer membrane, an immunogenic polysaccharide core, and an antigenic O-linked series of oligosaccharides (O-antigen) at the extracellular surface. Lipid A, a strong biological enhancer, can boost the immune system while the highly variable O-antigen often allows the pathogens to successfully evade the host immune response. LPS therefore was considered as one of the initially discovered virulence factors in bacterial pathogens (Moxon, et al., 1998).

Leptospires are classified into more than 200 serovars and 24 serogroups based on the structural diversity of LPS. LPS in the different serovar has the different sugar composition, number, repetitiveness, and ramification (Faine, et al., 1999). Therefore, changes in genes involved in the LPS biosynthesis may account for serovar diversity among leptospires (de la Pena-Moctezuma, et al., 1999). In *Leptospira*, as many other bacteria, at least part of the genes coding for enzymes of the polysaccharide biosynthesis pathway are found clustered in a region of chromosomes named O-antigen gene cluster (*rfb* locus) (de la Pena-Moctezuma, et al., 1999). In agreement with findings in other *rfb* loci of leptospires, almost all of the identified genes are encoded on the same strand.

In the *rfb* locus, 63, 61, 62, and 62 genes were identified in the genomes of *L. grippotyphosa*, *pomona*, *copenhageni*, and *lai*, respectively (Figure 3.3.4, Table 3.3.8). Among them, 35 genes are shared by all the four *Leptopira* species. 17 genes are shared only by *grippotyphosa* and *pomona*, 9 and 8 genes are species specificto *grippotyphosa* and *pomona*, respectively, while 25 genes are shared only by *copenhageni* and *lai*. Most of these genes are co-linear between *copenhageni* and *lai*, with the only exception being that one gene of *copenhageni*, LIC12159 (hypothetical protein), is absent in *lai*, while one gene of *lai*, LA1622 (galactoside O-acetyltransferase), is absent in *copenhageni*. These differences in the *rfb* loci of *L. grippotyphosa*, *pomona*, *copenhageni*, and *lai* may reflect their evolutionary adaptation to different animal hosts (Ren, et al., 2003; Nascimento, et al., 2004).

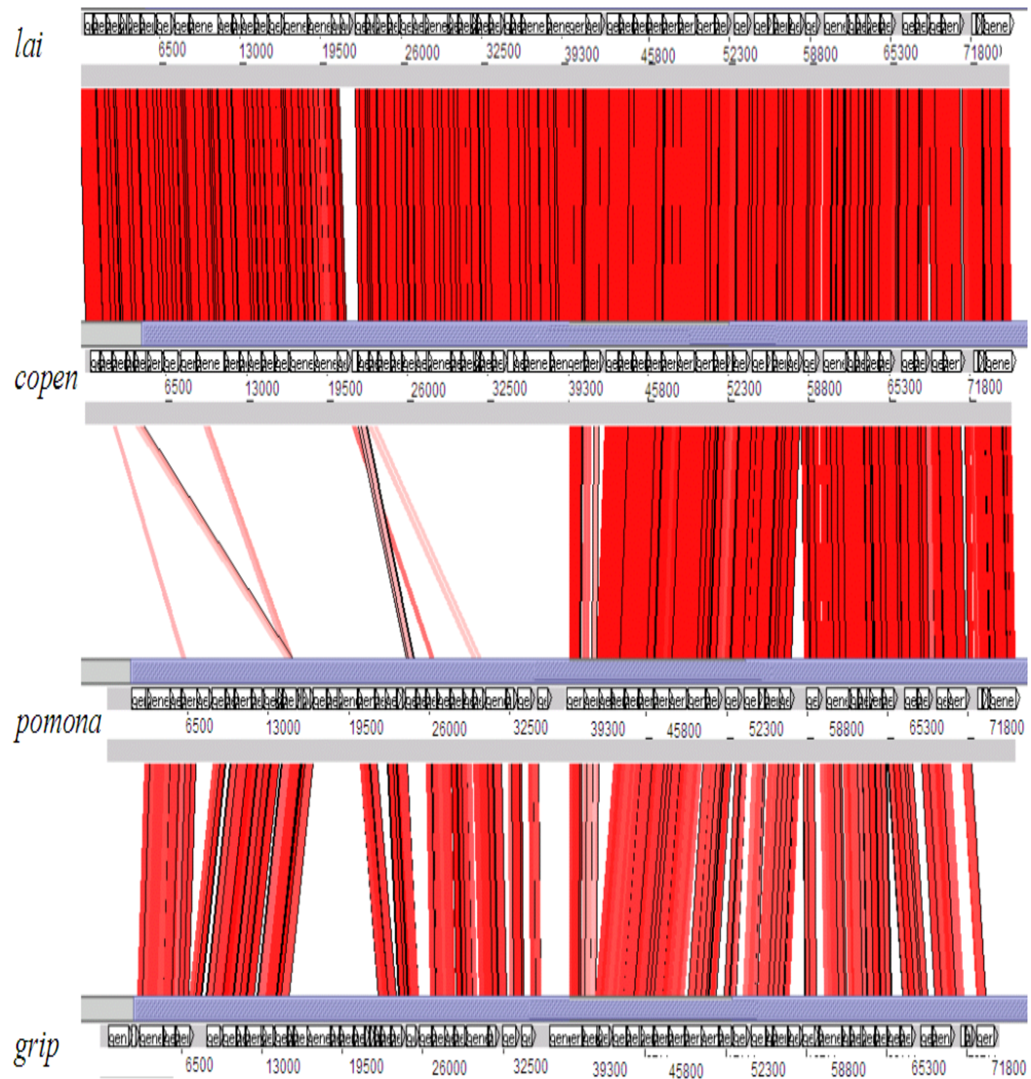


Figure 3.3.4 Genes involved in the *rfb* locus are compared among four *Leptospira* species: *grippotyphosa*, *pomona*, *copenhageni,* and *lai*.

Table 3.3.8 Genes in the *rfb* locus are compared among four *Leptospira* species: *grippotyphosa*, *pomona*, *copenhageni,* and *lai*.

grippotyphosa	***pomona***	***copenhageni***	***lai***	Genes in the *rfb* locus
●	●	●	●	35 genes 5 transferase 7 *rfb* genes 7 glycosyl transferase 16 others
●	○	●	●	1 gene 1 putative lipoprotein
●	●	○	○	17 genes 3 *rfb* genes 3 reductase 11 others
○	○	●	●	25 genes 3 epimerase 4 dehydratase 5 *rfb* genes 6 transferase 7 others
●	○	○	●	1 gene 1 galactoside O-acetyltransferase
○	●	●	○	1 gene 1 methyltransferase
●	○	○	○	9 genes 1 methyltransferase 1 TPR repeat 7 *rfb* genes
○	●	○	○	8 genes 1 transposase 1 methyltransferase 1 FAD dependent oxidoreductase 5 *rfb* genes

Note: ● present, ○ absent.

3.3.8 Antibiotic Resistance

In addition to the above mentioned virulence factors, numerous antibiotic resistance genes were identified in all four *Leptospira* genomes. These resistance genes appear to operate by catabolizing antibiotic, preventing an antibiotic from reaching its intracellular target, modifying the antibiotic's target site, or by producing an alternative metabolic pathway that bypasses the blockage of the antibiotic (Tenover, 2006). The antibiotic resistance genes discovered in four *Leptospira* species are listed and compared in Table 3.3.9, and it should be noted that different

Leptospira species harbors different numbers of multiple antibiotic resistance genes and acriflavine resistance genes.

Table 3.3.9 The putative antibiotic resistance genes identified in four *Leptospira* species.

Gene Category	**Genes in *pomona***	**Genes in *grippotyphosa***	**Genes in *lai***	**Genes in *copenhageni***
bacitracin resistance protein	1	1	1	1
small multidrug resistance protein	1	1	1	1
vancomycin resistance protein	1	1	1	1
tetracycline resistance protein	2	2	2	2
bleomycin resistance protein	5	5	5	5
multiple antibiotic resistance protein	4	3	3	3
acriflavine resistance protein	13	11	12	12

Chapter IV

Conclusion

Living organisms exist and prosper in unique environments because the genes in their genomes encode for macromolecules that provide the ability for that organism to have a selective advantage. Through genomic sequencing we can uncover the specific genotype components that yield the resulting successful phenotype. In the research leading to this dissertation, the complete nucleotide sequences of two animal pathogents, *Leptospira interrogan* serovar *pomona and Leptosrira kirschneri* serovar *grippotyphosa* were sequenced, their putative genomic features, including their predicted coding and non-coding regions, genomic organization, metabolic capabilities and virulence-related genes were determined and compared in detail to those predicted for other spirochetes, e.g. *Leptospira interrogans* serovar *lai* (Ren, et al., 2003) and *Leptospira interrogans* serovar *copenhageni* (Nascimento, et al., 2004), that also cause Leptospirosis in domestic and wild animals as well as secondary infections in humans, T. *pallidum* (Fraser, C. M. et al., 1998) , the causative agent of syphilis, and *B. burgdorferi* (Fraser, C. M. et al., 1997), that causes Lime's Disease. An additional comparison with *Actinobacillus actinomycetemcomitans,* an oral pathogen (Najar, F., 2002), reveals that *Leptospira* species likely have many fewer membrane-associated transporters because they infect tissues that do not provide fewer nutrients than are available to the oral pathogen.

The first major conclusion is that almost 50 of the central metabolic pathways identically, genomic and megaplasmid sizes are conserved in all four *Leptospira* species: *pomona*, *grippotyphosa*, *lai*, and *copenhageni*, while their genomic

organization and virulence-related genes differ in various degrees. In contrast, the spirochetes, *T. pallidum* and *B. burgdorferi* have relatively smaller ~1 Mbp genomes, and limited biosynthetic abilities while their metabolic pathways, including the TCA cycle, electron transport chain, fatty acid synthesis, LPS biosynthesis, nucleotide and amino acid biosynthesis, often are incomplete. The limited metabolic capacities of these two spirochetes may reflect convergent evolution by gene loss from more metabolically competent progenitors and may explain why both *T. pallidum* and *B. burgdorferi* are obligate parasite of humans, and cannot survive without a host.

One unique metabolic characteristic of *Leptospira* species is that they lack the asparagine synthetase that is involved in the synthesis of asparagine and since they do encode both aspartyl-tRNA synthetase and the three-subunit tRNA-dependent amidotransferase, the latter may be responsible for the production of Asn-$tRNA^{Asn}$ by the transamidation of Asp-$tRNA^{Asn}$, as has been observed for example in *Pseudomonas aeruginosa* (Akochy, et al., 2004) and *Bacillus megaterium* (Wilcox and Nirenberg, 1968).

Interestingly, all four *Leptospira* species do encode a HemR-type receptor containing the amino acid motif FRAP that is specific to receptors of haem-containing compounds (Bracken, et al., 1999). Since each *Leptospira* species possesses the genes for the heme biosynthetic enzymes, it is likely that leptospiries are capable of both heme uptake *and de novo* heme synthesis, consistent with the observation that heme is an essential growth factor for *Leptospira* species (Faine, 1959).

The second major conclusion from this dissertation research is that we have provided additional support that Leptospirosis virulence is attributed in part to the

presence of species specific Lipopolysacharides (LPS) on the surface of the microbial outer membrane (Levett, 2001). These genes synthesizing LPS are encoded in an O-antigen biosynthesis gene cluster (*rfb* locus) and differences in the *rfb* locus and subsequent synthesized LPSs and have been implicated in determining the diversity of *Leptospira* serovars (Levett, 2001). In addition, based on their O-antigen genes, the four sequenced *Leptospira* species can be divided into two groups, with one group including *grippotyphosa* and *pomona* and the other including *copenhageni* and *lai*. This is a grouping consistent with their unique host ranges as both *grippotyphosa* and *pomona* infect dogs, cows, sheep, pigs, and mice (Leptospirosis, 2000), while *lai* and *copenhageni* are closely related, and infect the striped field mice (*Apodemus agrarius*) (Ren, et al., 2003) and the domestic rats (*Rattus norvegicus*), respectively (Nascimento, et al., 2004) (Table 4.1).

Table 4.1 Strains of *Leptospira* species used in DNA sequence analysis studies.

Species and Serovar	**Isolation Country**[1]	**Host Range**
Leptospira interrogan serovar *pomona*	Australia	dogs, cows, sheep, pigs, horses, mice, and humans[2]
Leptosrira kirschneri serovar *grippotyphosa*	United States	dogs, cows, sheep, pigs, mice, rats, and humans[2]
Leptospira interrogans serovar *lai*	China	field mice and humans[3]
Leptospira interrogans serovar *copenhageni*	Brazil	domestic rats and humans[4]

Note: (1) Haake, et al., 2004.
(2) Leptospirosis, Control of Communicable Diseases Manual. 2000.
(3) Ren, et al., 2003
(4) Nascimento, et al., 2004.

The third major conclusion is that the sequence of *Leptospira interrogans* serovar *pomona* and *Leptosrira kirschneri* serovar *grippotyphosa* and our comparative study with the other two sequenced *Leptospira interrogans* serovar *lai* and *copenhageni* has revealed much information about the *Leptospira* lifecycle (see Figure 4.1). The primary hosts for *Leptospira* are wild and domestic animals, and

leptospirosis has a major economic impact in the meat and diary industry. However, humans can be accidental or secondary hosts for *Leptospira*, and when infection occurs by drinking contaminated water or the spirochete entering the blood stream through cuts in the skin, can cause symptoms that range from subclinical to fatal. As shown below in Figure 4.1, *Leptospira* enter through broken skin where it often causes a skin rash at the site of infection or enters the blood stream through the mucosal lining of the mouth, where it is spread throughout the body via the blood stream, and subsequently causes meningitis in the brain, hepatitis in the liver, and nephritis in the kidney and the appearance of *Leptospira* in the urine.

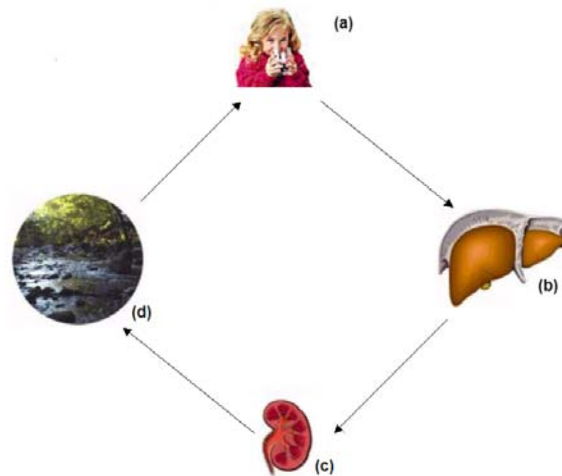


Figure 4.1 The lifecycle of *Leptospira* species in human.
(a) *Leptospira* species enter into the host through the mouth or cut in the skin.
(b) *Leptospira* species colonize in the liver, the major site for beta-oxidation of fatty acid. Using the fatty acid transporter, *Leptospira* species import the fatty acid from the environment.
(c) By adhesion, chemotaxis, and flagella, *Leptospira* species move to the kidney where they secret enzymes that digest kidney cells to produce their needed nutrients, such as fatty acid, nucleotides, and amino acids.
(d) After destroy the kidney, *Leptospira* species re-enter the environment via urine.

Note: The four images in Figure 4.1 were adopted/modified from the following websites, respectively:
(a) http://www.gordonwater.com/wt_residential.html
(b) http://www.hepcbc.ca/liver.htm
(c) http://www.kidney.ab.ca/kidneys/index.html
(d) http://www.pequannockriver.org/

Each of the *Leptospira* species studied occupies a unique niche in the ecosystem as shown in Table 4.1. Although both the glucose and long-chain fatty-

acid catabolic pathways are complete in all four *Leptospira* species, they often utilize beta-oxidation of long-chain fatty acids as the major energy and carbon source instead of the more common sugar oxidative pathway based on the host they infect (Henneberry and Cox, 1970). In addition, one fatty acid transporter (fadL) was identified from each *Leptospira* species. This observation is confirmed by detailed genomic analysis and supports the postulate that *Leptospira* species evolved the ability to fill a broad niche in organisms (or environments) that are rich in both fats and/or lipids as well as those rich in glycogen and/or simple sugars.

Another observation involves NADPH biosynthesis. Although all the genes needed to encode the enzymes for the non-oxidative arm of the pentose phosphate pathway are present, their oxidative arms are incomplete. Therefore leptospiries can use this pathway to produce robose-5-phosphate but require an alternative pathway to produce NADPH. As a result, with the oxidative pathway bypassed and with the requirement that the reducing potential of NADPH is critical for all microbial survival, NADPH is produced from NAD to NADH then NADPH via the fatty acid oxidation pathway rather than pentose phosphate pathway. In details, NADH dehydrogenase catalyzes NAD to produce NADH. Then NADH is converted to NADPH by NAD(P) transhydrogenase or malate dehydrogenase. In addition, there is no NADPH transporter identified from *Leptospira* species.

The Phosphotransferase System (PTS) is incomplete in all four *Leptospira* species since neither the genes nor the active site domains for permease enzyme II, IIB and IIC, could be identified in the four *Leptospira* genomes. However, because of the presence of several sugar transporters, it may be that the *Leptospira* species uptake

sugar through the transporter instead of phosphotransferase system as has been observed in *E. coli* (Kundig, et al., 1964).

Compared with *Actinobacillus actinomycetemcomitans,* an oral pathogen (Najar, F., 2002), *Leptospira* species have many fewer membrane-associated transporters (Figure 4.2a and 4.2b). For example, the transporters for pantothenate, nicotinate, nucleotides, and thiamine were not identified in any of the four *Leptospira* species. Since *A. actinomycetemcomitans* colonizes in the human oral cavity where nutrients usually are abundant, it is not surprising that this microbe has many transporters. In contrast, *Leptospira* species in acute infection colonizes the kidneys that have fewer nutrients available. Therefore, *Leptospira* species evolved fewer transporters, an observation in agreement with their metabolic pathways. For example, most central metabolic pathways of *Leptospira* species are complete compared with *A. actinomycetemcomitans* that uptakes numerous metabolic intermediates.

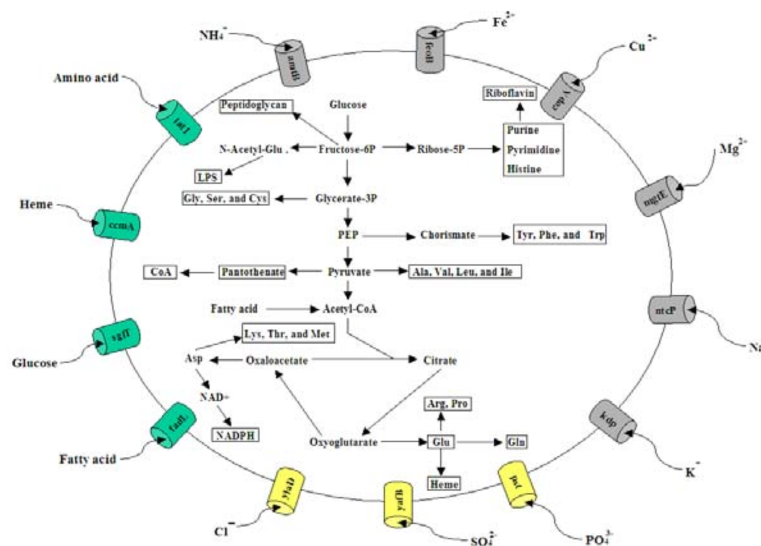


Figure 4.2a Overview of the metabolic scheme utilized by *Leptospira* species showing transporters and metabolic pathways.

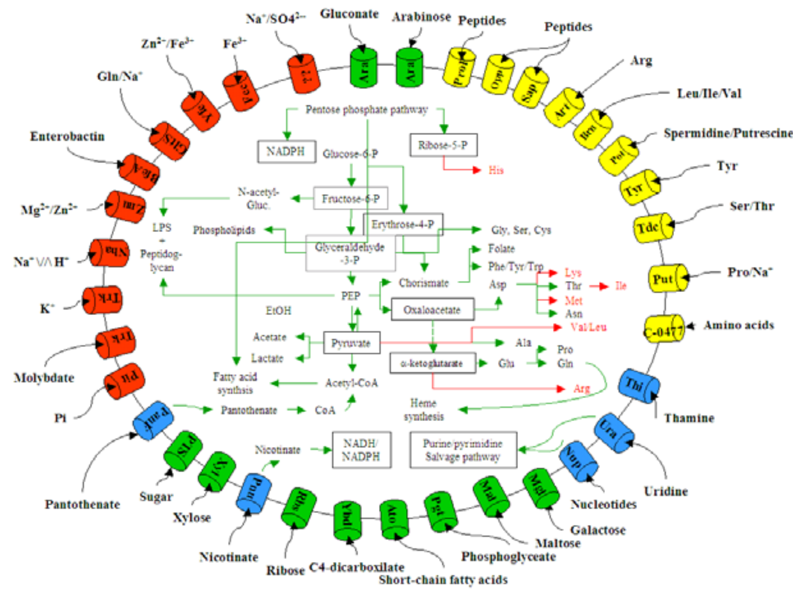


Figure 4.2b Overview of the metabolic scheme utilized by *A. actinomycetemcomitans* showing transporters and metabolic pathways (Adopted from Dr. Fares Najar's Ph. D. dissertation).

The genes for many virulence factors were observed in all four *Leptospira* genomes. Since chemotaxis and motility are critical for the virulence of pathogenic leptospires, it was interesting to discover at least 40 genes involved in the flagellar assembly and at least 20 genes involved in the chemotaxis system of each *Leptospira* species. The presence of these complicate chemotaxis and motility systems may account for the ability of leptospira to rapidly translocate across host cell monolayers during infection (Nascimento, et al., 2004).

The sequenced *Leptospira* species also possess at least three families of adhesins related to the attachment of eukaryotic cells. The first family includes three integrin alpha-like proteins (Springer, 1997). The second one consists of the Lig (leptospira immunoglobulin-like) proteins that include about 10 bacterial immunoglobulin-like repeat domains (Palaniappan, et al., 2002). The third family consists of only fibronectin-binding protein (cadF) (Merien, et al., 2000). It is a

36kDa outer-membrane protein with a conserved ompA functional domain (Pautsch and Schulz, 1998). The presence of these adhesins may account for the ability of leptospiries to invade and colonize host tissues and to establish the robust infection (Nascimento, et al., 2004).

In addition, at least three families of proteins involved in degrading host kidney cells were identified in the sequenced *Leptospira* species. They include hemolysin, phospholipase, and protease. Both hemolysin and phospholipase act on host cell membranes containing the substrate phospholipids, leading to cytolysis (Lee, et al., 2002), while the protease degrades extracellular matrix proteins to facilitate invasion of host tissues. These proteases may play a significant role in causing one of the fatal features of leptospira infections, acute renal failure (Levett, 2001).

In summary, sequencing two *Leptospira* species and the extensive comparative study enhances our understanding of the molecular mechanisms of leptospiral physiology, virulence, and pathogenesis. Our studies showed that nearly 50 central metabolic pathways identically are conserved in four *Leptospira* species: *pomona*, *grippotyphosa*, *lai*, and *copenhageni*. These findings may account for their similar phenotype and physiological characters. However, the virulence genes identified from four *Leptospira* species involved in chemotaxis and motility, attachment to the host cells, corruption of the membranes, and the defense against the host immune response may play the important roles in the pathogenesis in leptospirosis as well as contributing to each species adaptation to its specialized host environment. With these sequences completed, additional genetic transformation, proteomics and microarray studies aimed at further understanding these unique parasitic spirochetes

are likely to reveal additional virulence gene candidates and the development of new vaccines, which ultimately will result in the prevention and treatment of leptospirosis worldwide.

References

Akochy, P., Bernard, D., Roy, P., and Lapointe, J. (2004) Direct glutaminyl-tRNA biosynthesis and indirect asparaginyl-tRNA biosynthesis in *Pseudomonas aeruginosa* PAO1. *J. Bacteriol*. 186(3): 767-776.

Altshul, S. F. et al. (1990) Basic local alignment search tool. *J. Mol. Biol*. 215: 403-410.

Anderson, S., de Bruijn, M. H., Coulson, A. R., Eperon, I. C., Sanger, F., and Young, I. G. (1982) Complete sequence of bovine mitochondrial DNA. Conserved features of the mammalian mitochondrial genome. *J. Mol. Biol.* 156: 683-717.

Aoki, H., Dekany, K., Adams, S., and Ganoza, M. (1997) The gene encoding the elongation factor P protein is essential for viability and is required for protein synthesis. *J. Biochem*. 272(51): 32254-32259.

Austin, F. E., Barbieri, J. T., Corin, R. E., Grigas, K. E., and Cox, C. D. (1981) Distribution of superoxide dismutase, catalase, and peroxidase activities among *Treponema pallidum* and other spirochetes. *Infect. Immun.* 33(2): 372-379.

Baker, M. D., Wolanin, P. M., and Stock, J. B. (2005) Signal transduction in bacterial chemotaxis. *BioEssays* 28: 9-22.

Baltimore, J. (1970) Central dogma reversed. *Nature* 27: 1198-1199.

Baranov, P. V., Gesteland, R. F., and Atkins, J. F. (2002a) Recoding: translational bifurcations in gene expression. *Gene* 286(2): 187–201.

Baranov, P. V., Gesteland, R. F., and Atkins, J. F. (2002b) Release factor 2 frame shifting sites in different bacteria. *EMBO Rep.* 3(4): 373–377.

Barnes, W. M. (1994) PCR amplification of up to 35 kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc. Natl. Acad. Sci. USA*. 91: 2216-2220.

Barry, K. J. and Wainwright, N. R. (1997) Biosynthetic Induction of a Secondary Metabolite by a Marine Bacterium under Nutritional Stress: Potential Role of the Incomplete Oxidation of an Organic Acid *Biol. Bull*. 193: 274-275.

Baskaran, N, Kandpal, R. P., Bhargava, A. K., Glynn, M. W., Bale, A., and Weissman, S. M. (1996) Uniform amplification of a mixture of deoxyribonucleic acids with varying GC content. *Genome Res.* 7: 633-638.

Baxevanis, A. D. and Davison, D. B. (2004) Viewing and editing assembled sequences using consed, in *Current protocols in bioinformatics*. John Wiley and Co., 11.2.1-11.2.43.

Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol*. 340: 783-795.

Berks, B. C., Palmer, T., and Sargent, F. (2005) Protein targeting by the bacterial twin-arginine translocation (Tat) pathway. *Curr*. *Opin*. *Microbiol*. 8(2): 174-181.

Bernard, L. (1990) Cytochrome bc1 complexes of microorganisms. *Microbiological Reviews* 54(2): 101-129.

Bernheimer, A. W. and Bey, R. F. (1986) Copurification of *Leptospira interrogans* serovar *pomona* hemolysin and sphingomyelinase C. *Infect. Immun.* 54: 262-264.

Birch, M., Robson, G., Law, D., and Denning, D. W. (1996) Evidence of multiple extracellular phospholipase activities of *Aspergillus fumigatus*. *Infect. Immun*. 64: 751-755.

Birnboim, H. C. and Doly, J. (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucl. Acids Res*. 7: 1513-1523.

Björk, G. R. (1996) Stable RNA modification. In *Escherichia coli and Salmonella*: *Cellular and Molecular Biology*. 2nd edn. pp. 861-886. Neidhardt, F.C., Curtiss, III, R., Ingraham, J. L., et al. (Eds.). American Society for Microbiology Press, Washington, D.C.

Blaudeck, N., Kreutzenbeck, P., Muller, M., Sprenger, G.A., and Freudl, R. (2005) Isolation and characterization of bifunctional *Escherichia coli* TatA mutant proteins that allow efficient tat-dependent protein translocation in the absence of TatB. *J. Biol. Chem.* 280(5): 3426-3432.

Bodenteich, A., Chissoe, S., Wang, Y. F., and Roe, B. (1994) Shotgun cloning as the strategy of choice to generate templates for high throughput

dideoxynucleotide sequencing. *Automated DNA sequencing and analysis*. pp. 42-45. Adams, M., Fields, C., and Venter, J. (Eds.). Academic Press, New York.

Bracken, C.S., Baer, M.T., Abdur-Rashid, A., Helms, W., and Stojiljkovic, I. (1999) Use of heme-protein complexes by the *Yersinia enterocolitica* HemR receptor: histidine residues are essential for receptor function. *J Bacteriol* 181: 6063–6072.

Brand, L. A. and Strauss, E. (2005) Characterization of a new pantothenate kinase isoform from *Helicobacter pylori*. *J. Biol. Chem*. 280: 20185-20188.

Braun, V. (2001) Iron uptake mechanisms and their regulation in pathogenic bacteria. *Int. J. Med. Microbiol.* 291: 67-79.

Bulach, D. M., Kalambaheti, T., Pena-Moctezuma, A., and Adler, B. (2000) Lipopolysaccharide biosynthesis in *Leptospira*. *J. Mol. Microbiol. Biotechnol.* 2(4): 375-380.

Charon, N. W. and Goldstein, S. F. (2002) Genetics of motility and chemotaxis of a fascinating group of bacteria: the spirochetes. *Annual Review of Genetics* 36: 47-73.

Chissoe, S. L., Wang, Y. F., Clifton, S. W., Ma, N., Sun, H. J., Lobsinger, J. S., Kenton, S. M., White, J. D., and Roe, B. A. (1991) Strategies for rapid and accurate DNA sequencing. *Methods: Companion Methods Enzymol*. 3: 55-65.

Cinco, M., Cini, B., Perticarari, S., and Presani, G. (2002) *Leptospira interrogans* binds to the CR3 receptor on mammalian cells. *Microb. Pathog.* 33(6): 299-305.

Claros, M. G. and von Heijne, G. (1994) TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci*. 10(6): 685-686.

Claustres, M., Kjellberg, P., Desgeorges, M., Bellet, H., Sarda, P., Bonnet, H., and Boileau, C. (1989) Detection of deletions by the amplification of exons (multiplex PCR) in Duchenne muscular dystrophy. *J. Genet. Hum.* 37(3): 251-257.

Clayton, R. A., Gwinn, M. L., Dodson, R. J., et al. (2000) DNA sequence of both chromosomes of the cholera pathogen Vibrio. *Nature* 406: 469-470.

Crick, F. (1970) Central dogma of molecular biology. *Nature* 227: 561-563.

Cronan, J. E. and Rock, C. O. (1996) Biosynthesis of membrane lipids. In *Escherichia coli and Salmonella*: *Cellular and Molecular Biology*. 2nd edn. pp. 612-618. Neidhardt, F. C., Curtiss III, R., Ingraham, J. L., et al. (Eds.). American Society for Microbiology, Washington, D.C.

Curnow, A. W., Hong, K., Yuan, R., Kim, S., Martins, O., Winkler, W., Henkin, T. M., and Soll, D. (1997) Glu-tRNAGln amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. *Proc. Natl. Acad. Sci. USA*. 94: 11819-11826.

de la Pena-Moctezuma, A., Bulach, D. M., Kalambaheti, T., and Adler, B. (1999) Comparative analysis of the LPS biosynthetic loci of the genetic subtypes of serovar Hardjo: *Leptospira interrogans* subtype *Hardjoprajitno* and *Leptospira borgpetersenii* subtype *Hardjobovis*. *FEMS Microbiology Letters* 177: 319-326.

Deininger, P. L. (1983) Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Anal. Biochem.* 129(1): 216-223.

del Real, G., Segers, R. P., van der Zeijst, B. A., and Gaastra, W. (1989) Cloning of a hemolysin gene from *Leptospira interrogans* serovar *hardjo*. *Infect. Immun*. 57: 2588-2590.

Deutscher, M. P., Marshall, G. T., and Cudny, H. (1988) RNase PH: an *Escherichia coli* phosphate-dependent nuclease distinct from polynucleotide phosphorylase. *Proc. Natl. Acad. Sci. USA.* 85: 4710-4714.

Dierick, H., Stul, M., Kelver, W., Marynen, P., and Cassiman, J. J. (1993) Incorporation of dITP or 7-deaza dGTP during PCR improves sequencing of the product. *Nucl. Acids Res.* 21(18): 4427-4428.

Dressman, D., Yan, H., Traverso, G., Kinzler, K. W., and Vogelstein, B. (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci*. *USA.* 100: 8817-8822.

Eriani, G., Delarue, M., Poch, O., Gangloff, J., and Moras, D. (1990) Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature* 347: 203-206.

Ewing, B. and Green, P. (1998) Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* 8: 186-194.

Ewing, B., Hillier, L., Wendl, M., and Green, P. (1998) Base calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* 8: 175-185.

Faine, S. (1959) Iron as a growth requirement for pathogenic *Leptospira*. *J Gen Microbiol* 20: 246–251.

Faine, S. and Stallman, N. D. (1982) Amended descriptions of the genus *Leptospira Noguchi* 1917 and the species *L interrogans* (Stimson 1907) Wenyon 1926 and *L. biflexa* (Wolbach and Binger 1914) Noguchi 1918. *Int. J. Syst. Bacteriol*. 32: 461-463.

Faine, S., Adler, B., Bolin, C., and Perolat, P. (1999). *Leptospira and Leptospirosis*. MediSci., Melbourne, Australia.

Farabaugh, P. J. (1996) Programmed translational frameshifting. *Ann. Rev. Genetics* 30: 507–528.

Fekkes, P. and Driessen, A. J. (1999). Protein targeting to the bacterial cytoplasmic membrane. *Microbiol. Mol. Biol. Rev*. 63: 161-173.

Ferguson, A. D., Chakraborty, R., Smith, B. S., Esser, L., van der Helm, D., and Deisenhofer, J. (2002) Structural basis of gating by the outer membrane transporter FecA. *Science* 295: 1715-1719.

Fitzgerald, M. C., Skowron, P., Van Etten, J. L., Smith, L. M., and Mead, D. A. (1992) Rapid shotgun cloning utilizing the two base recognition endonuclease CviJI. *Nucl. Acids Res*. 20: 3753-3762.

Fraser, C. M., Gocayne, J. D., White, O., et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* 270(5235): 397-403.

Fraser, C., Casjens, S., Huang, W., et al. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 390(6660): 580-586.

Fraser, C. M. et al. (1998) Complete genome sequence of *Treponema pallidum*, the

syphilis spirochete. Science 281, 375-388.

Fukunaga, M. and Mifuchi, I. (1989) The number of large ribosomal RNA genes in *Leptospira interrogans* and *Leptospira biflexa*. *Microbiol. Immunol.* 33: 459-466.

Gordon, D., Abajian, C., and Green, P. (1998) Consed: a graphical tool for sequence finishing. *Genome Research* 8: 195-202.

Gentschev, I., Dietrich, G., and Goebel, W. (2002) The *E. coli* alpha-hemolysin secretion system and its use in vaccine development. *Trends. Microbiol.* 10(1): 39-45.

Green, M. L. and Karp, P. D. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 5: 1-16.

Guégan, R., Camadro, J., Girons, I., and Picardeau, M. (2003) *Leptospira* spp. possess a complete haem biosynthetic pathway and are able to use exogenous haem sources. *Mol. Microbiol.* 49(3): 745-754.

Haake, D. A. (2000) Spirochaetal lipoproteins and pathogenesis. *Microbiol*. 146: 1491-1504.

Haake, D. A., Suchard, M. A., Kelley, M. M., Dundoo, M., Alt, D. P. and Zuerner, R. L. (2004) Molecular Evolution and Mosaicism of Leptospiral Outer Membrane Proteins Involves Horizontal DNA Transfer. *J. Bacteriol*. 186:2818-2828.

Hammell, A. B., Taylor, R. C., Peltz, S. W., and Dinman, J. D. (1999) Identification of putative programmed –1 ribosomal frameshift signals in large DNA databases. *Genome Res.* 9(5): 417–427.

Heath, R. J., Su, N., Murphy, C. K., and Rock, C. O. (2000) The enoyl-[acyl-carrier-protein] reductases FabI and FabL from *Bacillus subtilis*. *J. Biol. Chem.* 275(51): 40128-40133.

Henneberry, R. C., and Cox, C. D. (1970) Beta-oxidation of fatty acids by *Leptospira*. *Can. J. Microbiol.* 16: 41-45.

Heurgué-Hamard, V., Karimi, R., Mora, L., et al. (1998) Ribosome release factor RF4 and termination factor RF3 are involved in dissociation of peptidyl-

tRNA from the ribosome. *EMBO J.* 17(3): 808-816.

Ikeda, T., Oosawa, K., and Hotani, H. (1996) Self-assembly of the filament capping protein, FliD, of bacterial flagella into an annular structure. *J. Mol. Biol.* 259: 679-686.

Iwanyshyn, W. M., Han, G. S., and Carman, G. M. (2004) Regulation of phospholipid synthesis in Saccharomyces cerevisiae by zinc. *J. Biol. Chem*. 279(21): 21976-21983.

Jack, R. L., Sargent, F., Berks, B. C., Sawers, G., and Palmer, T. (2001) Constitutive expression of *Escherichia coli* tat genes indicates an important role for the twin-arginine translocase during aerobic and anaerobic growth. *J. Bacteriol.* 183: 1801-1804.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3: 318-356.

Jett, J. H. et al. (1989) High-speed DNA sequencing: an approach based upon fluorescence detection of single molecules. *Biomol. Struct. Dynam*. 7: 301-309.

Jez, J. M. and Cahoon, R. E. (2004). Kinetic mechanism of glutathione synthetase from Arabidopsis thaliana. J Biol Chem 279(41):42726-31.

Johnson, R.C. and Gary, N. D. (1962) Nutrition of *Leptospira pomona*. 1. A chemically defined substitute for rabbit serum ultrafiltrate. *J. Bacteriol*. 83: 668-672.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* 28: 27-30.

Karger, A. E., Harris, J. M., and Gestland, R. F. (1991) Multiwavelength fluorescence detection for DNA sequencing using capillary electrophoresis. *Nucl. Acids Res.* 19, 4955-4962.

Kim, M., Bird, J. C., van Parys, A. J., Breuer, K. S., and Powers, T. R. (2003) A macroscopic scale model of bacterial flagellar bundling. *Proc. Natl. Acad. Sci. USA.* 100(26): 15481-15485.

Klenow, H., Overgaard-Hansen, K., and Patkar, S. A. (1971) Proteopytic cleavage of native DNA polymerase into two different catalytic fragments. *Eur. J.*

Biochem. 22: 371-381.

Kotrba, P., Inui, M., and Yukawa, H. (2001) Bacterial phosphotransferase system (PTS) in carbohydrate uptake and control of carbon metabolism. *J. Bios. Bioeng*. 92(6): 502-517.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E .L. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305: 567-580.

Kundig, W., Ghosh, S., and Roseman, S. (1964) Phosphate Bound to Histidine in a Protein as an Intermediate in a Novel Phospho-Transferase System. *Proc Natl Acad Sci USA.* 52:1067-1074.

Kunst, F., Ogasawara, N., Moszer, I., et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390: 249-256.

Lapointe, J., Duplain, L., and Proulx, M. (1986) A single glutamyl-tRNA synthetase aminoacylates tRNAGlu and tRNAGln in *Bacillus subtilis* and efficiently misacylates *Escherichia coli* tRNAGln in vitro. *J. Bacteriol*. 165: 88-93.

Leamon, J. H. et al. (2003) A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* 24: 3769-3777.

Lee, S. H., Kim, K. A., Park, Y. G., Seong, I. W., Kim, M. J., and Lee, Y. J. (2000) Identification and partial characterization of a novel hemolysin from *Leptospira interrogans* serovar *lai*. *Gene* 254: 19-28.

Lee, S. H., Kim, S., Park, S. C., and Kim, M. J. (2002) Cytotoxic activities of *Leptospira interrogans* hemolysin SphH as a pore-forming protein on mammalian cells. *Infect. Immun*. 70(1): 315-322.

Leptospirosis. In: Benenson AS (ed), Control of Communicable Diseases Manual. *American Public Health Association 2000*, Washington DC pp 293-296.

Levett, P. N. (2001) Leptospirosis. *Clin. Microbiol. Rev*. 14: 296-326.

Li, C., Motaleb, A., Sal, M., Goldstein, S. F., and Charon, N. W. (2000) Spirochete periplasmic flagella and motility. *J. Mol. Microbiol. Biotechnol*. 2: 345-354.

Lin, M., Bughio, N., and Surujballi, O. (1999) Expression in *Escherichia coli* of

flaB, the gene coding for a periplasmic flagellin of *Leptospira interrogans* serovar *pomona*. *J. Med. Microbiol.* 48(11): 977-982.

Lin, M., Dan, H., and Li, Y. (2004) Identification of a second flagellin gene and functional characterization of a sigma70-like promoter upstream of a *Leptospira borgpetersenii flaB* gene. *Curr. Microbiol.* 48(2): 145-152.

Louvel, H., Saint Girons, I., and Picardeau, M. (2005) Isolation and characterization of FecA- and FeoB-mediated iron acquisition systems of the spirochete *Leptospira biflexa* by random insertional mutagenesis. *J. Bacteriol.* 187(9): 3249-3254.

Lowe, T. M. and Eddy, S. R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.* 25: 955-964.

Maguire, B. A. and Wild, D. G. (1997) The roles of proteins L28 and L33 in the assembly and function of *Escherichia coli* ribosomes in vivo. *Mol. Microbiol*. 23(2): 237-245.

Mardis, E. R. and Roe, B. A. (1989) Automated methods for single-stranded DNA isolation and dideoxynucleotide DNA sequencing reactions on a robotic workstation. *BioTechniques* 7: 736-746.

Margulies, M., Egholm, M., Altman, W. E., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057): 376-380.

Marotto, P. C., Nascimento, C. M., Eluf-Neto, J., Marotto, M. S., Andrade, L., Sztajnbok, J., and Seguro, A. C. (1999) Acute lung injury in leptospirosis: clinical and laboratory features, outcome, and factors associated with mortality. *Clin. Infect. Dis.* 29: 1561-1563.

Matsunaga, J., Barocchi, M. A., and Croda, J. (2003) Pathogenic *Leptospira* species express surface-exposed proteins belonging to the bacterial immunoglobulin superfamily. *Mol. Microbiol*. 49: 929-945.

Maxam, A. M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA.* 74(2): 560-564.

Merien, F., Truccolo, J., Baranton, G., and Perolat, P. (2000) Identification of a 36-kDa fibronectin-binding protein expressed by a virulent variant of

Leptospira interrogans serovar *icterohaemorrhagiae*. *FEMS Microbiol. Lett.* 185(1): 17-22.

Metzker, M. L., Lu, J., and Gibbs, R. A. (1996) Electrophoretically uniform fluorescent dyes for automated DNA sequencing. *Science* 271(5254): 1420-1422.

Micklos, D. A. and Freyer, G. A. (1999) *DNA Science*. pp. 58-59. Cold Spring Harbor laboratory press, Cold Spring Harbor, NY.

Mirelman, D., Yashouv-Gan, Y., and Schwarz, U. (1976) Peptidoglycan biosynthesis in a thermosensitive division mutant of *Escherichia coli*. *Biochemistry* 15(9): 1781-1789.

Moon, S., Byun, Y., Kim, H. J., Jeong, S., and Han, K. (2004) Predicting genes expressed via -1 and +1 frameshifts. *Nucleic. Acids Res*. 32(16): 4884-4892.

Motaleb, M. A., Miller, M. R., Li, C., Bakker, R. G., Goldstein, S. F., Silversmith, R. E., Bourret, R. B., and Charon, N. W. (2005) CheX is a phosphorylated CheY phosphatase essential for *Borrelia burgdorferi* chemotaxis. *J. Bacteriol*. 187(23): 7963-7969.

Moxon, E. R., Hood, D., and Richards, J. (1998) Bacterial lipopolysaccharides: candidate vaccines to prevent *Neisseria meningitidis* and *Haemophilus influenzae* infection. *Adv. Exp. Med. Biol.* 435: 237-243.

Muller, M. and Klosgen, R. B. (2005) The Tat pathway in bacteria and chloroplasts. *Mol. Membr. Biol.* 22(1-2): 113-121.

Mullis, K. and Faloona, F. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* 155: 335-350.

Najar, F. (2002) Sequence and analysis of *Actinobacillus actinomycetemcomitans*. Ph. D. dissertation, University of Oklahoma.

Nascimento, A. L., Ko, A. I., Martins, E. A., et al. (2004) Comparative genomics of two Leptospira interrogans serovars reveals novel insights into physiology and pathogenesis. *J. Bacteriol.* 186 (7): 2164-2172.

Niviere, V., Wong, S. L., and Voordouw, G. (1992) Site-directed mutagenesis of the hydrogenase signal peptide consensus box prevents export of a beta-lactamase fusion protein. *J. Gen. Microbiol*. 138: 2173-2183.

Nyren, P., Pettersson, B., and Uhlen, M. (1993) Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal. Biochem*. 208: 171-175.

Oefner, P. J., Hunicke-Smith, S. P., Chiang, L., Dietrich, F., Mulligan, J., and Davis, R. W. (1996) Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. *Nucl. Acids Res.* 24(20): 3879-3886.

Ollis, D. L., Brick, P., Hamlin, R., Xuong, N. G., and Steitz, T. A. (1985) Structure of large fragment of *Escherichia coli* DNA polymerase I complexed with dTMP. *Nature* 313: 762-766.

Pace, N. R. and Smith, D. (1990) Ribonuclease P: function and variation. *J. Biol. Chem.* 265(7): 3587-3590.

Palaniappan, R. U., Chang, Y. F., and Jusuf, S. S. (2002) Cloning and molecular characterization of an immunogenic LigA protein of *Leptospira interrogans*. *Infect. Immun.* 70: 5924-5930.

Pautsch, A. and Schulz, G. E. (1998) Structure of the outer membrane protein A transmembrane domain. *Nat. Struct. Biol*. 5(11): 1013-1017.

Pawelek, P. D., Croteau, N., Ng-Thow-Hing, C., Khursigara, C. M., Moiseeva, N., Allaire, M., and Coulton, J. W. (2006) Structure of TonB in complex with FhuA, *E. coli* outer membrane receptor. *Science* 312(5778): 1399-1402.

Perna, T., Plunkett, G., Burland, V., et al. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409: 529-533.

Perona, J. J., Rould, M. A., and Steitz, T. A. (1993) Structural basis for transfer RNA aminoacylation by *Escherichia coli* glutaminyl-tRNA synthetase. *Biochemistry* 32: 8758-8771.

Phadnis, S. H., Huang, H. V., and Berg, D. E. (1989) Tn5supF, a 264-base-pair transposon derived from Tn5 for insertion mutagenesis and sequencing DNAs cloned in phage lambda. *Proc. Natl. Acad. Sci. USA.* 86(15): 5908-5912.

Pheiffer, B. H. and Zimmerman S. B. (1983) Polymer-stimulated ligation: enhanced blunt- or cohesive-end ligation of DNA or deoxyribooligonucleotides by T4 DNA ligase in polymer solutions. *Nucl.*

Acids Res. 11(22): 7853-7871.

Poltorak, A., He, X., Smirnova, I., et al. (1998) Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in *Tlr4* gene. *Science* 282(5396): 2085-2088.

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., et al. (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 238: 336-341.

Puente, X. S., Sanchez, L. M., Overall, C. M., and Lopez-Otin, C. (2003) Human and mouse proteases: a comparative genomic approach. *Nat. Rev. Genet*. 4: 544-558.

Ramadass, P., Jarvis, B. D., Corner, R. J., Penny, D, and Marshall, R. B. (1992) Genetic characterization of pathogenic *Leptospira* species by DNA hybridization. *Int. J. Syst. Bacteriol.* 42: 215-219.

Rapoport, T. A. (1992) Transport of proteins across the endoplasmic reticulum membrane. *Science* 258(5084): 931-936.

Rech, J., Cathala, G., and Jeanteur, P. (1980) Isolation and characterization of a ribonuclease activity specific for double-stranded RNA (RNase D) from Krebs II ascites cells. *J. Biol. Chem.* 255: 6700-6706.

Reizer, J., Hoischen, C., Reizer, A., Pham, T. N., and Saier, M. H. Jr. (1993) Sequence analyses and evolutionary relationships among the energy-coupling proteins enzyme I and HPr of the bacterial phosphoenolpyruvate: sugar phosphotransferase system. *Protein Sci.* 2: 506-521.

Reizer, J., Reizer, A., Lagrou, M. J., Folger, K. R., Stover, C. K., and Saier, M. H., Jr. (1999) Novel phosphotransferase systems revealed by bacterial genome analysis: the complete repertoire of pts genes in *Pseudomonas aeruginosa*. *J. Mol. Microbiol. Biotechnol.* 1: 289-293.

Ren, S. X., Fu, G., Jiang, X. G., et al. (2003) Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing. *Nature* 422: 888-893.

Richardson, C. C. (1981) Boyer, P. D. (editor), *The Enzyme.* pp. 229-314. Academic press, San Diego, CA.

Rock, C. O. and Cronan, J. E. (1996) *Escherichia coli* as a model for the regulation of dissociable (type II) fatty acid biosynthesis. *Biochim. Biophys. Acta.* 1302: 1-16.

Ronaghi, M. et al. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem*. 242: 84-89.

Ronaghi, M., Uhlen, M., and Nyren, P. (1998) A sequencing method based on real-time pyrophosphate. *Science* 281: 363-365.

Rosenblum, B. B., Lee, L. G., Spurgeon, S. L., et al. (1997) New dye-labeled terminators for improved DNA sequencing patterns. *Nucl. Acids Res.* 25: 4500-4504.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., and Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* 16(10): 944-945

Saier, M. H. Jr. (2000) Families of transmembrane sugar transport proteins. *Mol. Microbiolo*. 35: 699-710.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA.* 74(12): 5463-5467.

Santini, C. L., Ize, B., Chanal, A., Muller, M., Giordano, G., and Wu, L. F. (1998) A novel sec-independent periplasmic protein translocation pathway in *Escherichia coli*. *EMBO J.* 17: 101-112.

Schiemann, D. A. (1973) Thiosulfate and survival of *Leptospira autumnalis* Akiyami A. *Am. J. Vet. Res.* 34(10): 1345-1350.

Schumann, W. (2003) The *Bacillus subtilis* heat shock stimulon. *Cell Stress Chaperones* 8(3): 207-217.

Serres, M. H., Gopal, S., Nahum, L. A., Liang, P., Gaasterland, T., and Riley, M. (2001) A functional update of the *Escherichia coli* K-12 genome. *Genome Biol.* 2(9): 1-6.

Shimura, Y., Sakano, H., and Nagawa, F. (1978) Specific ribonucleases involved in processing of tRNA precursors of *Escherichia coli*. Partial purification and some properties. *Eur. J. Biochem*. 86: 267-281.

Simpson, W., Olczak, T., and Genco, C. A. (2000) Characterization and expression

of HmuR, a TonB-dependent hemoglobin receptor of Porphyromonas gingivalis. *J. Bacteriol.* 182(20): 5737-5748.

Smibert, R. M. (1977) The spirochaetales. In *CRC Handbook of Microbiology,* 2nd edn. vol. 1. pp. 195-228. Laskin, A. I. and Lechavelier, H. A. (Eds.), CRC Press, Cleveland, OH.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., et al. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321: 674-679.

Springer, T. A. (1997) Folding of the N-terminal, ligand-binding region of integrin α-subunits into a ß-propeller domain. *Proc. Natl. Acad. Sci. USA*. 94: 65-72.

Stephenson, K. (2005) Sec-dependent protein translocation across biological membranes: evolutionary conservation of an essential protein transport pathway. *Mol. Membr. Biol.* 22(1-2): 17-28.

Sugiura, I., Nureki, O., Ugaji-Yoshikawa, Y., et al. (2000) The 2.0 A crystal structure of *Thermus thermophilus* methionyl-tRNA synthetase reveals two RNA-binding modules. *Structure* 8: 197-208.

Swerdlow, H. and Gesteland, R. (1990) Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucl. Acids Res.* 18: 1415-1419.

Sylvia, H., Jan, F., Nasser, G., et al. (2001) Autotrophic CO2 fixation by *chloroflexus aurantiacus*: study of glyoxylate formation and assimilation via the 3-Hydroxypropionate cycle. *J. Bacteriol*. 183 (14): 4305-4316.

Szurmant, H., Bunn, M. W., Cannistraro, V. J., and Ordal, G. W. (2003) *Bacillus subtilis* hydrolyzes CheY-P at the location of its action, the flagellar switch. *J. Biol. Chem.* 278(49): 48611-48616.

Tabor, S. and Richardson, C. C. (1990) DNA sequence analysis with a modified bacteriophage T7 DNA polymerase: effect of pyrophosphorolysis and metal ions. *J. Biol. Chem.* 265: 8322-8328.

Tabor, S. and Richardson, C. C. (1995) A single residue in DNA polymerases of *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxyand dideoxynucleotides. *Proc. Natl. Acad. Sci. USA*. 2: 6339-6343.

Takahashi, Y., Akase, K., Hirano, H., and Fukunaga, M. (1998) Physical and

genetic maps of the *Leptospira interrogans* serovar *icterohaemorrhagiae* strain Ictero no. 1 chromosome and sequencing of a 19-kb region of the genome containing the 5S rRNA gene. *Gene* 215: 37-45.

Takeshita, S., Sato, M., Toba, M., Masahashi, W., and Hashimoto-Gotoh, T. (1987) High-copy-number and low-copy-number plasmid vectors for *lacZ* alpha-complementation and chloramphenicol- or kanamycin-resistance selection. *Gene* 61(1): 63-74.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4(41): 1-14

Tenover, F. C. (2006) Mechanisms of antimicrobial resistance in bacteria. *Am. J. Med.* 119(6 Suppl.1): S3-S10.

Thomas, J. D., Daniel, R. A., Errington, J., and Robinson, C. (2001) Export of active green fluorescent protein to the periplasm by the twin-arginine translocase (Tat) pathway in *Escherichia coli*. *Mol. Microbiol*. 39: 47-53.

Tumbula, D. L., Becker, H. D., Chang, W. Z., and Soll, D. (2000) Domainspecific recruitment of amide amino acids for protein synthesis. *Nature* 407: 106-110.

Umbarger, H. E. (1978) Amino acid biosynthesis and its regulation. *Annu. Rev. Biochem.* 47: 533-606.

van Heijenoort, J. (1998) Assembly of the monomer unit of bacterial peptidoglycan. *Cell Mol. Life Sci.* 54(4): 300-304.

Vinetz, J. M. (2001) Leptospirosis. *Current Opinion in Infectious Diseases*. 14: 527-538.

Voet, D. and Voet, J. G. (1995) *Biochemistry.* 2nd edn. pp. 967-971. John Wiley & Sons Inc., New York.

von Heijne, G. (1985) Signal sequences. The limits of variation. *J Mol Biol*. 184(1): 99-105.

Watson, J. and Crick, F. (1953) A structure for deoxyribose nucleic acid. *Nature* 171(4356): 737-738.

Werts, C., Tapping, R. I., Mathison, J. C., et al. (2001) Leptospiral lipopolysaccharide activates cells through a TLR2-dependent mechanism.

Nature Immuno. 2: 346-352.

Wilcox, M. and Nirenberg, M. (1968) Transfer RNA as a cofactor coupling synthesis with that of protein. *Proc. Natl. Acad. Sci. USA* 61:229–236.

Wilson, M., McNab, R., and Henderson, B. (2002) *Bacterial Disease Mechanisms: An Introduction to Cellular Microbiology.* pp. 125-157. Cambridge University Press, New York.

World Health Organization (1999). Leptospirosis worldwide, *Wkly. Epidemiol. Res*. 74: 237-242.

Wu, J. and Woodard, R. W. (2003) *Escherichia coli* YrbI is 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase. *J. Biol. Chem.* 278(20): 18117-18123.

Xu, H., Zhang, Y., and Guo, X., et al. (2004) Isoleucine biosynthesis in *Leptospira interrogans* serotype *lai* strain 56601 proceeds via a threonine-independent pathway. *J. Bacteriol.* 186(16): 5400-5409.

Yamane, K., Bunai, K., and Kakeshita, H. (2004) Protein traffic for secretion and related machinery of *Bacillus subtilis*. *Biosci. Biotechnol. Biochem*. 68(10): 2007-2023.

Yanagihara, Y., Kojima, T., and Mifuchi, I. (1982) Hemolytic activity of *Leptospira interrogans* serovar *canicola* cultured in protein-free medium. *Microbiol. Immunol.* 26: 547-556.

Yanagihara, Y., Taniyama, T., Misaki, H., Suzuki, Y., Matsumoto, M., and Mifuchi, I. (1984) Phospholipases of *Leptospira.* I. Presence of phospholipase A1 and lysophospholipase in *Leptospira biflexa*. *Microbiol Immunol.* 28(7): 747-756.

Zempleni, J. and Mock, D. M. (1999) Biotin biochemistry and human requirements. *Nutr. Biochem.* 10: 128-138.

Zuerner, R. L. (1991) Physical map of the chromosomal and plasmid DNA comprising the genome of *Leptospira interrogans*. *Nucl. Acids Res.* 19: 4857-4860.

Appendix A 454/Roche GS-20 instrument DNA Sequencing Protocol

A1 DNA Library Preparation

A1.1 DNA Nebulization

1. Pipette 3-5µg of DNA (in TE) into a nebulizer.
2. Add TE buffer to a final volume of 100ul.
3. Add 500ul of nebulization buffer and swirl the mixture completely.
4. Nebulize the mixture for 1 min. by 45 psi of nitrogen.
5. Add 2.5ml of Qiagen's buffer PB into the nebulizer.
6. The nebulized DNA is purified using two columns from a MinElute PCR Purification Kit (Qiagen) according to the manufacture's instructions.

A1.2 Fragment End Repair

1. Add the following reagents into a tube: 23 µl of purified, nebulized DNA fragments, 5 µl of 10× Polishing buffer, 5 µl of BSA, 5 µl of ATP, 2 µl of dNTPs, 5 µl of T4 PNK, and 5 µl of T4 DNA polymerase rather than Klenow DNA polymerase.
2. Incubate the end repair mixture for 15 min. at 12°C followed by 15 min. at 25°C.
3. Purify the end-repaired fragments as mentioned in Step 6 of A1.1 DNA Nebulization.

A1.3 Adaptor Ligation

1. Add the following reagents into a tube: 15 µl of nebulized, polished DNA, 20 µl of 2× Ligase buffer, 1 µl of adaptor, 4 µl of ligase.
2. Incubate the mixture for 15 min. at 25°C.

3. Purify the mixture using MinElute PCR Purification Column.

A1.4 Library Immobilization and Fill-in Reaction

1. Elute the ligated DNA from tge MinElute column with 25ul of Buffer EB into the tube containing washed 50ul Library Immobilization Beads.
2. Spin the mixture for 20 min. at 22°C.
3. Add the following reagents into a tube: 40 µl of Mol. Biol. Grade water, 5 µl of 10× Fill-in polymerase buffer, 2 µl of dNTPs mix, 3 µl of Fill-in polymerase.
4. Incubate the mixture for 20 min. at 37°C.
5. Wash the immobilized library twice with 100ul of Library Wash Buffer.

A1.5 ssDNA Library Isolation

1. Remove the 100ul of Library Wash Buffer from the library-carrying beads using Magnetic Particle Collector (MPC).
2. Add 50 µl of Melt Solution (0.125 ml of NaOH and 9.875 ml of Mol. Biol. Grade water).
3. Vortex the mixture and remove the beads from the supernatant.
4. Transfer the supernatant to the nebulization solution (500ul of Qiagen PB buffer and 3.8ul of 20% acetic acid).
5. Repeat steps 2-4 and combine the supernatant.
6. Purify ssDNA library using MinElute PCR Purification Column.

A2 ssDNA Amplification and Sequencing

A2.1 ssDNA Library Capture

1. Transfer 60ul of DNA capture beads per reaction to a tube.
2. Spin the tube for 10 sec., rotate the tube 180°, and spin again for 10 sec.

3. Pellet the beads, then remove the supernatant.

4. Wash beads in each tube twice with 200ul of 1× Capture Bead Wash buffer.

5. Resuspend the beads in 50ul of 1× Capture Bead Wash buffer.

6. Pellet the beads and discard 30ul of the supernatant.

7. Add the appropriate amount of ssDNA determined from titration experiment to the tube.

8. Anneal the ssDNA library to beads by running ssDNA annealing program on thermal cycler.

9. Pellet the beads and discard ~20ul of the supernatant. In total, 16 tubes of beads are prepared.

A2.2 Emulsification and Amplification

1. Prepare the Live Amplification Mix for 16 reactions, including 2905.92 µl of Amplification Mix, 160 µl of MgSO4, 33.28 µl of Amplification primer mix, 96 µl of Platinum HiFi polymerase, 4.8 µl of PPiasa.

2. Vortex 16 tubes of emulsion oil (500ul in each tube) for 10 sec.

3. Add 240 µl of Mock Amplification mix to each tube.

4. Put tubes into the TissueLyser and shake for 5 min. at 25/sec.

5. Add 160 µl of Live Amplification Mix to each of the 16 tubes prepared in **A2.1 ssDNA Library Capture**.

6. Add the bead mixture to the emulsion tube and shake in TissueLyser for 5 min. at 15/sec.

7. Split each emulsion reaction into 8 amplification reactions.

8. ssDNA is amplified in each emulsion by running the amplification program on thermal cycler.

A2.3 Emulsion Breaking and Bead Washing

1. Add 100 µl of isopropanol to each emulsion reaction tube.
2. Draw the emulsion/isopropanol mix from every 32 wells (4 reactions) into a syringe. In total, four syringes are needed.
3. Add 100 µl of isopropanol to each reaction wells.
4. After mixing, draw the content of the wells to corresponding syringe until the solution volume reaches 9 ml.
5. After inverting, draw some air into the syringe. Then, attach the SwinLock Filter between the blunt needle and the syringe.
6. After mixing, expel the content of syringe and draw 9 ml of fresh isopropanol into the syringe for a second wash.
7. Wash the beads with 6 ml of 1× Bead Wash Buffer and 6 ml of 1× Enhancing Fluid.
8. Draw 0.5 ml of 1× Enhancing Fluid into each syringe to resuspend beads.

A2.4 Bead Recovery

1. Expel the content of syringe into a 1.5ml tube after removing the Swinlock filter.
2. Spin the content of the tube to pellet beads. Then discard the supernatant.
3. Draw 0.5 ml of 1× Enhancing Fluid into each syringe to resuspend beads for a 2nd recovery.
4. Expel the content of the syringe into a corresponding 1.5 ml tube.

5. Spin the content of the tube to pellet beads, and discard ~100 µl of 1× Enhancing Fluid from each tube.

A2.5 Bead Enrichment

1. Resuspend the tube of Enrichment beads completely.
2. Add 1 ml of 1× Enhancing Fluid to each of four 1.5 ml tubes. Then add 20 µl of Enrichment beads to each tube and vortex for 5 sec.
3. Place the tubes on a MPC to pellet beads.
4. Discard ~400 µl of supernatant, then add 100 µl of Enrichment Fluid to each tube and resuspend the beads.
5. Combine 100ul of washed Enrichment Beads with 100ul of amplified DNA beads, and votex for 3 seconds.
6. Add 1ml of 1× Enhancing Fluid to each tube.
7. Leave the tube on a MPC for 2 min. Then discard the supernatant.
8. Add 1 ml of 1× Enhancing Fluid to each tube.
9. Repeat steps 7 and 8 twice.
10. Take the tube out of the MPC, and resuspend the bead pellet in 700 µl of Melt Solution (0.125 ml of 10N NaOH and 9.875 ml Mol. Grade water).
11. After vortexing for 5 min, put the tube back on theMPC.
12. Transfer the supernatant containing the enriched ssDNA into a separate 1.5 ml tube.
13. Pellet the enriched DNA beads by spinning for 10 sec, and after 180° rotating, additional 10 sec.

14. Discard the supernatant. Then wash the beads twice with 1 ml of 1× Annealing Buffer.

15. Spin and discard 900ul of the supernatant.

16. Transfer the remaining enriched bead suspension to a 0.2ml tube.

A2.6 Sequencing Primer Annealing

1. pellet the enriched beads and discard the supernatant.

2. Add 15 µl of 1× Annealing Buffer and 12 µl of Sequencing Primer to each of the four tubes.

3. After vortexing for 5 sec, run the Sequencing Primer Annealing program.

4. After the program is finished, add 100 µl of 1× Annealing Buffer to each tube, pellet the beads and discard the supernatant.

5. Wash the beads again with 200 µl of 1× Annealing Buffer, and then resuspend them in 100ul of 1× Annealing Buffer.

6. Count a 5ul aliquot of the beads in the Coulter Counter following the manufacturer's instructions.

Finally, load the sample in the wells of a PicoTiterPlate device, and perform the sequencing using the GS 20 sequencer.

Appendix B Fosmid DNA Isolation and End Sequencing Protocol

B1 Culture Growth

1. Prepare 1L of media including 900ml of TB media, 100ml of TB salt, and 1.5ml of Antibiotic Chloraphenicol (10mg/ml).
2. Fill 96 deep-well blocks with 1.5ml of media.
3. Incubate deep well blocks with 2µl of initial inoculation culture.
4. Grow in floor shaker for 15 hours.
5. Dilute the induction solution (1000×) 100 times to get 10× of induction solution, and add 15µl into each well.
6. Keep growing 5 hours more.

B2 fosmid DNA Isolation

1. Centrifuge the blocks at 25,000rpm for 10 minutes. Then discard the supernatant.
2. Immediately add 300µl of TE-RNase A/T1 to each well.
3. Incubate on the top shaker at the speed of 6 for 20 minutes or more until cells are fully resuspended.
4. Make the 400ml of fresh lysis buffer including 40ml of 10% SDS, 40 ml of 2N NaOH, and 320ml of ddH_2O.
5. Add 300µl of fresh lysis buffer to each well.
6. Incubate by shaking at a speed of 2 for 5 minutes or more until a clear lysate is developed.
7. Add 300µl of KOAc (3M, pH 4.5) to each well. Shake at a speed of 4 for 15 minutes.

8. Freeze the blocks at –80°C overnight.

9. Thaw the blocks and centrifuge at 4,250 rpm for 45 minutes.

10. Transfer 600µl of supernatant into the new deep-well blocks.

11. Add 1ml of 95% of ethanol into the each well.

12. Centrifuge at 3,000 rpm for 30 minutes.

13. Decant the supernatant. Wash the pellet by adding 1ml of 70% ethanol into each well.

14. Centrifuge the blocks at 3,000rpm for 30 minutes.

15. Decant the supernatant. And dry the DNA pellets on the bench for at least 30 minutes.

16. Prepare 24ml of RNase A solution including 240µl of Tris-HCl (50mM, pH 7.6), 4.8µl of 0.5M EDTA, 5µl of 40µg/ml RNase A, and X ml of H_2O.

17. Resuspend the DNA in 20µl of RNase A solution.

18. Shake on the table-top shaker for 30 minutes to dissolve DNA.

B3 End Sequencing

Sequencing mix includes 10µl of fosmid DNA, 2µl of 20× diluted ET, 1.5µl of DMSO, and 2µl of 7nM universal forward or reverse primer.