12-10-2010

Evolution Of The Protein-Coding Genes In The Genomes Of The Mycoplasmatales

Dipaloke Mukherjee

EVOLUTION OF THE PROTEIN-CODING GENES IN THE GENOMES OF THE MYCOPLASMATALES

By

Dipaloke Mukherjee

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Biological Sciences
in the Department of Biological Sciences

Mississippi State, Mississippi

December 2010

EVOLUTION OF THE PROTEIN-CODING GENES IN THE GENOMES OF THE MYCOPLASMATALES

By

Dipaloke Mukherjee

Approved:

Walter J. Diehl Professor of Biological Sciences & Associate Dean for Research and Graduate Studies, College of Arts and Sciences (Director of Dissertation)	Dwayne A. Wise Professor of Biological Sciences (Committee Member)
Lisa Wallace Assistant Professor of Biological Sciences (Committee Member)	Susan M. Bridges Professor of Computer Science and Engineering (Committee Member)
G. Todd Pharr Associate Professor of Basic Sciences (Committee Member)	Gary N. Ervin Associate Professor of Biological Sciences (Graduate Coordinator)
Gary L. Myers Professor & Dean, College of Arts and Sciences	

Name: Dipaloke Mukherjee

Date of Degree: December 10, 2010

Institution: Mississippi State University

Major Field: Biological Sciences

Major Professor: Dr. Walter J. Diehl, III

Title of Study: EVOLUTION OF THE PROTEIN-CODING GENES IN THE GENOMES OF THE MYCOPLASMATALES

Pages in Study: 86

Candidate for Degree of Doctor of Philosophy

The bacterial species belonging to the order Mycoplasmatales have highly truncated genomes, and are thus ideal for studying genome evolution patterns. Fourteen members (twelve species) of this order were selected for study of genome evolution based on gene function and phylogeny. A database was constructed that consisting of the set of genes that are common to all of these species, and these genes were further subdivided based on their functions. A Bayesian phylogenetic tree was also constructed from the 16 S ribosomal DNA sequences from these species and robust clades were identified for testing the influence of selection on gene evolution, from which the clades were selected and tested for evidence of natural selection. Two separate statistical techniques, namely the codon substitution models and McDonald-Kreitman tests were used to analyze the presence or absence of selection for genes in different functional categories..

The studies demonstrated that the set of genes associated with cellular processes show the highest percentage of selection and are likely to play a crucial role in

Mycoplasma evolution (for example, by altering the arrangement of antigens on the cell surface and thus enabling a particular *Mycoplasma* species to expand its host range). The presence of selection could only be identified at the earliest divisions of the phylogeny. Tests were also performed to detect the presence of a number of neutral genetic processes that can potentially confound detection of patterns of selection. None of these processes were found to affect the results of the analyses. The study has the potential to identify genes, gene complexes or even pathways that may be involved directly or indirectly in speciation.

DEDICATION

Dedicated to the memory of my beloved mother.

ACKNOWLEDGEMENTS

“ Nature, it seems, is popular name

For milliards and milliards and milliards

Of particles playing their infinite game

Of billiards and billiards and billiards”

--- Piet Hein (1969)

Hein illustrates the physics of molecule. Whole nature with us contains millions of particles. My inquisitive mind with nature has brought me here to study the science regarding the evolution, precisely speaking molecular evolution, more precisely the pattern of genome evolution. Mind is not totally saturated with all questions revolving around my mind and words are a few when to express saturated feelings. On a day when the years of run-up to my quest finally has taken shape in the form of this dissertation, all the agony and ecstasy, grins and grimaces, bouquets and brickbats seem relevant. They are all part of my endeavor; they are intertwined like the strands in a fabric. On this day my heart swells with emotion and gratitude to persons whose relentless help, encouragement, support and guidance have made possible my effort to pursue my goal.

At the onset, I pay my heartfelt gratitude, deep reverence and sincere acknowledgement to my major professor Dr. Walter J. Diehl whose constant vigil, round the clock

supervision, tireless guidance, lively encouragement and invaluable advice have made my run-up to this effort possible. No word in the universe will be adequate to express my emotions, regards and gratefulness to him.

I am also deeply thankful to my committee members, Dr.(s) Dwayne A. Wise, Lisa Wallace, Susan M. Bridges and G. Todd Pharr for their active help and cooperation during the entire period.

I would like to convey my heartiest thanks to Dr. Russell Stocker, Department of Mathematics and Statistics and Mr. Jonathan Harper, Department of Computer Science and Engineering for their active help regarding my research.

My sincere and grateful thanks are due to Dr. Nancy Reichert, Chair and Dr. Gary N. Ervin, Graduate coordinator, Department of Biological Sciences, Mississippi State University for their support in providing me with necessary facilities to conduct this work.

Special thanks are due to my friends, Dr. Amitava Moitra, Mr. Girish Jamnekar and Md. Shaheen Ahmed for their helps in various occasions. I sincerely thank Dr. Debarati Paul for her intellectual support during many of my lonely hours of brain storming. Thanks are due to the upcoming scientists Dr. Wael Badran, Ms Lashmi Narayanan and Ms. Sree Pramod, for their moral support during all these years.

I thank profusely all the faculty, staff and students working at various departments, laboratories and fields of Mississippi State University for their active help extended to me.

I profusely thank my wife Soma Mukherjee, for her support to allay the anxiety and stress associated with this work. I am also very proud of my little daughter Sanjhbati for her constant presence around me filling the void and loneliness in the desperate hours.

Finally, thanks and immense gratitude are due to my father Dr. Debdas Mukherjee, and my father-in- law Mr. Subhas Chandra Ray for their all-round support and constant encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Natural Selection Versus Neutrality

Natural selection is the process by which the traits that are favorable for the survival and fecundity of a species become more common in the successive generations (Darwin, 1859). It plays an important role in the evolution of some parts of the genome. If an amino acid substitution favors the survival and reproduction of a species, it is selected, and the phenomenon is known as positive Darwinian selection. If the substitution adversely affects the survival and/or reproduction of the organism, it is selected against. This process is known as purifying or negative Darwinian selection.

Other forces that may affect genome evolution at the molecular level are neutral with respect to selection. This concept, proposed by Motoo Kimura (1968), states that most of the molecular differences that exist among the genomes of different species do not affect their fitness, i.e. they are selectively neutral. Therefore natural selection does not affect their evolution. The theory suggests that genetic drift is the main factor responsible for most evolutionary changes. It is not clear, though, which parts of the genome evolve by one means or the other at the molecular level. This is an important question that needs to be answered.

Neutrality at the molecular level can be tested in a number of ways. The analysis involving the nonsynonymous/synonymous rate ratio (dN/dS) forms the basis of many

analyses. Two types of nucleotide substitution can take place in the protein-coding genes. In a synonymous substitution, a nucleotide replacement produces a codon that codes for the same amino acid (e.g., CUU replaced by CUC, both coding for leucine). In a non-synonymous substitution, a nucleotide replacement produces a codon that codes for a different amino acid (e.g., CUU replaced by CCU, latter coding for proline). In most of the cases, non-synonymous substitutions are selected against; though in some rare instances positive Darwinian selection can retain these substitutions in the population (Yang, 1998). The synonymous and non- synonymous nucleotide substitution rates are expressed as dS and dN respectively, and their ratio (dN/dS) is designated by the letter ω. A ratio of 1 indicates neutrality. In other words, the rate of fixation of a neutral amino acid mutation will be equivalent to that of a synonymous substitution. A ratio less than 1 indicates purifying selection and the substitution is eventually eliminated from the population. If the ratio is greater than 1, then positive Darwinian selection will retain the amino acid mutation in the population (Nei and Gojobori, 1986). It needs to be mentioned that these are not simply the ratios of single numbers, rather these are the ratios calculated from the numbers of all the synonymous and non synonymous substitutions across the entire gene.

A number of studies have used dN/dS ratios to evaluate selection. For example, Seibert et al. (1995) studied the natural selection acting on the principle genes of HIV-1, (namely *gag*, *pol*, *env* and gp4) in a number of related virus families. In the genes other than *env*, strong evidence of purifying selection was indicated, evident by dN/dS ratios less than one, but evidence of positive Darwinian selection was indicated in the V2, V3 and many other-variable regions of the *gp120* gene. These regions serve as antibody

recognition sites, and when such recognition does not occur, positive Darwinian selection can take place. The dN/dS ratio for the *env* gene was found to vary in different families of HIV strains, with some showing the indication of positive selection, whereas others were found to be subjected to purifying selection.

A number of variants of the dN/dS analyses have allowed refinement of hypothesis testing for selection. Yang (1997) developed several codon substitution models which can be used to test a number of hypotheses, such as:

1. The dN/dS ratio for a particular branch of a phylogenetic tree is different from the other branches,

2. The dN/dS ratio for a given branch of a tree is greater than 1 etc.

Yang (1997) used these models to detect positive selection in the evolution of primate lysozyme genes. He constructed a phylogenetic tree using lysozyme gene sequences from 24 primate species, and identified two main branches in the phylogeny: branch *h* (leading to hominoids) and *c* (leading to colobine monkeys). He found dN/dS ratio of the lysozyme gene sequence for the branch *h* to be significantly greater than one, indicating strong evidence of positive Darwinian selection. However, the other branch did not show any such indication.

Many authors have used the codon substitution models for selection studies. For example, Chen et al (2006) used these models to identify positively selected genes in uropathogenic strains of *E. coli*. Their study resulted in 29 such genes that were found to be associated with a number of biological processes such as nutrient uptake, nucleic acid metabolism, pathogenicity etc.

Another variant of dN/dS ratio analysis is the McDonald-Kreitman (1991) test. In this analysis, the pattern of evolution within a particular species can be compared to the pattern prevalent among more than one species (Eyre-Walker, 2002). This analysis tests whether the ratio of synonymous and non-synonymous nucleotide substitutions for fixed differences between species is same as the ratio of the above-mentioned substitutions for polymorphism occurring in more than one species. The value obtained from a McDonald-Kreitman test can be termed as a neutrality index (NI), which can be interpreted similarly to the dN/dS ratio (Rand and Kann, 1996). NI is calculated as follows,

(No. of polymorphic replacement sites / No. of fixed replacement sites) /

(No. of polymorphic silent sites / No. of fixed silent sites)

An NI of 1.0 indicates strict neutrality. An NI greater than 1.0 indicates greater amino acid variation among species than expected, which can be interpreted as stabilizing selection and may prevent species divergence, whereas a value less than 1.0 indicates adaptive fixation of amino acid variants between species and may favor divergence (Rand and Kann, 1996).

John McDonald and Martin Kreitman (1991) studied the adaptive protein evolution of the alcohol dehydrogenase (*Adh*) locus in three species belonging to the *Drosophila melanogaster* subgroup. They observed an increased rate of fixed replacement (non-synonymous) substitution differences between species than would be expected. They interpreted these results as indicating adaptive fixation of beneficial amino acid substitution.

In a particularly interesting study, Evans et al. (2004) studied the evolution of the *microcephalin* gene, which regulates the size of the human brain during developmental

stages. They compared the gene in humans and chimpanzees to the gene in their simian ancestral lineages. They found an about 45 non-synonymous amino acid substitutions more than expected from the simian ancestors to the human lineages (Evans et al., 2004). These results can be interpreted as an indication of adaptive fixation of amino acid substitution or diversifying selection, as the clades diverged.

1.2 An Overview on the Mycoplasmatales

The Mycoplasmatales is an order that belongs within the domain Eubacteria and includes single cell organisms that do not posses a cell wall. The members of this order lack the ability to synthesize peptidoglycan (the main component of the bacterial cell wall). The name is derived from two Greek words: mykes (fungus) and plasma (formed) (Waites et al., 2005). Initially the organism was considered to be some kind of virus. Later it was thought to be L-form bacteria (bacteria that have lost their cell walls). Not until the 1960s was *Mycoplasma* proven to be different from general L-form bacteria (Rogul et al., 1964). The first successful cultivation of a species of *Mycoplasma* (the bovine pleuropneumonia) was reported by Nocard and Roux (1898).

The genus *Mycoplasma* contains more than 100 recognized species (Fadiel et al., 2007). They are the simplest form of living organism and solely depend on the host cells to supply essential nutrients, such as amino acids, nucleotides, sterols etc. *Mycoplasma* species are responsible for causing a number of diseases of plants, animals and humans. For example, *Mycoplasma pneumoniae* is a common human pathogen that causes different types of respiratory tract infections (Razin, 1998). *Mycoplasma capricolum* is an example of animal pathogenic *Mycoplasma* that mainly infects goats and is the causal

agent of the disease caprine contagious agalactia (CCA) (Fe et al., 2005). Besides *Mycoplasma*, the genus *Ureaplasma* is also included in the order Mycoplasmatales. Organisms belonging to this genus (e.g. *Ureaplasma parvum*) are unique in their ability to hydrolyze and metabolize urea (Kenny and Cartwright, 1977).

The members of the order Mycoplasmatales have a small genome size, ranging between 0.6 and 1.35 million base pairs (Fadiel et al., 2007). The G+C content is usually very low (25-34 mol %) (Razin, 1998). They are Gram-negative bacteria and produce "fried egg" shaped colonies on agar. Because of their small genome size, *Mycoplasmas* provide researchers with an opportunity to study the minimal genome (i.e., at genome that consists only of the genes that are absolutely essential for an organism to survive, e.g., the genomes of the Mycoplasmatales which consist of mainly the functional genes with very little or no other genetic materials such as pseudogenes, transposons etc. [Razin, 1998; Fraser et al., 1995]). Genomes of 34 species had been partially or completely sequenced as of January, 2007 (Fadiel, 2007). When this study was initiated (April, 2008), the genomes of 12 species of *Mycoplasma* had been completely sequenced. At that time genome sequences were available for: *Mycoplasma agalactiae*, *M. capricolum,, M. gallisepticum, M. genitalium,* three strains belonging to the species *M. hyopneumoniae* (*M. hyopneumoniae* 232,7448 and J)*., M. penetrans, M. pneumoniae, M. pulmonis, M. mobile, M. mycoides, M. synoviae* and *Ureaplasma parvum.* This indicates that partial genome sequences were available for at least 22 species at the inception of the study.

Figure 1.1 shows the phylogenetic relationship among 15 *Mycoplasma* and *Ureaplasma* strains that are pathogenic to humans (Yoshida et al., 2001). Complete

genomes were sequenced for four of the 15 species used in the study (*U. parvum, M. penetrans, M. pneumoniae, M. genitalium*).

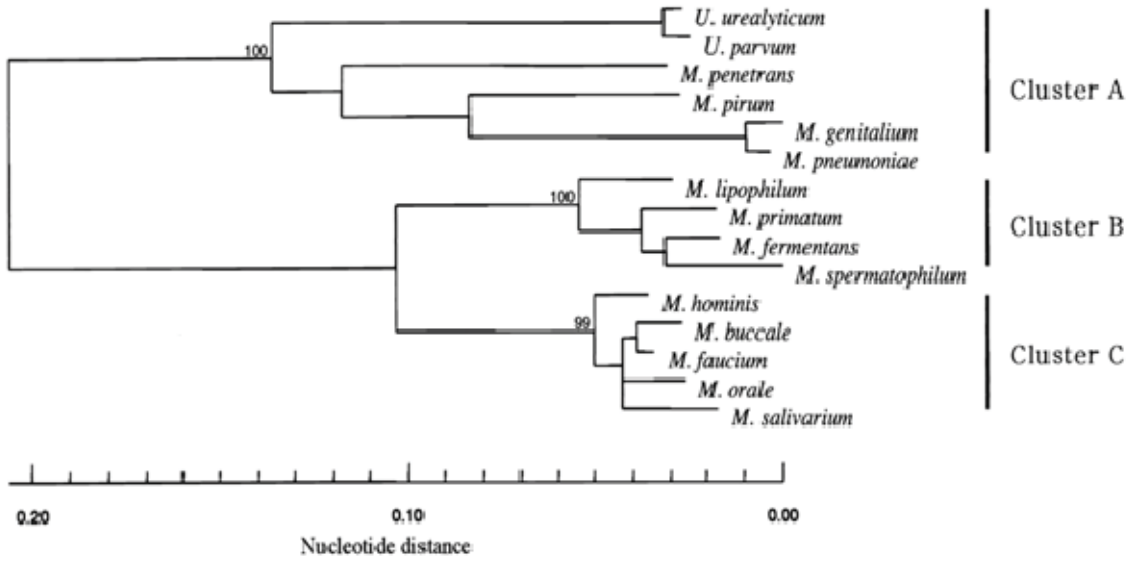


Figure 1.1: Phylogenetic tree based on neighbor-joining analysis of 16S rDNA sequences from 15 pathogenic *Mycoplasma* and *Ureaplasma* species (reproduced from Yoshida et al. 2001).

This tree is not only an example of the many existing *Mycoplasma* phylogenies that were constructed from the 16S rDNA sequences, but it also provides important information about the phylogenetic clustering pattern of a diverse array of *Mycoplasma* species. It consists of two primary clusters, one of which has been subdivided further into two minor clusters (B and C). Two strains belonging to the genus *Ureaplasma* (*U. parvum* and *U. urealyticum*) form a monophyletic group, which indicates their similarity in biochemical activity. The genus *Mycoplasma* is therefore a paraphyletic group within the Mycoplasmatales. All the members of the genus *Mycoplasma* require sterol for their

growth. *M. capricolum* is unique in its ability of accepting a diverse array of cholesterol derivatives (Odriozola, 1978).

The tree discussed above does not inform the current taxonomy of the Mycoplasmatales, which is based on host range and biochemical activity etc (e.g., Manso-Silvan et al., 2007). For example, it does not produce any information about the specific properties of the different species based on which they were clustered together. Some species may have unique metabolic properties which may be responsible for their clustering. Besides, the tree was limited to only human pathogenic species, and thus fails to yield any information about the phylogenetic relationship that can be observed among the *Mycoplasma* species that infect humans as well as other animal hosts.

1.3 Objectives

This study accomplished the following specific objectives:

Specific Objective 1: To determine whether patterns of selection/neutrality differ among functional groups of genes.

Because of their tiny genome size and small gene number, *Mycoplasma* species can be used as a model system to study a variety of aspects of genome evolution, including whether evolutionary patterns are consistent throughout a phylogeny, and whether natural selection affects the evolution of different functional classes of genes differentially. Genes can be broadly classified into three functional groups: cellular processes genes, information storage and processing genes and metabolism genes. In addition, each group is further sub-classified into 3-8 smaller groups based on the specific functions performed by the genes that are members of the functional group. There is also

a set of poorly characterized genes. Preliminary data suggest that cellular processes genes show more adaptive fixation than metabolism genes (Diehl and Perkins, 2004) at the clade separating cluster A from cluster B/C. It needs to be mentioned that, though previous works have been conducted to study this specific objective (Diehl and Perkins, 2004), it is worth further analysis, as the number of species used in the current study is considerably higher than the previous one.

This objective was tested by the following null hypothesis:

The pattern of selection does not depend on gene function.

Specific objective 2: To determine whether the patterns of selection/neutrality differ among clades of a phylogenetic tree.

A phylogenetic tree for the different species of *Mycoplasma* was constructed using Bayesian analysis of 16S ribosomal DNA (rDNA) sequences to determine their evolutionary relationships. The tree was used to identify robust clades about which selection was to be tested. Although virtually any nucleotide sequence can be used to construct a phylogenetic tree, 16S rDNA sequences were used in this study as these are ubiquitous and highly conserved in all species. The tree helped to compare the tree with other existing trees, a number of which have been constructed using 16S rDNA sequences. It was also important to understand the phylogenetic relationships among the sequenced *Mycoplasma* species and to analyze the variable pattern of selection that may exist across the clades of the tree, for example, from the earlier to the more recent clades. The previous study (Diehl and Perkins, 2004) that was conducted to study this specific objective was based on a tree constructed by neighbor-joining analysis. But the current

study followed a different approach (Bayesian method) with more species used for the analysis.

This specific objective was met by testing the following null hypothesis.

The pattern of selection/neutrality does not differ among clades in a phylogenetic tree.

Specific objective 3: To examine the effects of neutral processes on selection analyses.

Several neutral genetic phenomena can potentially affect the results of the analyses and cause a significant pattern of selection that may not be real, including genetic hitchhiking, codon usage bias, mutation saturation and relaxed constraint (Kreitman, 2000).

Genetic hitchhiking occurs when a gene is subject to natural selection and its linked genes are transmitted to the progeny (Barton, 2000). This can make results ambiguous, as it is difficult to determine if the linked gene was subject to natural selection. It can be a particular problem for the prokaryotes, as all of their genes are linked on a single molecule of circular DNA.

Mycoplasma species have circular chromosomes, so in theory, all the genes of a particular species are linked and are expected to be transmitted to the next progeny together. However, to detect genetic hitchhiking, the selection pattern has to be compared between the genes that are located next to each other on the chromosome. The objective of studying the effects of genetic hitchhiking on natural selection was tested by the following null hypothesis:

Genetic hitchhiking may confound detection of patterns of selection, as selection of the genes linked to a target gene might result in a false positive selection of the target gene itself.

Codon usage bias refers to the unequal usage of codons encoding the same amino acid (Powell and Moriyama, 1997). Previously, synonymous mutations were thought to be neutral in respect of natural selection (Juke and King, 1969), but in that case, considering mutations to be random, all codons coding for the same amino acid should be equally present in a large sample of genes (Powell and Moriyama, 1997). So, unequal usage of synonymous codons must be a result of selection, as mutation bias is not a cause of codon usage bias, as demonstrated by Powell and Moriyama's (1997) studies on *Drosophila*. Their study indicated that the genes showing high codon usage bias exhibit a lower rate of synonymous substitution than the genes that exhibit lower codon usage bias. The members of the order Mycoplasmatales exhibit some unique features associated with their genetic code (for example, the stop codon UGA is used to code for tryptophan in *Mycoplasma* [Razin, 1998]). It is important to see whether preferred usage of synonymous codons affect the pattern of selection in these organisms.

The effect of codon usage bias on selection analyses was tested by the following null hypothesis:

Codon usage bias can obscure the results of selection analyses and may produce an unrealistic pattern of natural selection.

Mutation saturation occurs when a particular base mutates to a different one, then mutates back to the original state (Henn et al., 2009) - for example, an A mutating to a T, then back to A. In such a case, it is difficult to determine if a particular base, though

being in the same form as the original base, underwent two successive mutations in its evolutionary history. The members of the order Mycoplasmatales have high rates of mutation (Razin, 1998), which is one of the primary driving forces behind their evolution.

The effect of mutation saturation on selection studies can be tested by the following null hypothesis:

Mutation saturation can significantly increase the dN/dS ratios or neutrality index values and thus can affect the pattern of selection acting on the genes in the order Mycoplasmatales because of the high mutation rates of the organisms .

Relaxed constraint occurs when a gene is subjected to reduced selection because of a change in the environment (Eng et al., 2010) or gene duplication (Wagner, 2002). The nonsynonymous sites of a gene may show a greater amount of polymorphism than would be expected relative to a gene that is under strong selection. Relaxed constraint, rather than positive selection is known to cause regulatory genes (e.g., plant transcription factors) to evolve at elevated rates (Streisfeld and Rausher, 2007).

The members of the order Mycoplasmatales are obligate parasites, and thus their hosts can be considered to be their environments. A host shift may represent a change in environment. It was important to study whether the genes were subjected to reduced selection across the clades of the tree, which represented the events of host shifts.

The effect of relaxed constraint on natural selection can be tested by the following null hypothesis:

Relaxed constraint may affect the pattern of selection which a gene may be subjected to across the clades of the *Mycoplasma* phylogeny.

In summary, this study aims at testing the pattern of natural selection and speciation occurring in the protein coding genes of the *Mycoplasmas*, based on gene function and phylogenetic relationships. The approach has the potential to identify genes, gene complexes or even pathways that may be involved directly or indirectly in speciation.

CHAPTER II

MATERIALS AND METHOD

Twelve species belonging to the order Mycoplasmatales, whose genomes have been sequenced as of April 15, 2008 were used in this study. This represented roughly 7% of the order Mycoplasmatales, since there are about 100 known species under this order (Fadiel et al., 2007). First, their phylogenetic relationships were deduced. Next, the sequences for genes that are common to all these species were obtained, and these genes were classified based on their specific functions. Finally, these datasets were analyzed using two different statistical tests to study the effect of natural selection on the evolution of protein-coding genes in these species.

The various steps followed in the current study are discussed in detail in the following sections:

2.1 Phylogeny of the Mycoplasmatales

A database containing the 16S ribosomal DNA sequences (obtained from the National Centre for Biotechnology Information [NCBI] database) from each of the 12 species was constructed and was used to develop a Bayessian phylogenetic tree. The 16S ribosomal DNA sequences from *Lactobacillus acidophilus* and *Escherichia coli* were used as outgroups. The program MrBayes version 3.1.2 (Huelsenbeck and Ronquist,

2001) was used to construct the tree. The sequences were aligned using the program ClustalX (Larkin et al., 2007) General Time Reversible (GTR) model with gamma-distributed rates (GTR + γ) was determined to be the best fit model for the run by Modeltest (version 3.7) (Posada and Crandall, 1998). The approach of likelihood ratio test was implemented for the purpose of model selection. A total of 100000 generations of Markov Chain Monte Carlo (MCMC) simulations were ran, with the first 2500 generations ignored as burnins, and the consensus tree was selected for further analyses. The Bayesian tree was found to converge after the run, indicated by a final standard deviation of split frequency of 0.006444.

Figure 2.1 shows the phylogeny that was developed following the above-mentioned approach. The phylogeny identified five clades (labeled A, B, C, D and E). Only four clades (A – D) had more than one species on either side and were used for further analyses. The Bayesian support values (Huelsenbeck and Ronquist, 2001) for the various nodes/clades (the posterior probability of observing the nodes/clades in the consensus tree), as well as the specific hosts for the different *Mycoplasma* species are also mentioned. The tree was not drawn to scale, as the variation in the evolutionary rate was not studied.

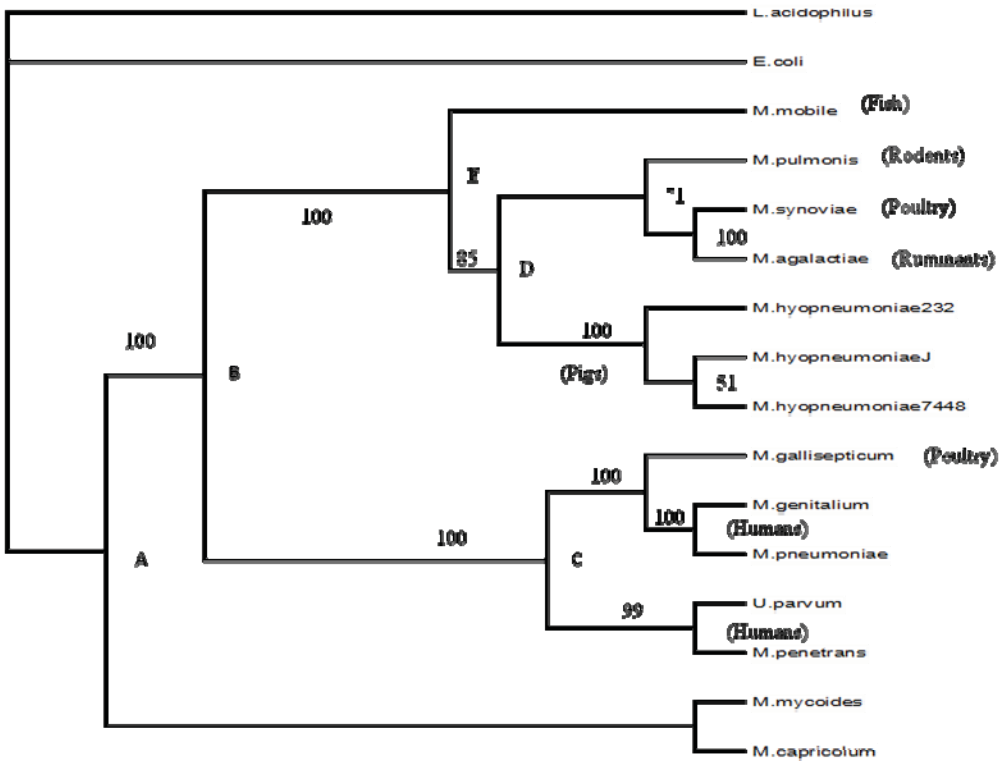


Figure 2.1: The Bayesian phylogenetic consensus tree constructed from 16S rDNA sequences showing the relationship of the 12 species belonging to the order Mycoplasmatales.The clades analyzed for selection studies, the Bayesian support values for the different nodes as well as specific hosts for different *Mycoplasma* species are shown.

2.2 Genomic Database

A database was constructed that contained the sequences of the 221 protein-coding genes common to each of the different *Mycoplasma* species under study.. These sequences were obtained from the NCBI database (http://www.ncbi.nlm.nih.gov/) (Appendix A). Each file contained information about species name, protein name, locus, gene I.D, locus tag and protein accession number.

The genes that had been annotated were identified by the locus. Orthologs for a particular gene were identified from the Kyoto Encyclopedia of Genes and Genomics

(KEGG) (http://www.genome.jp/kegg/) database. Finally, Clusters of Orthologus Groups (COG) (NCBI database; Tatusov et al., 2000) were used to sort these genes into the following three basic functional groups plus one poorly characterized group (according to the COG database), namely

1. Information storage and processing genes (with 3 subdivisions).
2. Cellular processes genes (with 6 subdivisions).
3. Metabolism genes (with 8 subdivisions).
4. Poorly characterized genes (with 2 subdivisions).

A few genes were found to be included in more than one functional groups. Such a gene was assigned to a particular functional group based on its biological functions that was documented in the NCBI database. For example, if a gene was found to be included in both the cellular processes and metabolism categories, its biological function was studied, and if it was found to be part of a biochemical pathway, it was assigned to the metabolism category. On the other hand, if the gene was found to be associated with controlling the physiological functions performed by the cell (e.g., cell division, signal transduction etc.), it was included in the cellular processes category.

The numbers of genes belonging to the different functional categories as well as the subgroups are shown in the Table 2.1.

Table 2.1: Number of genes belonging to the different functional categories as well as subgroups in which each category may be partitioned (Appendix A)

Functional Category	Subgroups	# Genes	Total
Information Processing and Storage Genes	Replication, recombination and repair	27	123
	Transcription	11	
	Translation	85	
Cellular Processes Genes	Post translational modification, protein turnover, chaperons	10	26
	Intracellular trafficking and secretion	5	
	Cell wall/membrane biogenesis	5	
	Inorganic ion transport and metabolism	3	
	Cell cycle control, mitosis and meiosis	2	
	Defense mechanisms	1	
Metabolism Genes	Carbohydrate transport and metabolism	10	51
	Amino acid transport and metabolism	12	
	Lipid transport and metabolism	3	
	Nucleotide transport and metabolism	12	
	Coenzyme transport and metabolism	2	
	Energy production and conversion	12	
Poorly Characterized Genes	General function production only	20	21
	Function unknown	1	

2.3 Sequence Alignment

Sequence alignment was performed with the software program package Data Analysis of Molecular Biology and Evolution (DAMBE) (Xia, 2000). First, the coding regions of each gene were translated into amino acid sequences. Multiple sequence alignments were conducted using ClustalW, with gap open and extension penalties of 10

and 0.1, respectively, and with BLOSUM as the amino acid substitution matrix. Next, the original nucleotide sequences were realigned against the aligned original amino acid sequences.

2.4 Data Analyses

The clades identified from the phylogenetic tree were analyzed using two different statistical approaches, namely, the codon substitution models test (Yang, 1997) and the McDonald-Kreitman test **(**McDonald and Kreitman, 1991).

2.4.1 Codon Substitution Models Test

The Codeml program of the software program package Phylogenetic Analysis Using Maximum Likelihood (PAML) (version 4.0) (Yang, 1997) was used to conduct these analyses. The branch models (the models that analyze the dN/dS (ω) ratios along the individual branches, nodes or clades) were used for the tests. This analysis would determine whether the dN/dS ratios vary among the different *Mycoplasma* species and lineages in the Mycoplasmatales phylogenetic tree.

First, all genes composing the original dataset were analyzed by comparing the free- and one-ratio models. The free-ratio model assumes different dN/dS ratios for the different branches of the phylogenetic tree, whereas the one-ratio model assumes a single ω value for the entire tree. These models were used to determine whether the dN/dS ratios were different among the various *Mycoplasma* lineages for each gene. The log-likelihood values for each of these two models were computed and their differences were calculated. Next, twice the log-likelihood differences for each gene was compared to a χ^2

distribution with degrees of freedom (df) = 24; 25 branches analyzed under the free ratio model plus one under the one ratio model minus two. The α value for these tests was determined to be 0.000056 (the same α value used for the McDonald-Kreitman test (described later), as it is a more stringent one). Genes that showed significant selection were tested further to identify where in the phylogeny selection occurred.

To determine whether the dN/dS ratios for the different clades under study are different from each other as well as from the background, a pair of branch models were used (a two-ratio (model D) and a three-ratio model (model E), respectively). The two-ratio model assumed an equal dN/dS ratio for the two clades that are being compared, while the ω value for the rest of the tree (the 'background') is assumed to be free. The three-ratio model, on the other hand, assumed different ω values for the two clades as well as for the background. These two models were applied only to the genes that showed significant selection from the free and one ratio models analyses (genes that failed to show significant selection in the previous analysis were assumed to be neutral for all subsequent analyses). All four nodes of the phylogenetic tree were analyzed this way. The log-likelihood values for these models were computed and twice the log-likelihood differences were compared to a χ^2 distribution with $\alpha = 0.05$ and df = 3 (three clades tested under the model E plus two under the model D minus 2). Bonferroni corrections were not applied because each of these genes had already been identified as showing selection in the previous broader and more conservative analysis. The branch models analyses were performed on a Novell SLES computer with 2 x Intel Xeon E5430, 2.66 GHz Quad Core and 16 GB RAM (courtesy, Dr. Susan M. Bridges and Mr. Jonathan W. Harper, Department of Computer Science and Engineering, Mississippi State University).

2.4.2 McDonald-Kreitman test

The software program package DNA Sequence Polymorphism (DnaSP) (version 4.20.2) (Rozas et al., 2003) was used to compute the neutrality indices (NIs) (Rand and Kann, 1996) for each gene for each pair of clades nested under the main clades under study. Two-tailed Fisher's exact tests were conducted to assess the significances of the computed NIs, with an α value of 0.000056 (after applying Bonferroni correction to reduce Type 1 error (Rice, 1989), $\alpha = (0.05 \div (221 \text{ X } 4))$.

2.4.3 Analyses of the results

Two three-way contingency tables (4 functional categories X 4 clades X selection or neutrality outcomes) were constructed to evaluate the outcomes from the codon substitution models and McDonald-Kreitman tests results, respectively. Log-linear analysis (King et al., 1988) ($\alpha = 0.05$) was conducted to determine whether the selection pattern varied with gene function and/or with clades of the phylogeny.

An online analytical tool available at the 'VassarStats website for statistical computation' (http://faculty.vassar.edu/lowry/VassarStats.html) (Lowry, 2010).was used to perform the log-linear analyses in both the above mentioned cases

2.5 Detection of Neutral Genetic Processes

The approaches for detecting the various neutral genetic processes that might have affected the results from the selection studies are described in the following sections:

2.5.1 Genetic Hitchhiking

The effects of genetic hitchhiking were evaluated by counting the number of different genes adjacent to the target gene (both upstream and downstream) and by measuring the frequency of the most common linked gene for each target gene. First a list was constructed that contained the titles for the genes linked upstream and downstream adjacent to the different target gene for all the *Mycoplasma* species. The gene that was found to be linked to a specific target gene from different organisms at least once was designated as a unique gene. The linked gene that was present in the maximum number of organisms was considered to be the most frequent gene. Theoretically all the genes in a *Mycoplasma* species may be assumed to be linked on a circular bacterial chromosome, but only the genes located immediately upstream and downstream of a target gene were used in the analysis, since linkage is expected to be strongest for these genes. This is the reason why the complete linkage of all genes on the circular molecule of DNA was not studied. These data were used to conduct a series of four t-tests (O'Mahony, 1986) to assess whether the numbers of unique and most frequent genes differed significantly for genes showing selection versus neutrality (as predicted by the codon substitution models tests and McDonald-Kreitman analyses) for clades with maximum variation (clades A and B). After applying a Bonferroni correction for the reduction of type I error (erroneously rejecting the null hypothesis when it is true) (Rice, 1989), the α value was determined to be $(0.05/4) = 0.0125$. If a significantly lower number of unique genes and/or significantly higher number of most frequent genes were found to be associated with the genes showing selection, it would indicate that the genes from the different *Mycoplasma* lineages were organized in similar orientation on the chromosomes. This

may support the hypothesis that genetic hitchhiking might have played a role in transmitting these sets of linked genes to the progeny.

It needs to be mentioned that groups of genes involved in a common biological function might be linked even though they may not be adjacent to one another. But due to the time consuming process of constructing the dataset, the current study was limited to just the most adjacent genes. This possibility of the above mentioned phenomenon will be studied in the future.

2.5.2 Codon Usage Bias

Analysis for detecting codon usage bias was performed by using the codon bias index (CBI) (Morton, 1993), which detects the deviation from the equal use of synonymous codons. CBI values range from 0 to 1, when a CBI value of 0 means uniform codon usage and 1 represents maximum bias.

The CBI values for the different genes were calculated using DnaSP (Rozas et al., 2003). Since the codon usage pattern is a property of the species/variant and is not expected to change with phylogeny as the species in a clade changes, the effect of codon usage bias on selection was tested only for the largest clade (clade A), which includes all the species in the phylogeny. The dN/dS ratios for each gene were computed using the software program package Molecular Evolutionary Genetic Analysis (MEGA) version 4. (Tamura et al., 2007).

Next two separate plots were constructed, in both of which CBI values were plotted as the independent variable. The dN/dS ratios were used as the dependent variable in one plot, and the NI values in the other. The dN/dS, NI, and CBI values were

transformed by previously published methods (dN/dS and NIs by square-root and CBIs by arcsine transformation) to convert the distribution to a normal one (Bland, 1996). These transformed values were used to perform two correlation tests to analyze the interaction between dN/dS ratios and CBIs, as well as NIs with CBIs respectively. The purpose of the analyses is to test that whether CBI has any effect on increasing or decreasing the values for the dN/dS ratios and/or NIs, or in other words, whether CBI makes selection more relaxed or stringent.

2.5.3 Mutation Saturation

Mutation saturation can be detected by analyzing the frequency and distribution of the complex codons. These are a group of codons in which the synonymous and non-synonymous substitutions cannot be analyzed by the program package DnaSP, and has been defined by Rozas et al. (2003) as “those triplet of sites segregating for several codons; i.e. in highly variable regions” It will be assumed that these complex codons have accumulated so many mutation saturation events during the their evolutionary history, that it is not possible to accurately determine the course of their evolution. Mutation saturation was analyzed by plotting the number of complex codons as a function of the number of codons in the gene. All genes from each clade were analyzed separately for the detection of this effect. If significant amount of mutation saturation occurred then the genes from a particular clade was expected to be present on or close to the zone of complete saturation on the plot (the location where all the codons in a certain gene has acquired mutation saturation and thus have been converted to complex codons).

2.5.4 Relaxed Constraint

To detect relaxed constraint, NIs were plotted as the function of dN/dS ratios for all genes at the four clades of the phylogenetic tree. Then both dN/dS ratios and NI values were transformed by the square-root transformation method (Bland, 1996), and these values were used for further statistical analyses to compute the slopes for the distribution of genes for each clade on the plot. A significantly negative slope is an indication of relaxed constraint, as it refers to a situation where an increase in selection by adaptive fixation or diversifying selection (greater NI) is associated with greater constraint against variation (lower dN/dS ratios). The occurrence of relaxed constraint can be confirmed if two conditions are met. First, the slopes for the clades A and B must be negative, and second, their slopes must be less than the slopes for the clades C and D.

CHAPTER III

RESULTS

3.1 Data Analyses

3.1.1 Codon Substitution Models

The genomic DNA dataset was first analyzed by the free and one-ratio models. Of the 221 genes, 153 showed selection at some place of the phylogeny of the Mycoplasmatales. When partitioned among functional groups, the cellular processes genes possessed the greatest percentage of genes showing selection. The results are summarized in the Table 3.1.

Table 3.1: Number of genes belonging to different functional categories showing selection or neutrality when tested by the free and one-ratio models. The asterisk refers to the genes that showed selection at an value of 0.000056 (after applying Bonferroni correction (Rice, 1989)).

Functional Categories	Information Processing and Storage	Cellular Processes	Metabolism	Poorly Characterized	Total
Selection*	74	24	39	16	153
Neutrality	49	2	12	5	68

The results from the comparisons of the free and one-ratio models were analyzed by Fisher's exact test (Fisher, 1922) (= 0.05) to determine whether the distribution of genes subjected to selection or neutrality varies depending on gene functional categories. The computed p value was found to be 0.0035, which indicate that the results are significant, meaning that the distribution of genes subjected to selection and neutrality differ depending on gene function, with a larger percentage of genes showing selection under the cellular processes category.

A three-way contingency table was constructed that consisted of the number of genes, partitioned by functional categories along with selection or neutrality at various clades in response to the two-ratio (model D) and three-ratio (model E) model tests (Table 3.2).

Table 3.2: Three-way contingency table summarizing the results of two and three-ratio models tests. The number of genes in different functional categories showing selection (S) or neutrality (N) at one or more clades are shown.

	Clade A		Clade B		Clade C		Clade D		Total	
	S	N	S	N	S	N	S	N	S	N
Information Processing and Storage	60	63	26	97	0	123	19	104	105	387
Cellular Processes	20	6	5	21	0	26	0	26	25	79
Metabolism	31	20	9	42	2	49	4	47	46	158
Poorly Characterized	10	11	6	15	0	21	3	18	19	65
Total	121	100	46	175	2	219	26	195	195	689

3.1.2 McDonald–Kreitman tests

Neutrality indices (NIs) and two-tailed Fisher's exact tests (Fisher, 1922) were calculated to assess the significance of the computed NIs for all of the 221 genes. A three-way contingency table (4 functional categories X 4 clades X 2 McDonald-Kreitman test results, namely adaptive fixation (diversifying selection [S] or neutrality [N]) was constructed that summarized the number of genes from various functional categories that showed selection or neutrality at various clades in response the McDonald-Kreitman tests (Table 3.4).

Finally, log-linear analyses ($\alpha = 0.05$) (King et al., 1988) were conducted on the number of genes partitioned according to functional categories that showed selection or neutrality at different clades in response to codon substitution models and McDonald-Kreitman tests (the data shown in the Table 3.2 and 3.3). The goal was to determine how the pattern of natural selection varies depending on gene function and/or clades of the phylogenetic tree. Since these genes have already been shown to be subjected to selection at some place in the phylogeny of the Mycoplasmatales by the previous more stringent tests (at $\alpha = 0.000056$), Bonferroni correction (Rice, 1989) was not applied in this analysis. The results from the log-linear analysis are summarized in the Table 3.4. The G^2 statistic (Turnbull and Weiss, 1978) for each test, the respective degrees of freedom and computed p-values are also shown in the table.

The significant interaction between functional categories, clades and selection pattern indicates that selection type varies with both gene function and phylogeny. The analyses further suggest that the dN/dS ratios and NI values indeed differ depending on gene function, as evident by the fact that a significant interaction exists between functional categories and selection pattern as the effects of clades were removed. A greater percentage of genes showing selection was found in the cellular processes category. Significant interaction between clades and selection pattern with the effects from the functional categories removed showed that the ω and NI values for the various clades of the phylogenetic tree to be indeed different from each other as well as from the background. A larger percentage of genes showing selection was identified in clades A and B.

Table 3.3: The three way contingency table summarizing results of the McDonald-Kreitman tests The number of genes belonging to different functional categories that show selection (S) or neutrality (N) at one or more clades are shown.

	Clade A		Clade B		Clade C		Clade D		Total	
	S	N	S	N	S	N	S	N	S	N
Information Processing and Storage	30	93	40	83	12	111	7	116	89	403
Cellular Processes	11	15	13	13	7	19	0	26	31	73
Metabolism	13	38	10	41	2	49	7	44	32	172
Poorly Characterized	7	14	4	17	4	17	0	21	15	69
Total	61	160	67	154	25	196	14	207	167	717

Table 3.4: Association among functional categories, clades (phylogeny) and selection as indicated by log-linear analysis. The G^2 statistic, degrees of freedom (df) and probability (P) values for each test are shown. Associations significant at 0.05. 0.01 and 0.001 are indicated by one, two and three asterisks respectively. NS indicates a non-significant association.

Source	G^2		DF		P	
	Codon substitution Models	MK Test	Codon substitution Models	MK Test	Codon substitution Models	MK Test
Functional Categories X Clades X Selection	243.96	97.22	24	24	<0.0001***	<0.0001***
Functional Categories X Clades	0	0	9	9	1 (NS)	1 (NS)
Functional Categories X Selection	0.42	8.88	3	3	0.9361 (NS)	0.0309 *
Clades X Selection	219.52	64.90	3	3	<0.0001 ***	<0.0001 ***
Functional Categories X Clades (effects from 'selection' removed)	24.02	23.44	18	18	0.1544 (NS)	0.1742 (NS)
Functional Categories X Selection (effects from 'clades' removed)	24.44	32.32	12	12	0.0177 *	0.0012 **
Clades X Selection (effects from 'functional categories' removed)	243.54	88.34	12	12	<0.0001 ***	<0.0001 ***

In summary, both the codon substitution models and McDonald-Kreitman tests identified the cellular processes genes to exhibit more diversifying selection compared to the other functional categories. It is also evident from the tests that such genes can be

identified at only the deepest clades (clades A and B) of the *Mycoplasma* phylogeny. In other words, clades A and B show more evidence of selection.

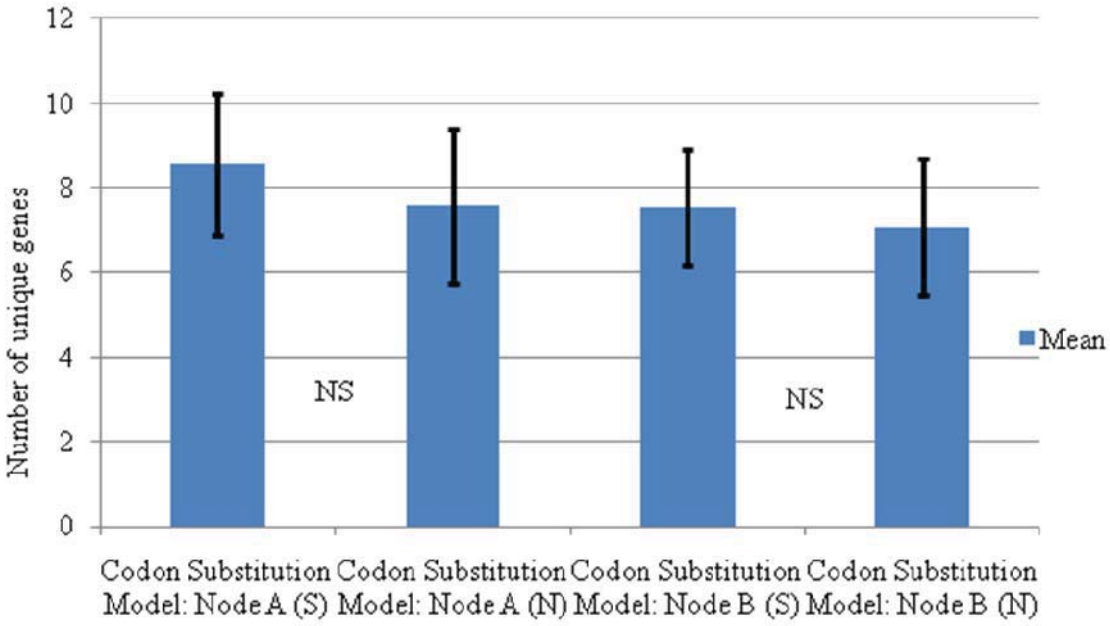
3.2 Detection of Neutral Genetic Processes

3.2.1 Genetic Hitchhiking

Figures 3.1 and 3.2 show the bar charts constructed by using the means and standard deviations of the numbers of unique genes (mean ± standard deviation) and the frequency of most frequent genes (mean ± standard deviation) linked adjacent to target genes, respectively. Each of the specific genes linked to a particular target gene in various organisms are termed as unique genes, whereas the linked genes present in maximum number of organisms are considered to be the most frequent genes. These data were analyzed using four separate t-tests (O'Mahony, 1986), $\alpha = 0.0125$ (after applying Bonferroni correction (Rice, 1989)). Results from the various t-tests (p-values) are summarized in the Table 3.5. Only the clades showing selection (A and B) were analyzed.

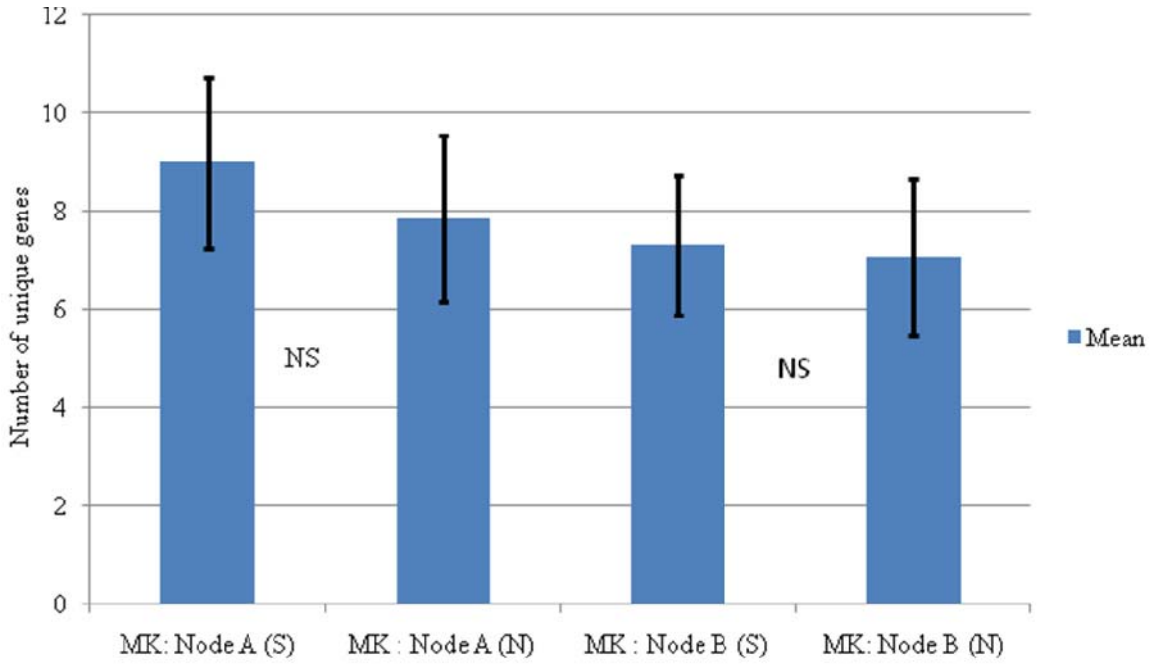
Result of only one of the four t-tests was found to be significant at $\alpha = 0.0125$ – the one conducted for the codon substitution models tests results for clade A, when the data from the frequency of most frequent genes were used for the analysis. It implies that in only this case, there is significant difference between the numbers of most frequent genes associated with the selected and neutral genes. The bar charts were used to indicate whether there was significant difference associated with the means of the number of unique genes and most frequent genes linked to the genes showing selection or neutrality

in response to the codon substitution models and McDonald-Kreitman tests. It was also evident from the bar chart from Figure 3.2a that the number of most frequent genes associated with the neutral genes is higher than those associated with the genes that are subjected to selection. If genetic hitchhiking affects selection, one would expect the number of most frequent genes linked to the genes subjected to selection to be higher in number. Since the observed significant result is associated with the neutral genes, genetic hitchhiking does not appear to have affected selection.



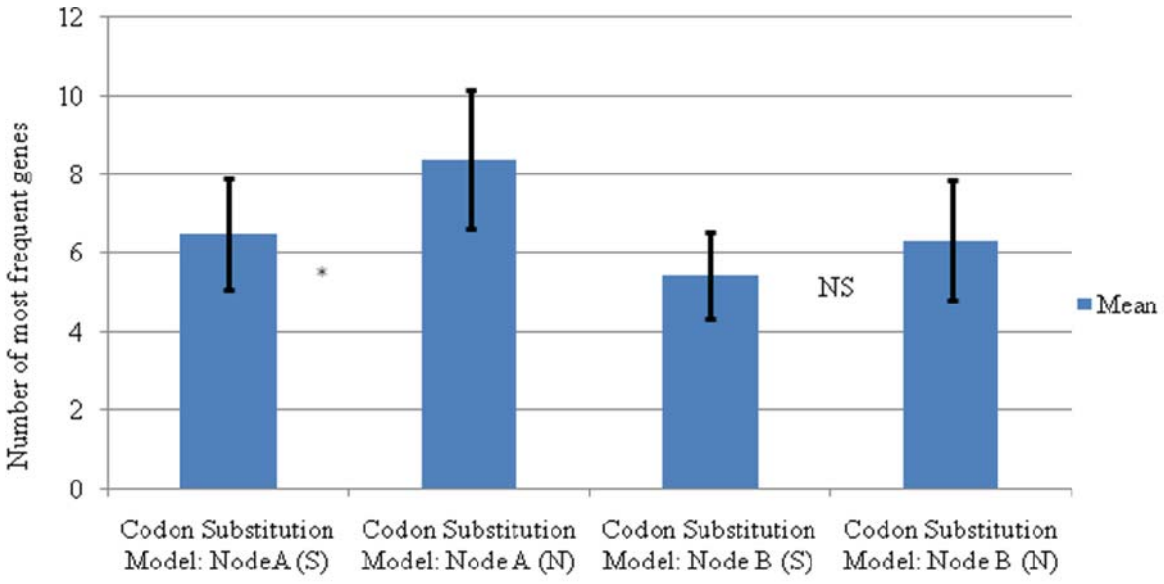
(a)

Figure 3.1: Number of unique genes (mean ± standard deviation) linked adjacent to the target genes from the clades A and B showing selection or neutrality (designated as S and N in the figure, respectively) in response to (a) the codon substitution models and (b) the McDonald-Kreitman (MK) test. NS indicates a situation where such numbers do not vary significantly with respect to the genes showing selection or neutrality from any given clade.



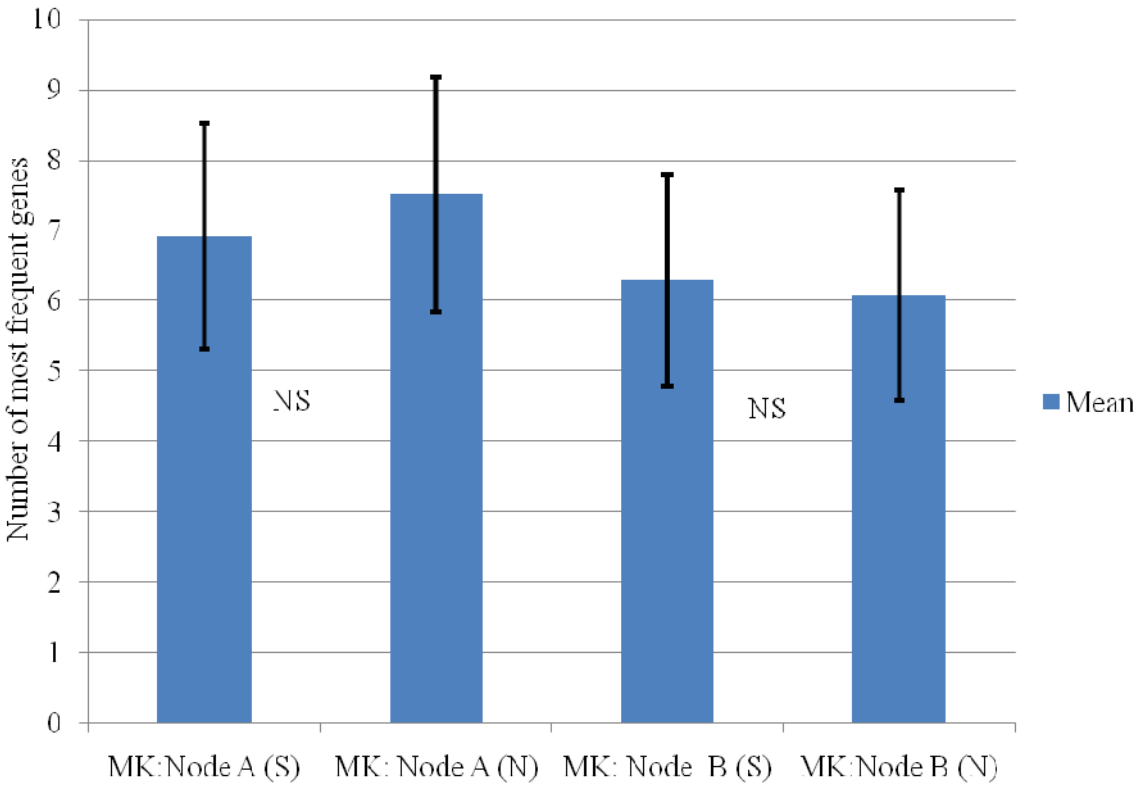
(b)

Figure 3.1 (continued)



(a)

Figure 3.2: Number of most frequent genes (mean ± standard deviation) linked adjacent to the target genes from the clades A and B showing selection or neutrality (designated as S and N in the figure, respectively) in response to (a) the codon substitution models and (b) the McDonald-Kreitman (MK) test. NS indicates a situation where such numbers do not vary significantly with respect to the genes showing selection or neutrality from any given clade. A * is used to designate the case in which the numbers do vary significantly at = 0.0125 (after applying Bonferroni correction (Rice, 1989)).



(b)

Figure 3.2 (continued)

Table 3.5: Results from the t-tests (computed p-values) (O'Mahony, 1986) conducted on the number of unique genes and most frequent geneslinked to the target genes from the clades A and B. A non-significant result (NS) indicated that the number of unique genes linked to the target genes showing selection or neutrality in response to a particular test from a given clade did not vary significantly.

Technique	Tests	Clades	p-values
Number of Unique Genes	Codon Substitution Model	A	0.038 (NS)
		B	0.301 (NS)
	McDonald-Kreitman Test	A	0.031 (NS)
		B	0.598 (NS)
Number of most frequent genes	Codon Substitution Model	A	< 0.001 ***
		B	0.023 (NS)
	McDonald-Kreitman Test	A	0.231 (NS)
		B	0.638 (NS)

3.2.2 Codon Usage Bias

Figures 3.3 and 3.4 show plots of the association of Codon Bias Index with dN/dS ratios and NIs, respectively. Since the same codon usage pattern is observed across the different clades of the phylogeny, only the deepest and largest clade (clade A) that includes all the organisms was analyzed.

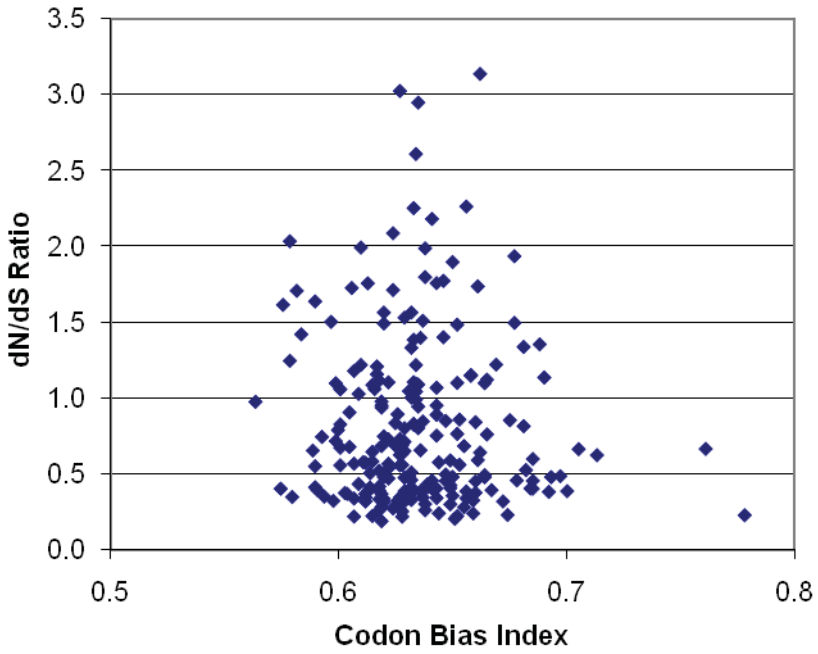


Figure 3.3: Effect of codon bias index on dN/dS ratios for clade A in the phylogeny of the Mycoplasmatales.

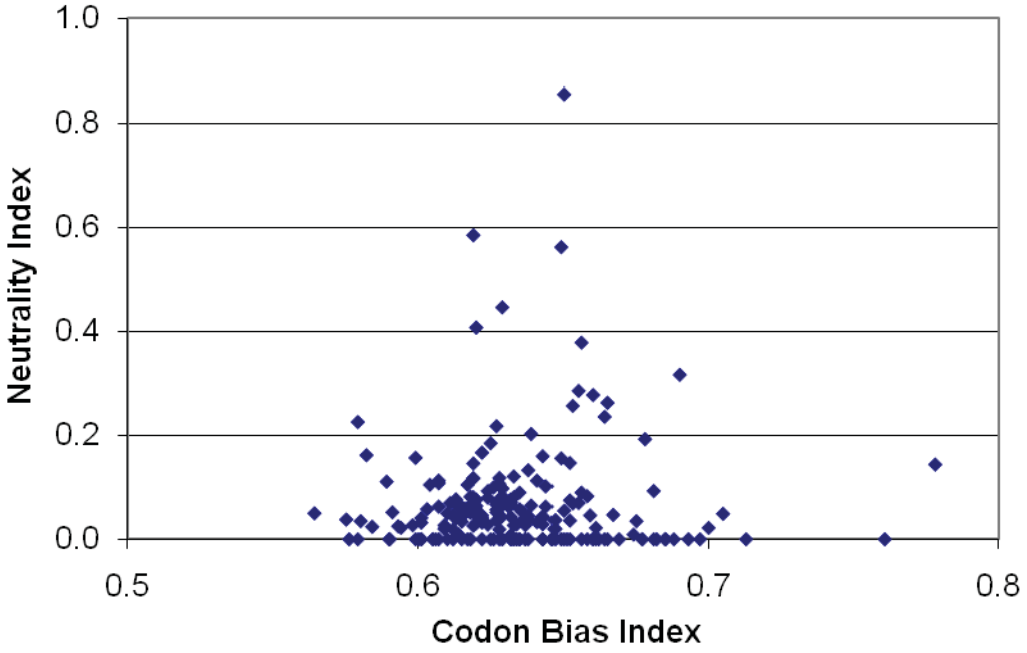


Figure 3.4: Effect of codon bias index on neutrality index for clade A in the phylogeny of the Mycoplasmatales.

The results in both figures indicate visually that there is no effect of codon bias index on either dN/dS ratios or neutrality index. A series of correlation tests were conducted to confirm these results. No significant relationship between codon bias index and dN/dS ratios or neutrality index could be determined by the correlation tests (Table 3.6).

Table 3.6: Results from the correlation tests conducted to determine the association between codon bias index (CBI) and either dN/dS ratios or neutrality index (NI). The degrees of freedom (df), correlation-coefficient (r) and computed probability (p) values for each test are shown. No significant correlation was indicated by NS.

Correlation Tests	DF	R	P-Values
CBIs and dN/dS ratios	1	0.0706	0.2957 (NS)
CBIs and NIs	1	0.0356	0.6125 (NS)

3.2.3 Mutation Saturation

Figure 3.5 shows the effect of the number of complex codons (Rozas, 2003) on the number of codons in a gene. The situation where all the codons from every gene have been subjected to mutation saturation and thus have been converted to complex codons can be termed as complete saturation. On the other hand, if no codon in a gene had been subjected to mutation saturation, then plotting the number of complex codons as a function of the number of codons for those genes will produce a straight line lying on the horizontal axis of the plot.

Because clades A+B showed the greatest evidence of selection, there is possibility that this could have been caused by mutation saturation. It is evident from the Figure 3.5 that in the course of evolution the genomes of the Mycoplasmatales have acquired some mutation saturation, though its level is a lot less compared to what would be required for complete saturation. The figure also indicates that the amount of mutation saturation is consistent with the phylogeny, since the deepest clade has acquired the highest level of saturation, followed by comparatively shallower ones.

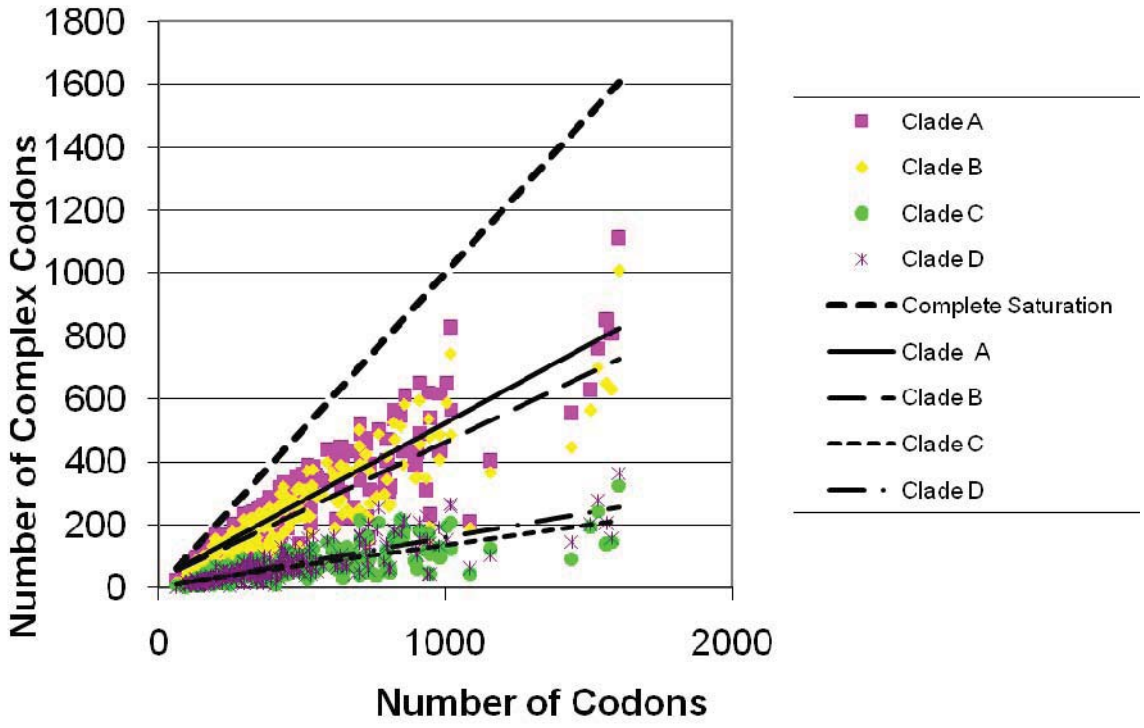


Figure 3.5: Effect of codons in a gene on the number of complex codons (Rozas et al., 2003) in the phylogeny of the Mycoplasmatales.

A series of regression tests were conducted to study the effects of the number of codons on the number of complex codons in the genes for the various clades. Since the clades A+B showed more evidence of mutation saturation compared to C+D, the r^2 values for the former clades are expected to be more than the latter ones. The p-values for all the tests are expected to be significant, as all the clades showed evidence of accumulation of mutation saturation.

The r^2 values and p-values for the different clades are shown in the Table 3.7.

Table 3.7: The r^2 and p-values calculated to study the effects of the numbers of codons on the numbers of complex codons in the genes for the different clades. The p-values that were significant at $\alpha = 0.0001$ are indicated with three asterisk marks.

Clades	r^2	p-values
A	0.7836	<0.001 ***
B	0.7670	<0.001 ***
C	0.5912	<0.001 ***
D	0.6517	<0.001 ***

The r^2 values also suggest that effects of the numbers of codons on the numbers of complex codons in a gene are considerably higher in the clades A+B compared to the clades C+D as expected. The p-values were also found to be significant, indicating accumulation of mutation saturation in every clade.

Mutation saturation has been further tested by examining the relationship between the number of codons and the number of complex codons for each claode separately, for the genes showing selection and neutrality to codon substitution models (Figure 3.6 to 3.9) and McDonald-Kreitman test (Figures 3.10 – 3.13).

The trendlines for the respective clades are indicated in each plot. If the points representing genes showing selection are found above the line (closer to the zone of complete saturation), it indicates accumulation of more mutation saturation compared to the ones below the line.

Figure 3.6 indicates that as the clade A is the deepest clade of the phylogeny, it may be predicted that over time some mutation saturation has been accumulated, though it is considerably less than complete saturation. In the figure, almost equal number of points showing selection is present below and above the trendline, which means that

about half of the genes showing selection in the clade A has accumulated mutation saturation.

In Figure 3.7 for clade B, higher number of points showing selection is located above the line compared to below indicating that higher number of genes showing selection has acquired mutation saturation compared to the ones that have not acquired it, although none of them has reached the level of complete saturation.

Only two genes show selection at the clade C. However, the entire distribution of the data points seen in the Figure 3.8 is located significantly lower than the level of complete saturation, indicating that mutation saturation has not taken place at considerable amount.

In Figure 3.9 clade D, greater number of genes showing selection are seen to ne present above the trendline. Thisindicates that significant amount of mutation saturation has not affected the genes from this clade.

Almost equal numbers of points are present above and below the trendline in Figure 3.10 for clade A, again suggesting the accumulation of greater amount of mutation saturation at the deepest clade of the phylogeny. This observation was consistent with the results from the codon substitution models.

In Figure 3.11 for clade B, higher number of points showing selection is seen to be present below the line, which suggests the absence of significant amount of saturation. The presence of a higher number of points below the line in Figure 3.12 for clade C is indicative of the absence of enough mutation saturation to affect selection.

As higher number of points is present below the trendline in Figure 3.13 for the clade D, absence of significant amount of mutation saturation can be assumed.

It is evident that the accumulation of mutation saturation is consistent with the phylogeny, as the clade A accumulated the greatest amount of mutation saturation, followed by shallower clades. Even though mutation saturation was detected in all the clades, based on the results it can be stated that no evidence was found that it causes selection events. If mutation saturation had significantly affected the selection pattern, then most, if not all of the genes showing selection would be detected above the trendlines and near or on the zone of complete saturation. Since this pattern was not observed in any of the clades, the hypothesis that mutation saturation may affect the pattern of the evolution of the protein-coding genes in the Mycoplasmatales can be rejected.

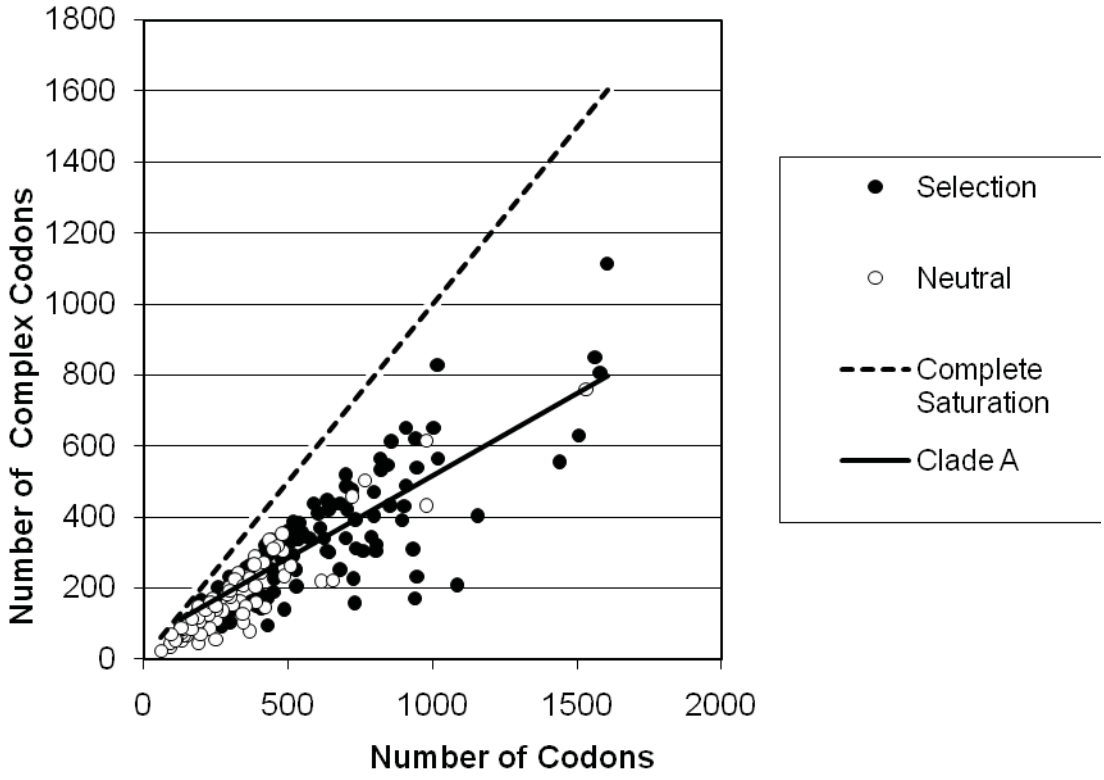


Figure 3.6: Effect of codons in the genes showing selection under the codon substitution models on the number of complex codons (Rozas et al., 2003) in the deepest clade (A) of the phylogeny of the Mycoplasmatales.

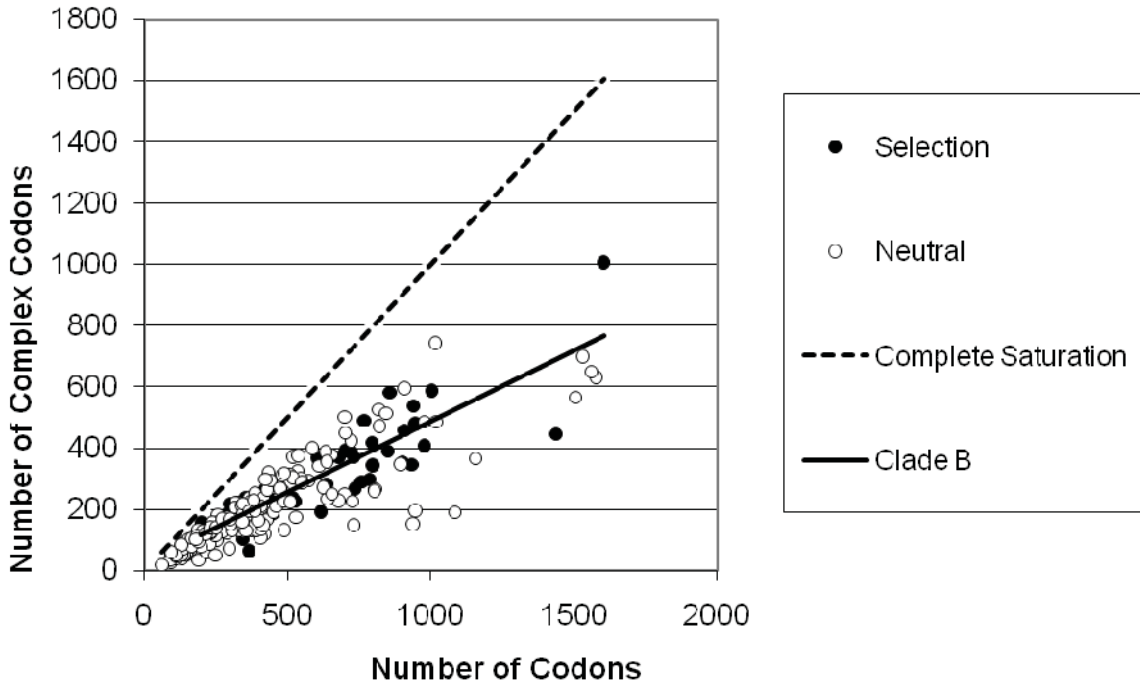


Figure 3.7: Effect of codons in the genes showing selection under the codon substitution models on the number of complex codons (Rozas et al., 2003) in the second deepest clade (B) of the phylogeny of the Mycoplasmatales.

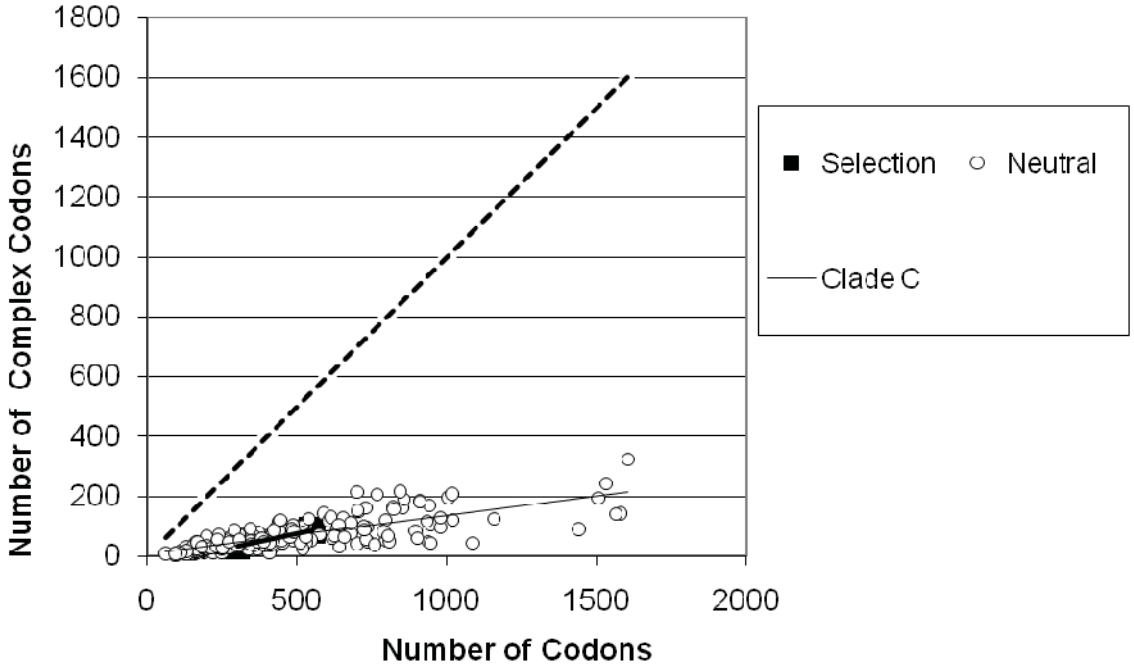


Figure 3.8: Effect of codons in the genes showing selection under the codon substitution models on the number of complex codons (Rozas et al., 2003) in one of the shallower clade (C) of the phylogeny of the Mycoplasmatales.

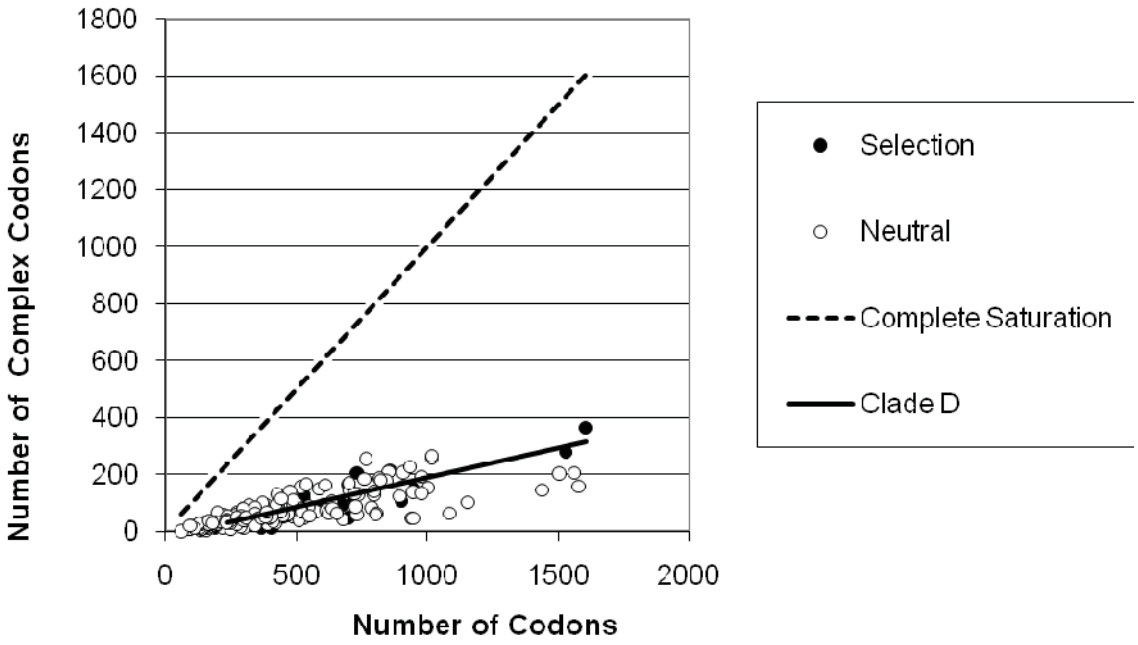


Figure 3.9: Effect of codons in the genes showing selection under the codon substitution models on the number of complex codons (Rozas et al., 2003) in one of the shallower clades (D) of the phylogeny of the Mycoplasmatales.

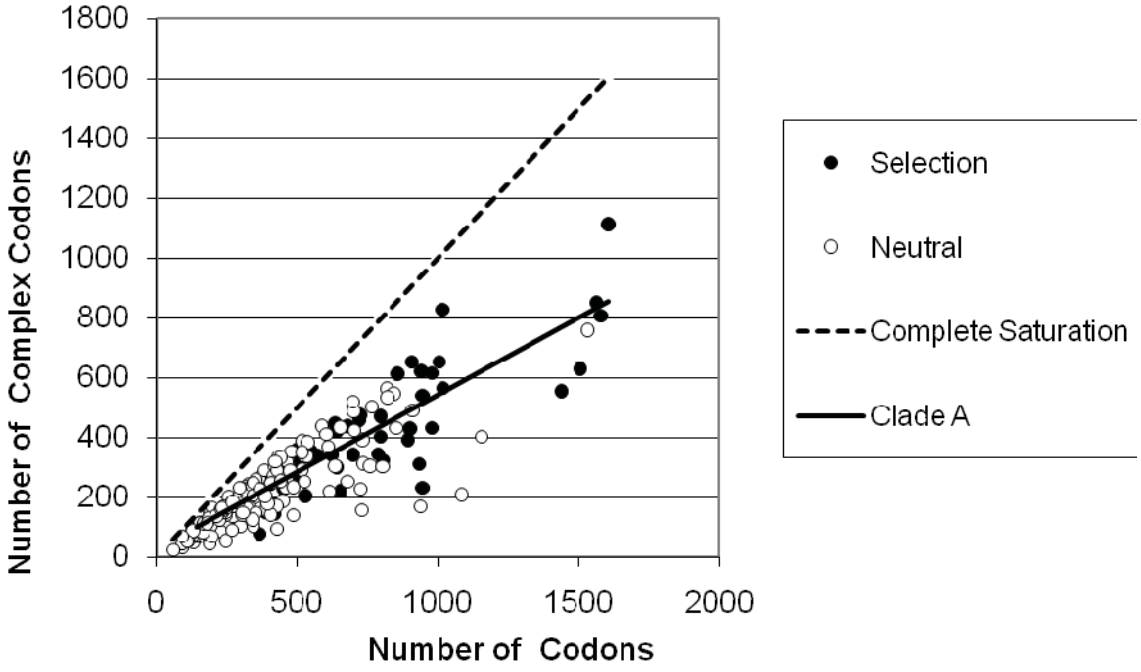


Figure 3.10: Effect of codons in the genes showing selection under the McDonald-Kreitman tests on the number of complex codons (Rozas et al., 2003) in the deepest clade (A) of the phylogeny of the Mycoplasmatales

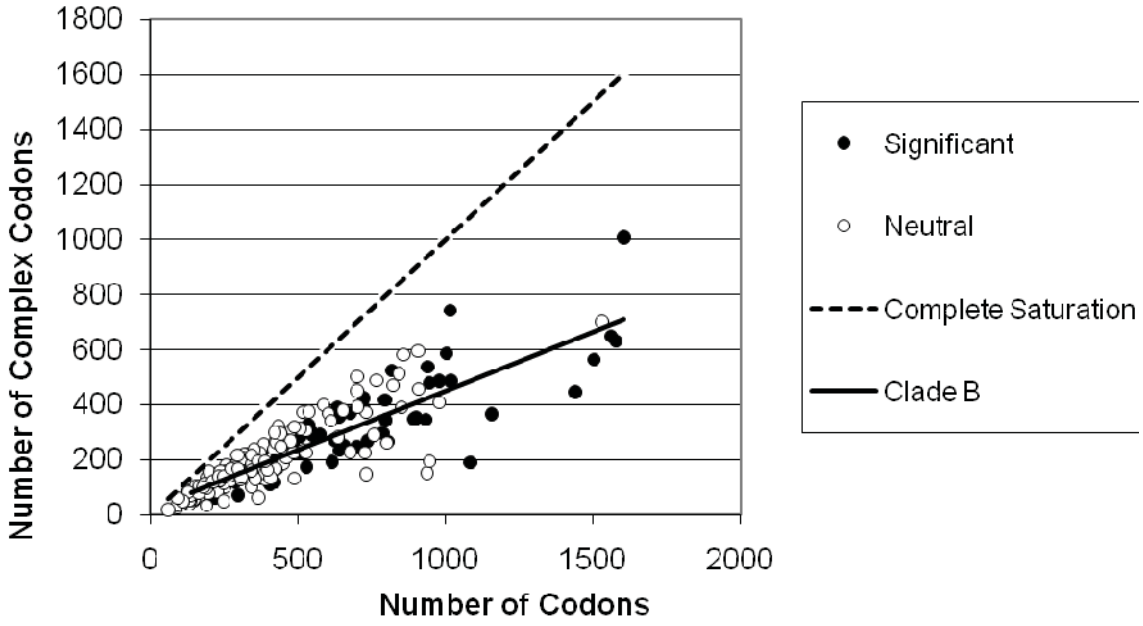


Figure 3.11: Effect of codons in the genes showing selection under the McDonald-Kreitman tests on the number of complex codons (Rozas et al., 2003) in the second deepest clade (B) of the phylogeny of the Mycoplasmatales.

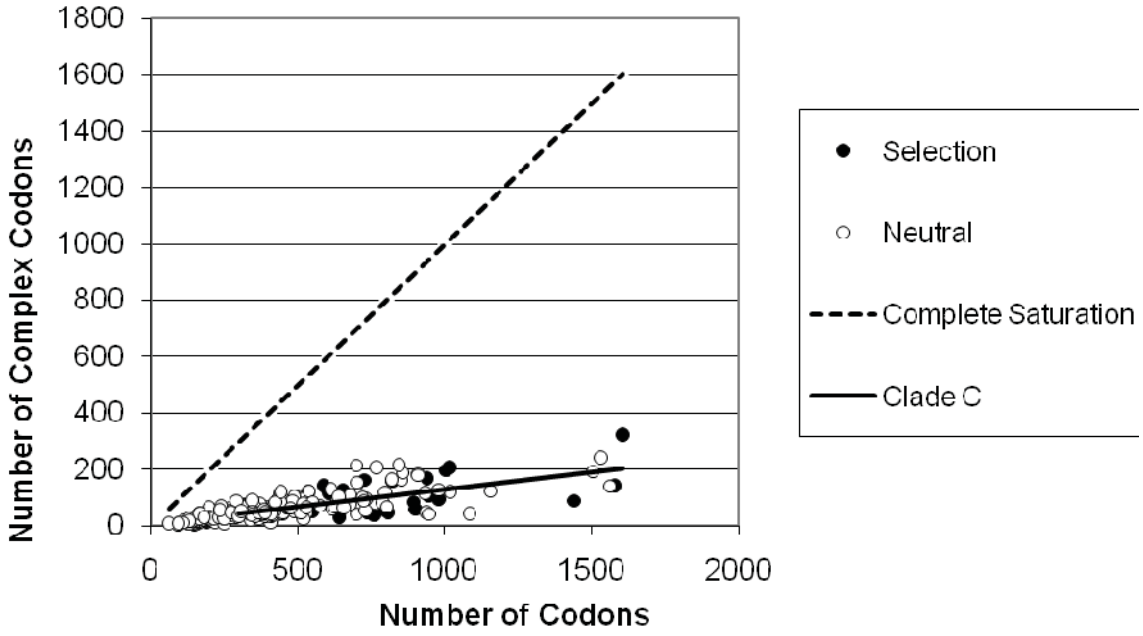


Figure 3.12: Effect of codons in the genes showing selection under the McDonald-Kreitman tests on the number of complex codons (Rozas et al., 2003) in one of the shallower clades (C) of the phylogeny of the Mycoplasmatales.

Figure 3.13: Effect of codons in the genes showing selection under the McDonald-Kreitman tests on the number of complex codons (Rozas et al., 2003) in one of the shallower clades (D) of the phylogeny of the Mycoplasmatales.

3.2.4 Relaxed Constraint

This was detected by plotting neutrality index values as functions of dN/dS ratios. Higher neutrality index associated with lower dN/dS ratio is the indication of relaxed constraint. The slopes of neutrality index values on dN/dS ratios for the distribution of genes from the different clades were computed (Table 3.8). The plot constructed to study the relationships between neutrality index and dN/dS ratios is shown in the Figure 3.14.

Table 3.8: Slopes of neutrality index on dN/dS ratios (both transformed by square-root transformation (Bland, 1996)) and their respective p-values calculated for gene distributions at each individual clades. The slopes that were significant at $\alpha = 0.05$ were denoted with a *, where as non-significant slopes were indicated by NS.

Clade	Slope	P-value
A	-0.0147	0.3522 (NS)
B	0.0565	$< 0.05^{*}$
C	0.0668	$< 0.01^{*}$
D	0.0398	0.1042 (NS)

If relaxed constraint caused genes to show significant selection, there should be an inverse relationship between neutrality index and dN/dS ratios at clades of substantial selection (clades A+B). In other words, in order for relaxed constraint to be a detectable force to affect the pattern of selection, the slopes for the clades A and B had to be negative, as well as less than those for the clades C and D. But since none of the slopes were significantly negative, the hypothesis that relaxed constraint might have influenced the pattern of selection can be rejected.

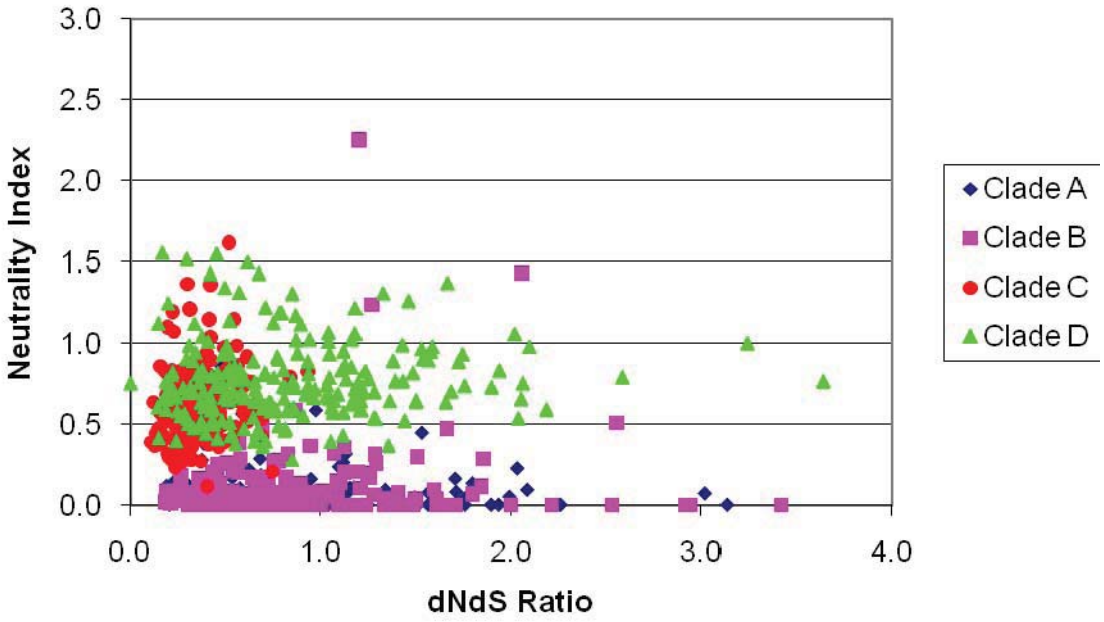


Figure 3.14: Relationships between neutrality index and dN/dS ratios for clades A – D in the phylogeny of the Mycoplasmatales.

CHAPTER IV

DISCUSSION

4.1 Phylogeny of the Mycoplasmatales

The clades about which selection was tested were found to be well-supported. The tree consists of mainly two lineages, one leading to a small clade consisting of only two species: *M. mycoides* and *M. capricolum*, both caprine pathogens. The second clade consists of rest of the species, which indicate a wide range of host specificity. This clade (clade B in the figure 2.1) is robust and subdivided into two nested clades (clades C+E). Under these two clades, some degree of consistency is maintained in terms of host specificities. For example, in the clade C, the human pathogenic species *M. genitalium* and *M. pneumonia* are clustered together. The same pattern can be observed for the human pathogens *U. parvum* and *M. penetrans*. The only fish pathogenic species that was used in the study, *M. mobile* forms an entirely separate group that is a sister taxa to clade D.

A number of features observed from the tree are consistent with the patterns found in the previously published *Mycoplasma* phylogenies. For example, the clustering of *M. genitalium*, *M. pneumonia*, *M. penetrans* and *U. parvum* was found to be in agreement with the clustering pattern seen in the tree constructed by Yoshida (2001). Similarly, *M. capricolum* and *M. mycoides* have been found to be sister taxa in other

Mycoplasma phylogenetic trees as well, for example, the tree constructed by Manso-Silvan et al. (2007). It needs to be mentioned that, though a number of *Mycoplasma* phylogenies have been constructed from 16S rDNA sequences, the current tree yields important information about the phylogenetic relationships among *Mycoplasma* lineages pathogenic to a wide range of hosts. Only about seven percent of the entire order was represented in the tree, since the analyses aimed to study the evolution of the protein-coding genes in only the species for which complete genome sequences were available, and there were only 12 species for which such information could be obtained when the study was initiated.

However, some inconsistencies in terms of host ranges can be observed in the tree, for instance, the clustering of *M. pulmonis*, *M. synoviae* and *M. agalactiae* (pathogens of rodents, poultry and ruminants, respectively). However, if a broader view is taken towards the phylogeny, a number of these inconsistencies can be resolved, at least, reduced. Most of the hosts of the *Mycoplasma* species used in this study are pathogenic to mammalian hosts, except for the fish pathogen *M. mobile* (clade D) and the poultry pathogenic species *M. gallisepticum* and *M. synoviae* (clades C and D, respectively). If host range is taken in account, the clades including the above-mentioned organisms can be considered to be paraphyletic.

The transition to various hosts (both humans and non-humans) can be the result of horizontal gene transfers. Horizontal gene transfer has been found to be associated with a number of pathogenicity-associated traits in bacteria, such as multi-drug resistance (Ochman et al., 2000). Except for *M. mobile*, all the other *Mycoplasma* species are pathogenic to hosts that, at least in a number of places around the world, live in very

close association with each other as well as with humans. Certain genes responsible for producing disease response may have been transferred directly or indirectly from a related or unrelated species of *Mycoplasma* pathogenic to a certain host species to a particular *Mycoplasma* species, resulting in its ability to infect new hosts.

It needs to be mentioned that the pathogenicity-associated genes were not used in the study (based on COG information). The genes that were used are basic housekeeping genes which are found in all bacteria species, and possibly in all life forms. It is unlikely that horizontal gene transfer has been responsible for their transmission. The instance of horizontal gene transfer was used as an example of how a few genes (pathogenicity-associated) may be involved in *Mycoplasma* speciation.

The host shift can also be a result of reductive evolution - a characteristic evolutionary pattern for *Mycoplasma* genomes. It has been mentioned earlier that like a number of other parasitic species, *Mycoplasma* genomes evolve by eliminating genes (Razin, 1998), since they obtain most of their essential nutrients and metabolites from their hosts, and thus do not need to synthesize them. So an evolutionary event in the past that may have resulted in loss of one or more pathogenicity-associated genes in a particular *Mycoplasma* species may have resulted in its shift to avian hosts from the mammalian ones – since the mammalian immune system is better developed compared to the birds (Razin, 1998).

The reason why the fish pathogen M. *mobile* forms a single group may have to do with the fact that except for fish, all the other hosts for the different *Mycoplasma* species used in this study are endothermic. The difference in metabolism may be the reason behind host specificity of *M. mobile*. In the phylogeny constructed by Jaffe et al. (2004)

M. mobile is found to be sister to *M. pulmonis*, which is to some extent congruent with the phylogeny developed in this study. But the other two species used in the current study (*M. agalactiae* and *M. synoviae*) were not used in the former phylogeny. It can be proposed that with the introduction of more *Mycoplasma* species in the phylogenetic tree, such questions can be answered more completely. However, due to the unavailability of the full genome sequences of such species they could not be incorporated in the current study.

4.2 Genomic Database

The genomic database yielded some interesting findings. The information processing and storage group contains the highest numbers of genes that are common to the different *Mycoplasma* species, possibly because this functional category is most highly conserved across the species (Razin, 1998). The *Mycoplasma* species can obtain major macromolecules from their hosts, but they need to synthesize their own protein, and any mutation in the genes associated with genetic information processing and storage (especially translation) would likely be deleterious (Razin, 1998). Besides, some of these genes are also associated with DNA damage repair mechanisms (for example, excinuclease ABC subunits A, B, and C and uracil-DNA glycosylase) and thus play critical roles in the survival of the organisms.

The genes associated with other functional categories were found to be significantly lower in number than those in the information processing and storage category. Their importance in the survival of the organism cannot be ignored, however, which is reflected by their conservation across the species. Most of the genes from the

cellular processes category were found to be associated with functions such as cell division, post translational modification and protein folding, secretion, heat shock response, and intracellular signaling.

Genes encoding a number of subunits for the F0F1 ATP synthase enzyme were detected under the metabolism functional category. Genes for many enzymes for glycolysis could also be detected. This functional category also includes many genes that are involved in transporting important biomolecules and inorganic ions from the extracellular environment to the inside of the cell, which may be essential for importing nutrients from the hosts' systems (Razin, 1998).

The poorly characterized functional category includes genes that are not well annotated and may be used as a 'control' group for the study of selection. Predicted protein functions for some of the genes belonging under this category include GTPase activities, transport etc, but these have not been confirmed.

Comparative genomics study indicates that genes associated with different functional categories in other bacteria are present in considerably higher number as compared to the different *Mycoplasma* species (Razin, 1998). One of the most important reasons behind this may be the pattern of *Mycoplasma* genome evolution. These genomes evolve by means of getting rid of genes, which is known as reductive evolution and is characteristic to a number of other parasitic organisms (Razin, 1998). The genes that are absolutely necessary for these organisms to survive are conserved in their genomes.

4.3 Data Analyses

The codon substitution models (Yang, 1997) and the McDonald-Kreitman test (McDonald and Kreitman, 1991) are essentially two different statistical techniques to analyze a more or less similar scenario. The codon substitution models compare the estimates of dN/dS ratios for two or more given branches (or clades) of a phylogenetic tree, and determine whether these ratios are different across the lineages, i.e. whether diversifying selection has operated. On the other hand, McDonald-Kreitman tests analyze a single clade, fractionate the non-synonymous and synonymous substitutions into fixed and polymorphic differences between the two clade nested within it, and determine whether the selection pattern operating at that node is diversifying or stabilizing, which the models do not. The fact that both these analyses yielded similar results (the cellular processes genes show more diversifying selection as compared to the other functional categories, and such genes can be identified at the deepest clades of the phylogeny) gives added credence to the conclusion.

There may be a number of ways that the cellular processes genes influence speciation of *Mycoplasmas.* One of the reasons why the cellular processes genes may play such a vital role might be related to the fact that some of these genes are involved in producing antigenic variation in the cell surface (Razin, 1998). This is an important biological process which allows the pathogenic microorganisms to evade the host immune responses Example of one such gene is the variable lipoprotein haemagglutinin (VIhA) (Noormohammadi, 2007) in *M. gallisepticum* and *M.* synoviae. But due to accumulation of new variation over time *Mycoplasma* species may expand their host ranges, as well as shift to entirely new host species, thus leading to speciation. Also, a

number of these genes are associated with cell division (*ftsH*, *ftsY* etc.), and a favorable substitution(s) in these genes may lead to increased rate of cell division, thus leading to higher fecundity rate and conferring an evolutionary advantage to the organism – which might ultimately lead to speciation.

Mycoplasmas have relatively simple signal transduction pathways, and so mutation in the genes encoding these proteins may enable a *Mycoplasma* species to be adapted to a new environment (i.e., a new host) than would be expected for a higher organism with much more complicated signal transduction pathways.

The cellular processes genes that showed selection under both codon substitution models and McDonald-Kreitman test (Table 4.1) may play the most vital role in *Mycoplasma* speciation (Appendix). One such gene is the GTP binding protein LepA, which belongs to the GTPase superfamily (Bourne, 1995). The members of this superfamily are involved in a number of important biological processes including the control of cell cycle and are believed to have evolved from a common ancestor (Bourne, 1995). The GTP binding protein LepA is involved in cell membrane biogenesis (Appendix). Variation in this gene may result in the alteration of the pattern of antigen arrangement on the cell surface, which may be an important contributing factor for host shift.

The proteins FtsH and FtsY (both showed evidence of selection under codon substitution models and Mcdonald-Kreitman tests) are involved in cell division (Seluanov and Bibi, 1997). These proteins may enable the *Mycoplasma* species to replicate in differential rates in various host species, depending on the immune response, metabolic rates and other biological factors associated with the hosts. FtsY is also involved in

membrane protein biogenesis (Seluanov and Bibi, 1997) which may be important for surface antigen variation.

The Sec dependent secretion pathway is ubiquitous in all biological organisms and the only secretory pathway that has been identified in the members of the order Mycopllasmatales (Stephenson, 2005). The components of this pathway (e.g., SecA, SecY etc.) show strong evidence of selection under both codon substitution models and McDonald-Kreitman tests. This pathway may enable the *Mycoplasma* species to secrete toxins specific to cause disease responses in particular host species.

4.4 Detection of Neutral Genetic Processes

It is clear from the study that the neutral genetic phenomena may not have affected the selection patterns of the *Mycoplasmas*. Even though all the *Mycoplasma* genes can be assumed to be linked on a circular chromosome, no significant indication of genetic hitchhiking could be found. Except for just one case, the results from the study indicated that there is no significant difference in number of unique or most frequent genes associated with the genes subjected to selection or neutrality. The only significant result also does not support the occurrence of genetic hitchhiking. One would expect a higher number of most frequent genes to be associated with the genes subjected to selection as an indication for genetic hitchhiking to affect selection. But in this case, more genes were found to be associated with the genes subjected to neutrality, which may be considered to be a random property of the *Mycoplasma* genomes.

Codon usage did not appear to have affected the selection of protein-coding genes in the Mycoplasmatales. It can be deduced from the study that the unequal usage of the synonymous codons was neutral in terms of selection.

The analysis showed that in the course of evolution, the *Mycoplasma* genomes have acquired some mutation saturation, and the amount in which it has accumulated is consistent with their phylogeny. Organisms belonging to the deepest clade have acquired the greatest amount of mutation saturation, followed by successively shallower clades. The distribution of mutation saturation for the genes showing selection vs neutrality indicated that though some saturation has occurred, it has not caused genes to show selection spuriously. *Mycoplasma* genomes consist of the set of genes absolutely essential in order for them to survive and replicate. Mutation in such genes would be deleterious. Besides, their obligate parasitic mode of life may have prevented them from the exposure of most of the environmental mutagens. This may be one reason why mutation saturation has not affected the genes under study. Mutation in the pathogenicity associated genes may have important contribution towards host shifts, but such genes were not included in the current study.

In order for relaxed constraint to affect selection, there has to be either a gene duplication event (Wagner, 2002) or a change in the environment (Eng et al., 2010). But most of the genes used in this study were present as single copies. Since the *Mycoplasmas* are obligate parasites, host shifts may be considered to be a change in the environment. But the analysis indicated no significant relaxation of one selection index, when the stringency of the other one is increased. This indicated that relaxed constraint was an unlikely factor to have affected selection.

Finally, none of the neutral phenomena discussed here are expected to affect the results, since the functionality of the genes have been taken under consideration in this study. These phenomena may skew the results in one direction or the other, but they are not expected to show a spurious pattern of selection belonging to a particular functional category.

Table 4.1: List of genes showing selection to both codon substitution models and McDonald-Kreitman tests

Protein Name	Subgroups
Molecular chaperon DnaK	Post translational modification, protein turnover, chaperons
ATP dependent heat shock protease La/LON	Post translational modification, protein turnover, chaperons
Cell division protein FtsH	Post translational modification, protein turnover, chaperons
Cell division protein FtsY	Intracellular trafficking and secretion
Preprotein translocase subunit SecA	Intracellular trafficking and secretion
Preprotein translocase subunit SecY	Intracellular trafficking and secretion
Inner membrane protein translocase (YidC)	Intracellular trafficking and secretion
Prolipoprotein diacylglyceryl transferase	Cell wall/membrane biogenesis
GTP binding protein LepA	Cell wall/membrane biogenesis
Chromosomal segregation protein SMC	Cell cycle control, mitosis and meiosis
tRNA-uracil-5-carboxymethylaminomethyl modification enzyme	Cell cycle control, mitosis and meiosis
Cobalt transporter ATP binding subunit	Inorganic ion transport and metabolism
Cobalt transporter ATP binding subunit	Inorganic ion transport and metabolism

CHAPTER V

CONCLUSION

The current study is an approach in which the entire genomes of the living organisms have been compared extensively for natural selection studies. Previously years (or even decades) long studies based on population genetics approaches yielded information about how a specific locus helped in the adaptation of a particular organism in a given environment. But molecular evolutionary studies, such as the one described here, are not only a lot higher throughput, but also have the potential of identifying an entire set of loci whose action (or interaction) may yield to the adaptive evolution of a whole group of organisms. At the same time, a number of future research approaches can be based on this study. Storz and Wheat (2010) argued that to achieve proper insight about the adaptation at a specific locus, information obtained from population genetics research should be coupled together with functional analyses. So it can be said that the molecular evolutionary analyses performed in this work should further be subjected to functional studies, which may yield to better understanding of adaptive evolution of an organism. The human pathogenic *Mycoplasma* species, *M. pneumonia* and *M. genitalium* can be used as examples to illustrate this. These species live in the respiratory and urogenital tracts respectively (Razin, 1998). The proteins coded by the genes used in this study may be subjected to pathway analyses to identify the differences in their pattern of

interaction. Such information can be valuable for better understanding their pathogenicity as well as tissue and organ specificity. Such analyses can be also conducted on the same set of genes from species with different host ranges, both close (*M. synoviae* and *M. gallisepticum* for instance, both poultry pathogens) as well as distantly related (for example, *U. parvum* and *M. mobile*, pathogenic to humans and fish, respectively). To broaden the understanding about *Mycoplasma* biology, such studies would be crucial. So it can be said that the current study sets the hypothesis for conducting such functional genomic analyses in the future.

The pathogenicity associated genes from a group of related *Mycoplasma* species can also be subjected to selection studies. It may lead to the identification of the gene(s) that show selection depending on the host species, indicating host specificity and disease response characteristic for the various *Mycoplasma* species. For example, analyses conducted on the pathogenicity associated genes from *M. agalactiae* (ruminant, especially bovine pathogen) and *M. capricolum* (caprine pathogen) can result in the proper identification of the genes responsible for causing the characteristic disease responses in their agriculturally important hosts. Such information may be important for drug design to treat various diseases caused by the *Mycoplasmas*, and prevent the considerable economic loss caused by them.

The genomes of the *Mycoplasmas* are composed of a set of genes that are entirely necessary in order for an organism to survive. The study described in this dissertation elucidates the basic pattern of evolution of the minimal genome. At the same time, the evolutionary pattern of the most primitive genomes can be perceived from similar approaches. It can be argued that the *Mycoplasma* genome cannot be considered to be a

model ancient genome, as all the known species of the *Mycoplasmas* are obligate parasites. But none of the genes used in this analysis are associated with causing disease responses (Seluanov and Bibi, 1997; Razin, 1998). The genes that an extremely primitive prokaryote would use in order to be metabolically active, replicate and adapt in its environment have been analyzed. Thus, this study has a potential to depict how the first footsteps of natural selection triggered the evolution of the most primordial organisms.

REFERENCES

Barton NH (2000) "Genetic Hitchhiking". Philosophical Transaction of the Royal Society of London. Series B, Biological Sciences 355: 1553 – 1562.

Bland JM (1996) Statistical Notes: Transforming Data. British Medical Journal 312:770.

Bourne HR (1995) GTPases: A Family of Molecular Switches and Clocks. Philosophical Transaction of the Royal Society of London. Series B, Biological Sciences 1329: 283-289.

Chambaud I, Hellig R, Ferns S, Samson D, Galisson F, Moszer I, Dybvig K, Wroblewski H, Viari A, Rocha EP, Blanchard A (2001) The Complete Genome Sequence of the Murine Respiratory Pathogen *Mycoplasma pulmonis*. Nucleic Acids Research 29: 2145 – 2153.

Chen SL, Hung C, Xu J, Reigstad CS, Magrin V, Sabo A, Blasiar D, Bieri T, Meyer RR, Ozersky P, Armstrong JR, Fulton RS, Latreille JP, Spieth J, Hooton TM, Mardis ER, Hultgren SJ, Gordon JI (2006) Identification of Genes Subject to Positive Selection in Uropathogenic Strains of *Escherichia coli*: A Comparative Genomics Approach. Proceedings of National. Academy of Sciences 103: 5977 – 5982.

Darwin CR (1859) On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life. London: John Murray. 1st Edition, 1st Issue, Chapter IV: pp 80-130.

Diehl WJ and Perkins JD (2004) Patterns of Natural Selection in the order Mycoplasmatales. Annual Meeting of the Society for the Study of Evolution. Colorado State University, Fort Collins.

Eng KH, Kvitek DJ, Keles S and Gasch AP (2010) Transient Genotype-By-Environment Interactions Following Environmental Shock Provide a Source of Expression Variation for Essential Genes. Genetics 184: 587 – 593.

Evans PD, Anderson JR, Vallender EJ, Choi SS and Watterson BT (2004) Reconstructing the Evolutionary History of *Microcephalin*, a Gene Controlling Human Brain Size. Human Molecular Genetics 13: 1139-1145.

Eyre-Walker A (2002) Changing Effective Population Size and the McDonald-Kreitman Test. Genetics 162: 2017-2024.

Fadiel A, Eichenbaum KD Semary EN and Epperson B (2007) Mycoplasma Genomics: Tailoring the Genome for Minimal Life Requirements through Reductive Evolution 12: 2020-2028

Fisher RA (1918) The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh 52: 399 – 433.

Fisher RA (1922) On the Interpretation of $\chi 2$ from Contingency Tables and the Calculation of P. Journal of Royal Statistical Society 85: 87-94.

Fisher RA (1947) The Analysis of Covariance Method for the Relation Between a Part and the Whole. Biometrics 3**:** 65–68.

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Doughterty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchinson CA 3rd, Venter JC (1995) The Minimal Gene Complement of *Mycoplasma genitalium.* Science 270: 397-403.

Fry AJ and Wernegreen JJ (2005) The Roles of Positive and Negative Selection in the Molecular Evolution of Insect Endosymbionts. Gene: 355: 1-10.

Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, Cassell GH (2000) The Complete Sequence of the Mucosal Pathogen *Ureaplasma urealyticum*. Nature 407 : 757 – 762.

Hein P and Arup J (1969) Grooks (Garden City, NY: Daubleday)in The River out of Eden, Richard Dawkins.pp ix.

Henn BM, Gignoux CR, Feldman MW and Mountain JL (2009) Characterizing the Time Dependancv of Human Mitochondrial DNA Mutation Rate Estimates. Molecular Biology and Evolution 26 : 217 – 230.

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R (1996) Complete sequence Analysis of the Genome of the Bacterium *Mycoplasma pneumonia.* Nucleic Acids Research 24: 4420-4449.

Holmes (1992) Convergence and Divergence Sequence Evolution in the Surface Envelope Glycoprotein of Human Immunodeficiency Virus Type 1 Within a Single Infected Patient. Proceedings of National. Academy of Sciences 89: 4835- 4839.

Huelsenbeck JP and Ronquist F (2001) MRBAYES: Bayesian Inference of Phylogenetic Tree. Bioinformatics 17: 754-755.

Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, Calvo S, Elkins T. FitzGerald MG, Hafez N, Kodira CD, Major J, Wang S, Wilkinson J, Nicol R, Nusbaum C, Birren B, Berg HC, Church GM (2004) The Complete Genome and Proteome of *Mycoplasma mobile*. Genome Research 14: 1447 – 1461.

J. Craig Ventor Institute, East Coast Campus, 9704 Medical Centre Drive, Rockville, MD 20850 USA and West Coast Campus, 10355 Science Centre Drive, San Diego, CA 92121 USA.

Jukes TH and Cantor CR (1969) Evolution in Protein Molecules. pp 21-123 in H.N.Munro ed. Mammalian protein metabolism Academic Press New York.

King JL and Jukes TH (1969) Non Darwinian Evolution. Science: 788-798.

King RJ, Plosser CI, and Rebelo ST (1988) Production, Growth and Business Cycles: 1. The Basic Neoclassical Model. Journal of Monetary Economics 31: 195 – 232.

Kenny G and Cartwright FD (1977) Effect of Urea Concentration on Growth of *Ureaplasma urealyticum* (T-Strain Mycoplasma). Journal of Bacteriology 132: 144-150.

Kimura M (1968) Evolutionary Rate at the Molecular Level. Nature 217: 624-626.

Kreitman M (2000) Methods to Detect Selection in Populations with Applications to Humans. Annual Review of Genomics and Human Genetics (2000): 539 – 339.

Kyoto Encyclopedia for Genes and Genomics (http://www.genome.jp/kegg/)

Lowry R (2010) Vassarstats Website for Statistical Computation. Vassar College, NY.

(http://faculty.vassar.edu/lowry/VassarStats.html)

Larkin MA, Blackshields C, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X Version 2.0. Bioinformatics 21. 2947 – 2948.

Manso–Silvan, L., Perrier, X., Thiaucourt, F.,(2007) Phylogeny of *Mycoplasma mycoides* Cluster Based on Analysis of Five Conserved Protein – Coding Sequences and Possible Implications for the Taxonomy of the Group. International Journal of Systematic and Evolutionary Microbiology 57: 2247 - 2258

McDonald JH and Kreitman M (1991) Adaptive Evolution at the Adh Locus in *Drosophila.* Nature 351: 652-654.

Minion PC, Lefkowitz EJ, Madsen ML, Cleary BJ, Swartzell SM, Mahairas GG (2004) The Genome Sequence of *Mycoplasma hyopneumoniae* strain 232, the Agent of Swine Mycoplasmosis. Journal of Bacteriology 186: 7123-33

National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/)

Nei M and Gojobori T (1986) Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitution. Molecular Biology and Evolution 3: 418-426.

Nielsen R and Zhang Y (1998) Likelihood Mmodels for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. Genetics 148: 929-936.

Odriozola JM, Waitzkin E, Smith TL and Bloch K (1978) Sterol Requirement of *Mycoplasma capricolum* (Growth Response/Sterol Structure Specificity/Membrane Structure) Proceedings of National. Academy of Sciences 75: 4107-4109.

O'Mahony M (1986) Sensory Evaluation of Food: Statistical Methods and Procedures. CRC Press: 487.

Papazist L, Gorton TS, Kulish G, Markham PF, Browning GF, Nguyen DK, Swartzell S, Madan A, Mahairas G, Geary SJ (2003) The Complete Genome Sequence of the Avian Pathogen *Mycoplasma gallisepticum* strain R(low). Microbiology 149: 2307 2316.

Noormohammadi, A.H. (2007) Role of phenotypic diversity in pathogenesis of avian mycoplasmosis. *Avian Pathology*; 36(6), 439 – 444.

Powell JR and Moriyama EN (1997) Evolution of Codon Usage Bias in *Drosophilla* Proceedings of National. Academy of Sciences 94: 7784-7790.

Rand DA and Kann LM (1996) Excess Amino Acid Polymorphism Among Mitochondrial Genes from *Drosophila*, Mice and Humans. Molecular Biology and Evolution 13: 735-748.

Razin S (1998) Molecular Biology and Genetics of Mycoplasmas (Mollicutes). Microbiology Review 49: 419.

Rice WR (1989) Analyzing Tables of Statistical Tests. Evolution 43: 223-225.

Rogul M, McGee ZA, Wittler RG and Falkow S (1965) Nucleic Acid Homologies of Selected Bacteria, L-Forms and Mycoplasma Species. Journal of Bacteriology 90:1200 – 1204.

Rozas J, Sanchez-DelBarrio JC, Messeguer X and Rozas R (2003) DnaSP: DNA Polymorphism Analysis by the Coalescent and Other Methods. Bioinformatics Application Notes 19: 2496-2497.

SAS Institute Inc., (2006) SAS Institute SAS 9.1.3 Language Reference Concepts 3rd Edition. Cary, NC.

Sasaki Y, Ishikawa J, Yamashita A, Oshima K, Kenri T, Furuya K, Yoshino C, Horino A, Shiba T, Sasaki T, Hattori M (2002) The Complete Genomic Sequence of *Mycoplasma penetrans*, an Intracellular Bacterial Pathogen in Humans. Nucleic Acids Research 30: 5293-300.

Seluanov A and Bibi E (1997) FtsY, The Prokaryotic Signal Recognition Particle Receptor Homologue, is Essential for Biogenesis of Membrane Proteins. Journal of Biological Chemistry 272: 2053-2055.

Sirand-Prugner P, Lartigue C, Merenda M, Jocob D, Barre A, Barbe V, Schenowitz C, Mangenot S, Couloux A, Segurens B, Daruver de A, Blanchard A, Citti C (2007) Being Pathogenic, Plastic, and Sexual While Living With a Nearly Minimal Bacterial Genome**.** Plos Genetics 3:e75.

Stell RGD and Torrie JH (1980) Principles and Procedures of Statistics: A Biometrical Approach, 2 nd Edition. McGraw-Hill, NY.

Storz JF and Wheat CW (2010) Integrated Evolutionary and Functional Approaches to Infer Adaptation at Specific Loci. Evolution 64: 2489 – 2509.

Swofford DL (2000) PAUP*. Phylogenetic Analysis Using Parsimony (*) and Other Methods) Version 4. Sinauer Associates. Sunderland, M.A.

Takeshi Y, Maeda SI, Takeshi D and Ishiko H (2002) Phylogeny-Based Rapid Identification of Mycoplasmas and Ureaplasmas from Urethritis Patients. Journal of Clinical Microbiology 40: 105-110.

Tamura K, Dudley J, Nei M and Kumar S (2007) MEGA4: Molecular Evolutionary Gene tic Analysis (MEGA) Software Version 4.0. Molecular Biology and Evolution 24: 1596-1599.

Tatusov RL, Galperin LY, Natale DA and Koonin EV (2000) The COG Database: a Tool for Genome Scale Analysis of Protein Functions and Evolution. Nucleic Acids Research 28: 33-36.

Turnbull and Weiss (1978) A likelihood Ratio Statistic for Testing Goodness of Fit with Randomly Censored Data. 34:367 – 375.

Vasconcelos AT, Ferreira HB, Bizarro CV, Carvalho MO, Pinto PM, Almeida DF, Almeida LG, Almeida R, Alves-Filho L, Assunoao EN, Azevedo VA, Bogo MR, Brigido MM, Broochi M, Burity HA, Camargo AA, Camargo SS, Carepo MS, Carraro DM, Mattos Cascardo de JC, Castro LA, Cavalcanti G, Chemalie G, Collevatti RG, Cunha CW, Dallagiovanna B, Dambros BP, Dellagostin OA, Faloao C, Fantinatti – Garboggini F, Felipe MS, Fiorentin L, Franco GR, Freitas NS, Frias D, Grangeiro TB, Grisard EC, Guimaraes CT, Hungria M, Jardim SN, Krieger MA, Laurino JP, Lima LF, Lopes MI, Loreto EL, Madeira HM, Manfio GP, Maranhao AQ, Martinkovios CT, Medeiros SR, Moreira MA, Neiva M, Ramalho-Neto CE, Nicolas MF, Oliviera SC, Paixao RF, Pedrosa

FO, Pena SD, Pereira M, Pereira-Ferrari L, Piffer I, Pinto LS, Potrich DP, Salim AC, Santos FR, Schmitt R, Schneider MP, Schrank IS, Schuck AF, Seuanez HN, Silva DW, Silva R, Silva SC, Soares CM, Souza KR, Souza RC, Staats CC, Steffens MB, Teixeira SM, Urmenyi TP, Vainstein MH, Zuccherato LW, Simpson AJ, Zaha A.(2005) Swine and Poultry Pathogens: the Complete Genome Sequences of Two Strains of *Mycoplasma hyopneumoniae* and a Strain of *Mycoplasma synoviae*. Journal of Bacteriology 187: 5568-5577.

Waites KB, Katz B and Schelonka RL (2005) Mycoplasmas and Ureaplasmas as Neonatal Pathogens. Clinical Microbial Reviews 18: 757 – 789.

Wagner A (2002) Selection and Duplication: A View From the Genome. Genome Biology 3: reviews 1012.1 – 1012.3.

Wang X, Stephanie DT and Zhang J (2004) Relaxation of Selective Constraint and Loss of Function in the Evolution of Human Bitter Taste Receptor Genes. Human Molecular Genetics 13(21): 2671-2678.

Westberg J, Persson A, Holmberg A, Goesmann A, Lundeberg J, Johansson KE, Pettersson B, Ulhen M (2004) The Genome Sequence of *Mycoplasma mycoides* subsp. *mycoides* SC Type Strain PG1T, the Causative Agent of Contagious Bovine Pleuropneumonia (CBPP). Genome Research 14: 221-227.

Xia X and Xie Z (2001) DAMBE: Software Package for Data Analysis in Molecular Biology and Evolution. Journal of Heredity 92: 371-373.

Yang Z (1997) PAML: a Program Package for Phylogenetic Analysis by Maximum Likelihood. Computer Applications in Biosciences 13: 555-556.

Yang Z and Bielawski JP (2000) Statistical Methods for Ddetecting Molecular Adaptation. Trends in Ecology and Evolution 15(12): 496-503.

APPENDIX A

LIST OF THE COMPLETE GENOMES USED IN THE STUDY

Organism	Accession Number	References
M. agalactiae PG2	CU179680	Sirand-Pugnet et al. (2007)
M. capricolum subsp *capricolul* ATCC 27343	CP000123	J. Craig Venter Institute
M. gallisepticum str R (low)	AE015450	Papazist et al., (2003)
M. genitalium G37	L43967	Fraser et al. (1995)
M. hyopneumoniae 232	AE017332	Minion et al. (2004)
M. hyopneumoniae 7448	AE017244	Vasconcelos et al. (2005)
M. hyopneumoniae J	AE017243	Vasconcelos et al. (2005
M. mobile 163K	AE017308	Jaffe et al. (2004)
M. penetrans HF-2	BA000026	Sasaki et al. (2003)
M. mycoides subsp *mycoides* PG1	BX293980	Westberg et al. (2004)
M. pneumonia M129	U00089	Himmelreich et al. (1996)
M. pulmonis UAB CTIP	AL445566	Chambaud et al. (2001)
M. synoviae 53	AE017245	Vasconcelos et al. (2005)
U. parvum serovar 3 str. ATCC 700970	AF222894	Glass et al. (2004)

APPENDIX B

LIST OF DN/DS RATIOS AND NEUTRALITY INDEX FOR THE DIFFERENT GENES THAT SHOWED SELECTION (INDICATED BY AN ASTERISK) OR NEUTRALITY AT DIFFERENT CLADES IN RESPONSE TO CODON SUBSTITUTION MODELS AND MCDONALD-KREITMAN TESTS. THE ACCESSION NUMBERS FOR THE DIFFERENT PROTEINS FROM MYCOPLASMA CAPRICOLUM (USED AS A SPECIES REPRESENTING THE DIFFERENT SPECIES USED IN THE STUDY) ARE ALSO SHOWN.

Protein Name	Clade A		Clade B		Clade C			Clade D	Accession Numbers (*M. capricolum*
	dN/dS	NI	dN/dS	NI	dN/dS	NI	dN/dS	NI	
30S ribosomal protein S2	1.401*	0.000	0.711	0.090	0.346	0.536	0.454	1.546	YP_424355
30S ribosomal protein S3	0.406	0.029*	0.377	0.041*	0.246	0.563	0.222	0.825	YP_424654
30S ribosomal protein S4	0.859	0.069	0.860	0.071	0.267	0.390	0.784	1.185	YP_424224
30S ribosomal protein S5	0.457*	0.033	0.451	0.018	0.243	0.743	0.216	0.808	YP_424643
30S ribosomal protein S6	0.684	0.286	0.603	0.000	0.445	0.599	0.522*	0.938	YP_424022
30S ribosomal protein S7	0.207	0.000	0.190	0.086	0.1538	0.571	0.189	0.666	YP_424143
30S ribosomal protein S8	0.408	0.156	0.435	0.208	0.267	0.264	0.379	0.841	YP_424646
30S ribosomal protein S9	0.388	0.022	0.481	0.000*	0.237	0.235	0.513	0.654	YP_424626
30S ribosomal protein S10	0.846	0.000	0.823	0.176	0.204	0.293	0.586	0.790	YP_424661
30S ribosomal protein S11	0.305	0.118	0.271	0.182	0.195	1.099	0.284	0.569	YP_424635
30S ribosomal protein S12	0.190	0.117	0.185	0.021*	0.110	0.392	0.148	0.418	YP_424142
30S ribosomal protein S13	0.231	0.009*	0.208	0.033	0.157	0.857	0.186	0.655	YP_424636
30S ribosomal protein S15	0.625	0.000	0.581	0.000	0.306	0.320	0.880	0.938	YP_424315
30S ribosomal protein S16	0.383	N/A	0.325	0.081	0.256	0.628	0.239	0.668	YP_424518
30S ribosomal protein S17	0.761	0.000	0.787	0.000	0.218	0.283	0.816	0.476	YP_424651
30S ribosomal protein S18	0.320	N/A	0.333	N/A	0.319	0.281	0.617	1.494	YP_424024
30S ribosomal protein S19	0.394	0.047	0.434	0.000	0.120	0.634	0.296	1.513	YP_424656
30S ribosomal protein S20	0.486	0.000	1.209	0.102	0.388	0.426	10.600	1.403	YP_424765
50S Ribosomal protein L1	0.765*	0.0750	0.633	0.0560	0.2393	0.3970	0.507	0.9860	YP_424061
50S ribosomal protein L2	0.221	0.063*	0.211	0.012*	0.156	0.387	0.189	0.763	YP_424657
50S ribosomal protein L3	0.346	0.047	0.351	0.033*	0.249	0.511	0.246	0.583	YP_424660
50S ribosomal protein L4	1.099	0.236	1.095	0.045	0.401	0.938	0.853	1.306	YP_424659
50S ribosomal protein L5	0.317	0.408	0.322	0.038*	0.196	0.808	0.332	0.480	YP_424648
50S ribosomal protein L6	0.341	0.0900	0.441	0.0580*	0.2583	0.6870	0.438	0.5020	YP_424645
50S ribosomal protein L7/L12	0.666	0.000	0.759	0.276	0.298	0.694	0.677	1.428	YP_424063
50S ribosomal protein L10	0.842	0.000	1.468	0.00*	0.407	0.458	0.813	0.463	YP_424062
50S ribosomal protein L11	0.558	0.067	0.612	0.073*	0.182	0.518	0.540	0.704	YP_424060
50S ribosomal protein L13	0.304	0.038	0.277	0.066*	0.208	0.578	0.143	0.608	YP_424627
50S ribosomal protein L14	0.326	N/A	0.306	0.000	0.144	0.376	0.458	3.185	YP_424650
50S ribosomal protein L15	0.371	0.105	0.368	0.000*	0.260	0.430	0.331	0.695	YP_424642

50S ribosomal protein L16	0.282	0.070	0.244	0.045*	0.129	0.366	0.178	0.671	YP_424653
50S ribosomal protein L17	0.404	0.000	0.377	0.045*	0.246	0.402	0.303	0.506	YP_424633
50S ribosomal protein L18	0.480*	0.000	0.506	0.000	0.227	0.506	0.910	0.556	YP_424644
50S ribosomal protein L19	0.664	0.049	0.695	0.057	0.253	0.339	0.764	0.597	YP_424515
50S ribosomal protein L20	0.459	0.193	0.593	0.000	0.178	0.540	0.522	0.845	YP_424190
50S ribosomal protein L21	0.952	0.160	0.927	0.044	0.261	0.469	0.697	0.365	YP_424387
50S ribosomal protein L22	0.347	0.027	0.346*	0.000	0.353	0.667	0.317*	0.698	YP_424655
50S ribosomal protein L23	0.642	N/A	0.48	0.000	0.520	1.617	0.329	0.892	YP_424658
50S ribosomal protein L24	1.100	0.147	0.916	0.000	0.300	0.421	1.206	0.696	YP_424649
50S ribosomal protein L27	0.330	N/A	0.340	0.064	0.180	0.828	0.424	0.589	YP_424389
50S ribosomal protein L28	0.480	0.000	0.535	0.000	0.227	1.074	0.500	1.341	YP_424461
50S ribosomal protein L29	1.354	0.000	0.535	0.000	0.405	0.115	0.755	0.789	YP_424652
50S ribosomal protein L31	0.426	0.000	0.366	0.000	0.316	0.802	0.485	0.931	YP_424131
50S ribosomal protein L34	0.229	0.144	0.197	N/A	0.176	0.431	0.167	1.552	YP_424812
alanyl-tRNA synthetase	1.003*	0.067*	1.028*	0.053*	0.456	0.512*	0.931	0.689	YP_424149
arginyl-tRNA synthetase	0.944*	0.090	0.930	0.000	0.451	0.439*	1.043	1.011	YP_424360
asparaginyl-tRNA synthetase	0.385*	0.077	0.379	0.098	0.289	0.586	0.376	1.040	YP_424771
aspartyl-tRNA synthetase	0.688*	0.185	0.687*	0.172	0.386	0.401*	0.964*	0.664	YP_424308
cysteinyl-tRNA synthetase	1.098*	0.000	1.289*	0.260	0.448	0.391	1.215	0.720	YP_424102
glutamyl-tRNA synthetase	0.568*	0.046*	0.790	0.022	0.332	0.671	0.810	0.659	YP_424122
glycyl-tRNA synthetase	0.635	0.019*	0.888	0.048*	0.259	0.681	0.533	0.444*	:YP_424478
histidyl-tRNA synthetase	0.890*	0.042	0.846	0.000	0.486	0.631	1.427*	0.989	YP_424309
isoleucyl-tRNA synthetase	0.395*	0.064*	0.366*	0.104*	0.342	0.539*	0.534*	0.776	YP_424373
leucyl-tRNA synthetase	0.506*	0.068	0.695	0.486*	0.325	0.569	1.325	1.308	YP_424624
lysyl-tRNA synthetase	0.337*	0.057*	0.320	0.000*	0.326	0.496	0.226	0.713	YP_424778
methionyl-tRNA synthetase	0.825*	0.121	0.798	0.022*	0.468	0.570	1.046	0.673	YP_424033
phenylalanyl-tRNA synthetase, alpha subunit	0.343*	0.000	0.332	0.000	0.294	0.734	0.241	0.400*	YP_424363
phenylalanyl-tRNA synthetase,beta subunit	1.060*	0.059*	1.006*	0.085	0.463	0.726	1.25*	0.781	YP_424364
prolyl-tRNA synthetase	0.325*	0.000	0.708	0.000*	0.419	0.438	0.598*	0.475*	YP_424303
seryl-tRNA synthetase	0.301*	0.000*	0.610	0.000*	0.387	0.557	0.867	0.588	YP_424781

threonyl-tRNA synthetase	0.713*	0.066	0.680	0.122	0.319	0.356*	0.657	0.718	YP_424209
tryptophanyl-tRNA synthetase	0.387*	0.036	0.369	0.000	0.396	0.583	0.448	0.739	YP_424326
tyrosyl-tRNA synthetase	0.831*	0.081	1.150*	0.000	0.422	1.361	0.936*	0.740	YP_424607
valyl-tRNA synthetase	1.028*	0.020*	1.005	0.000	0.431	0.592	0.855	0.769	YP_424244
translation initiation factor IF-2	0.6470*	0.046*	0.805	0.069*	0.348	0.521	0.824	0.582	YP_424317
translation initiation factor IF-3	1.218	0.000	0.893	0.131	0.308	0.446	1.073	0.647	YP_424192
translation elongation factor G	0.220*	0.047*	0.207	0.021	0.150	0.475	0.200*	0.690	YP_424144
translation elongation factor P	0.377	0.278	0.334	0.067*	0.227	0.560	0.299	0.695	YP_424491
translation elongation factor Ts	0.816	0.000	0.828*	0.316	0.362	0.452	0.794*	1.190	YP_424356
translation elongation factor Tu	0.223	0.036*	0.210	0.066*	0.159	0.384	0.147*	1.123	YP_424145
ribosome recycling factor	0.456	0.000	0.421	0.000	0.321	0.786	0.396	0.446	YP_424359
ribosome-binding factor A	0.600	0.000	1.418	0.000	0.398	0.738	1.232	0.829	YP_424310
modification methylase, HemK family	0.552	0.000	0.517	N/A	0.750	0.206	0.577	0.843	YP_424136
methionine aminopeptidase	1.245*	0.000	1.195	0.209	0.331	0.669	0.704	0.618	YP_424639
translation-associated GTPase	0.358*	0.055	0.349*	0.000	0.269	0.704	0.300	0.539	YP_424753
peptidyl-tRNA hydrolase	1.397*	0.053	0.404	N/A	0.349	0.899	1.131	0.865	YP_424097
tRNA-(5-methylaminomethyl-2-thiouridylate)-methyltransferase	0.569*	0.000	0.720	0.000	0.405	0.494	0.603	0.662	YP_424495
tRNA (guanine-N1)-methyltransferase	0.802*	0.000	0.928*	0.000	0.247	0.810	0.530	0.430	YP_424516
RNA methyltransferase	0.387*	0.379	0.362*	0.087	0.402	0.766	0.186	0.593	YP_424348
RNA methyltransferase	0.906	0.000	1.085	0.147	0.370	0.544	0.805	0.735	YP_424103
aspartyl/glutamyl-tRNA amidotransferase subunit A	0.990*	0.029*	0.959*	0.075*	0.352	0.576	0.972	0.717	YP_424670
aspartyl/glutamyl-tRNA amidotransferase subunit B	0.753*	0.000*	0.643	0.170	0.300	0.861	1.056*	0.394*	YP_424669
dimethyladenosine transferase	0.388	0.098*	0.384*	0.575*	0.411	0.556	3.640*	0.766	YP_424004

peptide chain release factor 1	0.674*	0.000	0.479	0.044	0.188	0.682	0.387	0.803	YP_424135
ribosomal large subunit pseudouridine synthase AYP_424375	0.340	0.113*	0.477	0.000	0.381	0.495	0.263	0.553	YP_424375
chromosomal segregation and condensation protein B	0.324*	0.000	0.416	0.060	0.362	0.762	0.000	0.754	YP_424569
transcription elongation factor GreA	1.152*	0.083	1.279	0.063	0.312	1.212	1.123	0.952	YP_424238
transcription elongation protein NusA	1.208*	0.000	1.842*	0.117	0.564	0.646	1.575*	0.949	YP_424320
DNA-directed RNA polymerase, alpha chain	1.043*	0.000*	0.857	0.073*	0.301	0.753*	0.780	0.902	YP_424634
DNA-directed RNA polymerase, beta subunit	0.406*	0.062*	0.398	0.043*	0.259	0.548	0.487	0.744*	YP_424065
DNA-directed RNA polymerase beta' subunit	0.435*	0.027	0.550	0.042*	0.218	0.640*	0.493	0.667	YP_424066
RNA polymerase sigma factor	0.749*	0.064	0.997*	0.041	0.371	0.269	0.870	1.172	YP_424476
ribonuclease III	1.614*	0.000	1.200	0.00*	0.405	0.582	1.429	0.768	YP_424465
ribonuclease R	1.725*	0.000	0.835	0.097	0.424	0.896	1.166	1.024	YP_424091
transcription antitermination protein NusG	0.522	0.000	0.555	0.000	0.609	0.555	0.409	1.012	YP_424106
heat-inducible transcription repressor HrcA	1.706	0.162	1.591	0.097	0.268	0.834	1.528	0.967	YP_424351
chromosomal replication initiator protein	2.032*	0.226	2.128	N/A	0.644	0.590	1.938	0.834	YP_424001
DNA topoisomerase IV, A subunit	1.384*	0.027*	1.397	0.048*	0.328	0.678	1.279	0.701	YP_424430
DNA topoisomerase IV, B subunit	0.324*	0.027*	0.484*	0.000*	0.237	0.796	0.665*	0.450*	YP_424431
single-strand DNA-binding protein	1.105	0.000	1.266	1.238	0.636	0.764	1.665	1.370	YP_424023
DNA gyrase, A subunit	0.574*	0.045*	0.540*	0.085*	0.272	0.532*	0.647	0.545*	YP_424037
DNA gyrase, B subunit	0.395*	0.052*	0.277*	0.030*	0.225	0.407*	0.355	0.812	YP_424038
DNA polymerase I	1.735*	0.000	2.056	1.431	0.540	0.832	2.019	1.057	YP_424604
5-3 exonuclease family protein	1.895	0.000*	1.705	0.000	0.308	1.215	2.062	0.754	YP_424044
DNA polymerase III PolC	0.583*	0.044*	0.802*	0.077*	0.455	0.628*	1.049*	0.934	YP_424323
DNA-directed DNA polymerase III subunit alpha	1.985	0.000	1.708	0.000	0.355	0.556	1.378*	0.893	YP_424605

DNA-directed DNA polymerase III subunit beta	2.945	0.000	2.941	0.000	0.692	0.538	2.097	0.979	YP_424002
DNA-directed DNA polymerase III subunit delta'	3.133	0.000	3.416	0.000	0.933	0.828	2.186	0.593	YP_424010
DNA-directed DNA polymerase III subunit gamma and tau	0.651*	0.000	0.927*	0.050	0.482	0.493	0.923	0.685	YP_424007
excinuclease ABC subunit A	0.320	0.019*	0.398	0.033*	0.245	0.466*	0.418	0.675	YP_424723
excinuclease ABC subunit B	0.375	0.058*	0.468*	0.135*	0.303	0.686	0.604*	0.778	YP_424722
excinuclease ABC subunit C	0.978*	0.586	1.853*	0.287	0.422	1.035	1.656	0.638	YP_424239
recombinase A	0.730	0.041	0.815	0.042	0.404	0.577	0.459	0.818	YP_424522
Holliday junction DNA helicase motor protein	0.655*	0.111	0.574	0.387	0.376	0.516	0.709	0.821	YP_424403
holliday junction resolvase protein	0.526	0.000	0.518	0.000	0.457	0.787	0.467	0.735	YP_424202
replicative DNA helicase	1.087*	0.000	1.071	0.318	0.417	0.555	1.188	0.686	YP_424099
uracil-DNA glycosylase	0.404*	0.113	0.344*	N/A	0.351	0.790	0.573	1.312	YP_424448
DNA topoisomerase I	1.058*	0.032*	0.970*	0.063*	0.350	0.708	1.049	0.789	YP_424741
endonuclease IV	1.179*	0.108	1.128	0.202	0.329	0.618	0.676*	0.469	YP_424056
ATP-dependent DNA helicase	1.491*	0.046	1.661	0.474	0.417	0.501	1.389	0.769	YP_424677
DNA primase	1.502	N/A	1.441*	0.000	0.598	0.744	0.523	1.140	YP_424477
DNA ligase	0.802*	0.098	1.103	0.139	0.380	0.470*	1.182	1.061	YP_424672
formamidopyrimidine-DNA glycosylase	0.361*	0.071	0.730*	0.119	0.257	0.708	0.939	0.813	YP_424603
heat shock protein GrpE	0.657*	N/A	0.677*	0.000	0.518	0.662	1.200	0.758	YP_424352
heat shock protein DnaJ	0.717*	0.157	1.363	0.000	0.382	0.487	0.337	1.122	YP_424354
molecular chaperone DnaK	0.241*	0.046*	0.219	0.070*	0.197	0.320*	0.198	1.246	YP_424353
SsrA-binding protein	1.147	0.000	0.866	0.583	0.402	0.437	1.286	0.535	YP_424090
trigger factor	0.626*	0.218	0.600	0.000	0.522	0.861	1.459	1.260	YP_424489
O-sialo glycoprotein endopeptidase	0.397	0.000	0.392	0.133	0.327	0.470	0.295	0.904	YP_424768
ATP-dependent heat shock protease	0.745*	0.025*	0.728	0.078*	0.222	0.576	0.935	0.888	YP_424488
thioredoxin reductase	1.115*	0.051	1.108	0.000	0.359	0.551	0.873	1.022	YP_424728
glycoprotease family protein	0.455	0.000	1.118	0.000	0.612	0.916	1.584	0.978	YP_424253
Cell division protein ftsH	0.352*	0.022*	0.407	0.054*	0.353	0.468*	0.364	0.761	YP_424015
Cell division protein ftsY	0.351*	0.039*	0.464	0.000*	0.196	0.553	0.548	0.854	YP_424455

preprotein translocase, SecA subunit	0.558*	0.041*	0.670*	0.051*	0.253	0.514*	0.667	0.685	YP_424042
inner membrane protein translocase component YidC	0.976*	0.050	1.252*	0.167	0.255	0.387*	1.183	0.720	YP_424810
signal recognition particle protein	0.284	0.146	0.266	0.036*	0.292	0.566	0.315	0.732	YP_424520
preprotein translocase, SecY subunit	0.404*	0.038*	0.380	0.046*	0.252	0.465*	0.368	0.636	YP_424641
signal peptidase II	0.537*	0.167	1.230	0.000	0.553	0.482	2.038	0.538	YP_424374
glucose-inhibited division protein B	1.484*	0.000	2.533*	0.000	0.608	0.765	2.053	0.658	YP_424755
prolipoprotein diacylglyceryl transferase	1.991	0.050	1.605*	0.037*	0.361	0.735	1.163	0.854	YP_424729
S-adenosyl-methyltransferase MraW	0.262	N/A	0.625	N/A	0.317	0.648	0.747	0.586	YP_424368
GTP-binding protein LepA	0.225*	0.034*	0.219	0.083*	0.240	0.562	0.208	0.743	YP_424306
chromosome segregation protein SMC	0.952*	0.026*	0.835	0.014*	0.454	0.538*	1.093	0.687	YP_424468
tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA	0.410*	0.008*	0.388	0.050*	0.317	0.558	0.343	0.944	YP_424798
ABC transporter, ATP-binding/permease protein	3.021*	0.071	2.918	0.000	0.615	0.521*	1.723	0.887	NP_073063
cobalt ABC transporter, permease protein	0.827*	0.000	0.817	0.045	0.384	0.611	0.866	0.727	YP_424630
cobalt transporter ATP-binding subunit	1.069*	0.044*	0.942	0.000*	0.2171	0.832	0.860	0.606	YP_424631
cobalt transporter ATP-binding subunit	0.836*	0.000*	0.899	0.048*	0.233	0.755	0.780	0.494	YP_424632
recombination protein U	1.563	0.000	0.878	0.000	0.545	1.148	1.120	0.432	YP_424542
metallo hydrolase	1.420*	0.024*	1.498	0.032*	0.469	0.536	1.048	0.659	YP_424241
triacylglycerol lipase	2.085*	0.093	2.124	N/A	0.558	0.694	1.564	0.897	YP_424574
DHH phosphoesterase	0.577*	0.102	0.476*	N/A	0.466	0.357*	0.944	1.023	YP_42413
ribosomal biogenesis GTPase	1.771	N/A	1.794*	0.068	0.321	0.593	1.208	0.646	YP_424514
Cof-like hydrolase	1.090	0.058	0.716	N/A	0.565	0.866	1.183	1.216	YP_424349

GTP-binding protein EngA	0.472*	0.032*	0.440	0.175	0.288	0.670	0.383	0.700	MCAP_0577
GTP-binding protein engB	0.395	0.000	0.391	0.000	0.317	0.455	0.686	0.639	YP_424232
GTP-binding protein Era	2.261	0.000	2.554	0.508	0.521	0.669	1.122	0.574	YP_424480
tRNA modification GTPase TrmE	0.731	0.036	0.686	0.151	0.409	0.529	0.729*	0.639	YP_424766
histidine triad protein	1.495	0.000	1.283	0.315	0.330	0.473	1.112	0.836	YP_424446
tRNA (guanine-N(7)-)-methyltransferase	1.530*	0.447	1.352	0.000	0.387	0.490	1.044*	1.066	YP_424711
GTP-binding protein EngC	0.371*	0.043	0.356	0.000	0.438	0.652	0.773	0.916	YP_424247
ABC transporter ATP-binding protein	0.269*	0.052*	0.368*	0.051*	0.244	0.634	0.388	0.610	YP_424482
ABC transporter, permease protein	1.636*	0.000	1.506	0.300	0.473	0.768	1.755	0.739	YP_424036
ABC transporter, permease protein	0.780	0.000*	0.582*	0.288	0.355	0.860	0.817	0.730	YP_424035
GTPase ObgE	0.692	0.060*	0.655	0.000*	0.348	0.458*	0.650	0.641	YP_424504
Hydrolase/ (metallo-beta-lactamase domain protein)	0.412*	0.009*	0.446*	0.060*	0.287	0.449*	0.503*	0.818	YP_424591
tetrapyrrole (corrin/porphyrin) methylase protein	0.498*	0.037	0.454	0.000	0.349	0.595	0.432	0.530	YP_424385
Ser/Thr protein phosphatase family protein	0.370	0.077*	0.324	0.076	0.324	0.287*	0.425	0.633	YP_424453
hypothetical protein	2.179	N/A	2.216*	0.000	0.535	0.412	1.534	0.904	YP_424730
Thioredoxin	1.095	N/A	1.440	N/A	0.495	0.395	1.684	0.705	YP_424777
F0F1 ATP synthase subunit A	1.755	0.040	1.608	0.000	0.4120	0.545	1.361	0.645	YP_424073
F0F1 ATP synthase subunit B	0.486*	N/A	1.128	0.356	0.691	0.426	1.897*	0.730	YP_424075
F0F1 ATP synthase subunit C	0.405	N/A	0.366	N/A	0.364	0.889	0.459	0.415	YP_424074
F0F1 ATP synthase subunit alpha	0.262*	0.045*	0.375	0.045*	0.252	0.360*	0.483	0.526*	YP_424077
F0F1 ATP synthase subunit beta	0.254*	0.107	0.241*	0.311	0.138	0.445	0.420	1.426	YP_424079
F0F1 ATP synthase subunit delta	2.251*	0.000	2.277*	N/A	0.599	0.596	1.744	0.933	YP_424076
F0F1 ATP synthase subunit gamma	1.046*	N/A	1.484*	0.000	0.524	0.635	1.499	0.649	YP_424078

inorganic pyrophosphatase	0.855	0.035*	0.818	0.146	0.221	1.194	0.852	0.283*	YP_424550
phosphotransacetylase	0.481*	0.857	0.459	0.250	0.445	0.785	0.695	0.600	YP_424216
acetate kinase	0.370*	0.000	0.357*	0.166	0.394	0.569	0.377*	0.496*	YP_424217
nitrogen fixation protein NifS	1.934	0.000	1.945	N/A	0.596	0.577	1.499	0.642	YP_424443
S-adenosyl-methyltransferase	0.335	0.066	0.311	N/A	0.293	0.499	0.277	0.727	YP_424368
nicotinate-nucleotide adenylyltransferase	0.476*	0.083	0.429	0.082	0.396	0.884	0.309	0.986	YP_424501
Adenylate kinase	1.331	0.071	1.278	N/A	0.419	0.767	1.353	0.371*	YP_424640
Cytidylate kinase	0.561*	0.257	0.539	0.211	0.494	0.814	0.711	0.808	YP_424547
Guanylate kinase	1.218	0.000	1.253	0.199	0.319	0.593	1.035	0.744	YP_424195
Thymidylate kinase	0.590	0.563	0.602	N/A	0.840	0.790	0.573	0.674	YP_424009
uridylate kinase	1.157	0.105	1.078	0.085	0.373	0.670	0.524	0.514	YP_424358
thymidine kinase	0.679*	0.000*	0.885	0.135	0.319	0.733	0.753	1.127	YP_424134
deoxyribose-phosphate aldolase	0.591*	0.022*	0.505	0.260	0.309	0.404	0.745	0.589	YP_424706
Adenine phosphoribosyltransferase	0.894*	0.100	0.948	0.367	0.354	0.471	0.711	1.220	YP_424470
Hypoxanthine-guanine phosphoribosyltransferase	1.221	0.000	1.200	2.250	0.321	0.650	1.033	0.677	YP_424203
uracil phosphoribosyltransferase	1.121	0.263	1.093	0.000	0.285	0.509	1.439	0.523	YP_424071
thymidine phosphorylase	0.511*	0.077	0.670	0.156	0.454	0.638	0.712	0.399*	YP_424708
purine nucleoside phosphorylase	0.850	0.020*	0.778	0.275	0.260	0.651	0.737	0.581	YP_4247
ribulose-phosphate 3-epimerase	1.757*	0.000	1.490*	0.038*	0.365	0.566	1.068	0.575	YP_424557
triosephosphate isomerase	0.453*	0.0000	0.422	0.165	0.416	0.519	0.221	0.697	YP_424701
phosphopyruvate hydratase	0.412	0.065*	0.446	0.070*	0.287	0.548	0.503	0.684	YP_424200
fructose-1,6-bisphosphate aldolase	0.274	0.028*	0.264	0.142	0.249*	0.568	0.189	0.773	YP_424127
phosphocarrier protein hpr	1.136	0.317	1.163	N/A	0.526	0.407	1.211	0.591	YP_424676
phosphoglycerate kinase	0.433*	0.203	0.565	0.000	0.368	0.648	0.519	0.653	YP_424599
pyruvate kinase	0.685*	0.118	0.882	0.046*	0.311	0.610	0.550	0.805	YP_424208
transketolase	0.462*	0.000	0.756	0.227	0.491	0.826	0.646	0.891	YP_424578
phosphoglycero mutase	0.579*	0.070*	0.524	0.044*	0.327*	0.505	0.438	0.598	YP_42470
ribose 5-phosphate isomerase	0.493*	N/A	0.478	0.000	0.355	0.525	0.554	0.382	YP_424069

spermidine/putrescine ABC transporter ATP-binding protein	0.418	0.082*	0.627	0.031*	0.300	0.506	0.655	0.763	YP_424189
spermidine/putrescine ABC transporter, permease protein	1.711*	0.080	1.402	0.080	0.411	0.807	0.806	0.835	YP_424188
spermidine/putrescine ABC transporter permease protein	1.337	0.093	1.356	0.040*	0.456	0.776	1.163	0.625	
ribose-phosphate pyrophosphokinase	0.480*	0.031	0.606	0.037*	0.356	0.419	0.654	0.399*	YP_424096
oligopeptide ABC transporter, permease protein	0.413*	0.000	0.463*	0.396	0.413	1.146	0.897*	1.117	YP_424109
oligopeptide ABC transporter, permease protein	1.797*	0.133	1.435	0.000	0.495	0.974	1.482	0.817	YP_424110
oligopeptide ABC transporter, ATP-binding protein	0.346*	0.057*	0.267	0.141	0.299	1.367	0.372	0.520	YP_424153
oligopeptide ABC transporter, ATP-binding protein	0.349*	0.035*	0.842	0.0000*	0.558	0.9870	1.277	0.541*	YP_424154
oligoendopeptidase F	0.938*	0.081	0.564	0.200	0.572	0.770	0.739*	0.635	YP_424181
XAA-Pro aminopeptidase	0.560	0.000*	0.727	0.080*	0.335	0.465	0.586	0.692	YP_424325
nitrogen fixation protein NifS	1.563*	0.076	1.333	0.000	0.438	0.652	1.137	0.780	YP_424443
leucyl aminopeptidase	0.242*	0.063*	0.458	0.154*	0.412	0.372	0.427	0.798	YP_424148
phosphatidate cytidylyltransferase	2.606*	N/A	2.868*	N/A	0.523	0.643	3.240	1.00	YP_424324
cdp-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase	1.104*	0.000	1.201*	N/A	0.408	0.591	1.060	0.609	YP_424756
1-acyl-SN-glycerol-3-phosphate acyltransferase	1.509	0.000	1.626	0.000	0.372	0.608	2.587	0.791	YP_424382