Sequence Analysis of the Bacterial Protein Elongation Factor P

by

Lynette Yee-Shee Lau

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Biology

Waterloo, Ontario, Canada, 2008

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

ABSTRACT

In 1975, Elongation Factor P (EF-P) protein was first discovered in the bacterium *Escherichia coli*. EF-P is believed to facilitate the translation of proteins by stimulating peptide bond synthesis for a number of different aminoacyl-tRNA molecules in conjunction with the 70S ribosome peptidyl transferase. Known eukaryotic homologs, eukaryotic translation initiation factor 5A (eIF-5A) of EF-P exist but with very low sequence conservation. Nevertheless, because of the high sequence similarities seen between bacterial EF-Ps and its low sequence similarity with eIF-5A, there is interest in the pharmaceutical industry of developing a novel antibacterial drug that inhibits EF-P. Of 322 completely sequenced bacterial genomes stored in GenBank, only one organism lacked an EF-P protein. Interestingly, sixty-six genomes were discovered to carry a duplicate copy of *efp*. The EF-P sequences were then used to construct a protein phylogenetic tree, which provided evidence of horizontal and vertical gene transfer as well as gene duplication. To lend support to these findings, EF-P GC content, codon usage, and nucleotide and amino acid sequences were analyzed with positive and negative controls. The adjacent 10 kb upstream and downstream regions of *efp* were also retrieved to determine if gene order is conserved in distantly related species. While gene order was not preserved in all species, two interesting trends were seen in some of the distantly related species. The EF-P gene was conserved beside Acetyl-CoA carboxylase genes, *accB* and *accC* in certain organisms. In addition, some *efp* sequences were flanked by two insertion sequence elements. Evidence of gene duplication and horizontal transfers of regions were also observed in the upstream and downstream regions of *efp*. In combination, phylogenetic, sequence analyses, and gene order conservation confirmed evidence of the complex history of the *efp* genes, which showed

incongruencies relative to the universal phylogenetic tree. To determine how *efp* is regulated, the upstream regions of *efp* were used to try to predict motifs *in silico*. While statistically significant motifs were discovered in the upstream regions of the orthologous *efp* genes, no conclusive similarities to known binding sites such as the sigma factor binding sites or regulatory protein binding sites were observed. This work may facilitate and enhance the understanding of the regulation, conservation, and role of EF-P in protein translation.

ACKNOWLEDGEMENTS

I would sincerely like to thank the following people…

My advisor Dr. Kirsten Muller for her patience, help, support, and not to mention all the pep talks she had to give me!

My committee members Dr. Bernard Glick, and Dr. Trevor Charles for all the long talks, discussions, guidance that were always associated with humour.

My friends at the University Health Network Microarray Center, Carl Virtanen, Zhibin Lu, the various co-ops, Mark Takahashi, and the gang for computational power, coffee breaks, humorous discussions, and for being my life coaches.

To Greg Hines for all his help, lunch breaks, and allowing me to 'borrow' a linux box.

To my close and supportive friends, Aline Chhun and Filomena Ng for all there listening skills and the food!

To Mike Lynch for all of his patience, teaching and guidance, you're the tree king!

To my friends and fellow lab-mates, 'Mikey' Kani, Mr. 'Robert' Young, Orietta Beltrami, and Adam Woodworth for all the C&D and Timmy's break, and chats.

A most special thank you to my Mom and my sister, Elaine for all there support and love, throughout the ages.

Thank you Michael Mandelzys for supporting me through the entire process! I could have never done it without you!

DEDICATIONS

I dedicate this thesis in memory of my Dad.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

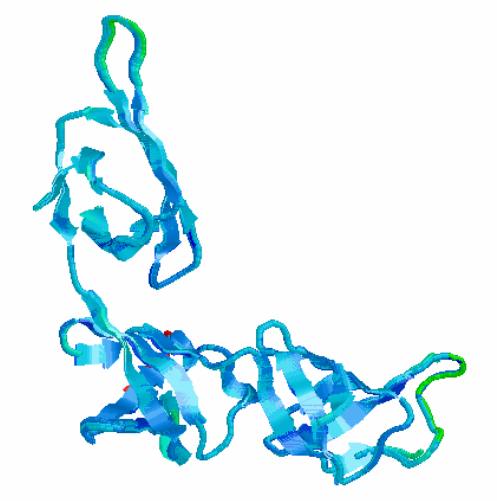
Chapter 1.0: GENERAL INTRODUCTION

1.1 The Bacterial Elongation Factor P (EF-P) Protein

One of the many auxiliary protein factors used in the elongation stage of bacterial translation is elongation factor P (EF-P, *efp*), which was first discovered in the bacterium *Escherichia coli* (Glick and Ganoza 1975a). Elongation Factor P is a soluble, acidic (pI = 4.6, 20.2 kDa) bacterial protein approximately 180 amino acid residues in length (Hanawa-Suetsugu et al. 2004). Interestingly, EF-P is structurally similar to tRNA and shares the same three-dimensional 'L' structure (Figure 1) and size as $tRNA^{Phe}$ from *Saccharomyces cerevisiae* (1EVV) (Aoki et al. 1997b; Jovine, Djordjevic, and Rhodes 2000; Ganoza, Kiel, and Aoki 2002; Hanawa-Suetsugu et al. 2004).

By mimicking the tRNA structure, EF-P may be able to attach to the tRNA-binding site(s) more readily in order for EF-P to function properly (Hanawa-Suetsugu et al. 2004). In fact, EF-P facilitates the translation of proteins by stimulating peptide bond synthesis for a number of different aminoacyl-tRNAs in conjunction with the 70S ribosome peptidyl transferase (Ganoza, Kiel, and Aoki 2002). Moreover, based on *in vitro* studies in *E. coli* without EF-P, incoming aminoacyl-tRNAs carrying small amino acid side chains have little or no ability to form a peptide bond. Hence, these small amino acid side chains may bind poorly to the ribosomal A site and are therefore weak substrates for the ribosomal peptidyl transferase (Ganoza, Kiel, and Aoki 2002). Due to the nearly ubiquitous nature of amino acids with small side chains and the suggested involvement of EF-P in the formation of the first peptide bond, it has been hypothesized that *E. coli* cells lacking EF-P are unable to survive (Glick, Chladek, and Ganoza 1979; Aoki et al. 1997a).

Figure 1. Structure of *efp* from Thermus Thermophilus HB8 determined by Hanawa-Suetsugu et al., (2004) x-ray crystallography from the RCSB protein data bank (Berman et al. 2000; Hanawa-Suetsugu et al. 2004).



In fact, Glass et al. (2006) constructed a minimal bacterium from the essential genes of *Mycoplasma genitalium* and had listed *efp* as an essential gene. This data is supported in two other studies by Mushegian and Koonin, (1996) and Gil et al. (2004), which list the *efp* gene as part of the minimal essential gene set. However, a recent study showed that using experimental and computational assessment to determine the essential genes of *E. coli* that *efp* is not needed for *E. coli* to grown on glycerol minimal medium. This is also contradictory to an earlier gene interruption study in *E. coli* in which it was demonstrated that the *efp* gene was essential for cell viability and protein synthesis (Aoki et al. 1997b; Joyce et al. 2006). Moreover, in the bacterium *Agrobacterium tumerfaciens* EF-P deletion mutants were able to survive but with a diminished growth rate (Peng et al. 2001). Hence, whether or not EF-P is essential for the cell to survive is still under debate.

The evolutionary history of the bacterial EF-P gene is not well understood and has not been examined extensively. However, Kyprides and Woese (1998) suggested that there are very distant eukaryote and archaeal homologs to EF-P, eukaryotic translation initiation factor 5A (eIF-5A) and archaeal translation initiation factor 5A (aIF-5A), respectively.

1.2 Archaeal and Eukaryotic EF-P Homologs

Translation is ubiquitous to all three domains of life and is integral in the central dogma of biology. The mechanism used in the initiation of translation is different between Eukaryote and Bacteria, while archaeal translation mechanisms are similar to those in Bacteria. Consequently, some of the orthologous bacterial and eukaryotic translation initiation factors have very low sequence identity and thus low functional similarity. Bacterial translation consists of only three initiation factors (IF-1, IF-2, and IF-3) while eukaryotes need numerous proteins such as eIF1, eIF1A, eIF2A, eIF2, eIF2B (or GEF), eIF3, eIF4A, eIF4B, eIF4E, eIF4G, eIF5, and eIF6

etc., which must work together in a complex process, to initiate the translation process (Kyrpides and Woese 1998; Weaver 2002). However, while initiation of translation between these domains are quite different, the translation elongation process is very similar (Weaver 2002). Sequence identity between EF-P and its homolog eIF-5A is approximately 20% (Kyrpides and Woese 1998). The archaeal homolog of EF-P, aIF-5A protein is known to have 26% - 32% overall amino acid sequence identity to eIF-5A (Kyrpides and Woese 1998). However, even with such low sequence identity, eukaryotic and archaeal translation initiation factor eIF-5A, formerly known as eIF-4D, is hypothesized to be a homolog of bacterial EF-P (Kyrpides and Woese 1998). Because of the high sequence similarity seen within bacterial EF-P and the low conservation between EF-P and eIF-5A, the pharmaceutical industry is using EF-P as a target to inhibit in Bacteria to develop a novel antibacterial drug for humans (Swaney et al. 2006).

Archaeal and eukaryotic translation IF-5A stimulate the synthesis of the initial peptide bond which is similar to the EF-P function of facilitating the synthesis of the peptide bond for most of the aminoacyl-tRNAs (Ganoza, Kiel, and Aoki 2002). In addition to initiating translation, the different isoforms of eIF-5A observed in *Arabidopsis* may be responsible for apoptosis by regulating p53 as well as cell division (Thompson et al. 2004). Eukaryotic IF-5A is also a cellular cofactor for human immunodeficiency virus type 1 (HIV) Rev and may play a role in mRNA degradation (Valentini et al. 2002; Li et al. 2004). Unfortunately, most of the research focuses on the eukaryotic initiation factor 5A rather than the archaeal homolog and hence little is known about the latter.

Eukaryotic IF-5A is also postulated to be an important, but not essential protein, for the viability of the cell (Kang and Hershey 1994). For instance, when there is no expression of eIF-5A in the cell due to a deletion, there is an approximately 25% decrease in protein synthesis (Li

et al. 2004). This result contradicts the suggested function of eIF-5A, in which it is hypothesized that eIF-5A stimulates the first peptide bond for all proteins as the absence of eIF-5A expression would seriously impede initialization of translation of all proteins causing a considerable decrease in protein synthesis [greater than the ~25% seen in Li et al. (2004) experiment] (Ganoza, Kiel, and Aoki 2002). Therefore, to account for such a low decrease in protein synthesis, some authors suggest that eIF-5A may only be responsible for stimulating the initiation of the first peptide bond for a subclass of mRNA that is required for cell growth (Kang and Hershey 1994).

1.3 Horizontal Gene Transfer

Horizontal gene transfer (HGT) is the transfer of genes from one organism to another and it has been hypothesized that this process is a major force in bacterial evolution, playing an important role in the diversification and speciation of this domain (Jain et al. 2003). Cohan (2002) estimated that 5 to 15% of genes located in the *E. coli* genome are derived from a foreign organism. The high frequency of HGT events in Bacteria is possibly caused by three (or a combination of three) different mechanisms in which Bacteria are able to acquire genomic or plasmid DNA from distantly or closely related organisms. These three mechanisms include i) transformation in which Bacteria are able to acquire genes by uptake of extracellular DNA, ii) phage-mediated transduction or iii) conjugation which can be mediated by mobile genetic elements such as insertion sequences, transposons, integrons, genomic islands (i.e. pathogenicity islands), and plasmids (Sorensen et al. 2005). Horizontal gene transfer may be advantageous for the bacterium if the transferred gene is maintained in the genome, does not disrupt any functioning genes, and confers selective advantages to the organism such as antibiotic resistance, increased virulence, or metabolism of other organic substances as an energy source (Barkay and

Smets 2005). In a recent study by Sorek et al. (2007) the genes which are transferred are only restricted by one factor which is whether or not the gene causes toxicity to the host by the increased expression of the gene.

Certain bacterial genomes are able to acquire and discard certain genes to their advantage (Goldenfeld and Woese 2007). For example, the rhodopsin gene, which is present in some marine microbes, is a 'cosmopolitan' gene able to wander in, and out of bacterial, and even archaeal genomes to the benefit of the organism (Frigaard et al. 2006). In addition, HGT causes the phylogeny to be inconsistent with the species tree because the horizontally transferred gene will reflect the evolutionary history of the organism from which the gene was transferred from and not the history of the recipient organism (Ciccarelli et al. 2006).

1.4 Gene Duplication

Gene duplication, which occurs in all three domains of life, provides 'new genetic material' for mutation and selection (Zhang 2003). These paralogs may be lost, become pseudogenes, subfunctionalized, persist in the genome which may result in a change of gene expression level, or acquire a novel function (Sankoff 2001; Hooper and Berg 2003). Although, many studies have hypothesized a paralog gaining a novel function, there has not been any documented case in Bacteria (Hughes 1994). However, in Eukaryotes there are many documented cases of functional divergence after gene duplication such as the Arabidopsis MADS-box gene family and the classic myoglobin/haemoglobin (Hughes 1994; Martinez-Castilla and Alvarez-Buylla 2003). The MADS-box genes regulate transcription under a variety of different developmental stages and have also been extensively duplicated to form a fairly large gene family (Martinez-Castilla and Alvarez-Buylla 2003). It is believed that positive Darwinian selection caused the acquisition of novel functions diversifying the MADS-box genes (Martinez-

Castilla and Alvarez-Buylla 2003). Another popular notion is that after the gene duplicates, the paralogs will divide up the function of the original gene (Sankoff 2001). For example, Tocchini-Valentini et al. (2005) discovered two cases of subfunctionalization in archaeal tRNA endonucleases. Nonetheless, the most common outcome of a gene duplication event in Bacteria is the silencing of one of the paralogs (Lynch 2002). In addition, duplicated DNA can be used as sites for homologous recombination causing chromosomal rearrangements and encouraging secondary rearrangements (Bailey et al. 2002; Samonte and Eichler 2002).

Gene duplications occur from unequal crossover events during homologous recombination, retroposition, horizontal transfer, chromosomal duplication, or genomic duplication through diploidization (Zhang 2003). Slippage during unequal homologous recombination tends to cause tandem gene repeats (Zhang 2003). However, the repeated region is composed of either gene fractions, the entire gene, or a segment encoding a number of different genes depending on where the sequences crossed over (Zhang 2003). After tandem duplication, chromosomal rearrangements and/or silencing of the duplicate genes in one of the duplicated segments may occur (Sankoff 2001). Duplicated gene which lack promoters are believed to develop into pseudogenes since they are silenced and no functional constraint is maintaining them (Lynch 2002). Most duplicated genes are believed to have been initially part of a larger duplicated region (Sankoff 2001).

Lynch and Conery (2000), suggested that the average duplication rate is approximately 1% per gene per million years based on calculations from analyses of duplicated genes within the genomes of multiple species. However, the average duplication rate for functioning duplicated genes is less than this due to the amount of gene loss (Lynch 2002). For example, DNA breakpoints which are quite common, randomly occur during the duplication method, and cause

some duplicated genes to lack regulatory regions such as promoter and even protein coding regions of the gene (Lynch 2002). Bacteria are more susceptible to duplication events and accordingly have a higher rate of duplication (Sankoff 2001). On the other hand, bacterial mechanisms are able to remove the duplicated genes at a faster pace (Sankoff 2001). Recent research shows that for certain groups of Bacteria from the Proteobacteria group and the *Bacillus/Clostridium* clade, horizontally transferred genes are more likely duplicated than indigenous genes (Hooper and Berg 2003). Possible explanations for this phenomenon include: genes that undergo HGT may be more mobile, HGT genes that are important for the organism to survive in a new environment may be needed at higher levels of expression, or genes may be horizontally transferred numerous times - indistinguishable from a HGT event with further duplications of the gene (Hooper and Berg 2003).

1.5 Evolutionary Biology and Bacterial Phylogenetics

In 1977, Carl Woese proposed using a single gene, the gene encoding the small subunit of the ribosomal RNA (SSU rRNA), to construct a phylogenetic tree that reclassified the five kingdoms of life into three domains: Bacteria, Archaea, and Eukaryotes (Woese and Fox 1977). Ribosomal RNA genes are ubiquitous in function and have highly conserved sequences due to their constrained roles in protein synthesis. This conservation allows for the classification of Bacteria by measuring the phylogenetic relationships between the organisms (Woese 1987; Gevers et al. 2005). The highly conserved nature of the SSU rRNA sequence renders it unable to resolve closely related organisms and unable to classify all Bacteria at a species level (Rossello-Mora and Amann 2001; Gevers et al. 2005). The tree of life has two basic assumptions to guarantee the validity of its hypothesis on the evolution of species, HGT has not occurred in the gene, and the rate of evolution or amount of dissimilarity between the gene is representative of

the genome (Goodfellow, Manfio, and Chun 1997). Unfortunately, using one gene as a phylogenetic marker is subject to simple stochastic variations in bacterial genomes (Gevers et al. 2005). An alarming study discovered that highly conserved genes that are involved in complex interactions and thus believed to be resistant to HGT are actually susceptible to HGT (Gogarten and Townsend 2005).

SSU rRNA sequence analysis is universally accepted amongst microbiologists to delineate and classify species (Rossello-Mora and Amann 2001). In addition, SSU rRNA trees have been instrumental in the clarification of many taxonomic issues, such as resolving the *Pseudomonas* clade (Rossello-Mora and Amann 2001; Garrity et al. 2005). Before the advent of molecular biology the *Pseudomonas* clade and all the other bacterial species were classified based on morphology and metabolic characteristics which for the most part worked well and agreed with the now current taxonomic classification using subunit rRNA trees with surprisingly small errors (Rossello-Mora and Amann 2001).

Well conserved elongation factors such as EF-Tu and EF-G have also been used to construct universal phylogenetic trees (Creti et al. 1994) and are hypothesized to have arisen from an ancient gene duplication event. This event may have occurred during the lifetime of the last common ancestor which makes these sequences excellent candidates to root and construct the universal phylogenetic tree (Baldauf, Palmer, and Doolittle 1996). Even though EF-Tu is ubiquitous in all three domains of life and therefore believed to be a good indicator of evolutionary relationships; there is evidence of HGT in EF-Tu in the *Enterococci* bacterial species and multi-copies of the EF-Tu genes can be located in other bacterial genomes those histories have not yet been researched (Ke et al. 2000). Hence, the universal phylogenetic trees

produced from elongation factor sequences are believe to be incorrect because of the complex history, which may include HGT and gene duplication (Woese 2000).

1.6 Conservation of Gene Order in Bacteria

Wolf et al. (2001) observed only a slight correlation between the conservation of gene order and evolutionary relationship, based on genomic analysis. Determining the genomic context of a gene facilitates the resolution of gene history and the possible selection pressures working upon the gene (Wolf et al. 2001). In addition, for recent duplicated or HGT genes, analysis of genomic context may help to determine how the recent duplication or HGT event occurred. For example, whether the gene was duplicated or transferred within a large region of genomic sequence or just the gene itself and the mechanisms involved. However, tracking evolutionary relationships and predicting gene function using gene order is deceiving due to the number of genomic rearrangment that occurs in Bacteria (Rogozin et al. 2004). In most part, gene order in Bacteria is not well conserved if it is not part of an operon (Koonin, Mushegian, and Rudd 1996). However, there are some cases of well conserved cluster of genes discovered in distantly related species such as the genes for ribosome proteins (Nikolaichik and Donachie 2000). Lathe, Snel, and Bork, (2000) discovered that in genomic regions where gene order is conserved, that even if some type of rearrangement occurs in this region, the genes are still kept closer together than other non conserved regions. Also certain regions in bacterial genome are more susceptible to genomic rearrangement of gene order such as at the terminus of replication (Sanderson and Liu 1998).

Only several operons such as the ribosomal protein operon and operons encoding proteins that interact with one another, are universally conserved amongst the bacterial and archaeal genome (Mushegian and Koonin 1996). Operons are believed to be favored by selection to be

horizontal transferred, over the transfer of a single gene, because operons are self sufficient, their regulation is maintained (Lawrence and Roth 1996; Lawrence 1999a). In addition, only 5 to 25% of genes in Bacteria and Archaea genomes belong to probable operons or gene strings which are similar in at least two genomes that are not closely related (Wolf et al. 2001). Because of the limited gene order conservation in Bacteria, the conservation of three genes in a row between distinctly related organisms, is statistically significant unless the genes are part of an operon (Wolf et al. 2001). The conservation of gene order may be maintained under selection pressures for the following reasons:

1. Species which have diverged relatively recently may not have had enough time for rearrangement of gene order (Tamames 2001).
2. Gene order maybe be maintained following a horizontal transfer event of a genomic region in distantly related species (Wolf et al. 2001).
3. If the integrity of the cluster of genes is essential or important for the viability of the cell (Tamames 2001)
4. The presence of key regulatory elements may maintain gene order (Koonin, Aravind, and Kondrashov 2000)

Both elongation factor Tu genes, *tufA* and *tufB* have been discovered in operon configurations. The streptomycin operon, which is one of the most well conserved operons in Bacteria, consists of ribosomal protein S12, ribosomal protein S7, elongation factor G (*fus*), and elongation factor Tu (*tufA*) in the *E. coli* organism (Koonin and Galperin 1997; Itoh et al. 1999). Interestingly, the streptomycin operon contains three different operons, the main promoter is upstream of ribosomal protein S12 gene, while the additional two promoters are located within the *fus* gene and is the promoter for *tufA*; these three promoters allows for the different levels of

expression of the genes in the operon (Post et al. 1978). Elongation Factor Tu (*tufB*) is located within the tRNA-tufB operon in *Thermus thermophilus*, and *Chlamydia trachomatis* (Satoh et al. 1991; Cousineau et al. 1992). Rogozin et al. (2002) has hypothesized that *efp* is within a gene neighbourhood, which is a group of genes that are not all present in any single genome but are linked to one another by gene arrays. Gene neighbourhoods are discovered by using orthologous genes which were transcribed in the same direction and only separated by zero or two genes (in the gene order of the orthologous genes) in at least three distantly related organisms (Rogozin et al. 2002). The conserved gene pairs were then merged with other gene pairs discovered in other species which contained similar or overlapping genes from the original gene pair (Rogozin et al. 2002). Then the gene arrays were clustered together if they contained at least two similar genes to form a gene neighbourhood (Rogozin et al. 2002). EF-P is hypothesized to be in an extended gene neighbourhood those primarily function was lipid metabolism and amino acid metabolism with the additional minor function of translation (including EF-P and proteins such as ribosomal proteins L32), transcription (transcriptional regulators), replication (DNA polymerase III sigma'), and coenzyme metabolism (O-succinylbenzoate synthetase) (Rogozin et al. 2002).

1.7 DNA Sequence Motifs

DNA motifs are short, repeating sequences of nucleic acid residues that are evolutionarily conserved and have a biological function (Wolf and Arkin 2003). Examples of DNA sequence motifs include restriction enzyme binding sites, and transcription factor binding sites (TFBS) such as TATAAT boxes, -10 and -35 promoter elements (D'haeseleer 2006). The consensus sequence of the TFBS may regulate the change in level of expression of the gene under different conditions (MacIsaac and Fraenkel 2006). The sequence of the TFBS is very important in the regulation of gene expression and therefore, are believed to be under selective pressure and

conserved (MacIsaac and Fraenkel 2006). Furthermore, in most cases, evolutionary conserved non-coding genomic sequences correspond to regulatory regions (MacIsaac and Fraenkel 2006).

Phylogenetic footprinting has been very effective in discovering putative motifs using many orthologous or co-regulated conserved regions (MacIsaac and Fraenkel 2006). Unfortunately orthologous genes may be regulated differently or use different transcription factors in some species (McCue et al. 2001). For example, genomic rearrangements may cause gene to be split up within an operon causing promoters to be upstream of different genes in different species (McCue et al. 2002). However, McCue et al., (2001) documented that using many orthologous sequences from many different species increases the chances that there will be enough sequences in the data that uses similar gene regulation and thus a similar motif to readily identify. Clustering orthologous sequences based on a number of different characteristics such as genomic size and the natural habitat of the organism increase the chances that the species are more likely to have the same regulatory mechanisms and thus enhance the motif discovery process *in silico* (McCue et al. 2001).

There have been at least 20 documented cases where a gene has been horizontally transferred with adjacent transcription factors (Price, Dehal, and Arkin 2007). Finding the putative TFBS may allow us to identify the sigma factor that is responsible for initiating transcription. Different species have different sigma factors; for example, *E. coli* is known to have eight different sigma factors. RegulonDB 5.0 (http://regulondb.ccg.unam.mx/) which documents all the predicted and known promoter sites for the organism *E. coli*, predicts that *efp* uses sigma 70 (Salgado et al. 2006). The predicted promoter of *efp*1 in *E. coli* is depicted in Figure 2. However, there is no record for the *YeiP* or *efp*2 promoter. The computational derived motifs or TFBS can then be experimentally tested for their ability *in vivo*.

1.8 Thesis Objectives

1. Confirm evidence of HGT, gene duplication, and the direction of the HGT for EF-P.

Not much is known about the conservation and role of the bacterial EF-P protein compared with its eukaryotic and archaeal homologs. Previous work on the bacterial EF-P noted numerous possible HGT and gene duplication events by unusual phylogenetic topologies. These results will be refined and compared with phylogenies constructed using SSU rRNA genes. Determination of possible HGT and gene duplication events are assessed using GC content, codon bias, and sequence identities as determined by BLAST in tandem with a positive and negative control for HGT and gene duplication in order to determine a threshold.

2. Determine if there is any conservation of the genomic context of EF-P.

Determine whether or not EF-P is within an operon configuration in any of the completely sequenced bacterial species in NCBI. In addition, determining the genomic context of EF-P may help lend support to the evidence of recent HGT or gene duplication of certain EF-P proteins and the history of the genes i.e. transferred or duplicated within a large genomic region or singularly.

3. Extract any DNA sequence motifs from the upstream regions of the orthologous EF-P.

DNA sequence motifs are discovered by analyzing overrepresented and/or conserved patterns upstream of orthologous *efp* genes. Ninety percent of known promoters occur in the 250 bp upstream region of the gene in *E. coli* (Huerta and Collado-Vides 2003). Since, EF-P is believed to have been present in the last common ancestor (an ancient gene) there maybe some regulatory regions conserved between the species.

Figure 2. EF-P promoter and transcription start site from *E. coli* K12 in RegulonDB 5.0 (Aoki et al. 1991; Salgado et al. 2006). The transcription start site is represented by the red capital letter and the -10 element is underlined. The -35 element has not yet been experimentally characterized.

aaagctttttggcgctgcgtccggctaacagtttttcctccgcgtctatattcaaaagacGcagaagttcatcaggatcgg

Chapter 2: METHODS

2.1 Sequences Retrieval

Nucleotide and amino acid sequences of EF-P were retrieved from the NCBI GenBank September 21, 2005 release of completely sequenced bacterial genomes using NCBI BLAST (available at http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) against the well characterized *E. coli* K-12 EF-P. The default BLASTP settings were used except that the filter option was turned off to find longer matches. The cutoff values used to find EF-P protein sequences was that the e-value must be less than 0.00001 to ensure that only orthologous *efp* sequences were retrieved and not just similar sequences. The EF-P sequences were subsequently updated on July 7, 2006 using the same technique, adding 94 EF-P sequences from newly sequenced bacterial genomes. A list of the EF-P sequences and their respective Genbank accession numbers used in the analyses is included in the supplementary files. Small subunit ribosomal (SSU rRNA) gene nucleotide sequences were retrieved from the corresponding bacterial (16S rRNA gene) and archaeal (16S rRNA gene) genomes in NCBI GenBank in conjunction with the Ribosomal Database Project-II Release 9 (Maidak et al. 2001). All SSU rRNA sequences used in the present study are included in supplementary files.

2.2 Positive and Negative Bioinformatic Controls for HGT

A positive and negative control establishes a putative threshold to determine which EF-P proteins may have been horizontally transferred using these sequence analysis methods. Elongation factor Tu (EF-Tu, *tuf*) was used as a positive control for horizontal gene transfer and gene duplication in GC and codon usage because it is known to have a history of gene

duplication and HGT (Sela et al. 1989; Ke et al. 2000). EF-Tu is hypothesized to have undergone HGT in the *Enterococci* species and certain bacterial species are known to have one to three copies of EF-Tu present in their genome (Sela et al. 1989; Ke et al. 2000). In addition EF-Tu is highly expressed, in *E. coli* there is a 1:10 EF-G/EF-Tu ratio (Young and Furano 1981). Recombinase A (*recA*) was used as the negative control for HGT. RecA is a common protein that has been utilized in resolving the universal species tree and is believed to have no known history of HGT (Eisen 1995). EF-Tu and RecA nucleotide sequences were retrieved from NCBI GenBank from the completely sequenced bacterial genomes on October 7, 2006. The positive and negative controls underwent the same analyses as EF-P for the GC content, codon bias, and sequence identity in order to determine EF-P genes that may have undergone HGT. This was done in order to establish the cutoff values that could be used to determine if a HGT event occurred.

2.3 Sequence Alignment and Phylogenetic Analyses

Alignments for all EF-P amino acid sequences were generated using MUSCLE 3.6 (Edgar 2004) and manually adjusted based on visual analysis using BioEdit 7.0.4 (Hall 1999). *Mycoplasma genitalium* G37 EF-P2 (ZP_00404835.1) was removed from the analyses as it was only 88 amino acid residues long compared to the mean length of EF-P (190 amino acid residues). The model of evolution for the EF-P proteins was determined using the program ProtTest 1.3 (Abascal, Zardoya, and Posada 2005). Identical sequences were removed from the analyses. Neighbor-joining (NJ) phylogenetic trees were constructed using PAUP* 4.10b (Swofford 2003) using a supplied rtREV (Dimmic et al. 2002) substitution matrix with the associated parameters. PHYML 2.4.4 (Guindon and Gascuel 2003) was used to approximate a maximum likelihood phylogenetic tree with the rtREV model and parameters chosen by ProtTest

1.3. A Bayesian tree using Mr. Bayes 3.1 (Ronquist and Huelsenbeck 2003)was attempted using 5 chains (2 heated and 3 cold), the rtREV model, and parameters chosen by ProtTest 1.3. However, the split deviations between two different runs did not converge using 1,000,000 samples. The average standard deviation of split frequencies were constantly fluctuating and the analysis was abandoned.

All SSU rRNA gene sequences were aligned based on secondary structure models (Cannone et al. 2002). These sequences were aligned and reduced to the universal core, regions I, II, and III, using BioEdit 7.0.4 (Hall 1999). Phylogenetic construction was generated similar to the EF-P tree except the best fit model of DNA substitution and parameters was chosen by Modeltest 3.7 (Posada and Crandall 1998).

2.4 Detecting Horizontal Gene Transfers

To confirm the putative HGT and gene duplication events the robustness of the EF-P and the SSU rRNA tree was tested. The SSU rRNA and EF-P NJ trees were bootstrapped using 1000 replicates in PAUP* 4.0b (Swofford 2003). Using Tree-Puzzle 5.2 (Schmidt et al. 2002) alternative EF-P tree topologies were compared using the Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa 1999). The Shimodaira-Hasegawa test is a statistical test which obtains the confidence limit of the topology and is based upon the Kishino-Hasegawa test with modifications to allow multiple tree topology tests (Shimodaira and Hasegawa 1999). In order to implement the rtREV model of evolution for EF-P, the source code for Tree-Puzzle 5.2 (Schmidt et al. 2002) was modified to include the rtREV model. For topology testing the following trees were constructed: a subset NJ tree that included all the multiple copies of EF-P and EF-P supported by high bootstrap values to represent the major clades; a similar subset NJ tree with the same organisms was used except all the multi-copy EF-Ps were made into sister taxa which is

consistent with a tree topology of gene duplication; an approximate maximum likelihood subset tree, and a maximum parsimony subset tree.

2.5 Sequence Similarities Analyses

Sequence identities of all the different *efp* nucleotide and EF-P amino acid sequences were determined using BLAST BL2SEQ (Tatusova and Madden 1999). Perl and BioPerl scripts were used in the facilitation of these sequences and to perform command line BL2SEQ. The filter parameter was disabled to ensure longer alignments for both nucleotides and amino acid sequences.

2.6 GC Content and Codon Bias Analyses

Genomic GC content information was gathered from the NCBI GenBank database for all the completely sequenced bacterial genomes (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). The *efp* gene GC content was calculated manually by adding the occurrences of G and C base and then dividing by the length of the gene. Genomic and *efp* codon frequency information was gathered from the Codon Usage Database (http://www.kazusa.or.jp/codon/). Some genomes were not included due to lack of genomic codon usage information. In the codon usage analysis the nonsense or termination codons (UAA, UAG, and UGA) were removed.

The Hamming distance calculation used is defined in Garcia-Vallve, Romeu, and Palau (2000):

$$d^{\mathrm{H}}(X, \overline{X}) = \sum_{i=1}^{61} |x_{\mathrm{i}} - \overline{x_{\mathrm{i}}}|,$$

where x_i is the frequency of the i^{th} codon for the *efp* gene and $\overline{x_i}$ is the mean frequency of the i^{th} codon for the organism (Garcia-Vallve, Romeu, and Palau 2000). The Hamming distance

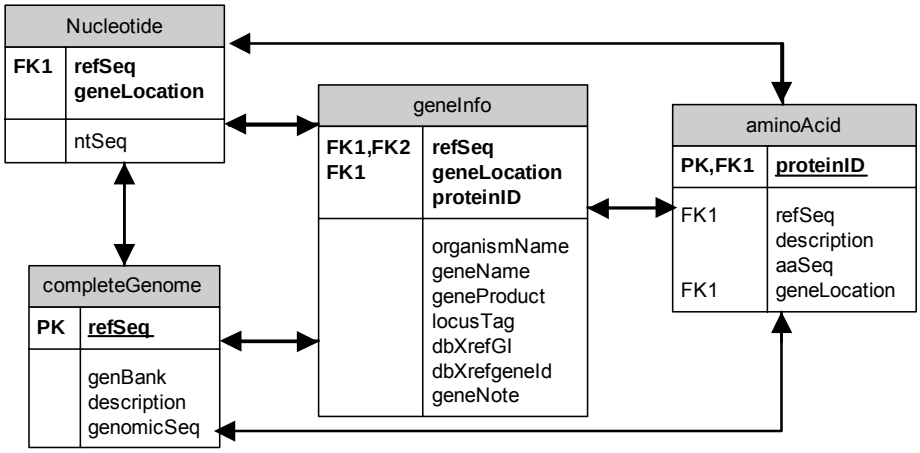
calculations were used to measure the distance between *efp* codon usage and the mean codon usage of its corresponding organism (Garcia-Vallve, Romeu, and Palau 2000). For example, if the Hamming distance calculated was small the codon usage of the gene most likely matches the codon usage of its genome (Garcia-Vallve, Romeu, and Palau 2000).

In codon usage clustering analysis, *tuf* which is a highly expressed gene in *E. coli* was used as a control to determine the separation based on different levels of expression compared with different codon usage due to HGT. The codon usage frequencies were first standardized, by subtracting the mean from the value and then dividing by the standard deviation. The appropriate distance calculation (i.e. Euclidean) that best measures the corresponding cluster tree made from the distance calculations, and how accurately it represents the original observations was determined using a cophenetic correlation coefficient (Matlab 6.5). The Jaccard distance was determined to be the best representative distance calculation and it was used to calculate the dissimilarity between the different codon frequencies. Hierarchical clustering was employed using the UPGMA algorithm to generate a linkage tree. A cutoff of 0.8 was used in the analysis to find natural divisions within the codon usage based on its calculated dissimilarity matrix.

2.7 Conservation of Gene Location Analysis

All the completely sequenced bacterial genome assembly and annotation projects which includes the complete genome/chromosome, reference nucleotide sequence, mRNA, and protein sequences (the .faa, .ffn, .fna, and .gbk files from each genome assembly) were downloaded from NCBI's ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) on May 2, 2007. The data were parsed into a tab delimited text file and imported into tables in an SQL database using appropriate relationship linkages that are similar to NCBI as seen in Figure 3. An SQL database allows for

Figure 3. The relational database including tables, and the information stored. PK represents primary key and FK represents foreign keys which can be used to determine corresponding data.


Nucleotide
FK1
refSeq
geneLocation
ntSeq
geneInfo
FK1,FK2
FK1
refSeq
geneLocation
proteinID
organismName
geneName
geneProduct
locusTag
dbXrefGI
dbXrefgeneId
geneNote
aminoAcid
PK,FK1
proteinID
FK1
refSeq
description
aaSeq
FK1
geneLocation
completeGenome
PK
refSeq
genBank
description
genomicSeq

the faster retrieval and organization of a large amount of data. The genomic location of EF-P coding regions were determined from the nucleotide sequences. The locations were used to retrieve 10 kb from the upstream and downstream adjacent genomic regions of all the EF-P genes. Using BL2SEQ encoded in a Perl script, these adjacent regions were blasted against each other to determine the percent similarity and thus any putative homologous genes or non-coding regions that are conserved. The cutoff I used for the BL2SEQ was quite stringent because non-coding and coding DNA regions were compared. The filter was turned on so repetitive elements would not result in spurious matches. Also, the e-values of any similar regions must be less than 0.00001. A program was then developed in the C# language to display the BL2SEQ results in an informative graphical manner which included gene information from NCBI to determine the regions and genes which are conserved.

2.8 Upstream Motifs of EF-P Analysis

Using the SQL genomic sequence table, 250 bp upstream of the start codon of the bacterial EF-P gene was retrieved. 90% of promoters occur in the 250 bp upstream region in *E. coli* (Huerta and Collado-Vides 2003). In addition, motif prediction algorithms have a higher accuracy when discovering motifs in smaller regions. The upstream regions were then reverse complemented depending on whether the EF-P genes are transcribed on the sense or anti-sense strand of the genome. The SSU rRNA distance positions, genomic size, natural habitat from NCBI microbial information, genomic GC content, and the number of annotated orthologs transcription factors discovered in GenBank using COGs databases were converted into numbers (if they were not already), and standardized due to their different measurements. The values were standardized by subtracting the mean from the value and then dividing by the standard deviation. I used the cophenetic correlation coefficient, as explained in the codon usage section, to calculate

the best distance measurement and linkage method to use in Matlab 6.5. The Euclidean distances are calculated for the different permutations of the above characteristics. Hierarchical clustering using the furthest neighbour clustering algorithm was then employed on the different permutation of the characteristic's distance measurements. The maximum number of clusters was defined as 10, 15, and 20 because using natural divisions in the data gave too many clusters (> 100 clusters) and most of these clusters contained only one sequence. All the upstream sequences clustered by different combination of characteristics and different maximum number of clusters were used to find motifs. Any clusters that only contained one sequence were removed from the dataset.

To discover motifs in the upstream region of EF-P a strategy similar to MacIsaac and Fraenkel (2006) outline was followed using the program TAMO 1.0 (Gordon et al. 2005; MacIsaac and Fraenkel 2006). TAMO 1.0 is able to use three independent motif discovery programs and score the motifs predicted. See Figure 4 on the method and the component of TAMO 1.0 used (Gordon et al. 2005). A python script was written in order to use all the different components of TAMO 1.0 such as using the different motif discovery algorithms and calculate the discovered motif's p-values. The following programs were used to discover motifs: AlignACE is based upon a Gibbs-sampling algorithm and calculates the most statistically significant alignments from the input data (Roth et al. 1998). AlignAce 2004 was downloaded from (http://atlas.med.harvard.edu/). AlignAce 2004 was used with the default settings, except that the seed used was based on a random number. MDScan uses word enumeration and position-specific weight matrix updating, using ChIP-array information to enhance motif discovery and decrease search time (Liu, Brutlag, and Liu 2002). ChIP-array probe sequences which represent protein-DNA interaction sites are used in MDScan to help search for DNA

motifs (Liu, Brutlag, and Liu 2002). MDScan 2004 was downloaded from (http://robotics.stanford.edu/~xsliu/MDscan/). MDScan 2004 was used with the default settings, the width of the motifs was specified to be from 16 to 24bps based on McCue et al. (2001) prokaryotic TFBS discovery suggestion. MEME is one of the more popular motif discovery programs (MacIsaac and Fraenkel 2006). MEME uses an expectation maximization algorithm to find motifs (Bailey and Elkan 1994). MEME 3.5.4 was downloaded from (http://meme.sdsc.edu/meme/intro.html). MEME 3.5.4 was used with the default values, and the motif widths did not need to be specified like AlignAce 2004. These two programs theorically discover the motif of any width that has the best score.

The discovered motifs were then scored by TAMO 1.0 using a hypergeometric distribution to calculate p-values. A hyper geometric distribution measures the probability that the motifs discovered from a chosen group of sequences would also be discovered if the group of sequences were chosen at random from the genome (Gordon et al. 2005). The p-value was scored using the following formula (Harbison et al. 2004):

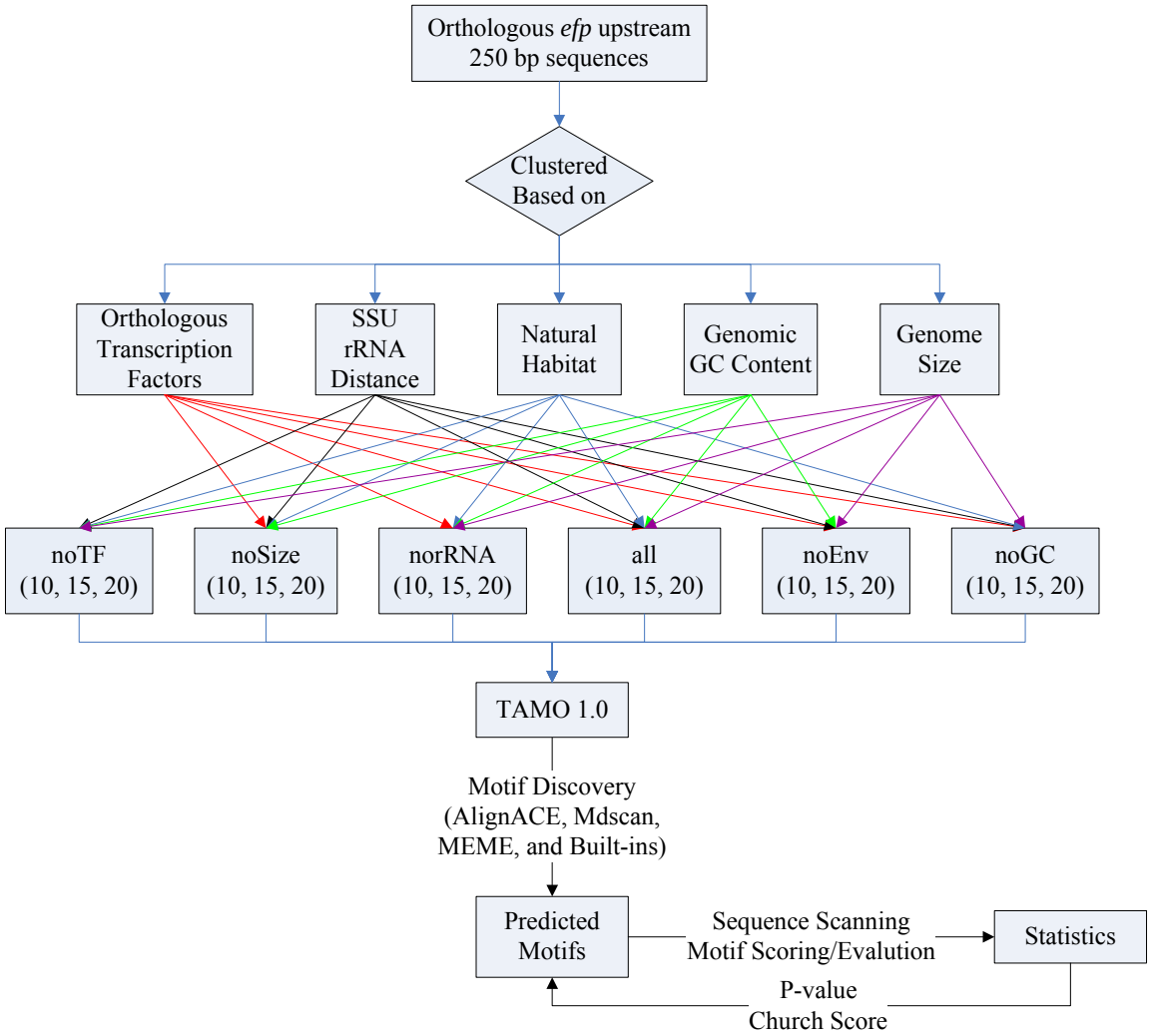
$$p = \sum_{i=b}^{\min(B,g)} \frac{\binom{B}{i}\binom{G-B}{g-i}}{\binom{G}{g}}$$

where B is the number of input sequences and G is the total number of sequences represented in the genome (Gordon et al. 2005). The variables b and g are a subset of B and G that contain the motif (Gordon et al. 2005). For each organism in the cluster, 250 bp upstream of all known annotated genes from each genome were retrieved using a Perl script in order to fulfill the requirements of G. The null hypothesis in this objective is to determine if the motifs discovered using TAMO 1.0 from our clustered upstream sequences are more significant than the

motifs discovered using TAMO 1.0 in randomly chosen upstream sequences from the organisms used. The active hypothesis in this method is the motifs discovered by TAMO 1.0 are better than random. Harbison et al. (2004) experimentally determined that a discovered motif with a p-value of less than 0.001 has the lowest probability of being a false positives and it is within this threshold that many motifs were correctly identified for known regulators.

A Perl script was used to parse through the TAMO 1.0 output and the top 5 motifs according to p-value were retrieved from the best cluster. The fasta sequences of known bacterial motifs were downloaded from the prokaryotic motif databases, RegTransBase and RegulonDB 5.0 (Salgado et al. 2006; Kazakov et al. 2007). Using a Perl script the position frequency matrices of these known bacterial motifs were calculated by counting the number of times a base occurs at a specific position and then dividing by the number of sequences used. The position frequency matrices of the top 5 discovered *efp* motifs were compared with the position frequency matrices of RegTransBase and RegulonDB 5.0 using a Perl script. When comparing the frequency matrices the score is obtained by taking the absolute value of the difference in frequencies for each position and base pair. The motifs were displayed using Weblogo 2.8.2 (Crooks et al. 2004).

Figure 4. Methods used in the prediction of motifs and the different components in TAMO 1.0 (Gordon et al. 2005) used to discover motifs in the upstream regions of orthologous *efp* sequences.


Orthologous *efp* upstream 250 bp sequences
Clustered Based on
Orthologous Transcription Factors
SSU rRNA Distance
Natural Habitat
Genomic GC Content
Genome Size
noTF (10, 15, 20)
noSize (10, 15, 20)
norRNA (10, 15, 20)
all (10, 15, 20)
noEnv (10, 15, 20)
noGC (10, 15, 20)
TAMO 1.0
Motif Discovery (AlignACE, Mdscan, MEME, and Built-ins)
Predicted Motifs
Sequence Scanning Motif Scoring/Evalution
Statistics
P-value Church Score

Chapter 3: RESULTS

3.1 Phylogenetic Analyses

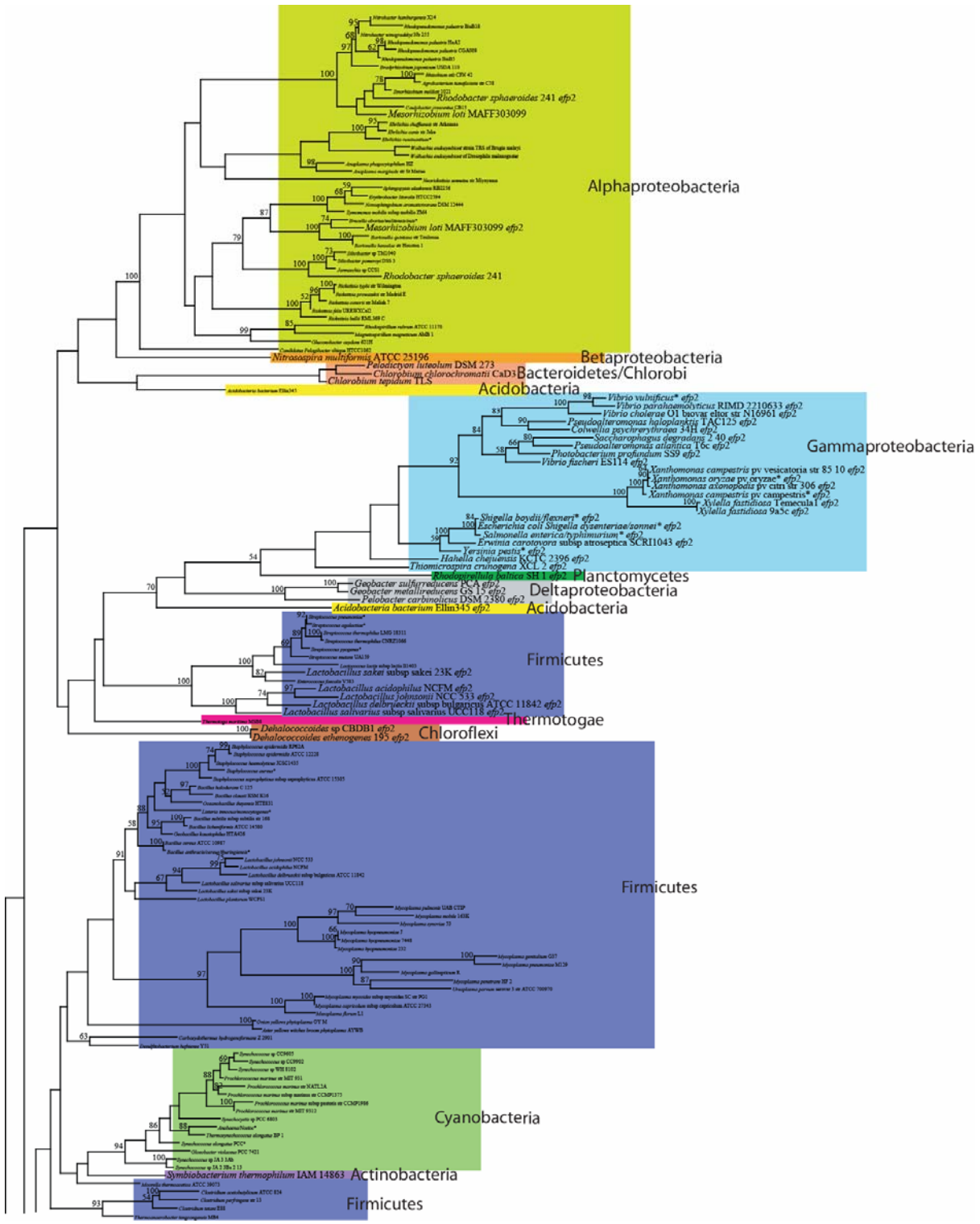
The presence of two divergent copies of the *efp* gene was observed in 69 out of the 322 completely sequenced bacterial genomes available at the time of this study. A single *efp* gene was observed in each of the remaining 253 bacterial genomes. In order to promote clarity, the EF-P duplicates are denoted as EF-P1 and EF-P2 (or *efp*1 and *efp*2), where EF-P2 denotes the EF-P protein with the lower amino acid sequence identity to *E. coli* K12 EF-P (NP_418571.1). *Leptospira interrogans serovar Copenhageni* str. *Fiocruz* L1-130 was not included in the study as its *efp* encoding region (NC_005824) contains what NCBI claims as an 'authentic frameshift mutation and not a sequencing artifact' and thus has no corresponding protein product. The mutations occurs in the first 52 amino acids of the translated *efp* encoding region from *Leptospira interrogans serovar Copenhageni* str. *Fiocruz* L1-130 and when compared with the *efp* from *Leptospira interrogans serovar Lai* str. 56601, 53 to 269 amino acid residues are identical. Interestingly, *E. coli* EF-P starts aligning with the *Leptospira interrogans serovar Copenhageni* str. *Fiocruz* L1-130 translated *efp* at amino acid residue 81 and thus *Leptospira interrogans serovar Copenhageni* str. *Fiocruz* L1-130 may contain three functional domains of EF-P (KOW, OB, and EF-P C-terminus). Table 1 contains a breakdown of the phylogenetic group of the genomes and the number of *efp* genes used in our analyses and the organisms that contain the duplicate *efp* genes.

In order to determine whether *efp*1 and *efp*2 sequences encode genuine EF-P, the deduced amino acid sequences of both genes were aligned to determine if the domains of EF-P were conserved. The multiple sequence alignment demonstrated that the OB domain, KOW-like

domain, and an EF-P C-terminus domain of EF-P are conserved (Pfam). Since the *efp*1 and the *efp*2 retrieved contain these three domains; it appears that both gene products could possibly fulfill the function of EF-P. The different variations of the domains are more strongly conserved within each bacterial group and the EF-P2 of each group. The OB domain is observed around residues 69 to 125 for *E. coli* K12 EF-P (Pfam). The KOW-like domain is observed around residues 3 to 62 for *E. coli* K12 EF-P (Pfam). An EF-P C-terminus domain is observed around residues 133 to 188 for *E. coli* K12 EF-P (Pfam)**.** Additionally, to ensure only orthologs and/or homologs are retrieved from NCBI, the EF-P and EF-P2 sequences used in this study have a greater than 30% amino acid residue identity with *E. coli* K12 EF-P and e-values of ≤ 0.00001.

The EF-P tree (Figure 5) was constructed from the EF-P amino acid sequence because the nucleotide sequence was too divergent as seen in Table 2. The original length of the MUSCLE aligned amino acid EF-P and SSU rRNA was 298 amino acid residues and 2299 nucleotides respectively, including gaps. After refinement the length of the EF-P and SSU rRNA alignment was 140 amino acid residues and 1148 nucleotides, respectively. These alignments are included as supplementary data files. The amino acid model selected using all 6 frameworks in protTest, including Akaike (AIC) and Bayesian Information Criterion framework, was the amino acid substitution matrix for the inference of retrovirus and reverse transcriptase, rtREV+I+G+F that calculated the amino acid residue frequencies (A = 0.058, C = 0.005, D = 0.059, E = 0.076, F = 0.047, G = 0.088, H = 0.010, I = 0.065, K = 0.063, L = 0.079, M = 0.033, N = 0.036, P = 0.044, Q = 0.025, R = 0.041, S = 0.036, T = 0.075, V = 0.109, W = 0.007, Y = 0.044), the gamma distribution shape parameter (a = 1.553), with a proportion of invariable sites (0.0070) (Dimmic et al. 2002). Modeltest 3.7 was used to examine 56 possible models of DNA substitution and identify the model that best fit the SSU rRNA data set (Posada and Crandall 1998).

Figure 5. The unrooted phylogram maximum likelihood EF-P phylogenetic trees with the bootstrap values from the NJ tree. Organisms are denoted as EF-P 2 if they have more than one copy of EF-P. Representative sequences are denoted with an asterisk and the groups and species they represent. The amino acid best-fit model of evolution used to construct this tree is rtREV+I+G+F (Dimmic et al. 2002), with a gamma shape of 1.553, and a proportion of invariable sites calculated at 0.0070 as determined all criteria implemented in ProTest v.1.3 [17]. / denotes removal of a large distance in order to present the tree. All possible HGT and/or gene duplication are highlighted excluding the ones seen in Figure 6 and 7. The organisms in large font are the putative HGT and/or gene duplication events. Organisms name followed by an asterisk denote representative sequences. / denotes removal of a large distance in order to present the tree. Each bacterial group is highlighted in a unique colour. Bootstrap values are denoted above the branch.


Alphaproteobacteria
Rhodobacter sphaeroides 241 efp2
Mesorhizobium loti MAFF303099
Mesorhizobium loti MAFF303099 efp2
Rhodobacter sphaeroides 241
Betaproteobacteria
Nitrosospira multiformis ATCC 25196
Bacteroidetes/Chlorobi
Pelodictyon luteolum DSM 273
Chlorobium chlorochromatii CaD3
Chlorobium tepidum TLS
Acidobacteria
Gammaproteobacteria
Vibrio vulnificus* efp2
Vibrio parahaemolyticus RIMD 2210633 efp2
Vibrio cholerae O1 biovar eltor str N16961 efp2
Pseudoalteromonas haloplanktis TAC125 efp2
Colwellia psychrerythraea 34H efp2
Saccharophagus degradans 2 40 efp2
Pseudoalteromonas atlantica 16c efp2
Photobacterium profundum SS9 efp2
Vibrio fischeri ES114 efp2
Xanthomonas campestris pv vesicatoria str 85 10 efp2
Xanthomonas oryzae pv oryzae* efp2
Xanthomonas axonopodis pv citri str 306 efp2
Xanthomonas campestris pv campestris* efp2
Xylella fastidiosa Temecula1 efp2
Xylella fastidiosa 9a5c efp2
Shigella boydii/flexneri* efp2
Escherichia coli Shigella dysenteriae/sonnei* efp2
Salmonella enterica/typhimurium* efp2
Erwinia carotovora subsp atroseptica SCRI1043 efp2
Yersinia pestis* efp2
Hahella chejuensis KCTC 2396 efp2
Thiomicrospira crunogena XCL 2 efp2
Rhodopirellula baltica SH 1 efp2
Planctomycetes
Geobacter sulfurreducens KA efp2
Geobacter metallireducens GS 15 efp2
Deltaproteobacteria
Pelobacter carbinolicus DSM 2380 efp2
Acidobacteria bacterium Ellin345 efp2
Firmicutes
Lactobacillus sakei subsp sakei 23K efp2
Lactobacillus acidophilus NCFM efp2
Lactobacillus johnsonii NCC 533 efp2
Lactobacillus delbrueckii subsp bulgaricus ATCC 11842 efp2
Lactobacillus salivarius subsp salivarius UCC118 efp2
Thermotogae
Dehalococcoides sp CBDB1 efp2
Dehalococcoides ethenogenes 195 efp2
Chloroflexi
Firmicutes
Cyanobacteria
Symbiobacterium thermophilum IAM 14863
Actinobacteria
Firmicutes

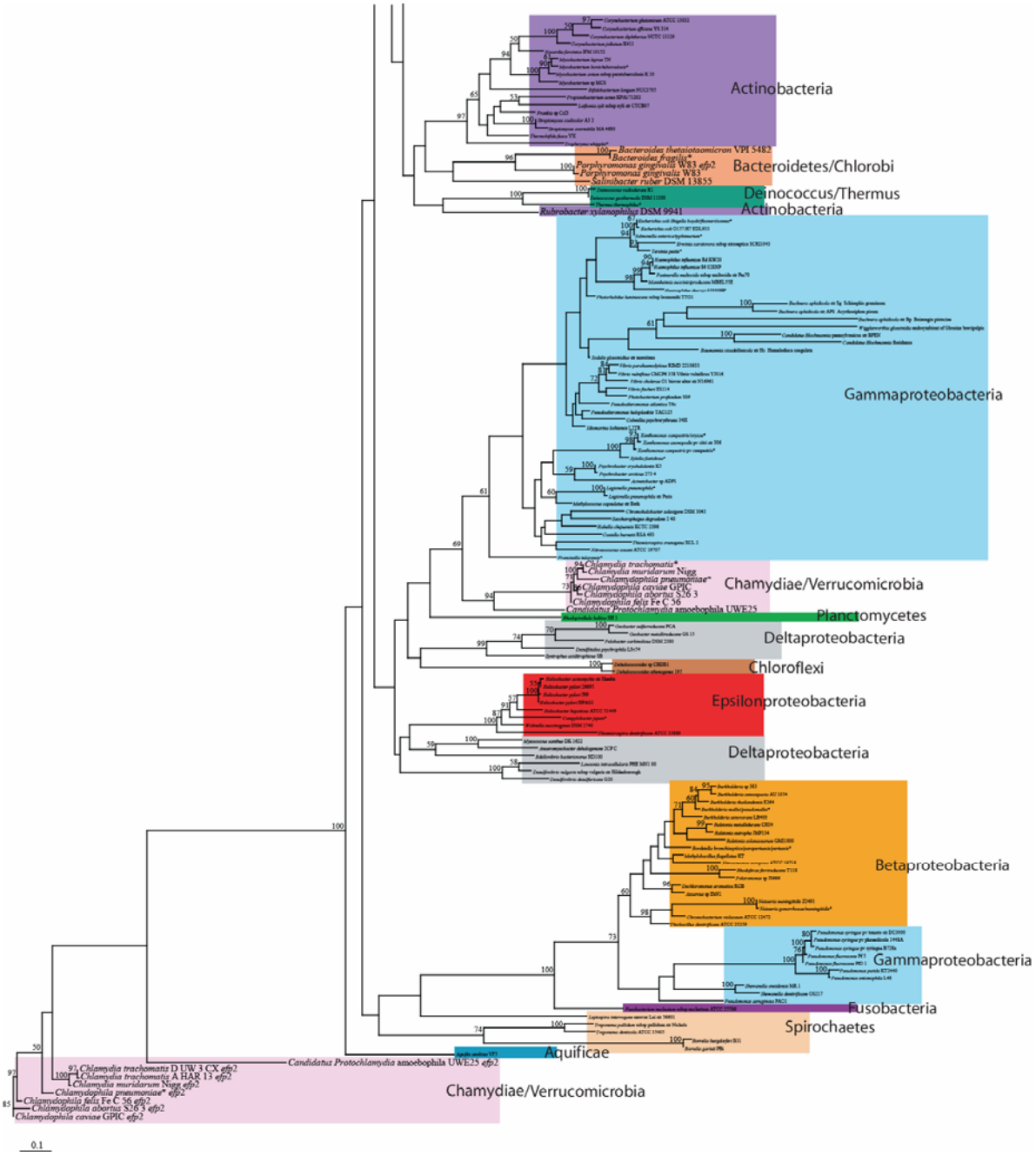

Actinobacteria
Bacteroides thetaiotaomicron VPI 5482
Bacteroides fragilis*
Porphyromonas gingivalis W83 efp2
Porphyromonas gingivalis W83
Salinibacter ruber DSM 13855
Bacteroidetes/Chlorobi
Deinococcus/Thermus
Rubrobacter xylanophilus DSM 9941
Actinobacteria
Gammaproteobacteria
Chlamydia trachomatis*
Chlamydia muridarum Nigg
Chlamydophila pneumoniae*
Chlamydophila caviae GPIC
Chlamydophila abortus S26 3
Chlamydophila felis Fe C 56
Candidatus Protochlamydia amoebophila UWE25
Chamydiae/Verrucomicrobia
Planctomycetes
Deltaproteobacteria
Chloroflexi
Epsilonproteobacteria
Deltaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Fusobacteria
Spirochaetes
Aquificae
Candidatus Protochlamydia amoebophila UWE25 efp2
Chamydiae/Verrucomicrobia
0.1

Table 1. Genomic and *efp* G+C content of the phylogenetic Bacterial groups. Values are given as mean ± standard deviation.

Phylogenetic Bacterial Group	**Genomic G+C content (%)**	***efp* G+C content (%)**	**# of organism**	**# of *efp* genes observed**
Acidobacteria	58.4	59 ± 0.6	1	2
Actinobacteria	63.4 ± 7.8	60.8 ± 6.4	22	22
Alphaproteobacteria	51.2 ± 14.8	50.7 ± 12.8	44	46
Aquificae	43	44.2	1	1
Bacteroidetes/Chlorobi	49.0 ± 10.4	50.6 ± 7.3	8	9
Betaproteobacteria	62.7 ± 6.4	56.2 ± 4.2	25	25
Chlamydia/Verrucomicrobia	39.6 ± 1.7	39.3 ± 3.2	11	22
Chloroflexi	46 ± 1.4	43.9 ± 1.4	2	4
Cyanobacteria	49.0 ± 10.8	49.6 ± 7.6	17	17
Deinococcus-Thermus	68 ± 1.9	64.0 ± 3.8	4	4
Deltaproteobacteria	55.7 ± 11.8	54.3 ± 8.6	11	14
Epsilonproteobacteria	37.5 ± 5.1	41.2 ± 4.6	9	9
Firmicutes	36.2 ± 6.2	37.1 ± 5.5	79	85
Fusobacteria	27	29.8	1	1
Gammaproteobacteria	47.5 ± 10.5	48.1 ± 8.2	80	122
Planctomycetes	55.4	53.2 ± 0.8	1	2
Spirochaetes	36.0 ± 8.8	39.1 ± 6.8	6	5
Thermotogae	45	44.1	1	1

The model selected by AIC and hierarchical likelihood ratio tests was the general time-reversal model GTR+I+G that calculated the base frequencies (A = 0.2091, C = 0.2522, G = 0.3038, T = 0.2349) and the gamma distribution shape parameter (a = 0.6482), with a proportion of invariable sites (0.2723) (Rodriguez et al. 1990).

The NJ (not shown) and the approximate maximum likelihood phylogenetic (Figure 5) EF-P protein trees were observed to be congruent with each other. Approximate maximum likelihood analyses resulted in one tree (Figure 5) with a log-likelihood value of -29526.80. The minimum evolution score of the NJ tree is 33.47. The major difference between the NJ and the maximum likelihood EF-P tree was dissimilar topology branching patterns seen within the organisms of the same bacterial groupings. Several bacterial groups (Firmicutes, Actinobacteria, Bacteroidetes/Chlorobi, Gammaproteobacteria, Betaproteobacteria, Deltaproteobacteria, and Chamydiae/Verrucomicrobia) within the tree (Figure 5) were paraphyletic with only the Cyanobacteria, Deinococcus/Thermus, Epsilonproteobacteria, Alphaproteobacteria, and Spirochaetes bacterial groups being monophyletic (excluding groups that contained only one or two organism, e.g., Fusobacteria). The phylogenetic groupings (Figure 5) did show similar topologies to the SSU rRNA tree (Figure 8). For example, the Chlamydiae/Verrucomicrobia EF-P group (Figure 5) has the exact same branching order in the SSU rRNA tree (Figure 7) and was supported in all cases by a greater than 70% bootstrap value. However, ignoring EF-P2, the Gammaproteobacteria group in the EF-P tree was split into two separate clusters, one consisting of mostly the *Pseudomonas* spp. and the *Shewanella* spp., while the other cluster consists of the rest of the Gammaproteobacteria used in this study. In comparison to the SSU rRNA tree, the Gammaproteobacteria group is also split into two separate clusters but comprised of different species than the EF-P tree. Also, the Firmicutes group is separated into three clusters in both the

SSU rRNA and the EF-P tree; however, again these clusters show no similarity in the organisms that they contain. Our protein tree suggests several possible HGT events of EF-P2. I tested this hypothesis using the SH test by analyzing an alternative tree topology that minimize the number of HGT events by constraining the topology to show all duplicate EF-P2 as sister taxa to its corresponding EF-P1. The topology of the unconstrained approximate maximum likelihood subset (SH = 0.000) EF-P tree was best supported by the Shimodaira-Hasegawa test compared to the constrained sister taxa subset (SH = 1.000) and the corresponding subset NJ (SH = 0.002) EF-P tree. From the topology testing results, the self-constructed tree with all the EF-P2 as sister taxa to their corresponding EF-P is less likely than the unconstrained tree. Therefore, HGT is a likely explanation for the complex phylogeny of EF-P.

The duplicated EF-Ps from *Porphyromonas gingivalis* W83 appear as sister taxa in Figure 6, with a bootstrap support of 100% and similar branch lengths (approximately less than 0.02 substitutions per site). In addition, the EF-P2 of *Mesorhizobium loti* MAFF303099 and *Rhodobacter sphaeroides* 2.4.1, are apart from their corresponding EF-P1 but are still located within the monophyletic Alphaproteobacteria clade (Figure 7). The branch lengths of EF-P1 and EF-P2 of *M. loti* MAFF303099 were very similar (approximately 0.078 substitutions per site difference) with EF-P2 having a shorter branch length then EF-P1 and vice versa for *R. sphaeroides* 2.4.1. with EF-P2 having a longer branch length than EF-P1 (approximately 0.156 substitution per site).

There were two main groups of EF-P2; one consisting of the Chlamydiae/Verrucomicrobia EF-P2 and the other containing a large cluster of EF-P2 from several different bacterial groups. The large clade of EF-P2 was comprised of EF-P2 from Gammaproteobacteria, Planctomycetes, Deltaproteobacteria, and Acidobacteria, and formed a

Figure 6. Portion of the EF-P tree (Figure 5) showing evidence of gene duplications in *Porphyromonas gingivalis* W83 of the Bacteroidetes/Chlorobi group. Bootstrap values are denoted above the branch.

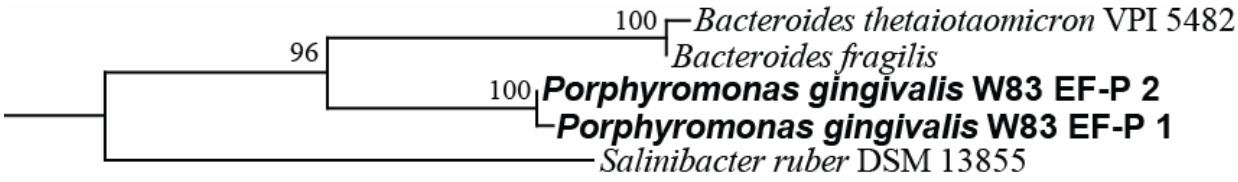
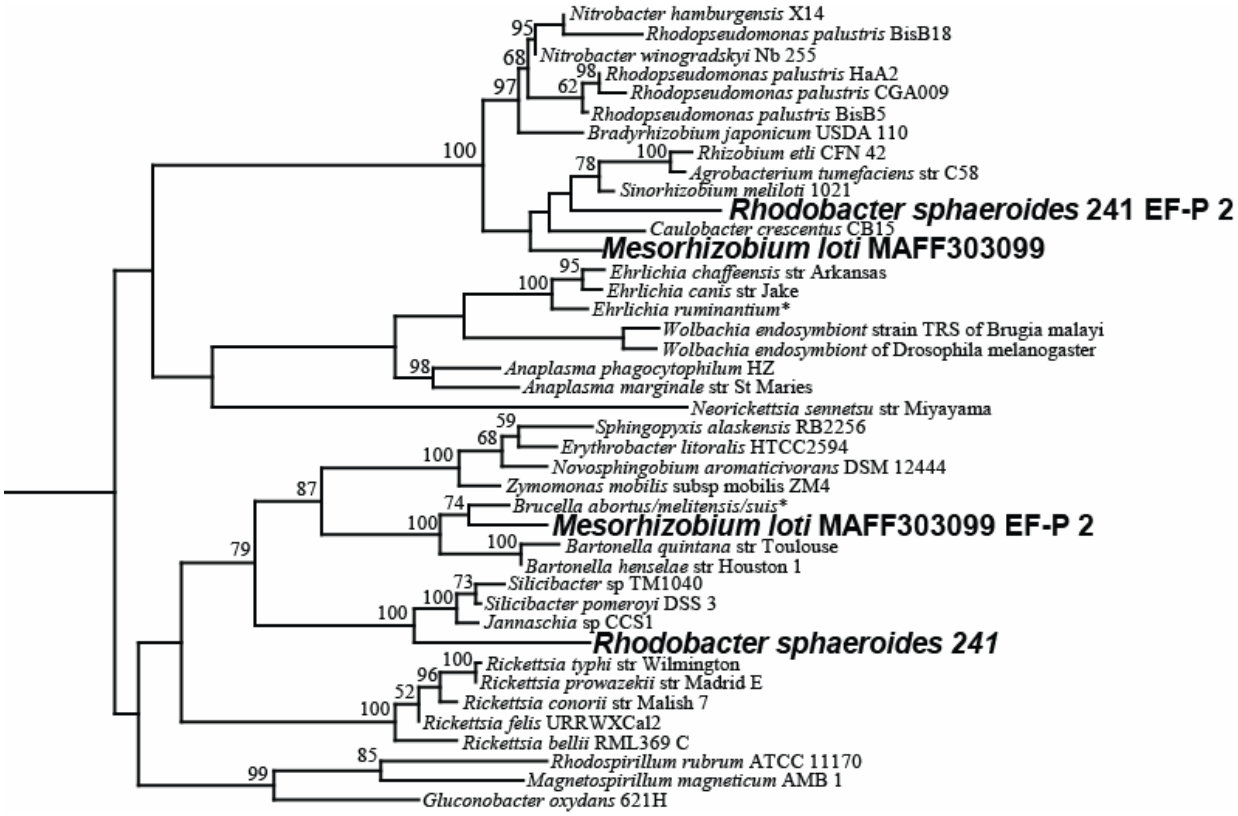
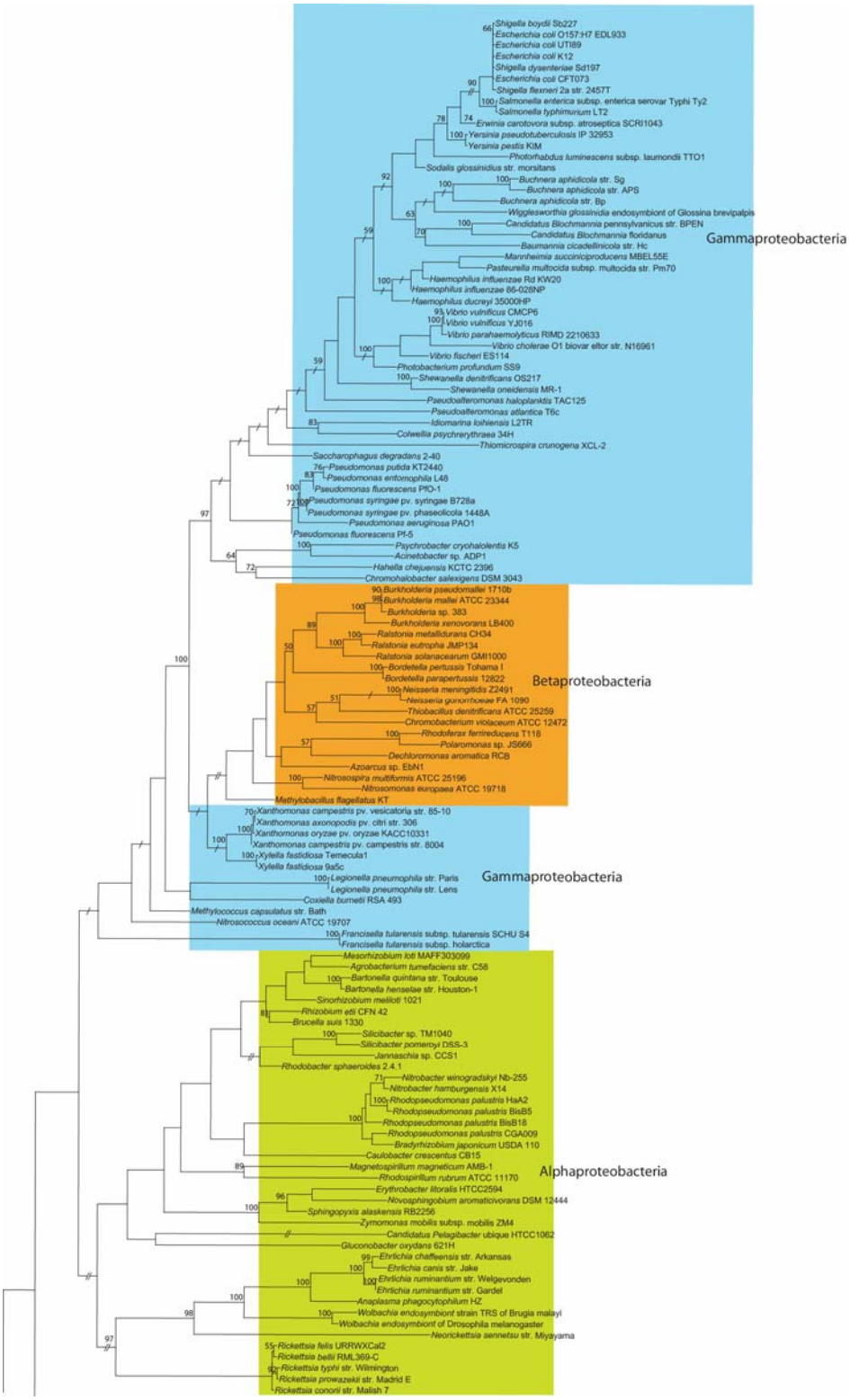

100
Bacteroides thetaiotaomicron VPI 5482
96
Bacteroides fragilis
100
Porphyromonas gingivalis W83 EF-P 2
Porphyromonas gingivalis W83 EF-P 1
Salinibacter ruber DSM 13855

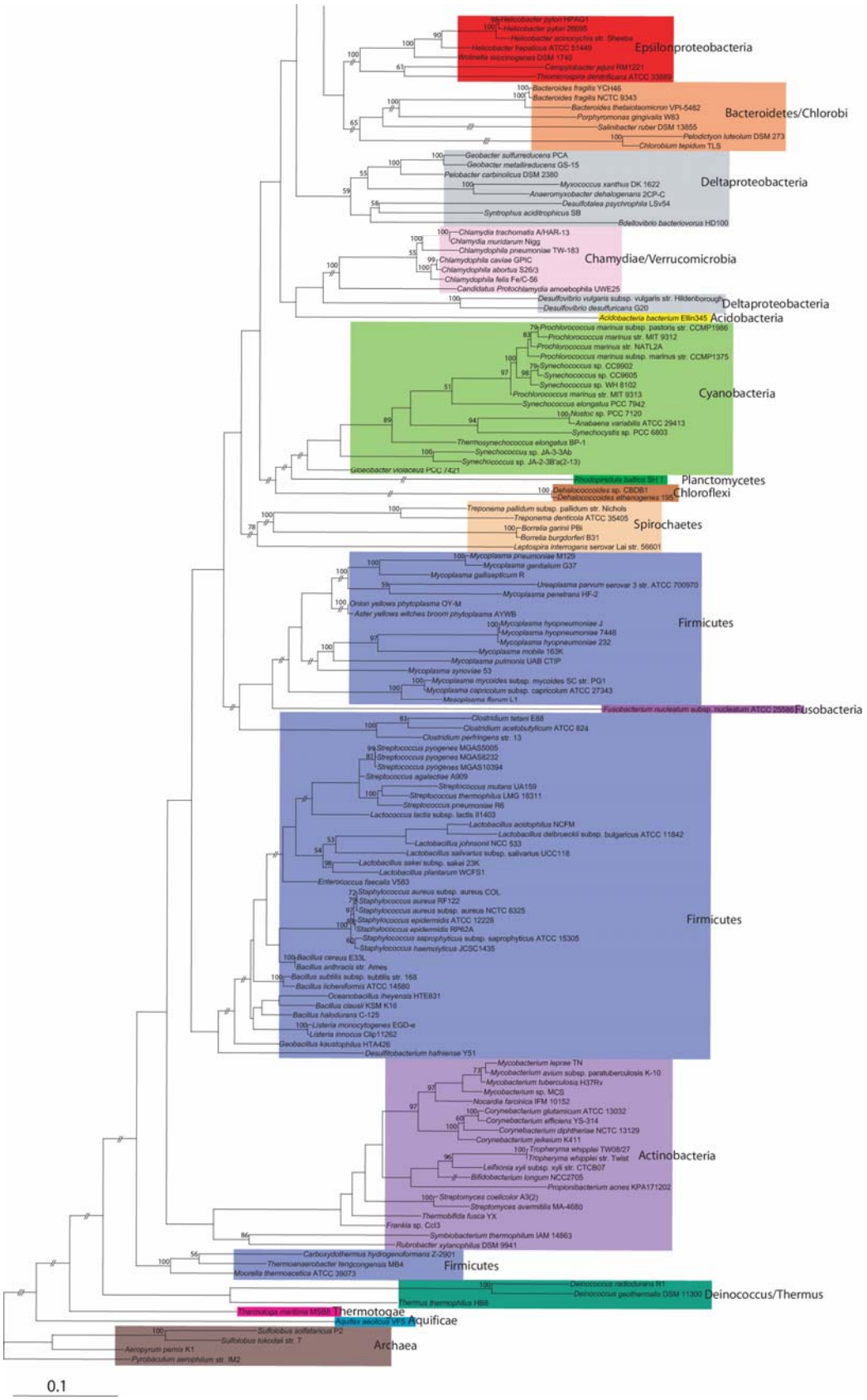
Figure 7. Alphaproteobacteria group from the EF-P tree (Figure 5) showing evidence of gene duplication or HGT. Organisms name followed by an asterisk denote representative sequence. Bootstrap values are denoted above the branch.


Nitrobacter hamburgensis X14
Rhodopseudomonas palustris BisB18
Nitrobacter winogradskyi Nb 255
Rhodopseudomonas palustris HaA2
Rhodopseudomonas palustris CGA009
Rhodopseudomonas palustris BisB5
Bradyrhizobium japonicum USDA 110
Rhizobium etli CFN 42
Agrobacterium tumefaciens str C58
Sinorhizobium meliloti 1021
Rhodobacter sphaeroides 241 EF-P 2
Caulobacter crescentus CB15
Mesorhizobium loti MAFF303099
Ehrlichia chaffeensis str Arkansas
Ehrlichia canis str Jake
Ehrlichia ruminantium*
Wolbachia endosymbiont strain TRS of Brugia malayi
Wolbachia endosymbiont of Drosophila melanogaster
Anaplasma phagocytophilum HZ
Anaplasma marginale str St Maries
Neorickettsia sennetsu str Miyayama
Sphingopyxis alaskensis RB2256
Erythrobacter litoralis HTCC2594
Novosphingobium aromaticivorans DSM 12444
Zymomonas mobilis subsp mobilis ZM4
Brucella abortus/melitensis/suis*
Mesorhizobium loti MAFF303099 EF-P 2
Bartonella quintana str Toulouse
Bartonella henselae str Houston 1
Silicibacter sp TM1040
Silicibacter pomeroyi DSS 3
Jannaschia sp CCS1
Rhodobacter sphaeroides 241
Rickettsia typhi str Wilmington
Rickettsia prowazekii str Madrid E
Rickettsia conorii str Malish 7
Rickettsia felis URRWXCal2
Rickettsia bellii RML369 C
Rhodospirillum rubrum ATCC 11170
Magnetospirillum magneticum AMB 1
Gluconobacter oxydans 621H

monophyletic group with moderate bootstrap support of 70% (Figure 5). These EF-P2 proteins also formed a larger monophyletic group with the Firmicutes EF-P1 and EF-P2, Thermotogae EF-P1, and Chloroflexi EF-P2. However, there was no bootstrap support for the higher order branching. For Chlamydiae/Verrucomicrobia there is similarity in the branching order of the EF-P1 and EF-P2 groups, which was supported by high bootstrap values in most cases. Also, the EF-P2 of the Chlamydiae/Verrucomicrobia had a longer branch length than EF-P1 with a difference range of 0.93 to 1.91 substitutions per site. There were also several cases of putative EF-P HGT events due to incongruencies with the SSU rRNA tree (Figure 5). For example the paraphyletic Betaproteobacteria group consisted of two clades: one comprising of just *Nitrosospira multiformis* ATCC 25196, while the other clade contained the remaining of the Betaproteobacteria species. These two groups were very distant from each other with low bootstrap support of 60% for the monophyletic grouping of the Betaproteobacteria group (without *N. multiformis* ATCC 25196) and *N. multiformis* ATCC 25196 monophyletic with the Alphaproteobacteria, Bacteroidetes/Chlorobi, and Acidobacteria with no bootstrap support (Figure 5); the Betaproteobacteria group on the SSU rRNA tree appears to be monophyletic. The amino acid sequence divergence among the monophyletic Betaproteobacteria group and *N. multiformis* ATCC 25196 for EF-P ranged from 1.3 to 4.2 substitutions per site as shown by the branch lengths (Figure 5). Another example occurs with the EF-P of the Bacteroidetes/Chlorobi paraphyletic group (Figure 5). One clade of the Bacteroidetes/Chlorobi consisted of three species - *Pelodictyon luteolum* DSM 273, *Chlorobium chlorochromatii* CaD3, and *Chlorobium tepidum* TLS and the other clade contained *Bacteroides thetaiotaomicron* VPI 5482, *Bacteroides fragilis*, EF-P and EF-P2 of *Porphyromonas gingivalis* W83, and *Salinibacter ruber* DSM 13855.

Figure 8. The rooted phylogram maximum likelihood SSU rRNA phylogenetic tree with the consensus bootstrap from the NJ tree. The log likelihood score of this tree is -55644.85984. The nucleotide substitution model, GTR+I+G was selected by AIC and hLRTs from 56 DNA models using ModelTest 3.7 (Posada and Crandall 1998) with the following calculated parameters: rate matrix 0.9652 3.1980 1.7968 1.0278 5.1441, gamma shape 0.6645, proportion of invariable sites 0.1179. / denotes removal of a large distance in order to present the tree. Each bacterial group is highlighted in a unique colour. Bootstrap values are denoted above the branch.


Gammaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Alphaproteobacteria


Epsilonproteobacteria
Bacteroidetes/Chlorobi
Deltaproteobacteria
Chamydiae/Verrucomicrobia
Deltaproteobacteria
Acidobacteria
Cyanobacteria
Planctomycetes
Chloroflexi
Spirochaetes
Firmicutes
Fusobacteria
Firmicutes
Actinobacteria
Firmicutes
Deinococcus/Thermus
Thermotogae
Aquificae
Archaea
0.1

However, these species clustered together to form a monophyletic group in the SSU rRNA tree (Figure 8).

The Actinobacteria also form a monophyletic clade in the SSU rRNA tree (Figure 8), which was incongruent with the paraphyletic Actinobacteria group seen in the EF-P tree (Figure 5). Two species from the Actinobacteria group, *Symbiobacterium thermophilum* IAM 14863 and *Rubrobacter xylanophilus* DSM 9941 can be seen clustering with the Firmicutes and Cyanobacteria (for *S. thermophilum* IAM 14863); and Deinococcus/Thermus (*R. xylanophilus* DSM 9941) away from the monophyletic Actinobacteria (bootstrap support of 97%) group with both isolated species having no bootstrap support.

The SSU rRNA tree constructed is the standard for evolutionary relationships in the present study and is depicted in Figure 8. The Deltaproteobacteria, Gammaproteobacteria, and the Firmicutes all appear to be paraphyletic. At the basal position of the SSU rRNA tree (Figure 8) the long branch of the Archaeal group in both the NJ and approximate maximum likelihood tree is likely due to large evolutionary distances (ranges of ~18.5 to ~21.4 substitution per site from the archaeal organisms to the nearest bacterial organism, *Aquifex aeolicus* VF5). However, the SSU rRNA tree had little bootstrap support for the higher order branches similar to the EF-P tree.

3.2 Nucleotide and Amino Acid Sequence Identities

No significantly similar nucleotide or amino acid sequence identity of *efp* was detected among distantly related species (i.e., different clades) when compared to each other. The same general trends can be observed for both the nucleotide and amino acid sequence identities for the *efp*1 and their corresponding *efp2* sequences (Table 2). Both *efp*1 and *efp*2 genes show similar nucleotide and amino acid identity to corresponding genes of closely related species.

Table 2. Nucleotide and amino acid sequence identities of a subset of the Gammaproteobacteria that have multiple *efp* genes. The data is in the form of a sequence identities percentage that covers more than half of the sequence. Any data with sequence identities percentage that does not cover more than half of the sequence is listed as NSS The data in the upper right triangle represent the deduced amino acid sequence identities of EF-P, while the data in the lower left triangle represent the DNA sequence identities of the corresponding *efp* genes. NSS represents no significant identity similarities covering more than half of the sequence. The sequence identities above 70% are highlighted.

Bacterial efp gene	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1 *E. carotovora subsp. atroseptica SCRI1043 efp1*	-	50	90	49	77	48	84	56	81	56	80	56	90	49	90	49	77	55	82	52	77	49	78	50	89	48	81	57
2 *E. carotovora subsp. atroseptica SCRI1043 efp2*	NSS	-	49	87	47	78	46	68	47	68	50	67	49	88	49	87	48	67	48	70	48	59	46	56	50	86	51	65
3 *E. coli K12 efp1*	NSS	NSS	-	51	79	50	86	56	84	57	82	56	99	51	100	50	76	54	87	54	75	52	76	57	95	50	84	57
4 *E. coli K12 efp2*	NSS	78	NSS	-	50	80	47	70	48	70	54	70	51	98	51	99	51	70	49	73	49	61	47	58	51	91	51	69
5 *H. chejuensis KCTC 2396 efp1*	NSS	NSS	NSS	NSS	-	50	82	51	81	53	88	54	80	50	79	49	77	53	80	52	75	49	78	49	78	48	83	53
6 *H. chejuensis KCTC 2396 efp2*	NSS	NSS	NSS	NSS	NSS	-	46	75	47	72	53	72	50	80	50	80	50	74	47	77	46	63	47	62	50	82	51	72
7 *P. profundum SS9 efp1*	NSS	NSS	NSS	NSS	NSS	NSS	-	50	83	52	76	50	81	45	81	45	75	49	86	47	71	50	70	45	78	44	82	50
8 *P. profundum SS9 efp2*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	54	87	58	85	56	70	56	69	49	66	54	77	50	63	51	61	54	73	57	86
9 *P. atlantica T6c efp1*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	60	80	56	84	48	84	47	80	51	89	53	75	48	75	55	83	47	89	56
10 *P. atlantica T6c efp2*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	57	82	54	71	57	69	53	66	58	78	51	64	50	64	53	71	58	84
11 *S. degradans 2 40 efp1*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	58	83	54	82	54	78	51	82	54	77	52	79	57	82	51	81	54
12 *S. degradans 2 40 efp2*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	56	71	56	69	55	63	54	75	49	66	48	62	56	70	58	84
13 *S. typhimurium LT2 efp1*	NSS	NSS	88	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	51	99	50	76	54	87	54	76	52	77	57	95	50	84	57
14 *S. typhimurium LT2 efp2*	NSS	NSS	NSS	87	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	51	98	51	70	49	74	49	61	47	58	51	91	51	69
15 *S. boydii Sb227 efp1*	NSS	NSS	99	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	88	NSS	-	50	76	54	87	54	75	52	76	57	95	50	84	57
16 *S. boydii Sb227 efp2*	NSS	78	NSS	99	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	87	NSS	-	50	69	49	73	48	60	47	57	51	90	50	69
17 *T. crunogena XCL 2 efp1*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	52	79	52	75	51	74	49	76	48	82	53
18 *T. crunogena XCL 2 efp2*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	50	67	52	60	51	58	53	69	52	64
19 *V. vulnificus YJ016 efp1*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	51	75	48	75	54	86	48	86	56
20 *V. vulnificus YJ016 efp2*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	49	62	49	62	55	76	54	77
21 *X. axonopodis pv. citri str. 306 efp1*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	43	85	45	69	44	68	46
22 *X. axonopodis pv. citri str. 306 efp2*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	45	89	51	62	47	65
23 *X. fastidiosa 9a5c efp1*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	46	76	45	75	49
24 *X. fastidiosa 9a5c efp2*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	50	58	52	62
25 *Y. pestis KIM efp1*	NSS	NSS	81	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	82	NSS	81	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	49	83	59
26 *Y. pestis KIM efp2*	NSS	NSS	NSS	80	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	80	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	51	73
27 *C. psychrerythraea 34H efp1*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-	57
28 *C. psychrerythraea 34H efp2*	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	NSS	-

As seen in the tree, the *efp*1 from closely related species have a higher percent identity with each other then with *efp*2 of the same species or in certain cases there is not a significant identity between the *efp*1 and *efp*2. *P. gingivalis* W83 is an exception to the general trends as *efp*1 and *efp*2 in this organism have a high nucleotide and amino acid sequence identity with each other, which is suggestive of a relatively recent gene duplication event. The two Alphaproteobacteria species that each contain two copies of *efp* have high amino acid sequence similarities to other species within the same group such as *Brucella suis* 1330 with *Mesorhizobium loti* MAFF303099 EF-P2 (93%), *Mesorhizobium loti* MAFF303099 EF-P 1 with *Rhodopseudomonas palustris* HaA2 (89%), *Rhodobacter sphaeroides* 2.4.1 EF-P2 with *Sinorhizobium meliloti* 1021 (87%), and *Rhodobacter spaeroides* 2.4.1 EF-P 1 with *Silicibacteria* species (87%) than with each other. These results are consistent with the trends seen in Table 2.

3.3 GC Content Analyses

The mean value of the total GC content of all the *efp* genes was 46.9% ± 10.4% as summarized in Table 1. The majority of the mean values (Figure 9, Table 1) and ranges of the genomic and EF-P GC content are similar to each other. The Betaproteobacteria group has the highest difference in mean genomic and *efp* GC content of 6.5% (Table 1, Figure 9). Their *efp* GC content was more similar to the Deltaproteobacteria genomic GC content. However, these are two closely related bacterial groups.

The difference in percent GC content of the *efp* gene with its corresponding genome is usually observed between the positive and negative controls for HGT, therefore showing a difference in GC content and perhaps evidence of HGT (Figure 10). This may indicate that there are some *efp* genes that have undergone a recent HGT, especially the organisms in the high percent GC difference ranges such as the 10 to 12, and the 12 to 14 ranges as seen in the

Figure 9. Comparison of the genomic and EF-P protein GC content. The red lines represent the range of GC content for the genome of the phylogenetic group denoted on the x-axis where as the green lines represent the range of the GC content for the *efp* proteins. Duplicated *efp* genes were included in this analysis. The blue horizontal line represents the mean value of the GC content for the genome of the phylogenetic group and the purple lines represents the mean value of the GC content for the EF-P protein in its corresponding GC content ranges.

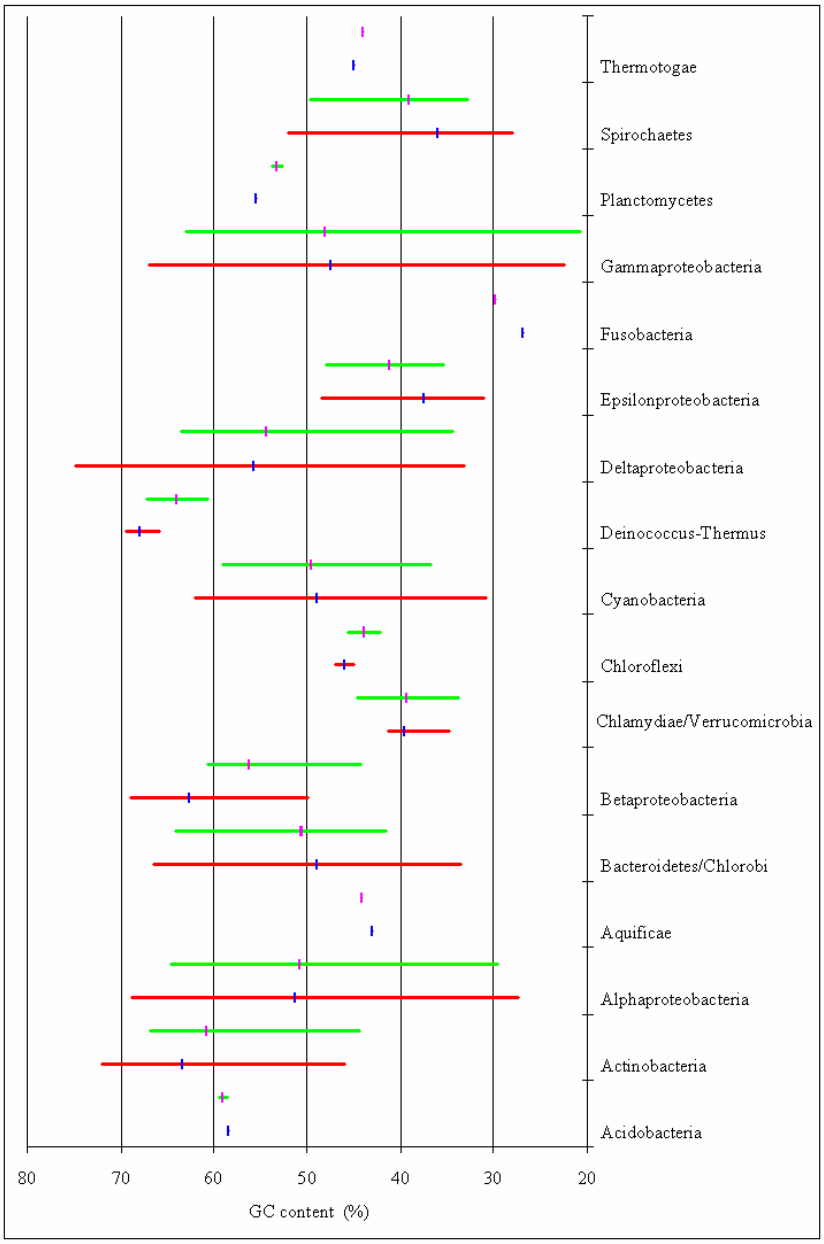
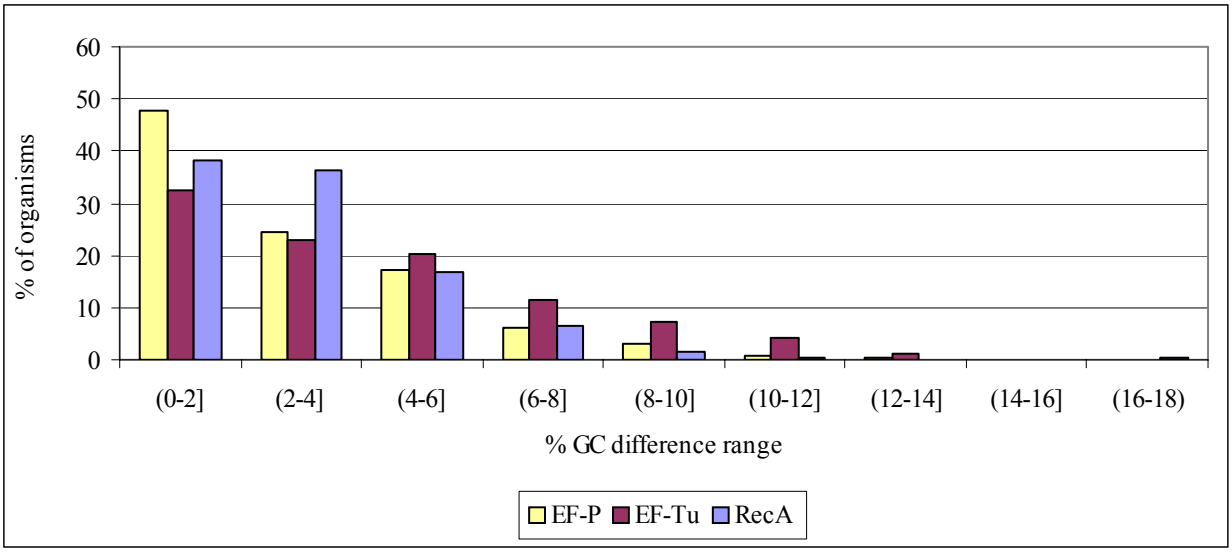

Thermotogae
Spirochaetes
Planctomycetes
Gammaproteobacteria
Fusobacteria
Epsilonproteobacteria
Deltaproteobacteria
Deinococcus-Thermus
Cyanobacteria
Chloroflexi
Chlamydiae/Verrucomicrobia
Betaproteobacteria
Bacteroidetes/Chlorobi
Aquificae
Alphaproteobacteria
Actinobacteria
Acidobacteria
80
70
60
50
40
30
20
GC content (%)

Figure 10. Histogram of the % GC content difference of the completely sequenced bacterial genomes compared with the negative control (recA), positive control (EF-Tu), and EF-P.


60
50
40
30
20
10
0
% of organisms
(0-2]
(2-4]
(4-6]
(6-8]
(8-10]
(10-12]
(12-14]
(14-16]
(16-18)
% GC difference range
EF-P
EF-Tu
RecA

high number of positive controls and low number of negative controls in these ranges (Figure 10). The negative control had only 2 species which have a >10% GC content difference and the positive control had 24 species with >10% GC content difference (Figure 10). Interestingly, the *Enterococcus faecalis* V583 positive control which is hypothesized to have undergone HGT has only a 3.8% GC content difference in comparison to *Pseudomonas fluorscens* Pf-5 which has a highly suspicious >12.5% GC difference compared with a closely related species, *Pseudomonas fluorescens* PfO-1 which has a <5.2% difference. The GC content of the EF-P sequences ranged from 67.2 to 20.7%, while that of the EF-P2 sequences ranged from 63.5 to 33.3%.

3.4 Codon Usage Analyses

The majority of the Hamming distance difference calculated means are in the lower range of values for most organisms (minimum Hamming distance calculated were less than 0.25 of the mean Hamming distance calculated from the *efp* gene's own genome), suggesting that the majority of *efp* genes have the same codon usage as the rest of the genome (Figure 11). The Gammaproteobacteria and the Firmicutes had the most difference between the minimum Hamming distance calculated and the Hamming distance calculated from the corresponding genome (Figure 11). EF-P Hamming difference calculations, compared with the positive and negative control differences shows evidence that some EF-P genes may have undergone HGT events (Figure 12). As expected, the negative control has the highest number of same genome and gene codon usage with EF-P, and the positive control has the least. Hierarchical clustering was employed to determine any atypical clusters of *efp* codon usage. As seen in Table 3 the Jaccard distance and the UPGMA or the unweighted average method created the best tree that represents the original data. Using the inconsistency coefficient to define our cutoff value of 0.8,

Figure 11. The Hamming distances calculated for each *efp* gene from all the bacterial genomes available are shown as ranges from each bacterial group. Only the *efp* genes that did not have a matching minimum Hamming distance with their corresponding genome are shown to determine any unusual Hamming distance. The red horizontal bar represents the mean Hamming distance calculated for the gene using the codon frequencies of its own genome for that bacterial phylogenetic group. Only 264 *efp* genes had a corresponding genomic codon usage observed in the Codon Usage Database (http://www.kazusa.or.jp/codon/).

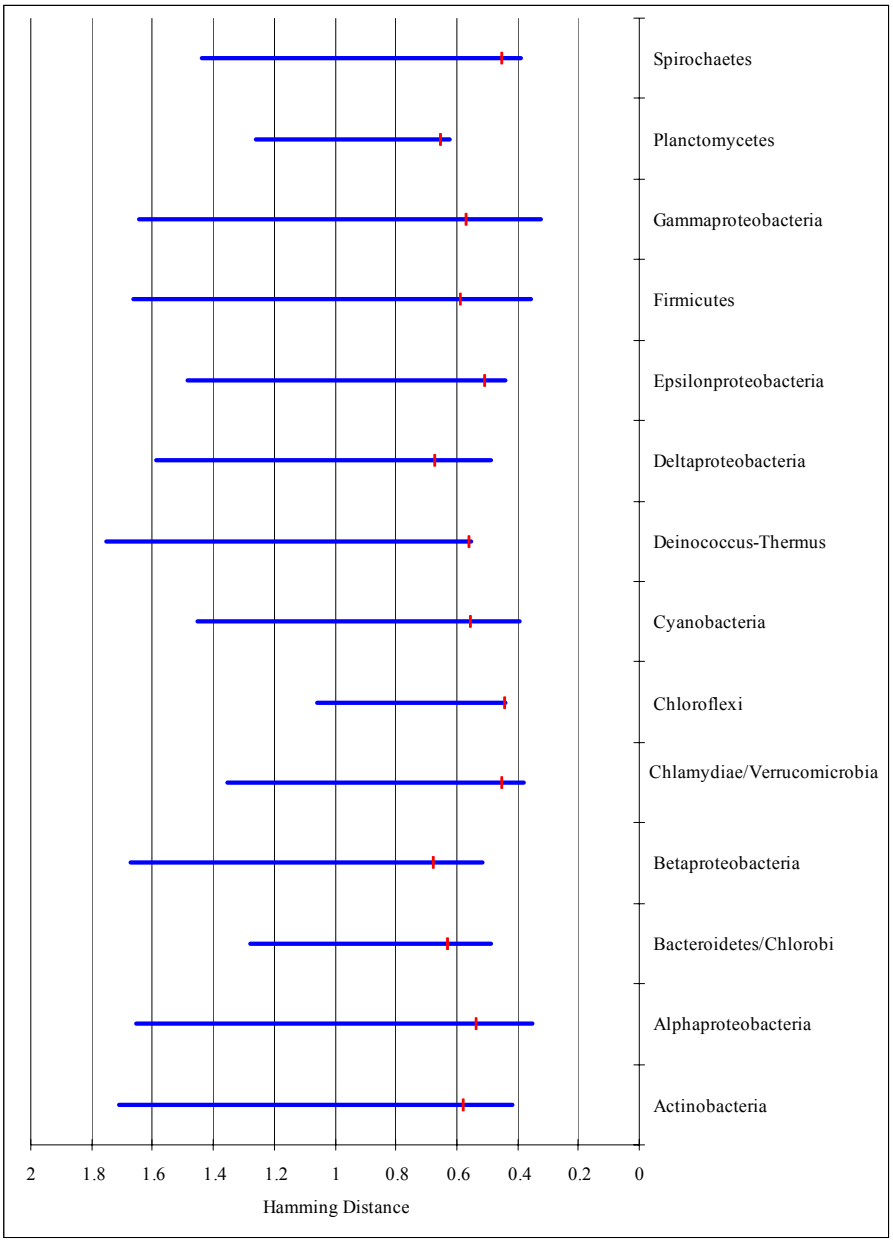

Spirochaetes
Planctomycetes
Gammaproteobacteria
Firmicutes
Epsilonproteobacteria
Deltaproteobacteria
Deinococcus-Thermus
Cyanobacteria
Chloroflexi
Chlamydiae/Verrucomicrobia
Betaproteobacteria
Bacteroidetes/Chlorobi
Alphaproteobacteria
Actinobacteria
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Hamming Distance

Figure 12. The histogram of the % range of organisms with Hamming distance difference. The Hamming distance differences are calculated from the minimum Hamming distance calculated with another genome compared with the Hamming distance of the gene's own genome. If there is a large difference in the two calculations (one using the gene's own genome codon usage vs. using another organism genome's codon usage) it is usually suggestive of a HGT event because the gene's codon usage is more similar to another organism genome than its own genome. EF-Tu is the positive control and recA is the negative control

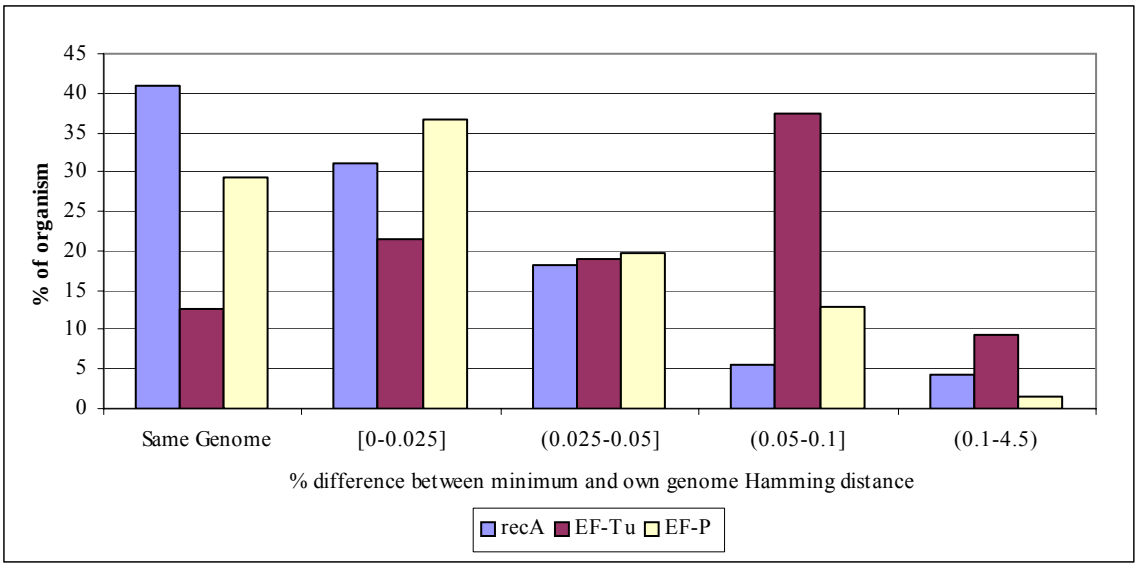

45
40
35
30
25
20
15
10
5
0
% of organism
Same Genome
[0-0.025]
(0.025-0.05]
(0.05-0.1]
(0.1-4.5)
% difference between minimum and own genome Hamming distance
recA
EF-Tu
EF-P

Table 3. The cophenetic correlation coefficient to discover the best representative distance measurement (on the top) and the linkage methods (on the side) of the original data, 61 codon frequencies of the *tuf* and *efp* genes. The shaded box denotes the best representation. Average denotes an unweighted average distance or UPGMA method of creating the tree; centroid distance or UPGMC must use Euclidean distances; and complete denotes the furthest distance method of creating the tree.

	Euclidean	Seuclidean	Cityblock	Mahalanobis	Minkowski	Cosine	Correlation	Hamming	Jaccard
Average	0.7315	0.7111	0.7053	0.8743	0.7315	0.6767	0.7513	0.8356	0.9757
Centroid	NM	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Complete	0.6579	0.5959	0.6079	0.5696	0.6579	0.6581	0.6918	0.6585	0.8716

I get a more natural division of the data set into different clusters. Using this cutoff, the 391 *efp* genes and 438 *tuf* genes are divided into 375 clusters. There are many clusters consisting of only one or two genes. These small clusters denote the outliers in our dataset and/or have unique codon usages. Using this small cutoff, gives us tight clusters which is more stringent and therefore, increases our certainty that the genes that are clustering together have similar codon usage. *Thermobifida fusca* YX *efp*1, *Dechloromonas aromatica* RCB *efp*1, *Xanthomonas oryzae* pv. Oryze KACC10331 *efp*1, *Buchneria aphidicola str*. APS *efp*1, *Salmonella enterica* subsp. enterica serovar Paratyphi A str. ATCC 9150 *efp*2, and *Vibrio cholerae* O1 biovar eltor str. N16961 *efp*2 are clustering with *tuf* genes; the rest of the clusters are exclusively *efp* or *tuf* genes. Most clusters contain species from the same bacterial groups; however, there are several clusters which contain a mix of bacterial groups. Most of the *efp*2 genes are not clustering with the *efp*1 gene from the same genome. Only two organisms, *Chlamydia muridarum* Nigg and *Porphyromonas gingivalis* W83 are discovered with there *efp*1 and *efp*2 in the same cluster.

The following is a list of all the organisms which did not cluster with there closely related Bacterial groups and thus may show evidence of atypical codon usage: *Rodobacter sphaeroides* 2.4.1 *efp*2 is seen clustered with *efp*1 of *Nitrobacter hambrugensis* X14 and *Nitrobacter wingradskyi* Nb-255 from Alphaproteobacteria. *Mesorhizobium loti* MAFF303099 *efp*2 clustered with *Metholococcus capsulatus* str. Bath *efp*1, a Gammproteobacteria (cluster 72). *Pelobacter carbinolicus* DSM 2380 *efp*2 clustered with Actinobacteria, *Mycobacteirum leprae* TN *efp*1; Deltaproteobacteria, *Geobacter metallireducen* GS-15 *efp*1 and *Geobacter sulfurreducens* PCA *efp*1; and Alphaproteobacteria, *Jannaschia* sp. CCS1 *efp*1 and *Silicibacter pomeroyi* DSS-3 *efp*1 (cluster 86). *Colwellia psychererythraea* 34H *efp*2 is clustered with *efp*1 of Alphaproteobacteria, *Bartonella henselae* str. Houston-1 and *Neorickettsia sennetsu* str. Miyayama (cluster 102).

Xylella fastidiosa Temecula1 *efp*2 is clustered with the *efp*1 of species from Alphaproteobacteria and Chloroflexi (cluster 105). *Xylella fastidiosa* 9a5c *efp*2 is clustered with *Lactobacillus delbrueckii* subsp. bulgaricus ATCC 11842 *efp*1 (cluster 113). *Lactobacillus delbrueckii* subsp. bulgaricus ATCC 11842 *efp*2 clustered with *Deinococcus geothermalis* DSM 1130 *efp*1 and *Thermotoga maritime* MSB8 *efp*1 (cluster 115). *Hahella chejuensis* KCTC 2396 *efp*2 is seen clustered with *efp*1 from Alphaproteobacteria (cluster 120). *Candidatus Protochlamydiz ameobophila* UWE25 *efp*2 clustered with the *efp*1 of *Mycoplasma capricolum* and *Mycoplasma mycoides* (cluster 127). *Thiomicrospria crunogena* XCL-2 *efp*2 is clustered with *Thiomicrospira dentirificans* ATCC 33889 (cluster 229). *Lactobacillus efp*2 is clustered with *Colwellia psychrerythraea 34H* and *Clostridum tetani* E88 (cluster 246). *Lactobacillus johnsonii* NCC 533 *efp* 2 is clustered with *efp*2 *Pseudoalteromonas haloplanktis* TAC125 (cluster 247). *Dehalococcoides ethenogenes* 195 *efp*2 and *Dehalococcoides* sp. CBDB1 *efp*2 are clustered with *efp*1 of *Prochlorococcus marinus* (cluster 251).

3.5 Conservation of Gene Order Analyses

Overall, the 10kb adjacent upstream and downstream regions of *efp* are not conserved amongst all the bacterial species used in this thesis. However, a small number of distantly related bacterial species which have a similar gene order was discovered (Figure 17). There were not many similar identity matches between the adjacent upstream and downstream region of *efp* which may be due to the stringency of the filters used in our BLAST BL2SEQ search. In addition, BL2SEQ may not be able to pick out relatively small alignments very well within large regions of sequence comparison. In addition, a number of the small matches (less than 75bps) are similar domains discovered in different non orthologous genes. However, some of the small matching regions (< 75bp) that have similar identities may actually represent orthologous genes

that have very low sequence similarities. In the following comparisons, the organisms that had significant similarities within the Proteobacteria group because they are known to have a close evolutionary relationship. There are also many organisms which have numerous small BL2SEQ hits all over their upstream and/or downstream region. I will define here that significant matches are greater than 75bps or species that are not in the same bacterial group and have more than 3 BL2SEQ results. Because of the limited gene order conservation in Bacteria, the conservation of three genes in a row between distinctly related organisms is statistically significant unless the genes are part of an operon (Wolf et al. 2001).

3.5.1 Gene Order of efp1 and efp2

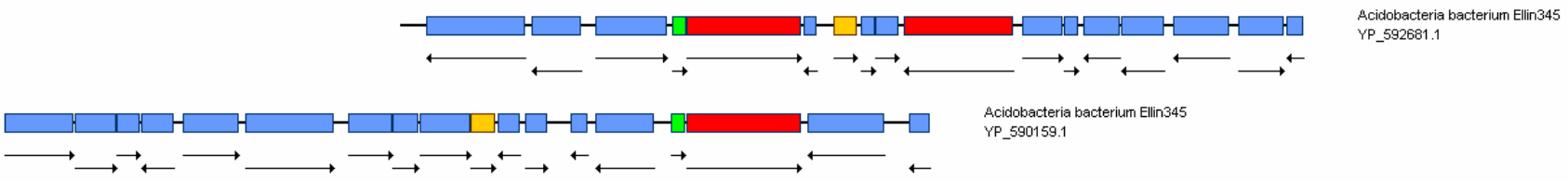
The general trend seen for the organisms that contain two copies of *efp* is that the gene order of closely related species for *efp*1 has a similar gene order. The *efp*2 of closely related species also have gene order similarities and most do not match with the gene order of their corresponding *efp*1 adjacent regions. Most of the *efp*1 and *efp*2 of the Gammaproteobacteria, Firmicutes, Chloroflexi and the Chlamydiae/Verrucomicrobia adhere to this trend. However, Table 4 contains the exceptions to the general trend, where the *efp*1 and *efp*2 of the same organism have similar gene order.

The *efp*1 of *Escherichia coli* CFT073 had similar gene order (upstream and downstream) with other *efp*1 organisms from Gammaproteobacteria. However, the *efp*2 of *E. coli* CFT073 has similar gene order to *efp*2 in most of the Gammaproteobacteria (upstream and downstream) such as *Erwinia* spp., *E. coli* spp., *Salmonella enterica* spp., *Shigella* spp., except for *Yersinia pestis*, and *Yersinia pseudotuberculosis* where the gene order only shares similarities in the upstream or downstream region but not both, considering that the *E. coli* and *Yersinia* spp. are closely related. Interestingly, *Hahella chejuensis* KCTC 2396 *efp*2 does not have similar gene order to *efp*2

Table 4. Organisms that have similar gene order in *efp*1 with *efp*2. The gene region denotes the gene to the left of it and the regions matching to the region of the gene to the right. For example, the *Shigella dysenteriae* Sd197 *efp*1 upstream region has a significant similarity to *efp*2 downstream region and *efp*1 downstream region has a significant similarity to *efp*2 downstream region.

Organism	Group	gene	Region matching	gene	Region matching	Strength of Similarity	Beside EF-P?
Acidobacteria bacterium Ellin345	Acidobacteria	*efp2*	downstream	*efp1*	upstream	Significant Similarity	No
Porphyromonas gingivalis W83	Bacteroidetes/Chlorobi	*efp1*	upstream, downstream	*efp2*	downstream, upstream	Significant Similarity, Significant Similarity	Yes
Shigella boydii Sb227	Gammaproteobacteria	*efp2*	downstream	*efp1*	upstream	Significant Similarity	No
Shigella dysenteriae Sd197	Gammaproteobacteria	*efp1*	upstream, downstream	*efp2*	downstream, downstream	Significant Similarity, Significant Similarity	No
Shigella flexneri 2a str. 2457T	Gammaproteobacteria	*efp1*	upstream, downstream	*efp2*	upstream, downstream	Significant Similarity, Significant Similarity	No
Shigella flexneri 2a str. 301	Gammaproteobacteria	*efp2*	upstream, upstream	*efp1*	upstream, downstream	Significant Similarity, Significant Similarity	No
Shigella sonnei Ss046	Gammaproteobacteria	*efp1*	upstream	*efp2*	upstream	Significant Similarity	No
Xanthomonas campestris pv. vesicatoria str. 85-10	Gammaproteobacteria	*efp2*	downstream	*efp1*	upstream	Small Similarity	No
Xanthomonas oryzae pv. oryzae KACC10331	Gammaproteobacteria	*efp1*	upstream	*efp2*	downstream	Significant Similarity	No
Xanthomonas oryzae pv. oryzae MAFF 311018	Gammaproteobacteria	*efp2*	downstream	*efp1*	upstream	Significant Similarity	No

Figure 13. *Acidobacteria bacterium* Ellin345 *efp*1 and *efp*2 gene order. The green represents PadR-like family (YP_590164.1, YP_592678.1), the red gene represents ABC efflux, inner membrane subunit (YP_590165.1, YP_592679.1, YP_592684.1), and the yellow gene represents *efp*.


Acidobacteria bacterium Ellin345
YP_592681.1
Acidobacteria bacterium Ellin345
YP_590159.1

from any organisms. It only has a small match with the *efp*1 of *Silicibacter* sp. TM1040 upstream region away from *efp*. The gene order that is conserved between *efp*1s and between *efp*2s in the Chlamydiae/Verrucomicrobia group is sometimes inverted. For example, the gene order for the adjacent regions of *efp*1 is inverted in *Chlamydophila caviae* GPIC and *Chlamydia trachomatis* D/UW-3/CX.

Acidobacteria bacterium Ellin345 *efp*1 (YP_590159.1) downstream region is similar to its *efp*2 (YP_592681.1) adjacent upstream region as seen in Figure 13. The area that they are similar (74% similarity identity) is in a region of a gene that is described as a transcriptional regulator called a PadR-like family in both *efp*1 and *efp*2 with an ABC efflux, inner membrane subunit, beside it. In addition, the ABC efflux, inner membrane subunit appears to be encoded twice, once in the upstream and once in the downstream adjacent regions of *efp*2.

Figure 14 compares the gene order in *efp*1 and *efp*2 of *Porphyromonas gingivalis* W83. They have a similar gene, DNA-binding protein, histone-like family in the upstream region of *efp*1 and the downstream region of *efp*2. In addition, the small gene (highlighted in green) in the upstream region of *efp*2 (NP_905456.1) is similar to the non-coding downstream region beside *efp*1 (NP_904857.1). Figure 15 shows the different regions of similarities between distantly related species and how they interact with each other.

3.5.2 Ribosomal Protein and Errors in GenBank

One of the most significant matches that had distantly related Bacteria with regions of similar identity covering greater than 100bps was between the organisms, *Idiomarina loihiensis* L2TR *efp*1 (YP_156657.1) from Gammaprotebacteria; *Ureaplasma parvum serovar* 3 str. ATCC 700970 *efp*1 (NP_078132.1) from Firmicutes; *Porphyromonas gingivalis* W83 *efp*2 (NP_905456.1) from Bacteroidetes/Chlorobi; and *Chlamydia trachomatis* D/UW-3/CX *efp*1

Figure 14. *Porphyromonas gingivalis* W83 *efp*1 and *efp*2 gene order. The red gene represents a DNA-binding protein, histone-like family (NP_905457.1, NP_904856.1), the red arrow denotes where the region of similarity, note that even the noncoding regions are similar, the green gene represents a hypothetical protein NP_905455.1, and the yellow gene represents *efp.*

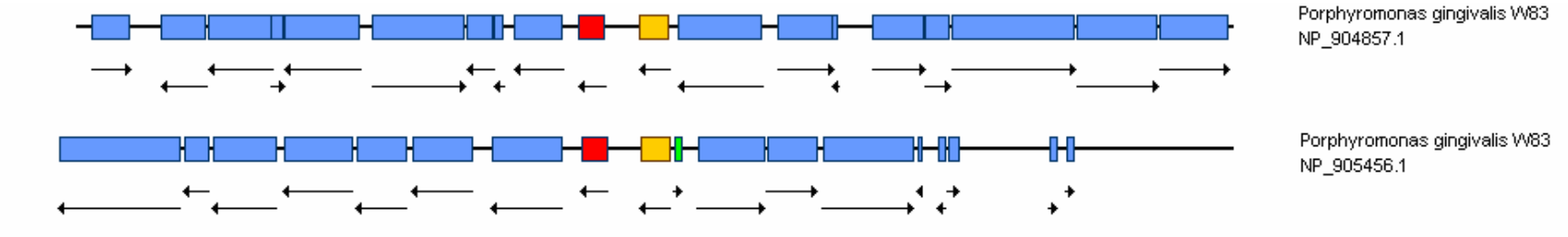
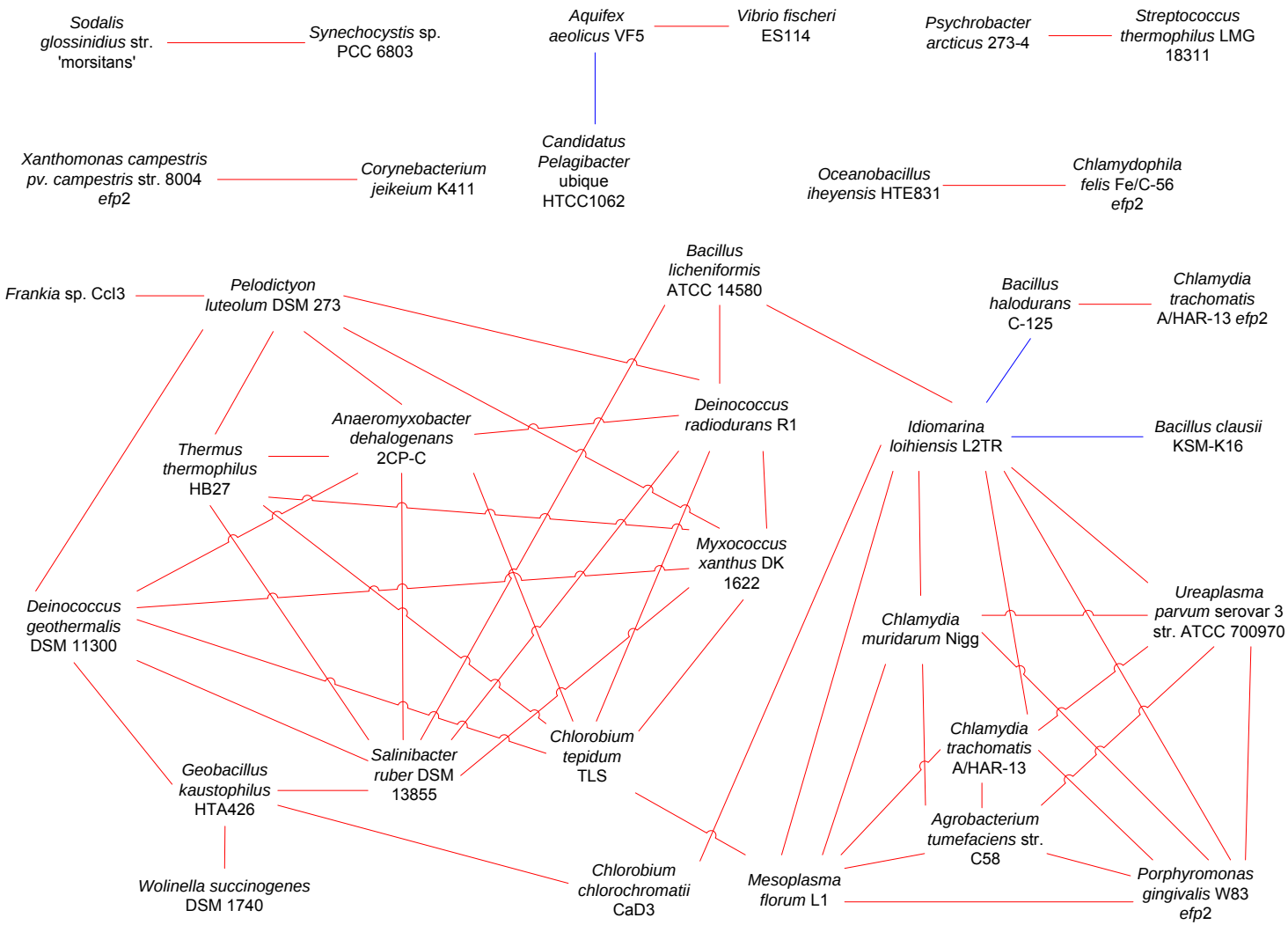

Porphyromonas gingivalis W83
NP_904857.1
Porphyromonas gingivalis W83
NP_905456.1

Figure 15. A network diagram overview of organisms that have significant similarities with distantly related bacterial organisms. These include similarities that are in any adjacent 10kb upstream and downstream of *efp*. Red lines denote more than or equal to 3 locations of similarity, and blue lines denote one match with length of similarity greater than 75bp. Organisms with one matches of less than 75bp similarity in length with distantly related species is not shown and organisms with significant similarities within the same group are also not shown.


Sodalis glossinidius str. 'morsitans'
Synechocystis sp. PCC 6803
Aquifex aeolicus VF5
Vibrio fischeri ES114
Psychrobacter arcticus 273-4
Streptococcus thermophilus LMG 18311
Xanthomonas campestris pv. campestris str. 8004 efp2
Corynebacterium jeikeium K411
Candidatus Pelagibacter ubique HTCC1062
Oceanobacillus iheyensis HTE831
Chlamydophila felis Fe/C-56 efp2
Frankia sp. CcI3
Pelodictyon luteolum DSM 273
Bacillus licheniformis ATCC 14580
Bacillus halodurans C-125
Chlamydia trachomatis A/HAR-13 efp2
Deinococcus radiodurans R1
Anaeromyxobacter dehalogenans 2CP-C
Thermus thermophilus HB27
Idiomarina loihiensis L2TR
Bacillus clausii KSM-K16
Myxococcus xanthus DK 1622
Deinococcus geothermalis DSM 11300
Ureaplasma parvum serovar 3 str. ATCC 700970
Chlamydia muridarum Nigg
Chlamydia trachomatis A/HAR-13
Chlorobium tepidum TLS
Salinibacter ruber DSM 13855
Geobacillus kaustophilus HTA426
Agrobacterium tumefaciens str. C58
Wolinella succinogenes DSM 1740
Chlorobium chlorochromatii CaD3
Mesoplasma florum L1
Porphyromonas gingivalis W83 efp2

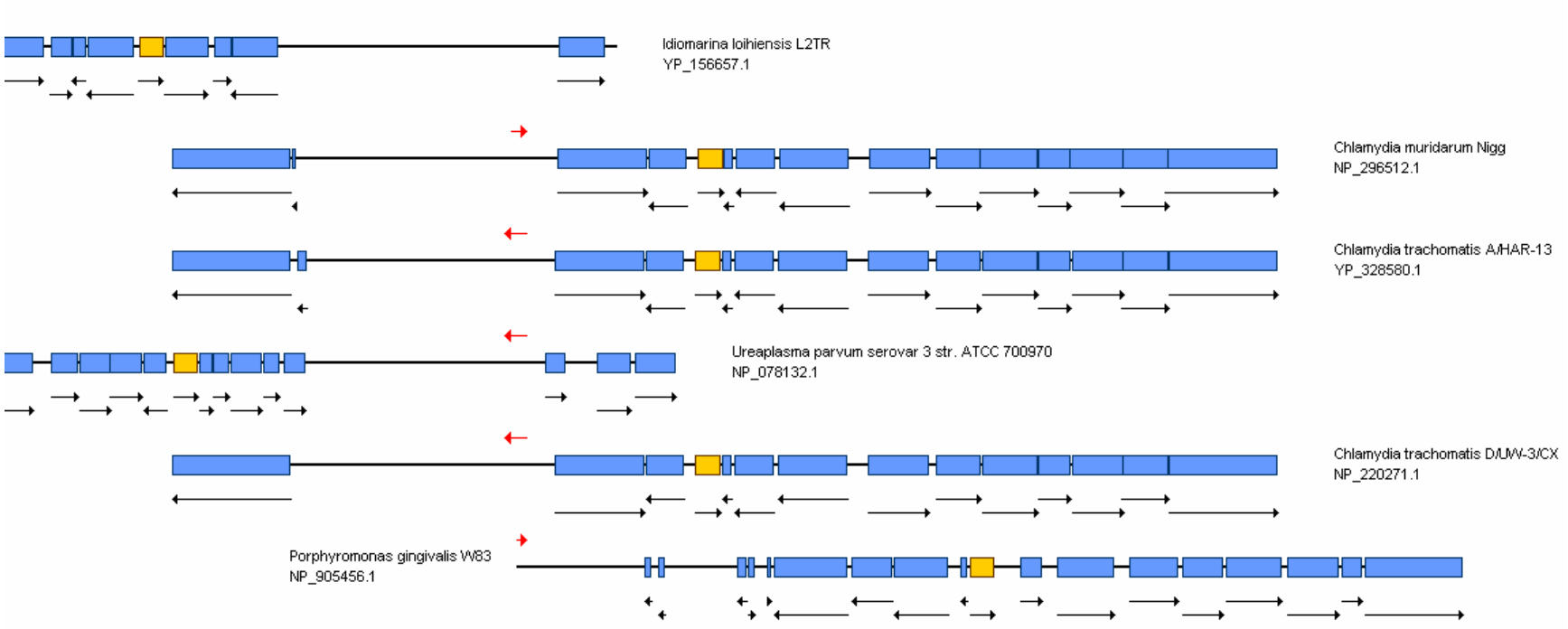
(NP_220271.1), *Chlamydia trachomatis* A/HAR-13 *efp*1 (YP-328580.1), and *Chlamydia muridarum* Nigg *efp*1 (NP_296512.1) from Chlamydiae/Verrucomicrobia. See Figure 16 for a graphical output of the regions of similarity.

Unfortunately within the downloaded GenBank information there are no genes annotated in the region of similarity seen in Figure 16 as denoted by the empty line. However, in the online NCBI gene browser there are three genes in this region. According to the NCBI's gene browser, the gene in the region of similarity is a 23S ribosomal RNA, which is flanked by 5S to the right and a 16S ribosomal RNA to the left. In the BL2SEQ results the other ribosomal RNA also has significant matches in these organisms. *Gluconobacter oxydans* 621H *efp*1 (YP_190675.1) and *Agrobacterium tumerfaciens* str. C58 *efp*1 (NP_533218.1) also have a 16S rRNA in common; however, the 23S and 5S rRNA ribosomal is out of range of the 10kb downstream region.

3.5.3 Conservation of efp, accB, and accC Gene Order

The acetyl-CoA carboxylase, biotin carboxyl carrier protein (*accC*) and acetyl-CoA carboxylase, biotin carboxylase (*accB*) genes were discovered within the 10 kb adjacent upstream and downstream region of *efp* amongst the organisms used in this thesis (Table 5). As seen in Figure 17, *accB*, *accC*, and *efp* are discovered in a similar gene order in organisms from distantly related Bacteria. The gene order observed in the majority of organisms is *efp*, *accB*, and *accC*. Figure 18 show some of the organisms that have the genes *accB* and *accC* in the adjacent regions of *efp* but not adjacent to the gene itself, including some species from Figure 17 for comparison. When comparing Figure 17 and Figure 18 we can see that the *accBC* operon is in the adjacent upstream and downstream region of *efp* gene in some species. In addition, there are some instances of just the *accB* gene being adjacent to the *efp* gene in Table 5.

Figure 16. Missing ribosomal genes in GenBank downloaded files. Each line represents the 10kb upstream and downstream region of *efp* for that organism. The blue squares denote genes, the yellow square denotes the *efp* gene, and the black lines represent non-coding regions. The black arrows underneath the genes denote the direction of transcription. The red arrows between each of the gene regions denotes where the similarity is.


Idiomarina loihiensis L2TR
YP_156657.1
Chlamydia muridarum Nigg
NP_296512.1
Chlamydia trachomatis A/HAR-13
YP_328580.1
Ureaplasma parvum serovar 3 str. ATCC 700970
NP_078132.1
Chlamydia trachomatis D/UW-3/CX
NP_220271.1
Porphyromonas gingivalis W83
NP_905456.1

3.5.4Presence of Insertion Sequence Elements

Interestingly, the *efp*1 and *efp*2 of the *Shigella* spp. are in most cases surrounded by insertion sequence (IS) elements from the IS1 family as seen in Figure 19. The only *Shigella* sp. that does not encode an IS element in the upstream and downstream adjacent regions of *efp* is *Shigella sonnei* Ss046 *efp*2. Other genomes that have IS elements in the 10kb upstream and downstream adjacent regions of *efp* can be seen in Table 6, not including the *Shigella* spp. in Figure 19.

3.6 Motif Prediction

Some general trends that can be seen in Figure 20 are that the two highest p-values (the worst scoring predicted motifs) occur for the defined maximum number of clusters of 15 and 20 when using no SSU rRNA distance information. The defined maximum number of cluster of 10 has the lowest p-values for all the different groups except for the clustering based on all the characteristics. However, the clustering not based on the genomic size and the clustering not based on environmental habitats have only slightly higher p-values for there defined maximum number of clusters of 10 compared with the defined maximum number of clusters of 15. There appears to be a rough correlation of the smaller the defined maximum cluster the lower the p-value. In addition, the following contains the lowest to highest (best to worst) p-values groups when averaging over the different defined maximum number of clusters and ignoring the different characteristic clustering: 10, 15, and 20. On average the no orthologous transcription factor characteristic group had the lowest p-values; therefore, it appears using orthologous transcription factors decrease the chances of clustering species that have similar regulation for *efp*. The following contains the lowest to highest (best to worst) p-values groups when averaging

Figure 17. Organisms which have a conserved gene order for *accC*, *accB*, and *efp*. *accC* is represented in red and *accB* is represented in green, and the *efp* is represented in yellow..

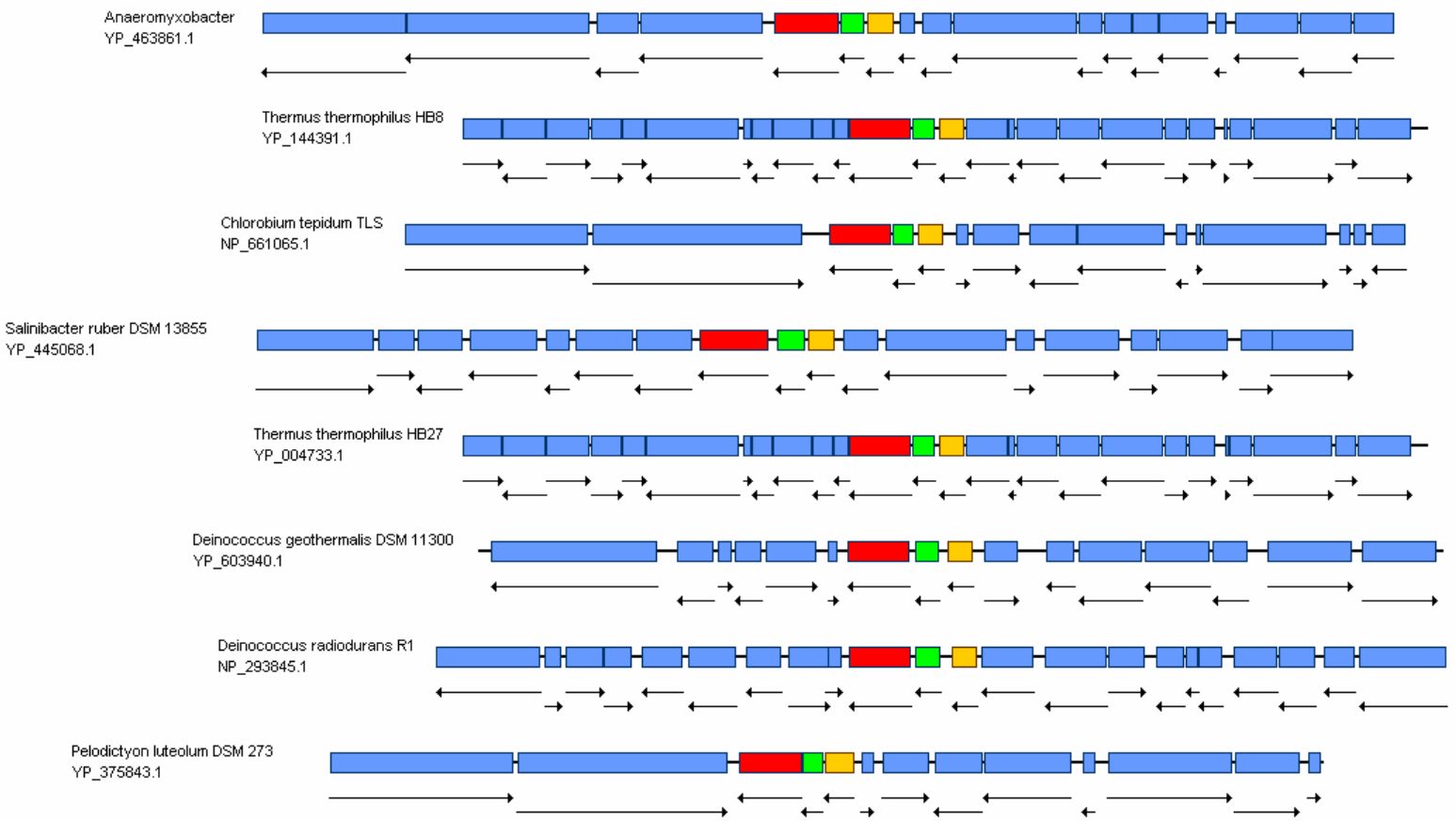

Anaeromyxobacter
YP_463861.1
Thermus thermophilus HB8
YP_144391.1
Chlorobium tepidum TLS
NP_661065.1
Salinibacter ruber DSM 13855
YP_445068.1
Thermus thermophilus HB27
YP_004733.1
Deinococcus geothermalis DSM 11300
YP_603940.1
Deinococcus radiodurans R1
NP_293845.1
Pelodictyon luteolum DSM 273
YP_375843.1

Figure 18. Organisms which contain acetyl-CoA carboxylase, biotin carboxyl carrier protein (*accC*) and acetyl-CoA carboxylase (*accB*) that is not adjacent to *efp*. This diagram includes some organism from Figure 15 that has the *accC*, *accB*, and *efp* gene together to give perspective. *accC* is represented in red and *accB* is represented in green, and the *efp* is represented in yellow

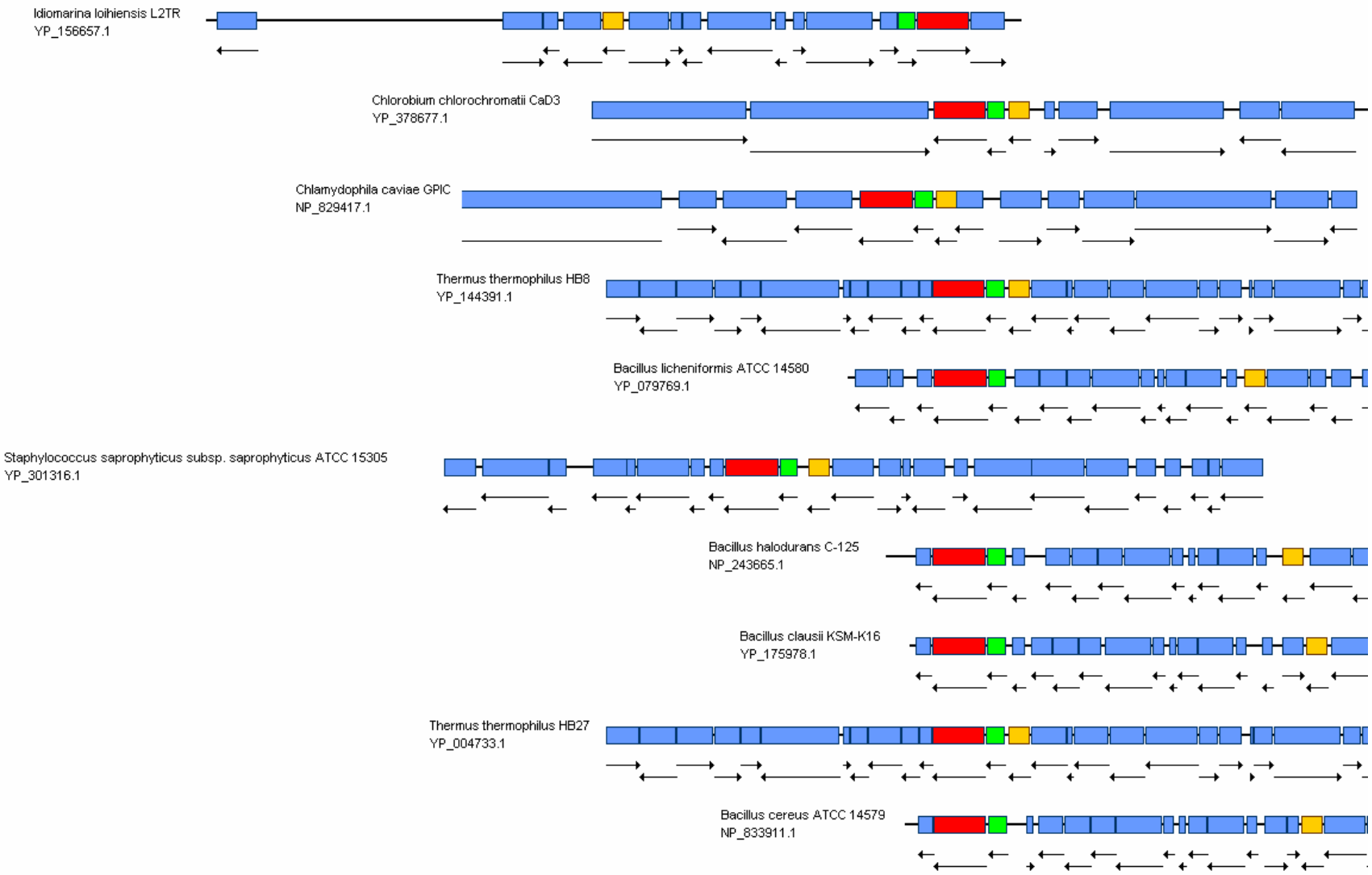

Idiomarina loihiensis L2TR
YP_156657.1
Chlorobium chlorochromatii CaD3
YP_378677.1
Chlamydophila caviae GPIC
NP_829417.1
Thermus thermophilus HB8
YP_144391.1
Bacillus licheniformis ATCC 14580
YP_079769.1
Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305
YP_301316.1
Bacillus halodurans C-125
NP_243665.1
Bacillus clausii KSM-K16
YP_175978.1
Thermus thermophilus HB27
YP_004733.1
Bacillus cereus ATCC 14579
NP_833911.1

Table 5. Organisms that have an annotated *accB* and/or *accC* in the 10kb adjacent upstream and downstream region of *efp*.

Organism	Organism Group	EF-P	*acc* Genes Discovered	Adjacent *acc* Upstream Gene	Adjacent *acc* Downstream Gene
Anabaena variabilis ATCC 29413	Cyanobacteria	*efp1*	*accB*	Transcriptional Regulator, ArsR family	elongation factor P
Anaeromyxobacter dehalogenans 2CP-C	Deltaproteobacteria	*efp1*	*accC, accB*	elongation factor P	tetratricopeptide repeat protein
Bacillus anthracis str. Ames	Firmicutes	*efp1*	*accC, accB*	hypothetical protein	stage III sporulation protein AH
Bacillus anthracis str. 'Ames Ancestor'	Firmicutes	*efp1*	*accC, accB*	hypothetical protein	stage III sporulation protein AH
Bacillus anthracis str. Sterne	Firmicutes	*efp1*	*accC, accB*	hypothetical protein	stage III sporulation protein AH
Bacillus cereus ATCC 10987	Firmicutes	*efp1*	*accC, accB*	hypothetical protein	stage III sporulation protein AH
Bacillus cereus ATCC 14579	Firmicutes	*efp1*	*accC, accB*	hypothetical Cytosolic Protein	hypothetical protein
Bacillus cereus E33L	Firmicutes	*efp1*	*accC, accB*	hypothetical protein	stage III sporulation protein AH
Bacillus clausii KSM-K16	Firmicutes	*efp1*	*accC, accB*	hypothetical protein	hypothetical protein
Bacillus halodurans C-125	Firmicutes	*efp1*	*accC, accB*	hypothetical protein	hypothetical protein
Bacillus licheniformis ATCC 14580	Firmicutes	*efp1*	*accC, accB*	conserved protein YqhY	stage III sporulation protein AH
Bacillus subtilis subsp. subtilis str. 168	Firmicutes	*efp1*	*accC, accB*	hypothetical protein	stage III sporulation protein AH
Bacillus thuringiensis serovar konkukian str. 97-27	Firmicutes	*efp1*	*accC, accB*	hypothetical protein	stage III sporulation protein AH
Chlamydia muridarum Nigg	Chlamydiae/Verruco microbia	*efp2*	*accC, accB*	elongation factor P	50S ribosomal protein L13
Chlamydia trachomatis A/HAR-13	Chlamydiae/Verruco microbia	*efp2*	*accC, accB*	elongation factor P	50S ribosomal protein L13
Chlamydia trachomatis D/UW-3/CX	Chlamydiae/Verruco microbia	*efp2*	*accC, accB*	elongation factor P	50S ribosomal protein L13
Chlamydophila abortus S26/3	Chlamydiae/Verruco microbia	*efp2*	*accC, accB*	elongation factor P	hypothetical protein
Chlamydophila caviae GPIC	Chlamydiae/Verruco microbia	*efp2*	*accC, accB*	elongation factor P	hypothetical protein
Chlamydophila pneumoniae AR39	Chlamydiae/Verruco microbia	*efp2*	*accC, accB*	elongation factor P	hypothetical protein

Organism	Organism Group	EF-P	*acc* Genes Discovered	Adjacent *acc* Upstream Gene	Adjacent *acc* Downstream Gene
Chlamydophila pneumoniae CWL029	Chlamydiae/Verrucomicrobia	*efp2*	*accC, accB*	elongation factor P	hypothetical protein
Chlamydophila pneumoniae J138	Chlamydiae/Verrucomicrobia	*efp2*	*accC, accB*	elongation factor P	hypothetical protein
Chlamydophila pneumoniae TW-183	Chlamydiae/Verrucomicrobia	*efp2*	*accC, accB*	elongation factor P	hypothetical protein
Chlorobium chlorochromatii CaD3	Bacteroidetes/Chlorobi	*efp1*	*accC, accB*	elongation factor P	DNA-directed RNA polymerase beta' subunit
Chlorobium tepidum TLS	Bacteroidetes/Chlorobi	*efp1*	*accC, accB*	elongation factor P	DNA-directed RNA polymerase beta' subunit
Deinococcus geothermalis DSM 11300	Deinococcus-Thermus	*efp1*	*accC, accB*	elongation factor P	hypothetical protein
Deinococcus radiodurans R1	Deinococcus-Thermus	*efp1*	*accC, accB*	elongation factor P	hypothetical protein
Geobacillus kaustophilus HTA426	Firmicutes	*efp1*	*accC, accB*	hypothetical protein	stage III sporulation protein AH
Gloeobacter violaceus PCC 7421	Cyanobacteria	*efp1*	*accB*	elongation factor P	hypothetical protein
Idiomarina loihiensis L2TR	Gammaproteobacteria	*efp1*	*accC, accB*	3-dehydroquinate dehydratase	ribosomal protein L11 methyltransferase
Listeria innocua Clip11262	Firmicutes	*efp1*	*accC*	hypothetical protein	hypothetical protein
Listeria monocytogenes EGD-e	Firmicutes	*efp1*	*accC*	hypothetical protein	hypothetical protein
Listeria monocytogenes str. 4b F2365	Firmicutes	*efp1*	*accC, accB*	elongation factor P	hypothetical protein
Moorella thermoacetica ATCC 39073	Firmicutes	*efp1*	*accC, accB*	PilT protein-like	protein of unknown function
Myxococcus xanthus DK 1622	Deltaproteobacteria	*efp1*	*accC, accB*	elongation factor P	TPR domain protein
Oceanobacillus iheyensis HTE831	Firmicutes	*efp1*	*accC, accB*	hypothetical protein	stage III sporulation protein AH
Pelodictyon luteolum DSM 273	Bacteroidetes/Chlorobi	*efp1*	*accC, accB*	elongation factor P	DNA-directed RNA polymerase beta' subunit
Prochlorococcus marinus str. MIT 9312	Cyanobacteria	*efp1*	*accB*	elongation factor P	4-hydroxythreonine-4-phosphate dehydrogenase

Organism	Organism Group	EF-P	*acc* Genes Discovered	Adjacent *acc* Upstream Gene	Adjacent *acc* Downstream Gene
Prochlorococcus marinus str. NATL2A	Cyanobacteria	*efp1*	*accB*	elongation factor P	4-hydroxythreonine-4-phosphate dehydrogenase
Salinibacter ruber DSM 13855	Bacteroidetes/Chlorobi	*efp1*	*accC*, *accB*	elongation factor P	aspartate aminotransferase
Staphylococcus aureus RF122	Firmicutes	*efp1*	*accC*, *accB*	elongation factor P	hypothetical protein
Staphylococcus aureus subsp. aureus MRSA252	Firmicutes	*efp1*	*accC*, *accB*	elongation factor P	hypothetical protein
Staphylococcus aureus subsp. aureus MSSA476	Firmicutes	*efp1*	*accC*, *accB*	elongation factor P	hypothetical protein
Staphylococcus aureus subsp. aureus Mu50	Firmicutes	*efp1*	*accC*, *accB*	elongation factor P	hypothetical protein
Staphylococcus aureus subsp. aureus MW2	Firmicutes	*efp1*	*accC*, *accB*	elongation factor P	hypothetical protein
Staphylococcus aureus subsp. aureus N315	Firmicutes	*efp1*	*accC*, *accB*	elongation factor P	hypothetical protein
Staphylococcus aureus subsp. aureus NCTC 8325	Firmicutes	*efp1*	*accC*, *accB*	elongation factor P	hypothetical protein
Staphylococcus epidermidis ATCC 12228	Firmicutes	*efp1*	*accC*, *accB*	hypothetical protein	acetylornithine aminotransferase
Staphylococcus epidermidis RP62A	Firmicutes	*efp1*	*accC*, *accB*	hypothetical protein	acetylornithine aminotransferase
Staphylococcus haemolyticus JCSC1435	Firmicutes	*efp1*	*accC*, *accB*	elongation factor P	hypothetical protein
Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305	Firmicutes	*efp1*	*accC*, *accB*	elongation factor P	hypothetical protein
Streptococcus pneumoniae R6	Firmicutes	*efp1*	*accB*, *accC*, *accD*, *accA*	(3R)-hydroxymyristoyl ACP dehydratase	hypothetical protein
Synechococcus elongatus PCC 6301	Cyanobacteria	*efp1*	*accB*	elongation factor P	hypothetical protein
Synechococcus elongatus PCC 7942	Cyanobacteria	*efp1*	*accB*	exonuclease SbcC	elongation factor P
Synechococcus sp. CC9605	Cyanobacteria	*efp1*	*accB*	elongation factor P	4-hydroxythreonine-4-phosphate dehydrogenase

Organism	**Organism Group**	**EF-P**	***acc* Genes Discovered**	**Adjacent *acc* Upstream Gene**	**Adjacent *acc* Downstream Gene**
Synechococcus sp. CC9902	Cyanobacteria	*efp1*	*accB*	elongation factor P	4-hydroxythreonine-4-phosphate dehydrogenase
Synechococcus sp. WH 8102	Cyanobacteria	*efp1*	*accB*	elongation factor P	4-hydroxythreonine-4-phosphate dehydrogenase
Synechocystis sp. PCC 6803	Cyanobacteria	*efp1*	*accB*	elongation factor P	carbon dioxide concentrating mechanism protein; CcmK
Thermosynechococcus elongatus BP-1	Cyanobacteria	*efp1*	*accB*	elongation factor P	acetolactate synthase III large subunit
Thermus thermophilus HB27	Deinococcus-Thermus	*efp1*	*accC, accB*	elongation factor P	hypothetical protein
Thermus thermophilus HB8	Deinococcus-Thermus	*efp1*	*accC, accB*	elongation factor P	hypothetical protein
Wolinella succinogenes DSM 1740	Epsilonproteobacteria	*efp1*	*accC*	hypothetical protein	putative undecaprenol kinase bacitracin resistance protein

Figure 19. The *Shigella* spp. where IS elements were discovered to be encoded in some of the adjacent 10kb upstream and downstream region of both *efp*1 and *efp*2. The green gene represents putative IS1 encoded protein (YP_408562.1), the red gene represents IS1 ORF2 (*insB*) (YP_408563.1, YP_408565.1, YP_410545.1, YP_405764.1, NP_708068.1, NP_710012.1, NP_710025.1), the dark blue gene represents IS1 ORF1 (YP_405765.1, NP_708069.1, NP_710011.1, NP_710026.1, YP_313053.1, YP_313065.1), the aqua gene iso-IS1 OF2 (YP_405762.1, YP_405766.1, YP_405778.1, YP_402577.1, YP_402580.1), the purple gene represents IS1 ORFA (NP_837784.1, NP_839704.1), the brown gene represents IS1ORFB (NP_837783.1, NP_839693.1), and the yellow gene represents *efp*. In *Shigella boydii* Sb227 *efp*1 there is another putative IS1 encoded protein within the IS1 ORF2 (*insB*). Note, the *efp*1 and the *efp*2 of these organisms have similarities at the IS element regions and in some cases the similarities include some of the non coding sequence beside the IS element.

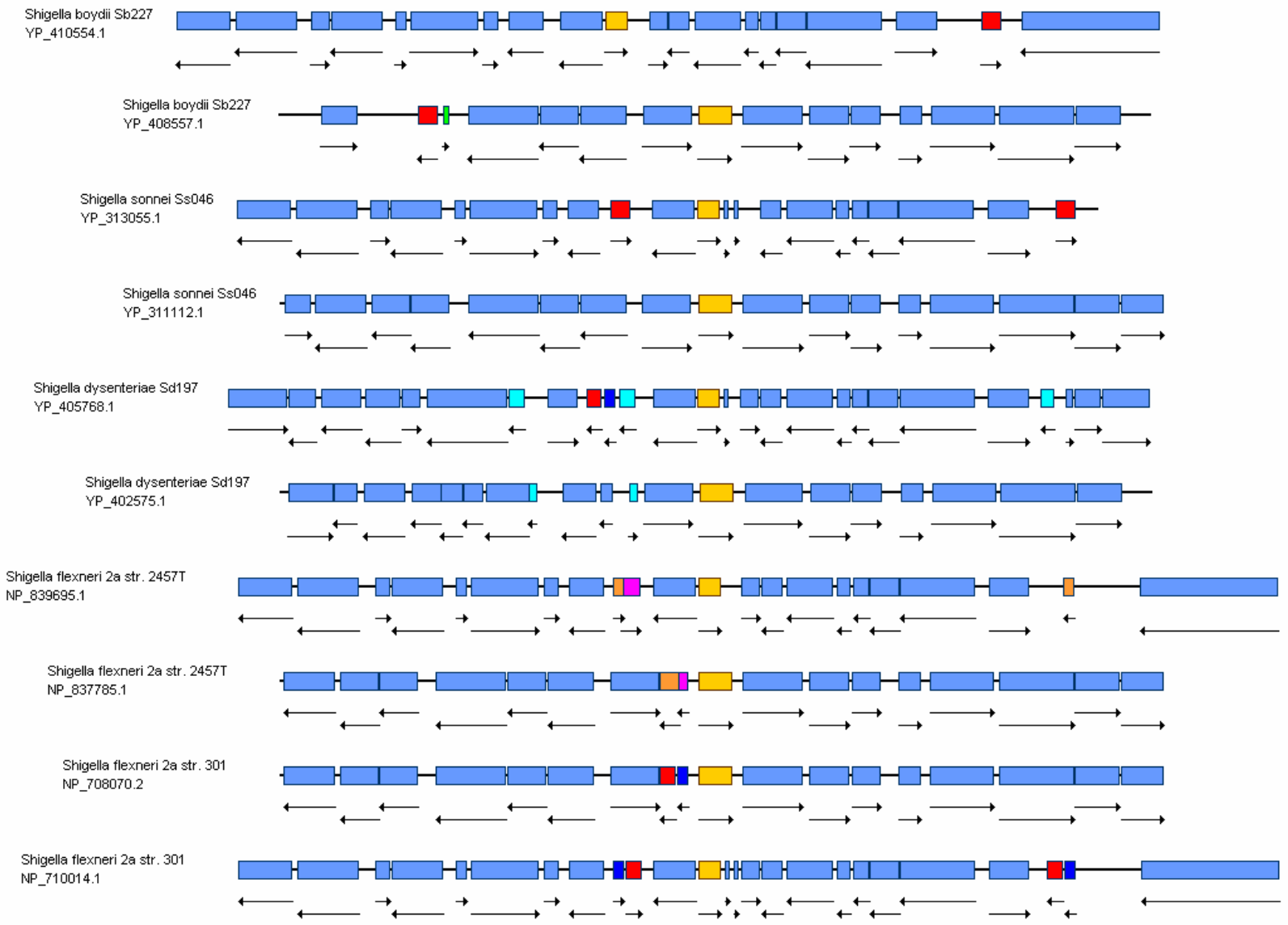
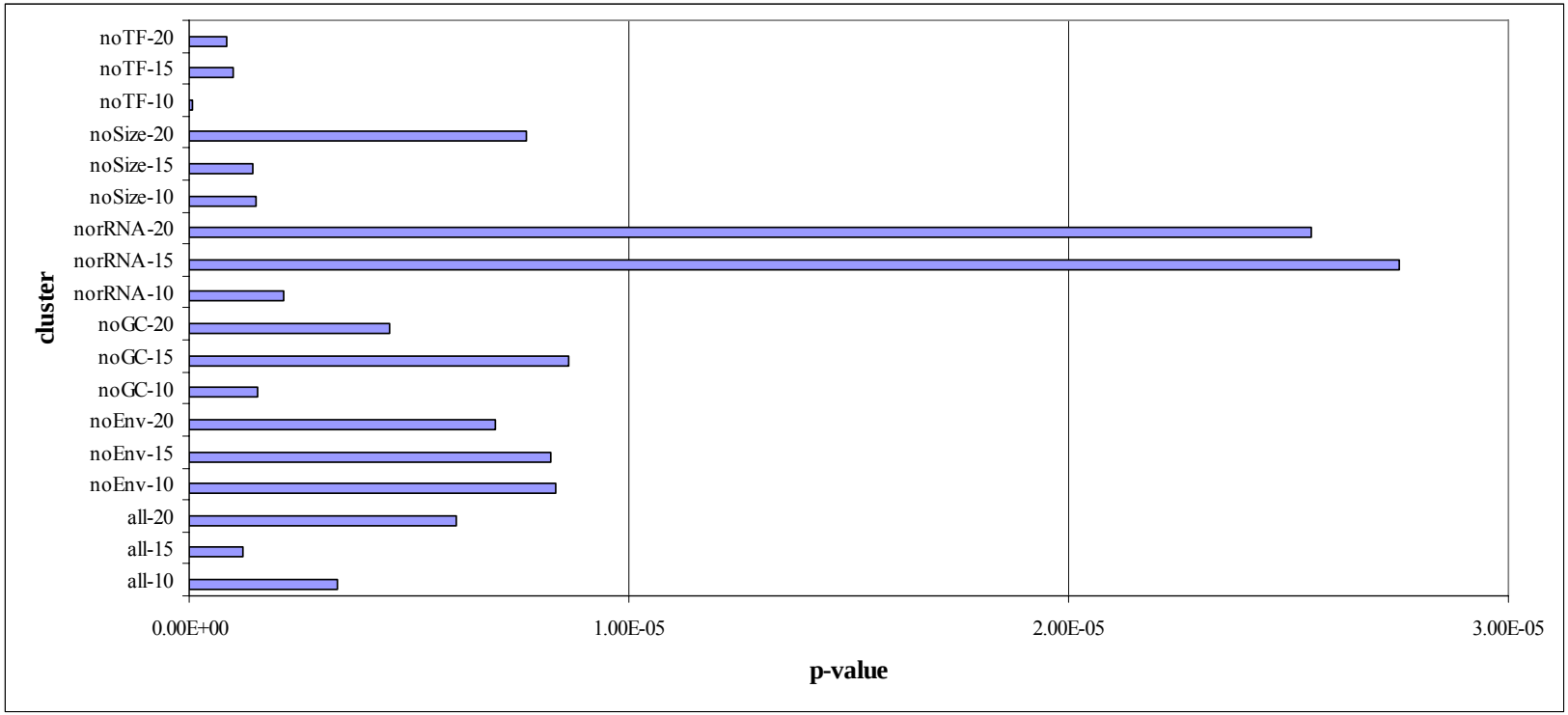

Shigella boydii Sb227
YP_410554.1
Shigella boydii Sb227
YP_408557.1
Shigella sonnei Ss046
YP_313055.1
Shigella sonnei Ss046
YP_311112.1
Shigella dysenteriae Sd197
YP_405768.1
Shigella dysenteriae Sd197
YP_402575.1
Shigella flexneri 2a str. 2457T
NP_839695.1
Shigella flexneri 2a str. 2457T
NP_837785.1
Shigella flexneri 2a str. 301
NP_708070.2
Shigella flexneri 2a str. 301
NP_710014.1

Table 6. Organisms with annotated IS elements excluding the *Shigella* spp. The EF-P, and IS element refseq ID is included to give limited positional location.

Organism	Bacterial Group	EF-P	RefSeq ID	# of IS elements	Names of IS element (IS family and/or name)	RefSeq ID of IS elements
Bacteroides thetaiotaomicron VPI-5482	Bacteroidetes/Chlorobi	*efp1*	NP_812620.1	1	unknown	NP_812617.1
Burkholderia thailandensis E264	Betaproteobacteria	*efp1*	YP_442272.1	1	Mutator	YP_442276.1
Dechloromonas aromatica RCB	Betaproteobacteria	*efp1*	YP_285250.1	1	IS116/IS110/IS902	YP_285256.1
Deinococcus geothermalis DSM 11300	Deinococcus-Thermus	*efp1*	YP_603940.1	1	IS4	YP_603936.1
Desulfovibrio desulfuricans G20	Deltaproteobacteria	*efp1*	YP_388445.1	1	putative	YP_388436.1
Francisella tularensis subsp. holarctica	Gammaproteobacteria	*efp1*	YP_512977.1	1	uknown	YP_512980.1
Francisella tularensis subsp. tularensis SCHU S4	Gammaproteobacteria	*efp1*	YP_169282.1	2	2 isftu1	YP_169273.1, YP_169279.1
Geobacter metallireducens GS-15	Deltaproteobacteria	*efp2*	YP_383073.1	1	IS111A/IS1328/IS1533/IS116/IS110/IS902	YP_383068.1
Idiomarina loihiensis L2TR	Gammaproteobacteria	*efp1*	YP_156657.1	1	Il-IS1	YP_156653.1
Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	Firmicutes	*efp2*	YP_619581.1	1	IS30	YP_619577.1
Lactobacillus johnsonii NCC 533	Firmicutes	*efp2*	NP_965635.1	1	IS30	NP_965633.1
Lactobacillus sakei subsp. sakei 23K	Firmicutes	*efp2*	YP_394863.1	3	orfB of IS1520 (IS3), orfA of IS1520 (IS3), orfB of ISLsa2 (IS150)	YP_394869.1, YP_394870.1, YP_394875.1
Lactococcus lactis subsp. lactis Il1403	Firmicutes	*efp1*	NP_266848.1	1	IS983A	NP_266837.1
Leptospira interrogans serovar Lai str. 56601	Spirochaetes	*efp1*	NP_714838.1	3	3 putative	NP_714839.1, NP_714847.1, NP_714848.1
Magnetospirillum magneticum AMB-1	Alphaproteobacteria	*efp1*	YP_420648.1	1	putative IS402	YP_420660.1
Methylococcus capsulatus str. Bath	Gammaproteobacteria	*efp1*	YP_113782.1	1	ISMca1	YP_113774.1
Nitrobacter hamburgensis X14	Alphaproteobacteria	*efp1*	YP_577554.1	1	IS116/IS110/IS902	YP_577558.1
Pelobacter carbinolicus DSM 2380	Deltaproteobacteria	*efp1*	YP_357898.1	1	putative	YP_357902.1
Photorhabdus luminescens subsp. laumondii TTO1	Gammaproteobacteria	*efp1*	NP_931320.1	2	ISPlu3G (IS630), ISPlu6D (IS982)	NP_931313.1, NP_931318.1

Organism	Bacterial Group	EF-P	RefSeq ID	# of IS elements	Names of IS element (IS family and/or name)	RefSeq ID of IS elements
Pseudomonas putida KT2440	Gammaproteobacteria	*efp1*	NP_744013.1	1	ISPpu8	NP_744020.1
Pseudomonas syringae pv. phaseolicola 1448A	Gammaproteobacteria	*efp1*	YP_275790.1	2	truncated ISPsy19, ISPsy18	YP_275781.1, YP_275782.1
Psychrobacter arcticus 273-4	Gammaproteobacteria	*efp1*	YP_264962.1	1	putative IS30	YP_264969.1
Rickettsia felis URRWXCal2	Alphaproteobacteria	*efp1*	YP_247065.1	2	2 Mutator	YP_247069.1, YP_247070.1
Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67	Gammaproteobacteria	*efp2*	YP_217214.1	1	Tn10	YP_217211.1
Shewanella oneidensis MR-1	Gammaproteobacteria	*efp1*	NP_717918.1	1	ISSod4	NP_717912.1
Staphylococcus epidermidis ATCC 12228	Firmicutes	*efp1*	NP_764768.1	1	truncated	NP_764774.1
Synechocystis sp. PCC 6803	Cyanobacteria	*efp1*	NP_442181.1	1	unknown	NP_442190.1
Vibrio vulnificus YJ016	Gammaproteobacteria	*efp1*	NP_935895.1	1	transposase and inactivated derivative	NP_935884.1
Xanthomonas axonopodis pv. citri str. 306	Gammaproteobacteria	*efp1*	NP_642696.1	2	IS1479	NP_642687.1, NP_642688.1
Xanthomonas oryzae pv. oryzae KACC10331	Gammaproteobacteria	*efp1*	YP_201344.1	3	2 putative ISXo8, putative transpoase	YP_201333.1, YP_201334.1, YP_201532.1
Xanthomonas oryzae pv. oryzae MAFF 311018	Gammaproteobacteria	*efp1*	YP_451580.1	2	ISXo8, IS1112	YP_451571.1, YP_451573.1
Xanthomonas oryzae pv. oryzae MAFF 311018	Gammaproteobacteria	*efp2*	YP_451784.1	4	ISXoo11, ISXoo12, 2 IS1112	YP_451771.1, YP_451772.1, YP_451787.1, YP_451788.1
Yersinia pestis Antiqua	Gammaproteobacteria	*efp2*	YP_650912.1	2	IS1661, IS1661 DNA-binding protein	YP_650918.1, YP_650919.1
Yersinia pestis CO92	Gammaproteobacteria	*efp1*	NP_404002.1	1	IS1541	NP_404004.1

Figure 20. The p-value averages of the motifs discovered using different clustering of natural habitat, 16S rRNA distance, genomic size, genomic GC content, and number of orthologous transcription factors and the maximum number of clusters defined. The statistically significant motifs have the lowest p-values. However, a lower p-value may not necessary correspond to the biologically functioning motif. The motif scores vary greatly from one group of bacterial sequences to another from the same cluster so standard deviation is very large and can be misleading. In addition, the p-values within the same clusters are variable because it depends upon the sequences and whether or not they have more than one lowly scoring motif. The cluster names are representing the following: all = all characteristics used in clustering, noEnv = no environmental characteristic used in defining the cluster, noGC = no genomic GC characteristic used in clustering, norRNA = no 16S rRNA distance measurement used, noSize = no genomic size characteristic used, and noTF = no orthologous transcription factor characteristic used. The number following the cluster name after the dash denotes the maximum number of clusters defined.


noTF-20
noTF-15
noTF-10
noSize-20
noSize-15
noSize-10
norRNA-20
norRNA-15
norRNA-10
noGC-20
noGC-15
noGC-10
noEnv-20
noEnv-15
noEnv-10
all-20
all-15
all-10
cluster
0.00E+00
1.00E-05
2.00E-05
3.00E-05
p-value

over all the different clustering characteristics and ignoring the defined maximum number of clusters: no orthologous transcription factor characteristic, no genomic size characteristics, all characteristics, no genomic GC content characteristics, no environmental habitat characteristics, and no 16S rRNA distances information.

3.6.1 *Top 5 Predicted Motifs*

Table 7 describes the top 5 motifs discovered in the average lowest scoring cluster noTF-10, based on a maximum of 10 clusters, the natural habitat of the organism, 16S rRNA distance, genomic size, and genomic GC content. The general trend in all the clusters, as seen in Table 7 is that the lowest scoring motif (the most statistically significant motif) discovered has a noticeable drop in p-value than the second lowest scoring motif. The top 5 motifs discovered using the noTF-10 clustering can be seen in Figure 21 to 30. Some of the discovered motifs are very similar; the only differences are the nucleotide frequencies. The different nucleotide frequencies for the similar predicted motifs are due to the different motif finding algorithms which may include or remove certain sequences during motif finding and the predicted motifs may be discovered in many instances in the sequences causing the frequencies of the motifs to be different if the programs are using the different motif instances. Clustering motifs can be used to remove these similarities. However, sometimes clustering motifs can obscure the true motif because it only shows the average motif that is represented in the cluster and not all the different possibilities of motifs in which one may be the biologically significant motif. Clustering is very useful when trying to find more than one unique motif. The null hypothesis is rejected because TAMO 1.0 discovered statistically significant motifs that had a p-value < 0.001.

Table 7. The lowest on average p-value for the motifs discovered was the cluster based on the organism's natural habitat, 16S rRNA distance, genomic size, and genomic GC content and specified to have the maximum number of clusters as 10. The top 5 scores are listed, with the average of each group and the total average p-value of the entire cluster.

Bacterial Groups in Cluster	Cluster	Number of sequences	p-values of the top 5 motifs					Average
Actinobacteria, (Alpha, Beta, Delta)proteobacteria, Cyanobacteria	1	27	8.66E-14	3.23E-09	7.21E-09	3.67E-08	7.13E-08	2.37E-08
Actinobacteria, (Alpha, Beta)proteobacteria	2	14	6.73E-16	6.77E-10	7.73E-09	1.09E-08	1.13E-08	6.11E-09
Acidobacteria, Actinobacteria, (Alpha, Beta, Delta)proteobacteria, Cyanobacteria	3	38	4.17E-20	6.03E-20	1.89E-19	7.98E-19	1.22E-18	4.62E-19
Deinococcus-Thermus, Gammproteobacteria	4	5	5.72E-09	3.97E-07	5.10E-07	7.14E-07	1.08E-06	5.42E-07
Alphaproteobacteria, Chlamydiae/Verrucomicrobia, Cyanobacteria	5	41	4.09E-21	5.32E-20	6.48E-20	6.48E-20	6.48E-20	5.03E-20
Bacteroidetes/Chlorobi, (Delta, Epsilon, Gamma)proteobacteria, Firmicutes, Fusobacteria, Spirochaetes	6	83	3.35E-24	2.40E-21	1.21E-20	1.71E-20	3.71E-20	1.37E-20
Cyanobacteria, Firmicutes, Gammaproteobacteria	7	48	1.66E-19	3.75E-17	4.10E-17	5.41E-17	3.16E-16	8.97E-17
Bacteroidetes/Chlorobi, Deinococcus-Thermus, Firmicutes, Gammaproteobacteria, Planctomycetes, Spirochaetes	8	62	7.23E-19	2.04E-18	2.23E-18	3.30E-18	3.46E-18	2.35E-18
Actinobacteria, (Alpha, Beta, Delta, Epsilon, Gamma)proteobacteria, Aquificae, Bacteroidetes/Chlorobi, Chloroflexi, Cyanobacteria, Firmicutes, Thermotogae	9	51	1.55E-16	1.55E-14	1.49E-13	1.49E-13	1.92E-13	1.01E-13
Actinobacteria, (Alpha, Beta, Delta)proteobacteria	10	13	7.28E-13	1.69E-08	2.54E-08	5.25E-08	1.64E-07	5.18E-08
							total:	6.24E-08

3.6.2 Comparison with Known DNA Binding Sites

There were no complete matches where both the -10 and -35 elements of the same sigma transcription factor binding sites from RegulonDB 5.0 were similar with the predicted motifs of the NoTF-10 group. See Figure 21 to 30 for *E. coli* sigma factor binding sites that have matching position frequency matrices with the top 5 predicted motifs for each cluster. Matching position frequency matrices are when the two matrices have a frequency for a nucleotide base in the same position and when a frequency is zero in one matrix the other is also zero for the base in the same position. When comparing the position frequency matrices of the known binding sites from RegTransBase, no known motifs matched the predicted motifs (Figure 21 to 30), perfectly. However, when allowing mismatches in one of the frequencies in the position matrix (mismatch where one position frequency matrix had no A, C, T, or G in any position and the other position frequency matrix had a known frequency for that nucleotide in that position) there was still no known motif matches. When allowing 9 mismatches, CRP motif matched with cluster 2's s..ss....sG.s..S..s..cG motifs. At ten mismatches, CRP motif matched with cluster 2's ss..G....s.ss..ss....SG motif. At eleven mismatches, cluster 4's motif s..SSas...sss..s..s..Gc matched with CRP motif, cluster 4's motif Ss.Gsts..s..s.ss.ksC matched with FadR motif, cluster 4's motif s..s.yGs.Gs.s..ss.s..s..Gs matched with CRP motif, and cluster 6's motif AAAwa..ww.wtAw.w.wAAAa matched with FNR motif (with a low score of 8.51 and thus high similarity) . At the cutoff value of 12 mismatches the FNR and the FadR motif similarities were discovered as seen in Table 8. The scores and the similar binding sites for regulatory proteins compared with the predicted motifs are summarized in Table 8.

Figure 21. Top 5 motifs for Cluster 1 based on environmental, 16S rRNA, genomic size, GC content, and specified to have the maximum number of clusters as 10. There are no similar *E. coli* TFBS for the top 5 motif of cluster 1.

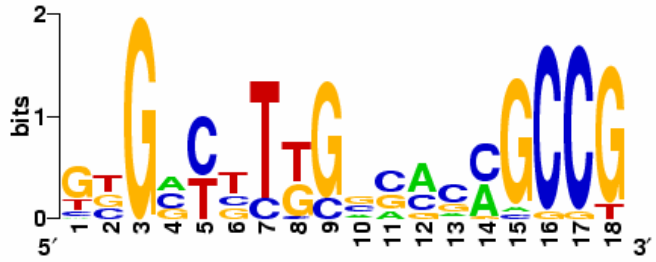
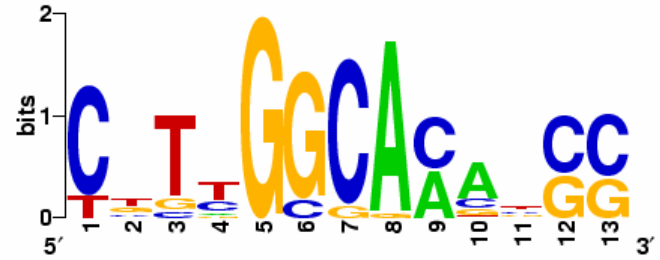
Motif sequence logos	p-value
	8.66E-14
	3.23E-09
	7.21E-09
	3.67E-08
	7.13E-08

Figure 22. Top 5 motifs and the TFBS they are similar to for Cluster 2 based on environmental, 16S rRNA, genomic size, GC content, and specified to have the maximum number of clusters as 10.

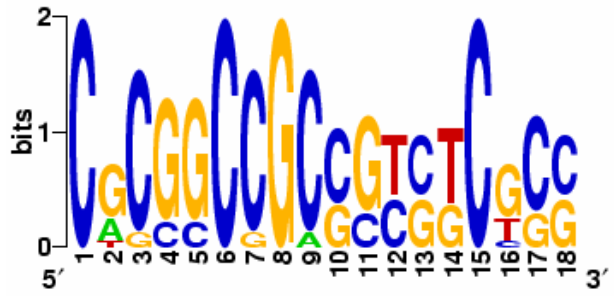
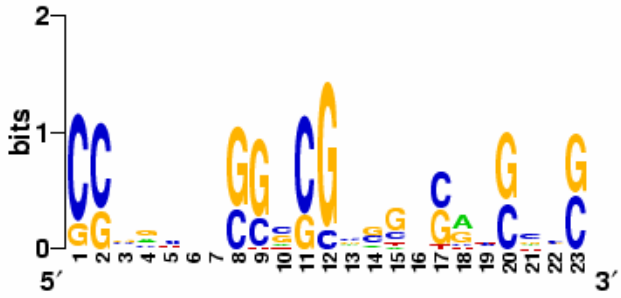
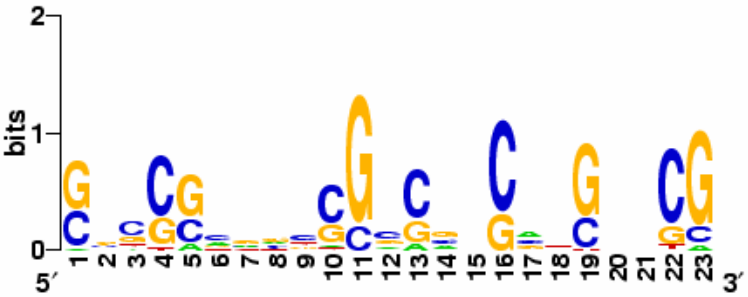
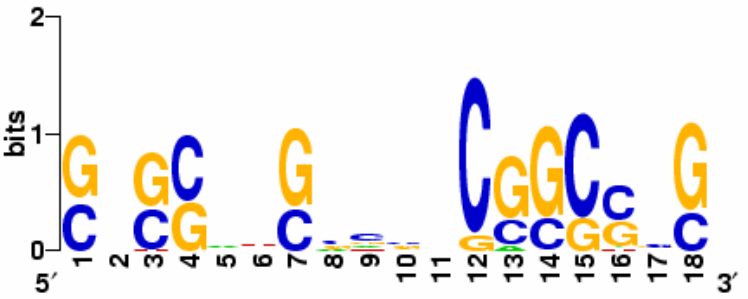
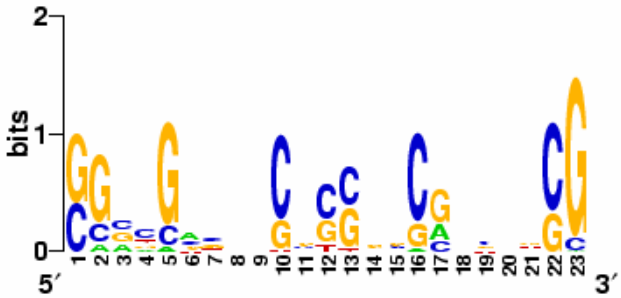
Motif sequence logos	p-value	Similar TFBS & score	
	6.73E-16		
	6.77E-10	Sigma24 -10	5.08
	7.73E-09		
	1.09E-08	Sigma24 -10	5.66
	1.13E-08		

Figure 23. Top 5 motifs and the TFBS they are similar to for Cluster 3 based on environmental, 16S rRNA, genomic size, GC content, and specified to have the maximum number of clusters as 10.

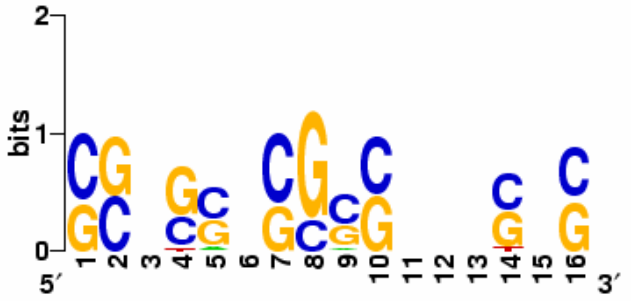
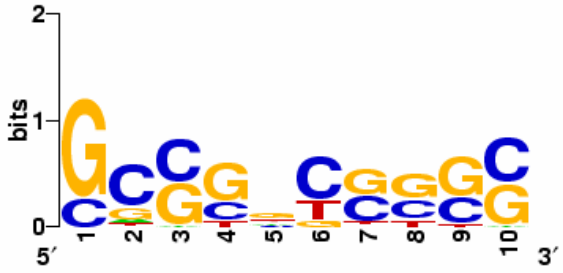
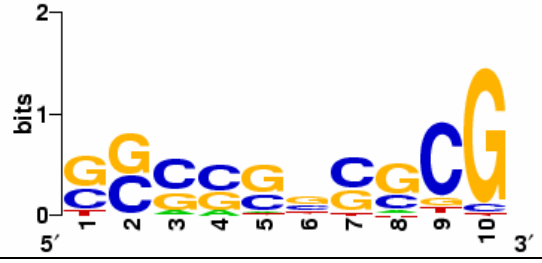
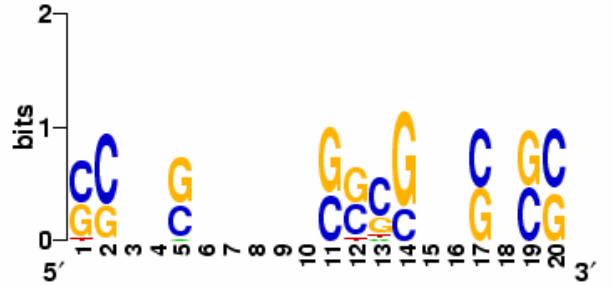
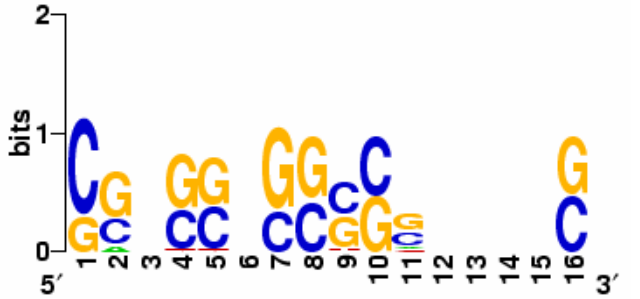
Motif sequence logos	p-value	Similar TFBS & score
	4.17E-20	
	6.03E-20	
	1.89E-19	Sigma24 -10 5.87 Sigma70 -35 5.18
	7.98E-19	Sigma24 -10 4.96
	1.22E-18	Sigma24 -10 5.02

Figure 24. Top 5 motifs and the TFBS they are similar to for Cluster 4 based on environmental, 16S rRNA, genomic size, GC content, and specified to have the maximum number of clusters as 10.

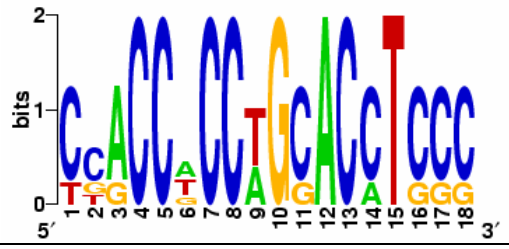
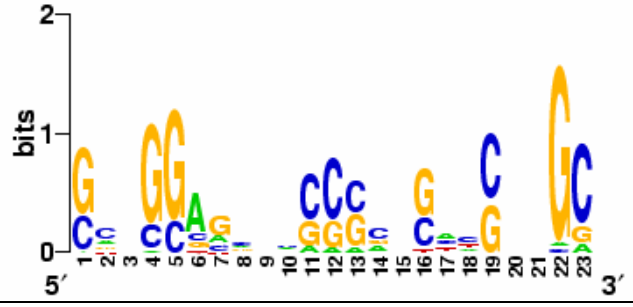
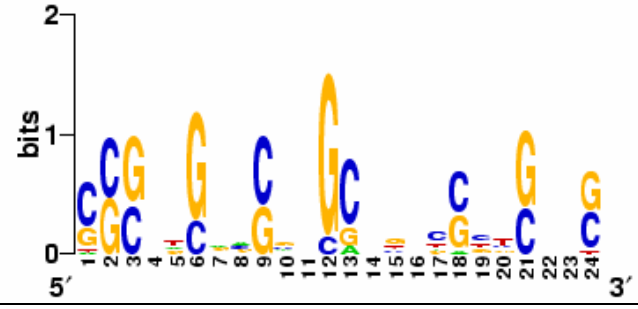
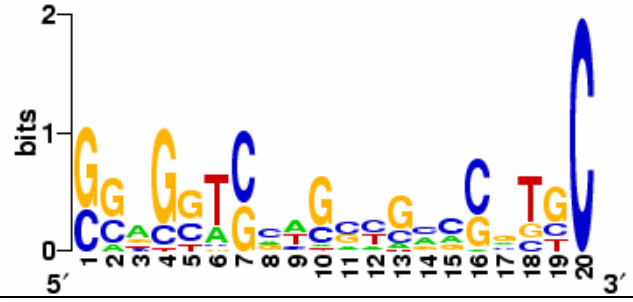
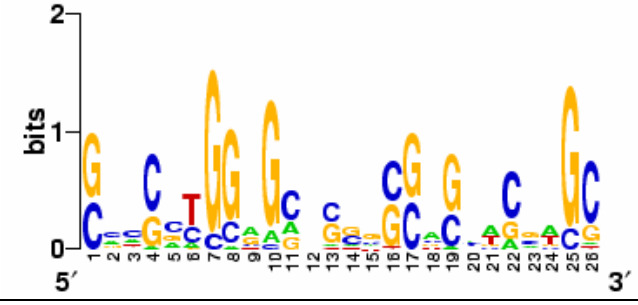
Motif sequence logos	p-value	Similar TFBS & score
	5.72E-09	
	3.97E-07	Sigma24 -10 5.35
	5.10E-07	
	7.14E-07	Sigma24 -10 5.97 Sigma70 -35 5.35
	1.08E-06	

Figure 25. Top 5 motifs and the TFBS they are similar to for Cluster 5 based on environmental, 16S rRNA, genomic size, GC content, and specified to have the maximum number of clusters as 10.

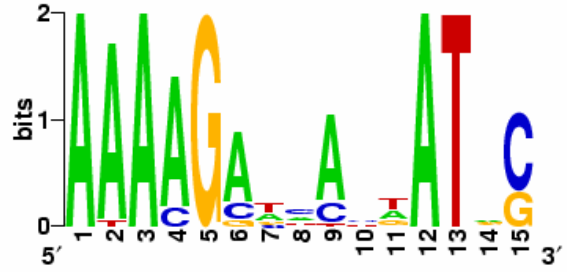
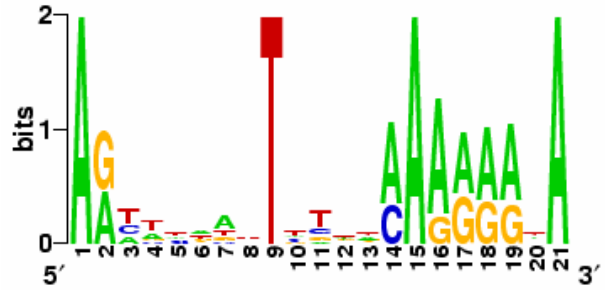
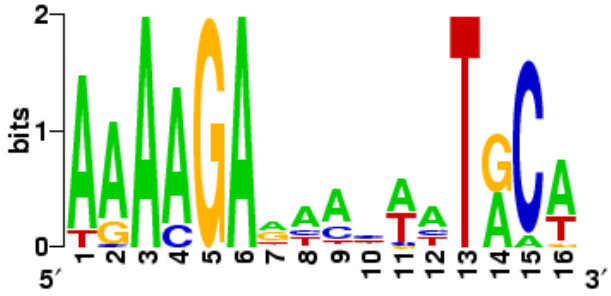
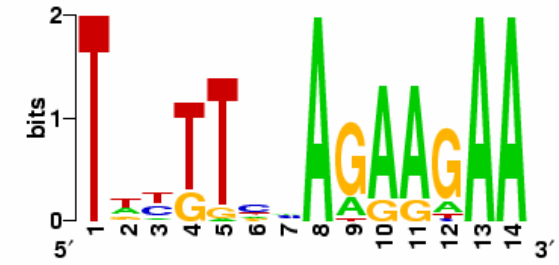
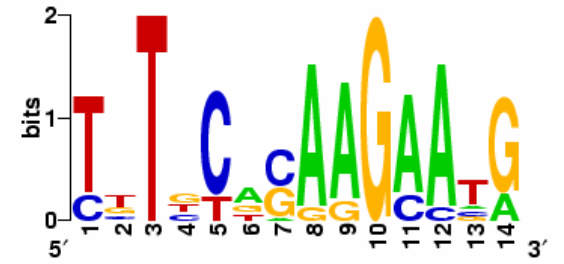
Motif sequence logos	p-value	Similar TFBS & score
	4.09E-21	
	5.32E-20	Sigma24 -10 5.03 Sigma70 -35 4.42
	6.48E-20	Sigma24 -10 4.59 Sigma70 -35 5.12
	6.48E-20	
	6.48E-20	

Figure 26. Top 5 motifs and the TFBS they are similar to for Cluster 6 based on environmental, 16S rRNA, genomic size, GC content, and specified to have the maximum number of clusters as 10.

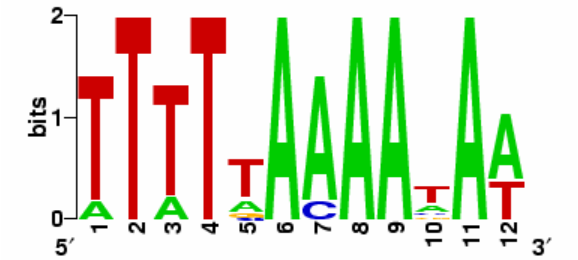
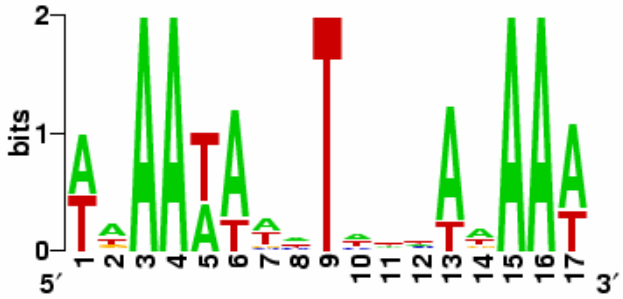
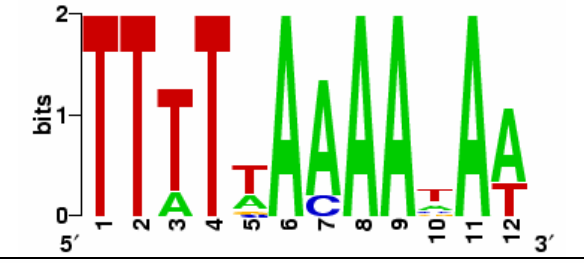
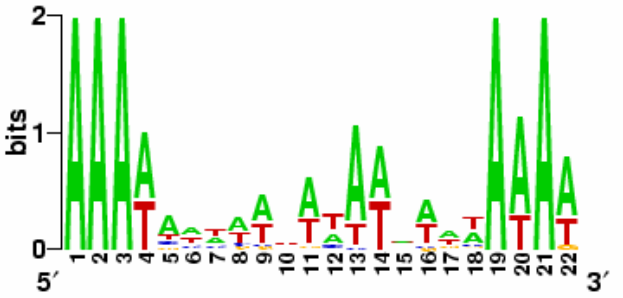
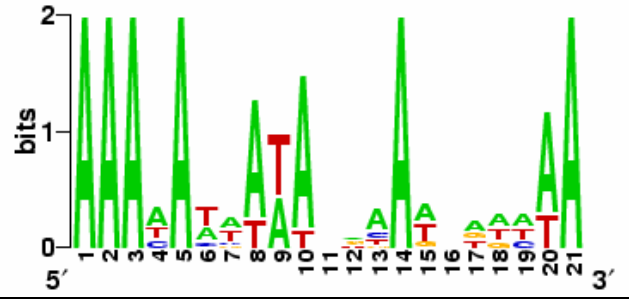
Motif sequence logos	p-value	Similar TFBS & score
	3.35E-24	
	2.40E-21	
	1.21E-20	
	1.71E-20	Sigma24 -10 5.14 Sigma70 -35 5.33
	3.71E-20	Sigma24 -10 4.59

Figure 27. Top 5 motifs and the TFBS they are similar to for Cluster 7 based on environmental, 16S rRNA, genomic size, GC content, and specified to have the maximum number of clusters as 10. There are no similar *E. coli* TFBS for the top 5 motif of cluster 7.

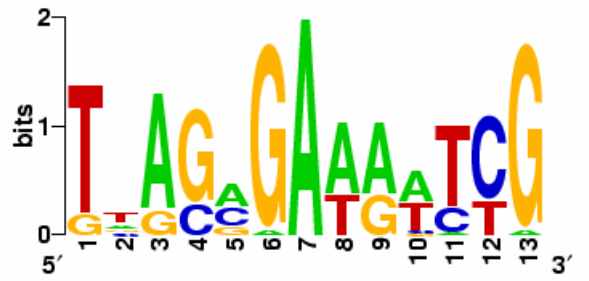
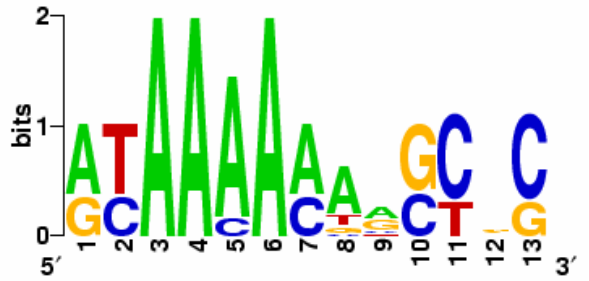
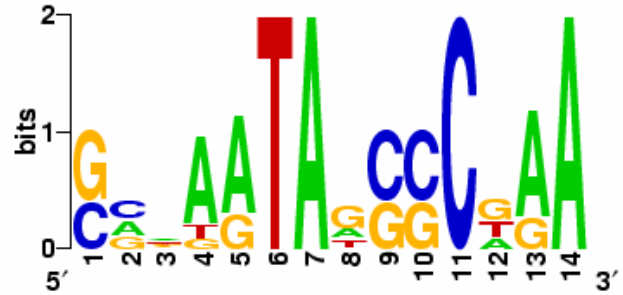
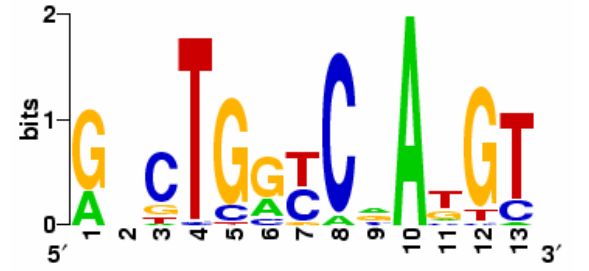
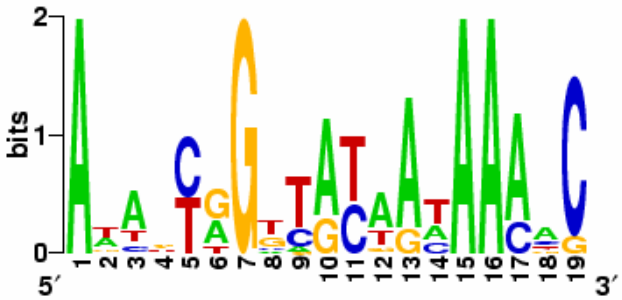
Motif sequence logos	p-value
	1.66E-19
	3.75E-17
	4.10E-17
	5.41E-17
	3.16E-16

Figure 28. Top 5 motifs and the TFBS they are similar to for Cluster 8 based on environmental, 16S rRNA, genomic size, GC content, and specified to have the maximum number of clusters as 10. There are no similar *E. coli* TFBS for the top 5 motif of cluster 8.

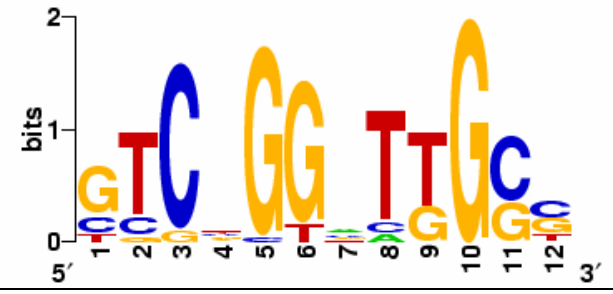
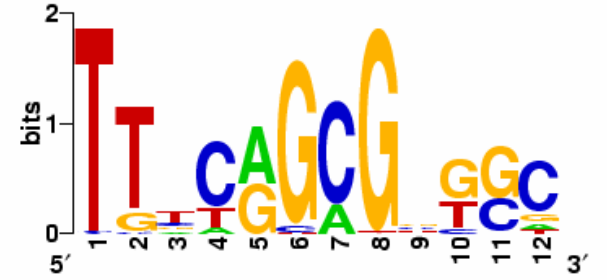
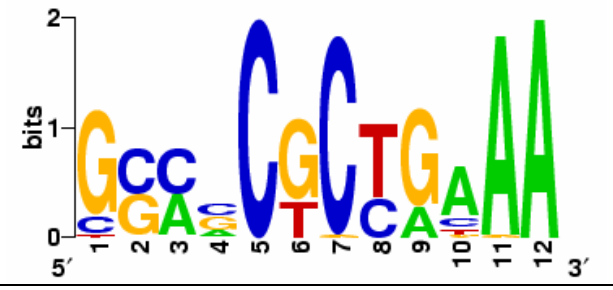
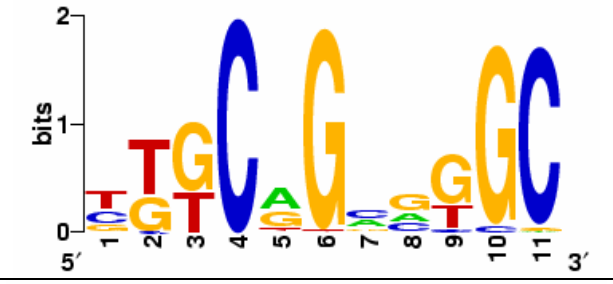
Motif sequence logos	p-value
	7.23E-19
	2.04E-18
	2.23E-18
	3.30E-18
	3.46E-18

Figure 29. Top 5 motifs and the TFBS they are similar to for Cluster 9 based on environmental, 16S rRNA, genomic size, GC content, and specified to have the maximum number of clusters as 10. There are no similar *E. coli* TFBS for the top 5 motif of cluster 9.

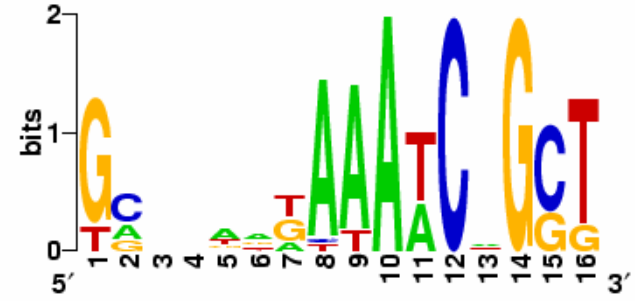
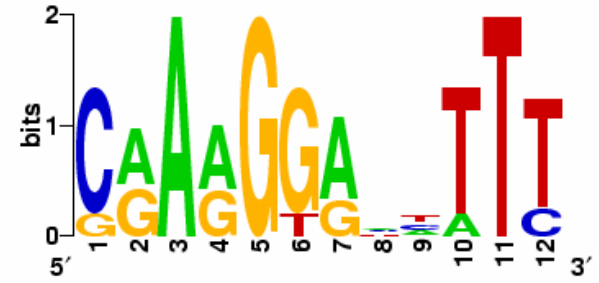
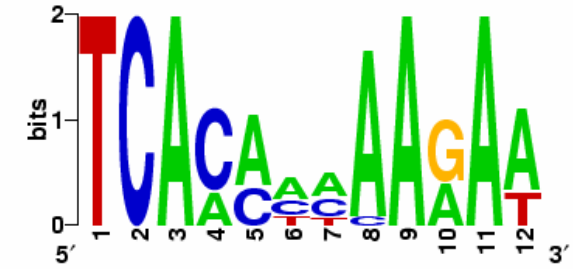
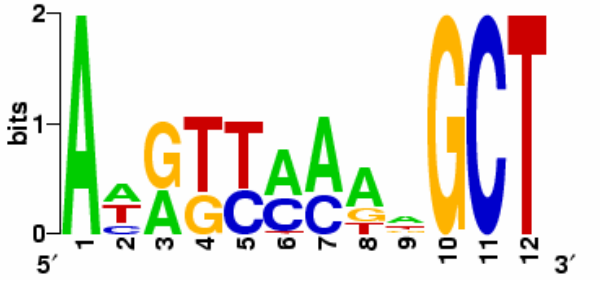
Motif sequence logos	p-value
	1.55E-16
	1.55E-14
	1.49E-13
	1.49E-13
	1.92E-13

Figure 30. Top 5 motifs and the TFBS they are similar to for Cluster 10 based on environmental, 16S rRNA, genomic size, GC content, and specified to have the maximum number of clusters as 10.

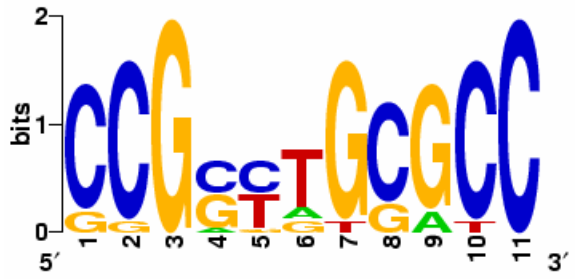
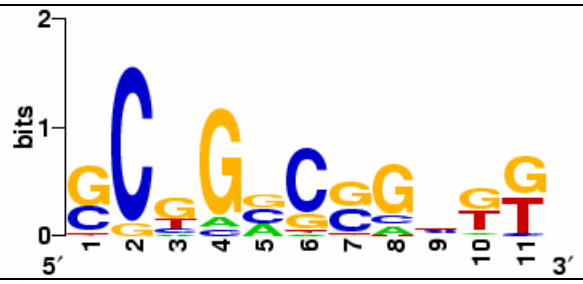
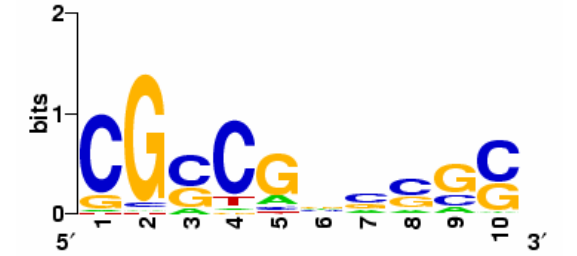
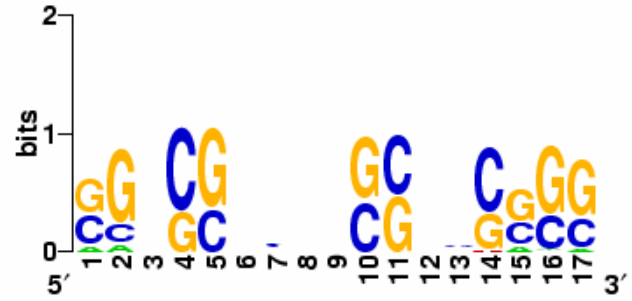
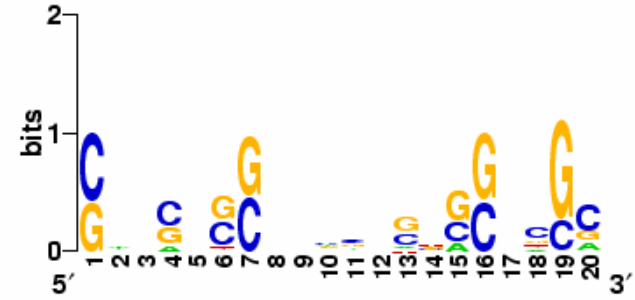
Motif sequence logos	p-value	Similar TFBS & score
	7.28E-13	
	1.69E-08	Sigma24 -10 5.13
	2.54E-08	Sigma24 -10 6.04 Sigma70 -35 4.06
	5.25E-08	
	1.64E-07	Sigma24 -10 5.58 Sigma70 -35 4.08

Table 8. The noTF-10 clusters predicted motifs with similarities to the binding sites from RegTransBase allowing a maximum of 12 mismatches. The lower the score the more similar the two position frequency matrices are.

Cluster	Motif	p-value	Binding site	Score	Binding Site	Score
2	ss..G....s.ss..ss....SG	7.73E-09	Crp Gammaproteobacteria	10.17	Fnr Gammaproteobacteria	14.47
2	s..ss....sG.s..S..s..cG	1.09E-08	Crp Gammaproteobacteria	10.95		
3	ss..s.....sssS..s.ss	7.98E-19	Fnr Gammaproteobacteria	13.10		
4	s..SSas...sss..s..s..Gc	3.97E-07	Crp Gammaproteobacteria	12.09		
4	Ss.Gsts..s..s.ss.ksC	7.14E-07	FadR Gammaproteobacteria	10.67	Fnr Gammaproteobacteria	14.35
4	s..s.yGs.Gs.s..ss.s..s..Gs	1.08E-06	Crp Gammaproteobacteria	16.50		
6	AAAwa..ww.wtAw.w.wAAAa	1.71E-20	Fnr Gammaproteobacteria	8.51	FadR Gammaproteobacteria	12.03
10	s..s.ss.....s.ss..Ss	1.64E-07	FadR Gammaproteobacteria	9.37	Fnr Gammaproteobacteria	13.43

Figure 31. Sequence logos of the binding site for the CRP group from Gammaproteobacteria. CRP group from Gammaproteobacteria is a DNA-binding site for CRP which is a cAMP sensitive transcriptional dual regulator (Kazakov et al. 2007).

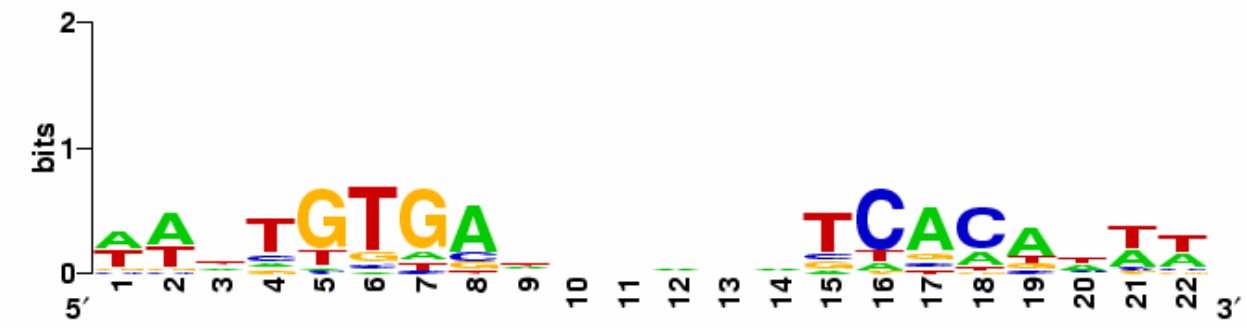

bits
2
1
0
5′
3′
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Figure 32. Sequence logos of the binding site for the FadR group from Gammaproteobacteria motif. The FadR motif from Gammaproteobacteria is a binding site for FadR which is a dual transcriptional regulator of fatty acids metabolism (Kazakov et al. 2007).

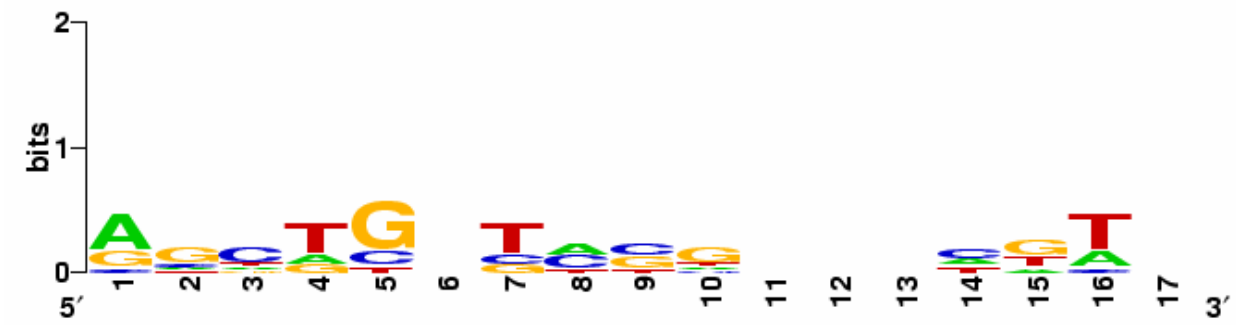
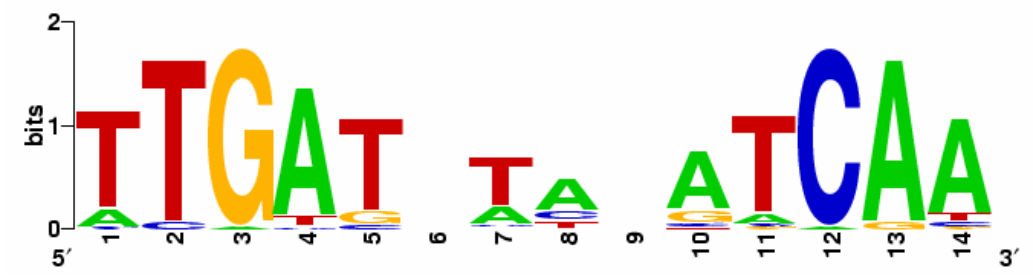

bits
2
1
0
5′
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
3′

Figure 33. Sequence logos of the binding site for the FNR group from Gammaproteobacteria. The FNR motif from Gammaproteobacteria is a binding site for FNR which is a oxygen sensitive transcriptional dual regulator (Kazakov et al. 2007).


bits
2
1
0
5′
1
2
3
4
5
6
7
8
9
10
11
12
13
14
3′

Chapter 4: DISCUSSION

4.1 EF-P Sequence Retrieval Analyses

In this study, the gene encoding EF-P was ubiquitous in all of the completely sequenced bacterial genomes except for *Leptospira interrogans* serovar *Copenhageni* str. Fiocruz L1-130. This notwithstanding, the *L. interrogans* serovar *Copenhageni* str. Fiocruz L1-130 genome appears to encode an *efp* pseudogene that contains a frameshift mutation. Whether or not this sequence is a pseudogene or encodes a functional EF-P protein needs to be experimentally verified. However, *L. interrogans* serovar *Copenhageni* str. Fiocruz L1-130 is an obligate pathogen of human and may not require a functional EF-P protein. It is hypothesized that some parasites may not require EF-P but rather rely on a host protein(s), which can function in a manner similar to EF-P. In fact, many mammalian host cell functions such as signal transduction pathways are known to be exploited by bacterial pathogens (Finlay and Cossart 1997). For example, *Bordetella pertussis* is able to bind to the host integrin protein causing the host cell signaling pathway to up-regulate the binding activity of more host integrin to increase the Bacteria's attachment to its host (Ishibashi, Claus, and Relman 1994). *L. interrogans* serovar *Copenhageni* str. Fiocruz L1-130 infects animal cells it may be able to use eIF-5A, the eukaryotic homolog of EF-P, to facilitate translation in the absence of EF-P. The N and C domains of eIF-5A correspond to *efp*'s KOW domain and OB domain (or C-terminus domain) (Hanawa-Suetsugu et al. 2004). Since eIF-5A and EF-P are very divergent from one another, *L. interrogans* serovar *Copenhageni* str. Fiocruz L1-130 may not be able to use eIF-5A as efficiently as EF-P which in turn may cause the organism to grow slowly. However, the slow growth of the parasite allows it to avoid inhibition by the animal's immune system; for example,

the slow growth of *Mycoplasm genitalium* in the human urogenital tract may help the bacterium avoid detection by the immune system (Glass et al. 2006). In addition, if *L. interrogans* serovar *Copenhageni* str. Fiocruz L1-130 *efp* encoding region was translated and the *efp* may still be functioning because it contains the three domains of *efp*, as long as the first 52 amino acids translated did not interfere with the folding of the domains.

4.2 EF-P Phylogenetic Tree Analyses

EF-P appears to have a complex evolutionary history of horizontal gene transfer and gene duplication events as seen in the phylogenetic tree (Figure 5). The distinctive largest clustering of EF-P2 from Gammaproteobacteria, Planctomyctes, Deltaproteobacteria, Firmicutes (including some EF-P1s), Chloroflexi, and Acidobacteria in the tree is suggestive of either HGT or an ancient gene duplication in the common ancestor of these bacterial groupings. Subsequent gene loss can explain why some closely related species do not contain a second copy of EF-P. In a genomic comparison between closely related species of yeast, *Saccharomyces pombe* and *Saccharomyces cerevisiae,* Aravind et al., (2000) noted that approximately 300 genes have been lost since these two organisms diverged from their common ancestor. Within the bacterial domain, *Enterococci* spp. appears to have duplicate copies of elongation factor Tu (*tufA* and *tufB*). The EF-Tu phylogenetic tree is similar to the *efp* phylogenetic tree in that *tufA* clusters together and the *tufB* from the corresponding species clusters together (Ke et al. 2000). Ke et al. (2000) and I were able to rule out a recent gene duplication in the evolutionary history of the duplicated gene as a cause of the topology of the tree, since the *tufA* and *tufB* genes from the same organisms are not clustering together as sister taxa in the phylogenetic tree. To explain the EF-Tu phylogenetic tree topology, Ke et al. (2000), proposed that these duplicates may have been horizontally transferred.

The similarities in the branching pattern of the Chlamydiae/Verrucomicrobia are suggestive of an ancient gene duplication or horizontal transfer event in a common ancestor of the Chlamydiae/Verrucomicrobia group which explains why the duplicated EF-P copies are seen distantly apart in the tree but still retain a similar branching order within the subclade. Also, the Chlamydiae/Verrucomicrobia group appears to exhibit vertical inheritance in both the EF-P1 and the EF-P2 clusters as they display the same topology as in the SSU rRNA tree (Figure 8). The duplicated copies of EF-P from *Mesorhizobium loti* MAFF303099 and *Rhodobacter sphaeroides* 2.4.1 are suggestive of a duplication event with subsequent divergence or possibly a HGT event from a closely related species in the same bacterial group. *Porphyromonas gingivalis* W83 EF-P1 and EF-P2 most certainly resulted from a recent gene duplication event because these two genes branch off from each other as sister taxa.

Some of the organisms with only one copy of the *efp* gene in their genomes also show evidence of HGT. For example, *efp* from the Actinobacteria, *Symbiobacterium thermophilum* IAM 14863 and *Rubrobacter xylanophilus* DSM 9941 both cluster alone, away from the rest of the Actinobacteria group which clusters as a monophyletic group in Figure 5. This is strong evidence for the HGT of *efp* in *S. thermophilum* IAM 14863 and *R. xylanophilus* DSM 9941. The *efp* from these two organisms depict the same atypical pattern as the bacterial intracellular parasites, *Rickettsiae* sp. and *Chlamydiae* sp. in phylogenetic analysis using the ubiquitous bacterial protein dCTP deaminase (Wolf, Aravind, and Koonin 1999). Like *S. thermophilum* IAM 14863 and *R. xylanophilus* DSM 9941, *Rickettsiae* sp. and *Chlamydiae* sp. are not clustering with their known bacterial relatives in the dCTP deaminase phylogenetic tree. This is incongruent with the SSU tree and therefore suggestive of a possible horizontal transfer event (Wolf, Aravind, and Koonin 1999).

The lone member of the Betaproteobacteria, *Nitrosopira multiformis* ATCC 25196, clusters beside the Alphaproteobacteria and Acidobacteria group with the rest of the Betaproteobacteria shown as a monophyletic group. This suggests a possible EF-P HGT in *N. multiformis* ATCC 25196 from either an Alphaproteobacteria, or an Acidobacteria, due to the close proximity of these groups, or possibly from another bacterial group with subsequent divergence as seen for an ancient HGT event. The paraphyletic Bacteroidetes/Chlorobi group shows evidence of HGT in either of the two Bacteroidetes/Chlorobi groups' common ancestors, with subsequent loss of the native EF-P gene, xenolog, or another possibility is that the *efp* genes within the two groups may have diverged from each other due to different selection pressures.

The Firmicutes and some of the Proteobacteria groups may be paraphyletic due to different evolutionary rates as shown by the Chlamydiae/Verrucomicrobia EF-P2 group which exhibits long branches compared with the rest of the species in this group (Figure 5). These species, in the SSU rRNA tree, do not exhibit such long branches (Figure 8). The variation in rates or differences in branch lengths seen for the Chlamydiae/Verrucomicrobia EF-P2s in Figure 5 maybe due to a difference in amino-acid replacement rates, which may be associated with the different protein structural environments (Thorne 2000) . For example, the sites on the surface of globular proteins have a 2-fold increase in rate heterogenity than compared with protein sites that are within the protein structure (Goldman, Thorne, and Jones 1998). Consequently, the Chlamydiae/Verrumicrobia EF-P2 may have a different protein structure than the other EF-Ps used in this analysis or it may be due to extreme differences of evolutionary pressures among the protein sequence because the evolutionary model used to construct Figure 5 incorporates different rate heterogeneity. In addition, the patchy distribution of the phylogenetic groups can also be attributed to HGT. Therefore, any HGT events within the bacterial groups used for the

EF-P tree and paraphyletic groups such as the Gammaproteobacteria, Deltaproteobacteria, and Firmicutes are difficult to hypothesize, since these groups are also paraphyletic in the SSU rRNA tree.

Unfortunately, it is impossible to tell the difference between a gene duplication event occurring in a common ancestor with subsequent divergence of the genes and a HGT event occurring in a common ancestor with both having subsequent gene loss. Both of these scenarios would show the distant branching patterns observed in the EF-P phylogeny. The EF-P2 clustering may also be explained by the possibility, albeit unlikely, of a number of independent HGT events. Gene loss in more than one ancestor and different rates of evolution or mutation may also explain the many incongruencies between the SSU rRNA tree and the EF-P tree (Kechris et al. 2006).

4.3 SSU rRNA Phylogenetic Tree Analysis

The SSU rRNA bacterial tree constructed by Olsen, Woese, and Overbeek (1994) has a similar topology to that of Figure 8. The Gammaproteobacteria group in the Olsen, Woese, and Overbeek (1994) and the present study display similar patterns. For example, both trees have a paraphyletic Gammaproteobacteria group separated by the Betaproteobacteria monophyletic clade (Olsen, Woese, and Overbeek 1994). Generally the higher branches or placement of the groups in our SSU rRNA tree (Figure 8) is congruent with the published tree of life such as the topology of the Aquificae, Thermotogae, and Deinococcus/Thermus bacterial group branching beside the archaeal outgroup species (Olsen, Woese, and Overbeek 1994). There are instances in Figure 8 where the bacterial groups are clustered differently such as the Spirochaetes, which are seen beside the Planctomyces and Chlamydia group in Olsen, Woese, and Overbeek (1994); however in Figure 8 the Spirochaetes appear to be grouping with the Chloroflexi. It is possible

that the different analyses in the present study compared to Olsen, Woese, and Overbeek (1994) could exhibit differences in topology due to the different species and number of taxons used in the construction of the two SSU rRNA trees. In addition, the published tree of life is a maximum likelihood tree; however, there is no mention of a use of a model of evolution (Olsen, Woese, and Overbeek 1994). Also, since this study by Olsen, Woese, and Overbeek (1994) there have been numerous changes in the taxonomy of Bacteria, mainly, some species of Bacteria's have been reclassified and their names changed.

In a study by Brochier et al. (2002), two different types of phylogenetic trees were constructed. One consisted of a concatenated alignment containing the SSU and LSU rRNA genes whereas the other tree was constructed from a concatenation of translational apparatus proteins such as recA and ribosomal proteins. Again, most of the topology is congruent with the SSU rRNA tree generated (Figure 8) in the present study. For example, the groups Cyanobacteria, Spirochaetes, Actinobacteria, Chlamydiae/Verrucomicrobia, Spirochaetes, and Bacteroidetes/Chlorobi group appear to be monophyletic. In addition, the hyperthermophiles, *Aquifex aeolicus* and *Thermotoga martima*, appear to be branching together. In both trees the Firmicutes are paraphyletic, however, the Fusobacteria and Actinobacteria is separating the Firmicutes clades in our phylogeny while in the rRNA fusion tree it is separated by the Bacteroidetes/Chlorobi group. The Gammaproteobacteria in the fusion tree is monophyletic; though, the fusion tree is based on a smaller number of taxa then our SSU rRNA tree. Incongruence seen between the phylogenetic trees in the present thesis and published trees may be due to different data sets, different reconstruction methods, or different models of evolution used.

4.4 EF-P Nucleotide and Amino Acid Sequence Identity Analyses

Nucleotide and amino acid sequence identity between distantly related species, as determined by BLAST hits, is another common method for finding putative horizontally transferred genes (Koonin, Makarova, and Aravind 2001). For example, using sequence analysis Gupta (2004) was able to detect a recent HGT between various Bacteroidetes species and *Brucella* spp. in the GyrB protein. There is an insertion in the conserved regions of GyrB that is shared by various Bacteroidetes species but usually not discovered in other groups of Bacteria (Gupta 2004). However, using sequence analysis of the conserved GyrB protein region from many different bacterial species, Gupta (2004) discovered this insertion in the Alphaproteobacteria, *Brucella* sp.. Since the insertion was not observed in another Proteobacteria and thus was more similar to Bacteroidetes species CyrB protein, it is hypothesized that this insertion may have occurred independently or was acquired by a HGT event (Gupta 2004).

A disadvantage of sequence analysis is that it can only detect relatively recent HGT events, whereas rate variation or convergent evolution may in some instances explain the presence of similar sequences in distantly related species (Koski and Golding 2001). Having noted no similar EF-P sequences between distantly related species suggests that the HGT seen in the phylogenetic tree may be a result of an ancient HGT event allowing enough time to pass for the foreign gene sequence to evolve and acquire its recipient genomic sequence characteristics thus avoiding detection. Also, the HGT may have occurred between genomes that are closely related and thereby hard to verify. The sequence analyses were consistent with the topology of the EF-P tree. The amino acid sequence identity trends concerning EF-P1 and EF-P2 also supported the clustering of most of the EF-P2 sequences together.

4.5 EF-P Codon Usage and Genomic GC Content Analyses

Genomic codon preference and GC content is known to vary from genome to genome and from gene to gene. Unusual codon usage and GC content of a gene in an organism may also be due to convergent evolution (Medigue et al. 1991a). Both codon usage and GC content may also be related to the level of expression of the gene. For example, a gene that is known to be expressed at a high level often has a higher GC content and uses more abundant tRNA codons (Garcia-Vallve et al. 2003). In contrast, low GC content in a gene may indicate structural constraints; for example, ribosomal proteins favor the AAA codon for lysine because it is preferred for RNA protein interactions (Medrano-Soto et al. 2004). Within an *E. coli* cell there is typically about one EF-P molecule per ten ribosomes or roughly one-tenth of that of EF-G (An et al. 1980). Another factor affecting codon usage in Bacteria is the selection pressure among synonymous codons, and thus highly expressed genes are encoded using the optimal synonymous codon for that species (Sharp and Matassi 1994). Evidence shows that codon usage typically correlates with evolutionary distance of the species; however, distantly related species may have similar codon usage due to convergent evolution from living in similar environments (McInerney 1998). Codon usage analysis have been used to recognize two types of genes; highly expressed genes and horizontally transferred genes (Gharbia et al. 1995). Unfortunately, codon usage and GC content analysis is unable to detect relatively old HGT due to amelioration and HGT between genomes with similar GC content and codon usage (Medigue et al. 1991b; Medigue et al. 1991a; Kechris et al. 2006).

In a previous study, Medigue et al. (1991a) observed that recently acquired foreign genes located in *E. coli* displayed a different codon usage than native *E. coli* genes. Using factorial correspondence analysis, Medigue et al. (1991a) discovered three different clusters of codon

usage in *E. coli* genes. The first two clusters are comprised of genes that are expressed lowly/rarely and highly. The third class contains genes which have atypical codon usage and are hypothesized to be horizontally transferred (Medigue et al. 1991a). Medigue et al. (1991a) was unable to detect atypical codon usage for *efp* in *E. coli* supporting our results that the horizontal transfer may have been an ancient event. In addition, when the complete genome of *E. coli* strain MG1655 was first sequenced, a codon analysis was performed to determine the presence of horizontally transferred genes within the genome and this study did not detect HGT of EF-P2, which again is consistent with our own codon analysis findings (Lawrence and Ochman 1998).

Using clustering techniques, most of the *tuf* genes did not cluster with the *efp* genes, which was expected due to their different levels in gene expression. The *tuf* control was used to try to determine the difference in codon usage for a highly expressed gene and a HGT event. The following organism's *efp* genes clustered with the highly expressed *tuf* control: *Thermobifida fusca* YX EF-P1, *Dechloromonas aromatica* RCB EF-P1, *Xanthomonas oryzae* pv. Oryze KACC10331 EF-P1, *Buchneria aphidicola* str. APS EF-P1, *Vibrio cholerae* O1 biovar eltor str. N16961 EF-P2, and *Salmonella enterica* subsp. *enterica serovar Paratyphi* A str. ATCC 9150 EF-P2. Lynn, Singer, and Hickey, (2002) analyzed the pattern of synonymous codon usage in more than 80 000 genes from 40 completely sequenced prokaryotic genomes. They determined that codon usage can be affected by two factors; genomic GC content and the growth of the organisms at high temperature (Lynn, Singer, and Hickey 2002). In addition, Lynn, Singer, and Hickey (2002) discovered cases of convergent evolution in codon usage amongst GC-rich species such as the GC-rich gram-positive *M. tuberculosis*, GC-rich gram-negative *P. aeruginosa*, and GC-rich archaeal species *Halobacterium*. The *efp* genes in these organisms may be expressed at a higher rate having more similar optimal codons like the *tuf* genes due to natural

high temperature habitats and higher genomic GC content. However, none of these organisms have been isolated from high temperature environments except for *Thermobifida fusca* YX which is a moderate thermophilic soil bacterium and has an optimal growth temperature of 55°C (Lykidis et al. 2007). The GC content of *Thermobifida fusca* YX EF-P1, *Dechloromonas aromatica* RCB EF-P1, *Xanthomonas oryzae* pv. Oryze KACC10331 EF-P1, *Buchneria aphidicola* str. APS EF-P1, *Vibrio cholerae* O1 biovar eltor str. N16961 EF-P2, and *Salmonella enterica* subsp. *enterica serovar Paratyphi* A str. ATCC 9150 EF-P2 are 67.5, 59.3, 63.7, 26, 47, and 51.4 respectively. The first three organisms and *S. enterica* subsp. enteric serovar Paratyphi A str. ATCC9150 have higher GC content, which may skew the codon usage, compared with the average GC content, 47.6 of all the organisms used in this study. In addition, *Buchneria aphidicola* str. APS *efp*1 has a very low GC content, which is surprising as it grouped with *tuf* genes which are highly expressed and thus usually have a higher GC content. The *efp* and the highly expressed *tuf* genes may have clustered together for *B. aphidicola* str. APS because the optimal codons may not be GC rich in this organism. In addition, *Vibrio cholerae* O1 biovar eltor str. N16961 *efp*2 may be expressed at a higher rate than its *efp*1 gene.

Many clusters contain species from the same bacterial groups; however, there are several clusters which contain a mix of bacterial groups. This may be the result of convergent evolution or show the atypical codon usage pattern that may denote a HGT event. Again, *Porphyromonas gingivalis* W83 *efp*2 or *efp*1 shows codon usage evidence of being a recent gene duplication as the *efp*1 and the *efp*2 of this species have similar codon usage and are clustering together. *Chlamydia muridarum* Nigg *efp*1 and *efp*2 genes have similar codon usage but they show no evidence of being a gene duplication event in the EF-P phylogenetic tree. *Chlamydia muridarum* Nigg is the only Chlamydiae/Verrumicrobia organisms to have their *efp*1 and *efp*2 clustering

together based on codon usage. Nevertheless, *C. muridarum* Nigg's *efp*1 and *efp*2 in the phylogenetic tree are acting in concert with the *efp*1 and *efp*2 from the Chlamydiae/Verrumicrobia organisms. In addition, all the *efp*2 from the Chlamydiae/Verrumicrobia group are clustering together based on codon usage except for *C. muridarum* Nigg. *C. muridarum* Nigg's *efp*2 may have evolve faster than the other *efp*2 from Chlamydiae/Verrumicrobia causing some of the codon usage to be similar to its corresponding *efp*1 but still retain its distinct composition in order to cluster together with the *efp*2 from the Chlamydiae/Verrumicrobia group in the phylogenetic tree.

The GC content analysis supports the phylogenetic (Figure 5) analysis in the putative HGT of the Bacterioidetes/Chlorobi species *Bacteroides fragilis* YCH46, but not the rest of its clade (Table 9). The Hamming distance supports the phylogenetic case that the *Chlamydophilia abortus* S26/3 *efp*2 may be horizontally transferred and is not a gene duplication as hypothesized (Table 10). However, it is unlikely that a HGT event occurred (and not a gene duplication event) in the common ancestor of this group and then the same vertical inheritance pattern occurred for both *efp* and *efp*2. In *P. gingivalis* W83, a recent gene duplication event was supported by having similar GC content to each other (*efp*1, 53.8% and *efp*2, 54.3%).

4.5.1 Optimal Codon Usage Analyses

Optimal codons for *E. coli*, *B. subtilis*, *S. cerevisiae*, *S. pombe*, and *D. melanogaster* are UUC, UAC, AUC, AAC, GAC and GGU (Sharp and Devine 1989). Rare codons for *E. coli* are AGG, AGA, AUA, CUA, CGA, CGG, CCC, and UCG (Ikemura 1981; Zhang, Zubay, and Goldman 1991). As seen in Table 11, *E. coli* spp. *efp*2 uses more rare codons and fewer optimal codons when compared with *efp*1 and the highly expressed *tuf* genes. Both *efp*1 and *efp*2 use the rare codon UCG quite frequently when compared with the *tuf* gene. The usage of rare and sub-

Table 9. A list of the putative HGT genes detected using a 10% G+C content difference between *efp* and the genome.

Organism/Group	Genomic G+C %	*efp* G+C %
Anaeromyxobacter dehalogenans 2CP-C Deltaproteobacteria *efp*1	74.8	63.4
Bacteroides fragilis YCH46 Bacterioidetes/Chlorobi *efp*1	33.5	45.1
Burkholderia xenovorans LB400 Betaproteobacteria *efp*1	68	55.9
Pseudomonas fluorescens Pf-5 Gammaproteobacteria *efp*1	67	54.6
Ralstonia metallidurans CH34 Betaproteobacteria *efp*1	67	56.9

Table 10. A list of the putative HGT genes detected using a Hamming distance difference of 0.08 between the smallest Hamming distance calculated and *efp*'s corresponding genome.

Organism/Group	Minimum Genome Depicted by Hamming distance	Hamming distance of its own genome	Hamming distance of minimum genome
Chlamydophila abortus S26/3 Chlamydiae/Verrucomicrobia *efp*	*Oceanobacillus iheyensis* HTE831 Firmicutes	0.50418	0.3675
Porphyromonas gingivalis W83 Bacteroidetes/Chlorobi *efp*2	*Chlorobium tepidum* TLS Bacteroidetes/Chlorobi	0.58856	0.49915
Prochlorococcus marinus subsp. pastoris str. CCMP1986 Cyanobacteria *efp*	*Streptococcus agalactiae* 2603V/R Firmicutes	0.52563	0.42385
Synechococcus elongatus PCC 7942 Cyanobacteria *efp*	*Chlorobium tepidum* TLS Bacteroidetes/Chlorobi	0.67607	0.54562
Synechococcus sp. WH 8102 Cyanobacteria *efp*	*Bifidobacterium longum* NCC2705 Actinobacteria	0.51773	0.43771

Table 11. The difference in codon frequency in synonymous optimal and rare codons for *E. coli efp* and *tuf*. Optimal codons are highlighted in yellow and rare codons are highlighted in blue.

		Phe		Tyr		Ile			Asn		Asp		Gly				Arg						Leu						Pro				Ser					
Organism	**gene**	**UUC**	**UUU**	**UAC**	**UAU**	**AUC**	**AUA**	**AUU**	**AAC**	**AAU**	**GAC**	**GAU**	**GGU**	**GGC**	**GGA**	**GGG**	**AGG**	**AGA**	**CGU**	**CGC**	**CGA**	**CGG**	**CUA**	**CUU**	**CUC**	**CUG**	**UUA**	**UUG**	**CCC**	**CCU**	**CCA**	**CCG**	**UCG**	**UCU**	**UCC**	**UCA**	**AGU**	**AGC**
Escherichia coli CFT073	efp1	3.7	1.06	2.65	0.53	3.17	0	0.53	3.7	0.53	1.06	5.82	4.76	4.23	0	0	0	0	1.59	2.12	0	0	0	0.53	0	6.88	0.53	0	0	0.53	0	3.7	3.7	2.65	1.06	0	0.53	0.53
Escherichia coli CFT073	efp2	1.45	2.17	2.17	1.81	1.81	1.09	3.62	1.81	1.09	3.26	3.62	2.17	3.62	0.72	1.45	0.36	0.36	3.26	1.81	0.36	0.72	0	1.45	0	5.07	0.36	1.81	0.72	0.36	1.45	3.26	3.26	1.45	0.72	0	0.36	1.45
Escherichia coli K12	efp1	3.17	1.59	2.65	0.53	3.17	0	0.53	3.7	0.53	1.59	5.29	5.29	3.7	0	0	0	0	1.59	2.12	0	0	0	0.53	0	6.88	0.53	0	0	0.53	0	3.7	3.7	2.65	1.06	0	0.53	0.53
Escherichia coli K12	efp2	1.45	2.17	2.17	1.81	1.81	1.09	3.62	1.81	1.09	2.9	3.99	1.81	3.62	0.72	1.81	0.36	0.36	3.26	1.81	0.36	0.72	0	1.45	0	5.43	0.36	1.45	0.72	0	1.81	3.26	3.26	1.45	0.72	0	0.36	1.45
Escherichia coli O157:H7 EDL933	efp1	3.72	1.06	2.66	0.53	3.19	0	0.53	3.72	0.53	2.13	4.79	4.26	4.79	0	0	0	0	1.6	2.13	0	0	0	0.53	0	6.91	0.53	0	0	0.53	0	3.72	3.72	2.66	1.06	0	0.53	0.53
Escherichia coli O157:H7 EDL933	efp2	1.45	2.17	2.17	1.81	1.81	1.09	3.62	1.81	1.09	3.26	3.62	1.45	3.62	1.09	1.81	0.36	0.36	3.26	1.81	0.36	0.72	0	1.45	0	5.43	0.36	1.45	0.72	0	1.81	2.9	2.9	1.45	0.72	0	0.36	1.45
Escherichia coli O157:H7 str. Sakai	efp1	3.7	1.06	2.65	0.53	3.17	0	0.53	3.7	0.53	2.12	4.76	4.23	4.76	0	0	0	0	1.59	2.12	0	0	0	0.53	0	6.88	0.53	0	0	0.53	0	3.7	3.7	2.65	1.06	0	0.53	0.53
Escherichia coli O157:H7 str. Sakai	efp2	1.57	1.57	2.09	2.62	2.62	0	3.14	2.09	0.52	3.66	4.19	2.09	3.14	0.52	1.57	0	0.52	4.19	1.57	0	0	0	0.52	0	7.33	0.52	1.05	0.52	0	2.09	2.62	2.62	1.05	0.52	0	0	1.57
Escherichia coli UTI89	efp1	3.17	1.59	2.65	0.53	3.17	0	0.53	3.7	0.53	2.12	4.76	4.23	4.76	0	0	0	0	1.59	2.12	0	0	0	0.53	0	6.35	0.53	0.53	0	0.53	0	3.7	3.7	2.65	1.06	0	0.53	0.53
Escherichia coli UTI89	efp2	1.45	2.17	2.17	1.81	1.81	1.09	3.62	1.81	1.09	3.26	3.62	2.17	3.62	0.72	1.45	0.36	0.36	3.26	1.81	0.36	0.72	0	1.45	0	5.07	0.36	1.81	0.72	0.36	1.45	3.26	3.26	1.45	0.72	0	0.36	1.45
Escherichia coli W3110	efp1	3.17	1.59	2.65	0.53	3.17	0	0.53	3.7	0.53	1.59	5.29	5.29	3.7	0	0	0	0	1.59	2.12	0	0	0	0.53	0	6.88	0.53	0	0	0.53	0	3.7	3.7	2.65	1.06	0	0.53	0.53
Escherichia coli W3110	efp2	1.57	1.57	2.09	2.62	2.62	0	3.14	2.09	0.52	3.14	4.71	2.62	3.14	0	1.57	0	0.52	4.19	1.57	0	0	0	0.52	0	7.33	0.52	1.05	0.52	0	2.09	2.62	2.62	1.05	0.52	0	0	1.57
Escherichia coli CFT073	tuf	3.17	0.24	2.2	0.24	6.34	0.49	0.73	1.71	0.24	4.88	0.98	4.88	4.88	0	0.49	0.24	0	5.12	0.49	0	0	0	0.24	0.24	6.59	0	0	0	0.24	0.24	4.63	0	1.95	0.73	0	0	0.24
Escherichia coli CFT073	tuf	3.29	0.25	2.03	0.51	6.58	0	0.76	1.77	0	5.06	1.01	5.32	4.56	0	0.25	0	0	5.06	0.76	0	0	0	0.25	0	6.84	0	0	0	0	0.25	4.81	0	1.52	1.01	0	0	0.25
Escherichia coli K12	tuf	3.29	0.25	2.03	0.51	6.58	0	0.76	1.77	0	5.06	1.01	4.81	5.32	0	0.25	0	0	5.32	0.51	0	0	0	0.25	0	6.84	0	0	0	0	0.25	4.81	0	1.77	0.76	0	0	0
Escherichia coli K12	tuf	3.29	0.25	2.28	0.25	6.58	0	0.76	1.77	0	5.06	1.01	4.81	5.06	0	0.25	0	0	5.06	0.76	0	0	0	0.25	0	6.84	0	0	0	0	0.25	4.81	0	1.77	0.76	0	0	0.25
Escherichia coli O157:H7 EDL933	tuf	3.29	0.25	2.03	0.51	6.33	0	1.01	1.77	0	5.06	1.01	4.81	5.32	0	0.25	0	0	5.32	0.51	0	0	0	0.25	0	6.84	0	0	0	0	0.25	4.81	0	1.52	1.01	0	0	0
Escherichia coli O157:H7 EDL933	tuf	3.32	0.26	2.3	0.26	6.38	0	1.02	1.79	0	5.1	1.02	4.85	4.85	0	0.51	0	0	4.85	0.77	0	0	0	0.26	0	6.38	0	0.51	0.26	0	0.26	4.34	0	1.53	1.02	0	0	0.26
Escherichia coli O157:H7 str. Sakai	tuf	3.29	0.25	2.28	0.25	6.58	0	0.76	1.77	0	5.06	1.01	4.81	5.06	0	0.25	0	0	5.06	0.76	0	0	0	0.25	0	6.58	0	0.25	0	0	0.25	4.81	0	1.52	1.01	0	0	0.25
Escherichia coli O157:H7 str. Sakai	tuf	3.29	0.25	2.03	0.51	6.33	0	1.01	1.77	0	5.06	1.01	4.81	5.32	0	0.25	0	0	5.32	0.51	0	0	0	0.25	0	6.84	0	0	0	0	0.25	4.81	0	1.52	1.01	0	0	0
Escherichia coli UTI89	tuf	3.17	0.24	2.2	0.24	6.34	0.49	0.73	1.71	0.24	4.88	0.98	4.88	4.39	0	0.73	0.24	0	5.12	0.49	0	0	0	0.24	0.24	6.59	0	0	0	0.24	0.24	4.63	0	1.71	0.98	0	0	0.49
Escherichia coli UTI89	tuf	3.29	0.25	2.28	0.25	6.58	0	0.76	1.77	0	4.81	1.27	4.81	5.06	0	0.51	0	0	4.81	1.01	0	0	0	0.25	0	6.84	0	0	0	0	0.25	4.81	0	1.52	1.01	0	0	0
Escherichia coli W3110	tuf	3.29	0.25	2.28	0.25	6.58	0	0.76	1.77	0	5.06	1.01	4.81	5.06	0	0.25	0	0	5.06	0.76	0	0	0	0.25	0	6.84	0	0	0	0	0.25	4.81	0	1.77	0.76	0	0	0.25
Escherichia coli W3110	tuf	3.29	0.25	2.03	0.51	6.58	0	0.76	1.77	0	5.06	1.01	4.81	5.32	0	0.25	0	0	5.32	0.51	0	0	0	0.25	0	6.84	0	0	0	0	0.25	4.81	0	1.77	0.76	0	0	0

optimal codons could cause *efp*2 to be more lowly/rarely expressed than *efp*1. In addition, using rare codons may be one of the mechanisms of the organism to regulate the expression *efp*2.

4.6 Gene Order Conservation Analyses

The gene order for *efp* does not appear to be conserved amongst all the completely sequenced Bacteria. This differs from other elongation factors such as elongation factor Tu (*tufA* and *tufB*), and Elongation Factor G (*fus*) that have been observed to be within an operon configuration (Post et al. 1978; Koonin and Galperin 1997; Itoh et al. 1999).The streptomycin operon consists of two elongation factors; elongation factor G (*fus*) and elongation factor Tu (*tufA*) and ribosomal proteins in the *E. coli* organism (Koonin and Galperin 1997; Itoh et al. 1999). Elongation Factor Tu (*tufB*) is located within the tRNA-tufB operon in *Thermus thermophilus*, and *Chlamydia trachomatis* (Satoh et al. 1991; Cousineau et al. 1992). In addition, EF-Ts (*tsf*) was discovered in an operon which also codes for the ribosomal protein S2 (An et al. 1981). It appears that most of the bacterial translation elongation factors are expressed in an operon configuration.

In general, gene order is poorly conserved amongst distantly related species if the genes are not in an operon configuration because bacterial genomes are prone to genomic rearrangements (Tamames 2001; Rogozin et al. 2002). Gene order, including operons, are easily lost during the course of bacterial evolution (Tamames 2001). In 2001, Wolf et al., determined that only 5 to 25% of genes that make up an operon are discovered in at least two genomes, when closely related bacterial and archaeal species are excluded. However, comparative analysis of gene order is problematic due to errors in annotation, falsely predicted genes, missing genes, and sequencing errors (Rogozin et al. 2004). To add to this confusion many parasitic bacterial

genomes, such as *Mycobacterium leprae* and *Rickettsia prowazekii,* have a higher than normal amount of pseudogenes (Rogozin et al. 2004).

4.6.1 General Trends seen in the Gene Order of efp1 and efp2

Most of the *efp*1 upstream and downstream regions have a similar gene order to the upstream and downstream adjacent regions of other *efp*1 from closely related species of the same bacterial group. Usually *efp*1 adjacent regions do not have any similarity in gene order with their corresponding *efp*2 except for the exceptions seen in Table 4. The *efp*2 adjacent upstream and downstream regions show the same trends. For example, *Shigella sonnei* Ss046 *efp*1 has similar gene order for the entire length of the upstream and downstream 10kb region with most of the *E. coli* spp. *efp*1 adjacent regions due to their close evolutionary relationship; the same trend is observed with the *efp*2 adjacent regions of these species. This general trend is consistant with the phylogenetic and sequence analysis done for *efp* in this thesis.

Only two of the organisms, *Acidobacteria bacterium Ellin* 345 and *Porphyromonas gingivalis* W83 have significant similarities between the adjacent upstream and downstream regions of *efp*1 and *efp*2, that does not include IS elements and/or partial gene similarities and/or domain similarities. In Figure 13, *Acidobacteria bacterium* Ellin345 *efp*2's adjacent upstream and downstream region shows evidence to support an ancient duplication event such as an internal repeat because the adjacent *efp*2's region contain two copies of the ABC efflux, inner membrane subunit (one in the upstream region and the other in the downstream region marked in red) while the *efp*1 adjacent regions only has one copy of this gene in its downstream region. In addition, both regions have the same configuration of the PadR-like family beside (one of) the ABC efflux, inner membrane subunit(s). An ancient internal repeat may have once been located here with enough time passing by to allow for gene loss and rearrangement in the internal

repeats, leaving behind only one *efp* sequence but two ABC efflux, inner membrane subunit proteins in *efp*2's adjacent region. Nevertheless, the two copies of the ABC efflux membrane protein flanking *efp*2 may also be caused by other genomic rearrangements such as gene duplication, homologous recombination, or a HGT event.

Porphyromonas gingivalis W83 *efp*1 and *efp*2 adjacent region gene order conservation shows evidence of being a gene duplication event. In Figure 14, *efp*1 and *efp*2 are seen sharing a similar adjacent gene, DNA-binding protein, histone-like family, and the non-coding region upstream and downstream of both *efp*1 and *efp*2. Again, this evidence supports the theory that *efp*1 or *efp*2 of *P. gingivalis* W83 may have undergone a recent gene duplication event because even the non-coding regions are conserved and have high sequence similarity with each other even though no functional constraints are believed to be acting upon these non-coding regions. In Figure 14, the similar protein and non-coding regions are flipped showing evidence of this region including *efp* were duplicated or inserted back into the genome in an inverted manner.

4.6.2 Evidence of Horizontal Gene Transfer in the Gene Order of efp

Figure 15 shows the different interactions (red lines) of the significant similarities in the adjacent regions of *efp* between distantly related organisms. All these matches show signs of possible horizontal transfers of the regions near *efp* or including *efp*. It is interesting to note which organisms are similar to each other and which are in their own isolated networks. For example, *Xanthomonas campestris pv. Campestris* str. 8004 *efp*2 is only similar to *Corynebacterium jeikeium* K411 *efp*1. However, *Chlorobium chlorochromatii* CaD3 is similar to *Idiomarina loihiensis* L2TR and *Geobacillus kaustophilus* HTA426, which have significant similarities with other organisms that are not similar to *C. chlorochromatii* CaD3. This large interacting network diagram (Figure 15) may show the remnants of the gene order from the

common ancestor before the species diverged or the genes that are similar may be involved in complex interactions making the gene order more likely conserved. In addition, the preservation of gene order may be due in part to ancient horizontal transfer of a large region allowing for gene loss and divergence to explain the differences in the regions seen today or even HGT of a single gene. However, one must also take into account sequence similarities due to convergent evolution and similar protein domains such as the common DNA-binding domains.

4.6.3 Conservation of efp, accB, and accC gene order

Table 5 lists the organisms which contain the Acetyl-CoA carboxylase genes, *accC* and *accB* beside *efp* or within the 10kb adjacent upstream and downstream region of *efp*. Acetyl-CoA carboxylase is an enzyme that catalyzes the ATP-dependent carboxylation of acetyl-CoA to malonyl-CoA in the fatty acid synthesis in plants, animals, and Bacteria (Sloane and Waldrop 2004). Acetyl-CoA carboxylase in *E. coli* and most bacterial species is assembled from four different subunits; biotin carboxyl carrier protein (*accB*), biotin carboxylase (*accC*), and carboxyl transferase which is a tetramer composed of alpha and beta subunits (*accA* and *accD*) (Marini et al. 1995). However, it has been reported that in *Mycobacterium leprae* and *M. tuberculosis*, biotinylated proteins have the functions of both *accB* and *accC* (Norman et al. 1994). The genes *accB* and *accC*, in some bacterial species such as *E. coli, Pseudomonas aeruginosa,* and *Bacillus subtilis* form a two gene operon, called the *accBC* operon (Li and Cronan 1992; Best and Knauf 1993; Marini et al. 1995). Conversely, in the Cyanobacteria species, *Anabaena*, the *accB* and *accC* genes do not form an operon and are not discovered beside each other (Gornicki, Scappino, and Haselkorn 1993). Usually, genes that encode the subunits of multiprotein complexes, in this case Acetyl-CoA carboxylase, are conserved as

operons in distantly related Bacteria (Rogozin et al. 2002). In addition, it is unknown whether or not the *accBC* operon has ever undergone horizontal transfer in its evolutionary history.

The conservation of the gene order of *efp*, *accB,* and *accC* or the preservation of the order of these three genes in distantly related Bacteria is significant since the evolution of Bacteria is detrimental to gene order conservation. The *efp* gene may be part of the *accBC* operon in only these organisms, since the *accBC* operon is not discovered extensively within Bacteria (Gornicki, Scappino, and Haselkorn 1993). In addition, Wolf et al., (2001) believed that if a conserved gene string or gene order occurs in two or three evolutionarily distantly related bacterial species there is little doubt that the genes are part of an operon again because of the limited gene conservation in Bacteria.

On the other hand, *efp* may be located beside the *accBC* operon by chance. Because of *efp*'s close proximity to an operon it may have increased the chances that *efp* hitchhiked along with the 'selfish' operon during a horizontal transfer event of the operon. This would explain the preservation of the *efp* and *accBC* operon gene order in distantly related species. Using sequence similarity analysis to determine predicted operons and then phylogenetic analysis, Omelchenko et al., (2003) discovered evidence for the horizontal transfer of entire operons such as the acquisition of the Sulfate/molybdate transport from *Bacillus halodurans* BH3128-BH3130 from a gram-negative Bacteria and a DNA repair SbcDC xenologous operon in *Vibrio cholerae* from a gram-positive Bacteria (Omelchenko et al. 2003). In addition, Rogozin et al. (2002) proposed that genomic hitchhiking is prominent in gene neighbourhoods that consist of genes coding for translational machinery. Rogozin et al. (2002) believed that genomic hitchhiking was responsible for the minority of genes that had no obvious functional connection to the main coherent function of the majority of genes that comprised the gene neighbourhood. The efp gene may

associate with the *accBC* operon due to the advantage of the hitchhiker being expressed at a particular level and have a similar regulation pattern of the operon (Rogozin et al. 2002).

In Table 5, the *efp*2 of the Chlamydiae/Verrumicrobia group are the only *efp*2 showing signs of horizontal transfer in the phylogenetic tree that has conserved gene order with the *accBC* operon. This may show evidence of an ancient horizontal transfer from one of the organisms that have the conserved gene order of *efp* beside the *accBC* operon in Table 5 to a common ancestor of the Chlamydiae/Verrumicrobia group causing the *efp*2 to cluster together in the EF-P tree (Figure 5) and preserve the *efp*, *accBC* operon gene order. However, I am unable to discern which organisms the *efp*2 region from Chlamydiae/Verrumicrobia may have originated from due to the wide spread pattern of the organisms that contain the conserved gene order of *efp* and the *accBC* operon in the phylogenetic tree and the *efp*2 group from Chlamydiae/Verrumicrobia are not clustering with any other organisms in the phylogenetic tree (Figure 5). Our results are in agreement with the discovery by Rogozin et al. (2002) that *efp* was in the same gene neighbourhood as *accC* and *accB*. In addition, *efp* is part of a gene neighbourhood whose principal functions include lipid metabolism, amino acid metabolism, and those secondary function include translational components (Rogozin et al. 2002).

4.6.4 Insertion Sequence Elements and Composite Transposons

Insertion sequence (IS) elements were discovered in the adjacent upstream and/or downstream region of the orthologous *efp* gene in Table 6 and Figure 19. Bacterial transposon or IS elements are know for being able to move in and out of genomes or replicons in a process called transposition (Snyder and Champness 2003). Plasmids are often discovered with IS elements which may allow for the formation of high-frequency recombination strains and plasmids containing chromosomal DNA thereby allowing horizontal transfer by conjugation

(Snyder and Champness 2003). Having IS elements shows evidence of transposition events occurring near *efp* and perhaps including *efp*. In addition, the presence of IS elements could mean a number of different chromosomal reshufflings may have occurred such as genomic rearrangements (Saedler et al. 1980); inversions and translocations, increased homologous recombinations (Lieb 1980), horizontal transfers (Chandler, Clerget, and Caro 1980), gene activations (Glandsdorff, Charlier, and Zafarullah 1980), gene repressions (Saedler et al. 1974), and deletions (Chow and Broker 1981). In addition, IS elements have been horizontally transferred by all known mechanisms of HGT; transformation, conjugation, and transduction (Frost et al. 2005). One such cases involves the IS1 family, where the difference in transposition and sequence provided evidence that the IS1 elements discovered in *E. coli* K12 and *E. coli* O157:H7 originally came from *Shigella boydii* (Hsu and Chen 2003).

Interestingly, the following organisms were discovered to have their *efp* genes flanked by IS elements: *Shigella sonnei* Ss046 *efp*1, *Shigella dysenteriae* Sd197 *efp*1, *Shigella flexneri* 2a str. 2457T *efp*2, *Shigella flexneri* 2a str. 301 *efp*1, *Xanthomonas oryzae* pv. Oryzae KACC10331 *efp*1, and *Xanthomonas oryzae* pv. *Oryze* MAFF 311018 *efp*2. Two copies of the same IS elements that are flanking a DNA region are sometimes able to 'act in concert' causing the region between the IS elements and including the IS elements to become mobile and act as a large transposon (Berg and Howe 1989). These large transposons composed of two IS elements and the DNA region between them are called composite or compound transposons and are usually denoted as Tn (Mahillon and Chandler 1998). The *Shigella* and *Xanthomonas* spp. may have a composite transposon containing the *efp* gene providing evidence of the mechanism behind the duplication or horizontal transfer of *efp* in these organisms. An ancient replicative transposition event may have allowed the duplication of *efp* in these Gammaproteobacteria

organisms with subsequent gene loss, divergence, and genomic rearrangments to explain the differences in gene order. In addition, plasmids can be assembled from the segment of DNA bracketed by the two IS elements (Snyder and Champness 2003), thereby allowing for the horizontal transfers of the *efp* gene by conjugation. The plasmid that contains *efp* and the IS elements, if conjugation was successful, can be integrated into the recipient chromosome by homologous recombination between similar IS elements. Elongation factor P may be more mobile and prone to genomic rearrangments due to the presence of IS elements in the adjacent upstream and downstream regions of *efp*.

The *Shigella* genomes are rich in IS elements, for example the *S. flexneri* genome contains 314 IS elements alone. This is a 7 fold increase over the number of IS elements that are observed in its close relative, *E. coli* K12 (Jin et al. 2002). The species, *S. dysenteriae*, *S. sonnei*, and *S. flexneri* also have a high copy numbers of IS1, (more than 20 copies) in their genomes (Hsu and Chen 2003). In addition, some of the *Shigella* species in Figure 19 were discovered to have their *efp* genes flanked by IS1 family elements, which is the predominant IS family in these genomes, followed by IS600, IS2, and IS4 (Jin et al. 2002). *S. dysenteriae efp*1 and *efp*2 also contain iso-IS1 elements which are iso-insertion sequences of IS1 and are homologous to IS1 elements (Ohtsubo et al. 1981). Bacterial IS1 elements are one of the smallest self-sufficient IS elements (Mahillon and Chandler 1998). In a standard mating assay, Mahillon and Chandler, (1998) discovered that IS1 natural transposition occurs at a low frequency, approximately 10^{-7}. IS1 elements may be responsible for the duplicate copies of *efp* in the *Shigella* genomes because IS1 can transpose using both mechanisms; replicative and conservative (insertion with duplication and insertion without duplication, respectively) (Galas and Chandler 1982). IS1 elements have been known to comprise many different composite transposons such as Tn9 and

Tn1681 (So, Heffron, and McCarthy 1979; Mahillon and Chandler 1998). The IS1 elements in Tn9 and Tn1681 are either in direct or inverted orientation (So, Heffron, and McCarthy 1979). The composite transposon Tn1681 contains a heat-stable toxin gene and a region of DNA that is approximately 8.9kb in length, which is approximately the same length of the DNA segment that contains the *efp* gene flanked by the IS1 elements (So, Heffron, and McCarthy 1979). The analysis of IS1 orthologous from distantly related species reveals a complex evolutionary relationship of horizontal gene transfer and multiple recombination events (Lawrence, Ochman, and Hartl 1992). The IS element in *Xanthomonas oryzae* pv. Oryze MAFF 311018 *efp*2 and *Xanthomonas oryzae* pv. Oryzae KACC10331 *efp*1 have been only identified by sequence homology and some of the suspected IS elements are labeled only as suspected or putative transpoases. However, the IS elements in *Xanthomonas oryzae* pv. Oryze MAFF 311018, which is discovered downstream from *efp*2, is identified as an IS1112 element.

An ancient replicative transposition event by IS1 elements may have been the cause of the duplication of *efp*2 or the *efp*1 genes in the common ancestor of the *Shigella* spp. or even help facilitate the horizontal transfer of these genes. However, none of the other Gammaproteobacteria *efp*2 organisms that are grouping with the above organisms (that contain IS elements) in the EF-P tree (Figure 5), have IS elements. Therefore, it may be that a composite transposon duplicated the *efp* gene or helped with the ancient horizontal transfer of the *efp* gene in a common ancestor of the Gammaproteobacteria organisms (the ones that have *efp*2) with enough time passed by to allow for gene loss, divergence, genomic rearrangments, and some of the IS elements to transpose away by the conservative mechanism leaving no trace.

Leptospira interrogans serovar Lai str. 56601 contains a putative transpoase upstream of *efp*. This putative transposase contains domains from transpoases 9 which is usually discovered

in members of the IS111A/IS1328/IS1533 family and transposase 20 from the IS116/IS110/IS902 family. Interestingly, IS elements not only cause genomic rearrangements but are able to activate and influence the expression of neighboring genes (Mahillon and Chandler 1998). Future work could include experimentally determining if the putative IS element upstream of *efp* in *Leptospira interrogans serovar* Lai str. 56601 is effecting the level of gene expression of *efp*.

4.7 Motifs Analyses

The motifs predicted from the 250bp upstream region of *efp* using TAMO 1.0 were statistically significant (p-value ≤ 0.001). However, experimental verification will be needed in order to prove that these are indeed the biologically functioning binding sites. Future work would include using a longer upstream region (>250bp) to predict the motif. Using orthologous sequences reduces the number of false-positive predictions of TFBS (Zhang and Gerstein 2003). Prediction of true motifs is often erroronus because the biologically significant binding site is usually not the best scoring motif observed. By clustering orthologous sequences based on the different characteristics of the organisms (natural habitat, 16S rRNA distance, genomic GC content, genomic size, and number of orthologous transcription factors) similar binding sites will be clustered together to improve the chances of identifying the biologically significant motif. Future work would include a full permutation of all the different combinations of characteristics to define the clusters of upstream regions in order to predict more biologically significant motifs and including physiology characteristic to cluster with. Additional work should also include data-mining databases which contain bacterial microarray experiments to determine if any genes are co-expressed with *efp* and thus may have the same TFBS as *efp* to increase the chances of

predicting a biologically significant motif. The top 5 motifs are listed because there maybe more than one binding site in the upstream region of *efp*.

4.7.1 Comparing TFBS from RegulonDB 5.0

When comparing the position frequency matrices of known TFBS from *E. coli*'s regulonDB 5.0 (Salgado et al. 2006) only 2 similar TFBS were discovered out of the 5 clusters that contained Gammaproteobacteria upstream sequences. Cluster 8 which are comprised of all the *E. coli* spp. *efp*1 and *efp*2 upstream sequences had no putative motifs that were similar to any of the known TFBS from regulonDB 5.0. EF-P maybe transcribed using a novel transcription factor or the TFBS may have a weak consensus making it incomparable to known *E. coli* TFBS. In addition, *efp*1 and *efp*2 may have different TFBS and so in clustering the duplicated *efp* genes together may have caused an erroneous motif prediction. As phylogenetic distances between the species increases, so does the probability that there are different regulation for the orthologous gene (McCue et al. 2002). For example, the arginine biosynthetic pathway for *E. coli* is repressed by a different protein than in *Pseudomonas aeruginosa* due to non-orthologous gene displacement (McCue et al. 2002).

Since all of the clusters contained Proteobacteria *efp* upstream sequences, the motifs from known *E. coli* TFBS that matched with the predicted motifs may still be biologically significant because the Proteobacteria group is closely related. For example, cluster 10 has the most similar motifs to Sigma70's -35 binding site for two of its predicted motifs, which are very similar to one another. The absence of a match with sigma70's -10 element binding site may be explained by: the -10 element is only partially predicted within the discovered motif, the TFBS has a different -10 binding site than what is known, or the -10 consensus sequence is very weak. In

addition, the organisms that were used in the prediction are from a wide variety of bacterial groups, while all the TFBS from RegulonDB 5.0 came from *E. coli*.

4.7.2 Comparing Experimentally Discovered Binding Sites of Regulatory Proteins

The experimentally discovered binding sites of regulatory proteins from RegTransBase were compared with the top 5 predicted motifs. The position frequency matrices were calculated from RegTransBase binding site sequence alignments from organisms within the same or different bacterial group depending on the species used to experimentally identify these sites. The fatty acid degradation regulator (FadR) protein is responsible for repressing transcription of genes needed in the transportation, activation, and β-oxidation of fatty acid such as *fadL*, *fadD*, *fadE*, *fadBA*, *fadH*, *fadI*, and *fadJ* (Klein et al. 1971; Clark 1981; Nunn 1986; Campbell, Morgan-Kiss, and Cronan 2003). In addition, FadR activates genes responsible for encoding essential enzymes of the unsaturated fatty acid biosynthesis (*faba* and *fabb*) and *iclR* which is responsible for repressing the glyoxylate operon (Gui, Sunnarborg, and LaPorte 1996; Campbell and Cronan 2001; Iram and Cronan 2005) . So far, the FadR protein is known to regulate the series of reactions which convert fatty acids to acetyl-CoA and the utilization of acetyl-CoA by the citric acid cycle (Iram and Cronan 2006). The FadR motif binding site imperfectly matched cluster 10's s..s.ss.....s.ss..Ss motif and cluster 4's Ss.Gsts..s..s.ss.ksC motif keeping in mind that *accB* and *accC* genes are adjacent to *efp* in certain organisms. As mentioned before, acetyl-CoA carboxylase catalyzes the ATP-dependent carboxylation of acetyl-CoA to malonyl-CoA in the fatty acid synthesis in Bacteria and interestingly, the binding site is for the FadR protein, which is a dual transcriptional regulator of fatty acids metabolism (Sloane and Waldrop 2004; Kazakov et al. 2007). FadR protein may also regulate the *accBC* operon due to there similar functions. However, only *Myxococcus xanthus* DK 1622 is known to contain the *accBC* operon beside *efp*

in cluster 10 and the *accBC* operon is in the downstream region of *efp*. Cluster 4, contains the upstream sequences of the Deinococcus-Thermus group, *Thermus thermophilus* HB27 *efp*1, *Thermus thermophilus* HB8 *efp*1, and *Deinococcus radiodurans* R1 *efp*1 which have the *accBC* operon again downstream of *efp*. Not all the sequences in both clusters 10 and 4 contained the *accBC* operon adjacent to *efp* and thus, it maybe by pure chance that the position frequency matrices of the predicted motifs were similar to the FadR binding site. Conversely, the non-coding regions may have retained a FadR binding site because perhaps at one time the *accBC* operon may have been present in most of the Bacterial species used to define the cluster before being lost due to divergence or genomic rearrangements. In addition, the regulation of the *accBC* operon may have a binding site in the upstream region of *efp* causing the regulation of these genes to be linked as an operon.

In addition to FadR protein, EF-P may also be regulated by CRP protein or the FNR protein because their position frequency matrices were similar to some extent to the predicted motifs upstream of *efp*. CRP and FNR may have another unknown function related to the regulation of *efp* expression. The cyclic AMP receptor (CRP) protein from Gammaproteobacteria is a cAMP sensitive transcriptional dual regulator (Kazakov et al. 2007). In *E. coli*, CRP is a well known global regulatory protein (Balsalobre, Johansson, and Uhlin 2006). FNR is an oxygen sensitive transcriptional dual global regulator (Kazakov et al. 2007). FNR activates genes encoding enzymes used in the anaerobic oxidation of carbon sources and the anaerobic reduction of alternative terminal electron acceptors; and proteins needed to transport the carbon sources or electron acceptors (Kang et al. 2005). In addition, FNR is known to repress the transcription of genes which encode enzymes that are essential for aerobic metabolism such as NADH dehydrogenase II and cytochrome oxidases (Kang et al. 2005). The CRP and FNR proteins are

also structurally related transcriptional factors; both consist of one DNA-binding domain, and a sensory function domain, FNR binds oxygen while CRP binds cAMP (Crack et al. 2007). The difference or similarities between the predicted motifs and the known motifs may correlate to the strength or weakness of the consensus sequence at the binding site which correlates to how strongly or tightly the protein will bind to the site. In a recent genome-wide expression analysis to determine the genes that FNR regulates in Bacteria, *efp* was not mentioned (Kang et al. 2005).

Chapter 5: GENERAL CONCLUSIONS

Bacterial translation uses multiple protein factors to assist in the elongation stage. An auxiliary protein factor, elongation factor P, was first discovered in 1975 in the bacterium *E. coli* (Glick and Ganoza 1975b). EF-P may facilitate the translation of proteins by stimulating peptide bond synthesis for a number of different aminoacyl-tRNA molecules in conjunction with the 70S ribosome peptidyl transferase (Ganoza, Kiel, and Aoki 2002). Despite having an important role in the translation process, not much is known about the evolutionary history and conservation of *efp*. The present study was initiated to facilitate a better understanding of the conservation of EF-P and address the following objectives:

1. Confirm evidence of HGT and gene duplication of EF-P using comparative phylogenies, sequence similarities, genomic GC content, and codon usage.
2. Determine the gene order of the orthologous EF-P genes.
3. Discover the motif(s) present in the upstream regions of the orthologous EF-P genes.

In this study, EF-P amino acid and nucleotide sequences were retrieved from all completely sequenced bacterial genomes stored in the GenBank database. Using NCBI genomic BLAST, 322 genomes studied had an EF-P gene (*efp*). Sixty-nine Bacteria carried an EF-P duplicate in their genome. The ubiquitous *efp* suggests that this is an important, if not essential, protein in the functioning of the bacterial cell. The EF-P sequences were then used to construct a protein phylogenetic tree, which provided evidence of horizontal and vertical gene transfer as well as gene duplication. A SSU rRNA tree was also constructed to determine congruency of our protein tree and accuracy of our construction process. For each EF-P, GC content, codon bias, nucleotide sequence, and amino acid sequence were analyzed to confirm any of the suspected HGT and/or

gene duplication events; however, the results of these analyses suggest that most of the suspected HGT is undetectable using sequence analysis and thus may be due to ancient events.

Most of the putative HGT events that can be deduced from the EF-P phylogenetic tree are believed to have occurred in the common ancestor of the bacterial groups and thus are probably ancient events. The methods employed; sequence, GC content, and codon usage analysis are unable to detect such ancient events because the unique signature of the foreign gene had already been lost or ameliorated. Some of the GC content and codon usage results are inconsistent with the EF-P tree (Figure 5; Table 9; Table 10). These inconsistencies can be easily explained by a difference in gene expression levels, convergent evolution, and/or pure chance that there is another genome more similar in GC and codon usage to the *efp* genes.

Codon analysis can only detect recent HGT event. Lawrence and Ochman (1998) estimated the oldest HGT event they could detect, in their study before amelioration of the HGT gene, is after the *E. coli* species diverged from the *Salmonella* species 100 million years (Myr). However, since most of the horizontally transferred genes were subsequently deleted, the transferred genes that have persisted are believed to have conferred beneficial characteristics allowing *E. coli* to exist in different ecological niches (Lawrence and Ochman 1998). Therefore, positive selection pressures acting on *E. coli* may have allowed two copies of EF-P to persist in the genome. Since codon analysis was unable to detect that EF-P2 was horizontally transferred which is clearly shown in the phylogenetic tree, that it may have been an ancient HGT event occurring over 100Myr in a common ancestor of the Gammaproteobacteria.

The HGT seen in the EF-P tree led to a complex phylogeny with topology at odds with our standard of evolutionary relationships, the SSU rRNA tree. Horizontal gene transfer is believed to be one of the main driving forces behind the evolution of cellular life and

subsequently, the existence of the three domains of life (Brown 2003). The estimates of foreign genes observed in all organism are in the range of 0% to 17% with a mean of 6% in bacterial genomes (Ochman, Lawrence, and Groisman 2000). There are examples of HGT that have been observed in nature such as the horizontal transfer of genes that are responsible for pathogenicity and antibiotic resistance (Brown 2003). The complexity hypothesis states that genes such as those encoding translation factors which are involved in numerous and complex interactions are less likely to be transferred (Jain, Rivera, and Lake 1999). Also, the newly transferred gene had to have overcome disadvantages such as a decrease in gene expression because the translation of non-optimal codons is slower, less efficient, and less accurate due to poor interactions with host tRNA (Lawrence 1999b). The newly transferred gene must have a large enough selective advantage to circumvent all the disadvantages.

Analysis of the 10kb upstream and downstream region of *efp* using BLAST BL2SEQ detected no similar gene order conservation for all bacterial species. However, I discovered more evidence to support possible HGT and gene duplication events in some bacterial organisms. In certain *Shigella* spp. (Figure 19) and *Xanthomonas* spp. *efp* is flanked by IS elements. EF-P maybe part of a composite transposon and thereby, one of the mechanisms driving an ancient duplication or horizontal transfer of *efp*. In addition, IS elements were discovered in other species listed in Table 6. While the IS elements in these species are not flanking the *efp* gene, just the presence of IS elements may have increased the chances of genomic rearrangement by homologous recombination. Homologous recombination may be another mechanism that caused an ancient duplication or horizontal transfer of *efp.*

Surprisingly, *efp* was discovered in some species to be adjacent to *accB* and *accC* genes which form a 2-gene operon in certain bacterial species. The genes *efp*, *accB*, and *accC* are a

conserved formation seen in distantly related species and therefore, is statistically significant because of the limited gene order conservation in Bacteria (Wolf et al. 2001). It may be that *efp* is part of the *accBC* operon however different their functions are. In addition, *efp*, *accB,* and *accC* may be linked in some undiscovered and unknown related function. Another explanation for the unusual preservation of the gene order of *efp* with the *accBC* operon maybe due to the 'selfish' operon theory. The close proximity of *efp* to the *accBC* operon may have allowed *efp* to hitchhike along with the operon as it was being horizontally transferred. Horizontal transfer of a region of genome may explain the conservation of gene order between distantly related bacterial species.

There were no complete matches (matches including both the same sigma factors -10 and the -35 elements binding site) amongst the known TFBS from RegulonDB 5.0 with the predicted motifs. A large factor in this case is that not all TFBS have been discovered and only in certain limited species have been determined. In addition, the predicted motifs were based on many distantly related bacterial species and not just *E. coli* sequences like in RegulonDB 5.0. When comparing the predicted motif position frequency matrices with those of the known binding sites of proteins from RegTransBase, there were matches; however, these matches were not perfect and contained some disagreements. The binding sites of CRP, FadR, and FNR were discovered to be similar to some of the predicted motifs, which denote that *efp* may be regulated using one of these proteins.

In this study, I provide the first account of a number of ancient horizontal transfers and recent gene duplications of EF-P. I have cases of *efp* genes being duplicated or horizontally transferred in Figure 5 highlighting the *efp* which shows incongruencies with our SSU rRNA tree; Table 9 highlighting the organisms which have GC content differences of the *efp* gene

compared with genomic GC content; Table 10 organism which show signs of atypical codon usage; and Figure 15 highlighting the organisms which are distantly related and have significant similarities in gene conservation. Only one organism, *Porphyromonas gingivalis* W83 has conclusively shown that either *efp*1 or *efp*2 occurred due to a recent gene duplication event. These two *efp* genes from *P. gingivalis* W83 are discovered as sister taxa on the phylogenetic tree, have similar amino acid and nucleotide sequences, similar GC content, similar codon usage, and have similar adjacent regions upstream and downstream of each other including non-coding sequences. The complex evolutionary history of EF-P shows that this gene is another exception to the complexity hypothesis joining similar genes such as *tuf*, the gene encoding elongation factor protein Tu, and some aminoacyl-tRNA synthetases (Ke et al. 2000).

Future research should concentrate on determining the expression levels of both EF-P proteins in order to verify the precise nature of the duplicates and any selective pressures that would cause certain bacterial organisms to keep one verus two *efp* genes. In addition to experimentation to determine whether or not *efp* is able to hitchhike along with horizontal transfers of the *accBC* operon, discover whether *efp* is part of the *accBC* operon in species where the *efp* and *accBC* gene order is conserved. Additional work would include resolving whether or not the IS elements that are flanking the *efp* gene in some of the *Shigella* spp. and *Xanthomonas* spp. make up composite transpositions. The motifs discovered using the program TAMO 1.0 should also be experimentally verified. However, before experimental verification of the predicted motifs, microarray data should be included to discover genes which are co-expressed with *efp* in order to facilitate in the discovery of the biologically significant motifs of the orthologous *efp* genes.

REFERENCES

Abascal, F., R. Zardoya, and D. Posada. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. **21**:2104-2105.

An, G., D. S. Bendiak, L. A. Mamelak, and J. D. Friesen. 1981. Organization and nucleotide sequence of a new ribosomal operon in Escherichia coli containing the genes for ribosomal protein S2 and elongation factor Ts. Nucleic Acids Res **9**:4163-4172.

An, G., B. R. Glick, J. D. Friesen, and M. C. Ganoza. 1980. Identification and quantitation of elongation factor EF-P in Escherichia coli cell-free extracts. Can J Biochem. **58**:1312-1314.

Aoki, H., S. L. Adams, D. G. Chung, M. Yaguchi, S. E. Chuang, and M. C. Ganoza. 1991. Cloning, sequencing and overexpression of the gene for prokaryotic factor EF-P involved in peptide bond synthesis. Nucl Acids Res. **19**:6215-6220.

Aoki, H., S. L. Adams, M. A. Turner, and M. C. Ganoza. 1997a. Molecular characterization of the prokaryotics efp gene product invovled in a peptidyltransferase reaction. Biochimie **79**:7-11.

Aoki, H., K. Dekany, S.-L. Adams, and M. C. Ganoza. 1997b. The Gene Encoding the Elongation Factor P Protein Is Essential for Viability and Is Required for Protein Synthesis. J Biol Chem. **272**:32254-32259.

Bailey, J. A., Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler. 2002. Recent segmental duplications in the human genome. Science. **297**:1003-1007.

Bailey, T. L., and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol **2**:28-36.

Baldauf, S. L., J. D. Palmer, and W. F. Doolittle. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc Natl Acad Sci USA. **93**:7749-7754.

Balsalobre, C., J. Johansson, and B. E. Uhlin. 2006. Cyclic AMP-dependent osmoregulation of crp gene expression in Escherichia coli. J Bacteriol **188**:5935-5944.

Barkay, T., and B. F. Smets. 2005. Horizontal Gene Flow in Microbial Communities. ASM News. **71**:412-419.

Berg, D. E., and M. M. Howe. 1989. Mobile DNA. American Society for Microbiology, Washington, D.C.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. Nucleic Acids Res **28**:235-242.

Best, E. A., and V. C. Knauf. 1993. Organization and nucleotide sequences of the genes encoding the biotin carboxyl carrier protein and biotin carboxylase protein of Pseudomonas aeruginosa acetyl coenzyme A carboxylase. J Bacteriol **175**:6881-6889.

Brown, J. R. 2003. Ancient Horizontal Gene Transfer. Nat Rev. **4**:121-131.

Campbell, J. W., and J. E. Cronan, Jr. 2001. Escherichia coli FadR positively regulates transcription of the fabB fatty acid biosynthetic gene. J Bacteriol **183**:5982-5990.

Campbell, J. W., R. M. Morgan-Kiss, and J. E. Cronan, Jr. 2003. A new Escherichia coli metabolic competency: growth on fatty acids by a novel anaerobic beta-oxidation pathway. Mol Microbiol **47**:793-805.

Cannone, J. J., S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Muller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell. 2002. The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics. **3**.

Chandler, M., M. Clerget, and L. Caro. 1980. IS1-promoted events associated with drug resistance plasmids. Cold Spring Harbor Symp. Quant. Biol. **45**:157-165.

Chow, L. T., and T. R. Broker. 1981. Adjacent insertion sequences IS2 and IS5 in bacteriophage Mu mutants and IS5 in lambda darg bacteriophage. J. Bacteriol. **133**:1427-1436.

Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. 2006. Toward Automatic Reconstruction of a Highly Resolved Tree of Life. Science. **311**:1283-1287.

Clark, D. 1981. Regulation of fatty acid degradation in Escherichia coli: analysis by operon fusion. J Bacteriol **148**:521-526.

Cousineau, B., C. Cerpa, J. Lefebvre, and R. Cedergren. 1992. The sequence of the gene encoding elongation factor Tu from Chlamydia trachomatis compared with those of other organisms. Gene **120**:33-41.

Crack, J. C., J. Green, M. R. Cheesman, N. E. Le Brun, and A. J. Thomson. 2007. Superoxide-mediated amplification of the oxygen-induced switch from [4Fe-4S] to [2Fe-2S] clusters in the transcriptional regulator FNR. Proc Natl Acad Sci U S A **104**:2092-2097.

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner. 2004. WebLogo: a sequence logo generator. Genome Res **14**:1188-1190.

D'haeseleer, P. 2006. What are DNA sequence motifs? Nature Biotechnology **24**:423 - 425.

Dimmic, M. W., J. S. Rest, D. P. Mindell, and R. A. Goldstein. 2002. rtREV: An Amino Acid Substitution Matrix for Inference of Retrovirus and Reverse Transcriptase Phylogeny. J Mol Evol. **55**:65-73.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res. **32**:1792-1797.

Eisen, J. A. 1995. The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. J Mol Evol. **41**:1105-1123.

Finlay, B. B., and P. Cossart. 1997. Exploitation of mammalian host cell functions by bacterial pathogens. Science **276**:718-725.

Frigaard, N. U., A. Martinez, T. J. Mincer, and E. F. DeLong. 2006. Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. Nature **439**:847-850.

Frost, L. S., R. Leplae, A. O. Summers, and A. Toussaint. 2005. Mobile Genetic Elements: The Agents of Open Source Evolution. Nat Rev Microbiol. **3**:722.

Galas, D. J., and M. Chandler. 1982. Structure and stability of Tn9-mediated cointegrates. Evidence for two pathways of transposition. J Mol Biol **154**:245-272.

Ganoza, M. C., M. C. Kiel, and H. Aoki. 2002. Evolutionary Conservation of Reactions in Translation. Microbiol Mol Biol Rev. **66**:460-485.

Garcia-Vallve, S., E. Guzman, M. A. Montero, and A. Romeu. 2003. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucl. Acids Res. **31**:187-189.

Garcia-Vallve, S., A. Romeu, and J. Palau. 2000. Horizontal Gene Transfer of Glycosyl Hydrolases of the Rumen Fungi. Mol Biol Evol. **17**:352-361.

Garrity, G. M., D. J. Brenner, N. R. Krieg, and J. T. Staley. 2005. Bergey's Manual of Systematic Bacteriology: The Proteobacteria. **2nd**.

Gevers, D., F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. Van de Peer, P. Vandamme, F. L. Thompson, and J. Swings. 2005. Opinion: Re-evaluating prokaryotic species. Nat Rev Microbiol **3**:733-739.

Gharbia, S. E., J. C. Williams, D. M. Andrews, and H. N. Shah. 1995. Genomic clusters and codon usage in relation to gene expression in oral Gram-negative anaerobes. Anaerobe **1**:239-262.

Glandsdorff, N., D. Charlier, and M. Zafarullah. 1980. Activation of gene expression by IS2 and IS3. Cold Spring Harbor Symp. Quant. Biol. **45**:153-156.

Glass, J. I., N. Assad-Garcia, N. Alperovich, S. Yooseph, M. R. Lewis, M. Maruf, C. A. Hutchison, III, H. O. Smith, and J. C. Venter. 2006. Essential genes of a minimal bacterium. Proc Natl Acad Sci USA. **103**:425-430.

Glick, B. R., S. Chladek, and M. C. Ganoza. 1979. Peptide Bond Formation Stimulated by Protein Synthesis Factor EF-P Depends on the Aminoacyl Moiety of the Acceptor. European Journal of Biochemistry **97**:23-28.

Glick, B. R., and M. C. Ganoza. 1975a. Identification of a soluble protein that stimulates peptide bond synthesis. Proceedings of the National Academy of Science USA **72**:4257-4260.

Glick, B. R., and M. C. Ganoza. 1975b. Identification of a soluble protein that stimulates peptide bond synthesis. Proc Natl Acad Sci USA. **72**:4257-4260.

Gogarten, J. P., and J. P. Townsend. 2005. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol **3**:679-687.

Goldenfeld, N., and C. Woese. 2007. Biology's next revolution. Nature **445**:369.

Goldman, N., J. L. Thorne, and D. T. Jones. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics **149**:445-458.

Goodfellow, M., G. P. Manfio, and J. Chun. 1997. Towards a practical species concept for cultivable bacteria. In:Species: the Units of Biodiversity (Claridge, M.R. and Dawah, H.A., Eds).

Gordon, D. B., L. Nekludova, S. McCallum, and E. Fraenkel. 2005. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics **21**:3164-3165.

Gornicki, P., L. A. Scappino, and R. Haselkorn. 1993. Genes for two subunits of acetyl coenzyme A carboxylase of Anabaena sp. strain PCC 7120: biotin carboxylase and biotin carboxyl carrier protein. J Bacteriol **175**:5268-5272.

Gui, L., A. Sunnarborg, and D. C. LaPorte. 1996. Regulated expression of a repressor protein: FadR activates iclR. J Bacteriol **178**:4704-4709.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. **52**:696-704.

Gupta, R. S. 2004. The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes. Crit Rev Microbiol **30**:123-143.

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser. **41**:95-98.

Hanawa-Suetsugu, K., S.-i. Sekine, H. Sakai, C. Hori-Takemoto, T. Terada, S. Unzai, J. R. H. Tame, S. Kuramitsu, M. Shirouzu, and S. Yokoyama. 2004. Crystal structure of elongation factor P from Thermus thermophilus HB8. Proc Natl Acad Sci USA. **101**:9595-9600.

Harbison, C. T., D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. 2004. Transcriptional regulatory code of a eukaryotic genome. Nature **431**:99-104.

Hooper, S., and O. Berg. 2003. Duplication is more common among laterally transferred genes than among indigenous genes. Genome Biol. **4**:R48.

Hsu, W. B., and J. H. Chen. 2003. The IS1 elements in Shigella boydii: horizontal transfer, vertical inactivation and target duplication. FEMS Microbiol Lett **222**:289-295.

Huerta, A. M., and J. Collado-Vides. 2003. Sigma70 Promoters in Escherichia coli: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals. Journal of Molecular Biology **333**:261-278.

Hughes, A. L. 1994. The Evolution of Functionally Novel Proteins after Gene Duplication. Proc Biol Sci. **256**:119-124.

Ikemura, T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol **151**:389-409.

Iram, S. H., and J. E. Cronan. 2005. Unexpected functional diversity among FadR fatty acid transcriptional regulatory proteins. J Biol Chem **280**:32148-32156.

Iram, S. H., and J. E. Cronan. 2006. The beta-oxidation systems of Escherichia coli and Salmonella enterica are not functionally equivalent. J Bacteriol **188**:599-608.

Ishibashi, Y., S. Claus, and D. A. Relman. 1994. Bordetella pertussis filamentous hemagglutinin interacts with a leukocyte signal transduction complex and stimulates bacterial adherence to monocyte CR3 (CD11b/CD18). J Exp Med **180**:1225-1233.

Itoh, T., K. Takemoto, H. Mori, and T. Gojobori. 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol Biol Evol **16**:332-346.

Jain, R., M. C. Rivera, and J. A. Lake. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci USA. **96**:3801-3806.

Jain, R., M. C. Rivera, J. E. Moore, and J. A. Lake. 2003. Horizontal Gene Transfer Accelerates Genome Innovation and Evolution. Mol Biol Evol. **20**:1598-1602.

Jin, Q., Z. Yuan, J. Xu, Y. Wang, Y. Shen, W. Lu, J. Wang, H. Liu, J. Yang, F. Yang, X. Zhang, J. Zhang, G. Yang, H. Wu, D. Qu, J. Dong, L. Sun, Y. Xue, A. Zhao, Y. Gao, J. Zhu, B. Kan, K. Ding, S. Chen, H. Cheng, Z. Yao, B. He, R. Chen, D. Ma, B. Qiang, Y. Wen, Y. Hou, and J. Yu. 2002. Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. Nucleic Acids Res **30**:4432-4441.

Jovine, L., S. Djordjevic, and D. Rhodes. 2000. The crystal structure of yeast phenylalanine tRNA at 2.0 A resolution: cleavage by Mg(2+) in 15-year old crystals. J Mol Biol. **301**:401-414.

Joyce, A. R., J. L. Reed, A. White, R. Edwards, A. Osterman, T. Baba, H. Mori, S. A. Lesely, B. O. Palsson, and S. Agarwalla. 2006. Experimental and Computational Assessment of Conditionally Essential Genes in Escherichia coli. J Bacteriol. **188**:8259-8271.

Kang, H. A., and J. W. Hershey. 1994. Effect of initiation factor eIF-5A depletion on protein synthesis and proliferation of Saccharomyces cerevisiae. J. Biol. Chem. **269**:3934-3940.

Kang, Y., K. D. Weber, Y. Qiu, P. J. Kiley, and F. R. Blattner. 2005. Genome-wide expression analysis indicates that FNR of Escherichia coli K-12 regulates a large number of genes of unknown function. J Bacteriol **187**:1135-1160.

Kazakov, A. E., M. J. Cipriano, P. S. Novichkov, S. Minovitsky, D. V. Vinogradov, A. Arkin, A. A. Mironov, M. S. Gelfand, and I. Dubchak. 2007. RegTransBase--a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. Nucleic Acids Res **35**:D407-412.

Ke, D., M. Boissinot, A. Huletsky, F. J. Picard, J. Frenette, M. Ouellette, P. H. Roy, and M. G. Bergeron. 2000. Evidence for Horizontal Gene Transfer in Evolution of Elongation Factor Tu in Enterococci. J Bacteriol. **182**:6913-6920.

Kechris, K. J., J. C. Lin, P. J. Bickel, and A. N. Glazer. 2006. Quantitative exploration of the occurrence of lateral gene transfer by using nitrogen fixation genes as a case study. Proc Natl Acad Sci USA. **103**:9584-9589.

Klein, K., R. Steinberg, B. Fiethen, and P. Overath. 1971. Fatty acid degradation in Escherichia coli. An inducible system for the uptake of fatty acids and further characterization of old mutants. Eur J Biochem **19**:442-450.

Koonin, E. V., L. Aravind, and A. S. Kondrashov. 2000. The impact of comparative genomics on our understanding of evolution. Cell **101**:573-576.

Koonin, E. V., and M. Y. Galperin. 1997. Prokaryotic genomes: the emerging paradigm of genome-based microbiology. Curr Opin Genet Dev **7**:757-763.

Koonin, E. V., K. S. Makarova, and L. Aravind. 2001. Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. Ann Rev Microbiol. **55**:709-742.

Koonin, E. V., A. R. Mushegian, and K. E. Rudd. 1996. Sequencing and analysis of bacterial genomes. Curr Biol **6**:404-416.

Koski, L. B., and G. B. Golding. 2001. The closest BLAST hit is often not the nearest neighbor. J Mol Evol. **52**:540-542.

Kyrpides, N. C., and C. R. Woese. 1998. Universally conserved translation initiation factors. Proc Natl Acad Sci USA. **95**:224-228.

Lawrence, J. 1999a. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Curr Opin Genet Dev **9**:642-648.

Lawrence, J. G. 1999b. Gene transfer, speciation, and the evolution of bacterial genomes. Curr Opin Microbiol. **2**:519-523.

Lawrence, J. G., and H. Ochman. 1998. Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci USA. **95**:9413-9417.

Lawrence, J. G., H. Ochman, and D. L. Hartl. 1992. The evolution of insertion sequences within enteric bacteria. Genetics **131**:9-20.

Lawrence, J. G., and J. R. Roth. 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics **143**:1843-1860.

Li, A.-L., H.-Y. Li, B.-F. Jin, Q.-N. Ye, T. Zhou, X.-D. Yu, X. Pan, J.-H. Man, K. He, M. Yu, M.-R. Hu, J. Wang, S.-C. Yang, B.-F. Shen, and X.-M. Zhang. 2004. A Novel eIF5A Complex Functions As a Regulator of p53 and p53-dependent Apoptosis. J. Biol. Chem. **279**:49251-49258.

Li, S. J., and J. E. Cronan, Jr. 1992. The gene encoding the biotin carboxylase subunit of Escherichia coli acetyl-CoA carboxylase. J Biol Chem **267**:855-863.

Lieb, M. 1980. IS5 increases recombination in adjacent regions as shown for the repressor gene of coliphage lambda. Gene **12**:277-280.

Liu, X. S., D. L. Brutlag, and J. S. Liu. 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol **20**:835-839.

Lykidis, A., K. Mavromatis, N. Ivanova, I. Anderson, M. Land, G. DiBartolo, M. Martinez, A. Lapidus, S. Lucas, A. Copeland, P. Richardson, D. B. Wilson, and N. Kyrpides. 2007. Genome sequence and analysis of the soil cellulolytic actinomycete Thermobifida fusca YX. J Bacteriol **189**:2477-2486.

Lynch, M. 2002. Genomics. Gene duplication and evolution. Science. **297**:945-947.

Lynn, D. J., G. A. Singer, and D. A. Hickey. 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res **30**:4272-4277.

MacIsaac, K. D., and E. Fraenkel. 2006. Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol **2**:e36.

Mahillon, J., and M. Chandler. 1998. Insertion sequences. Microbiol Mol Biol Rev **62**:725-774.

Maidak, B. L., J. R. Cole, T. G. Lilburn, C. T. Parker, Jr., P. R. Saxman, R. J. Farris, G. M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje. 2001. The RDP-II (Ribosomal Database Project). Nucl Acids Res. **29**:173-174.

Marini, P., S. J. Li, D. Gardiol, J. E. Cronan, Jr., and D. de Mendoza. 1995. The genes encoding the biotin carboxyl carrier protein and biotin carboxylase subunits of Bacillus subtilis acetyl coenzyme A carboxylase, the first enzyme of fatty acid synthesis. J Bacteriol **177**:7003-7006.

Martinez-Castilla, L. P., and E. R. Alvarez-Buylla. 2003. Adaptive evolution in the Arabidopsis MADS-box gene family inferred from its complete resolved phylogeny. Proc Natl Acad Sci U S A **100**:13407-13412.

McCue, L., W. Thompson, C. Carmack, M. P. Ryan, J. S. Liu, V. Derbyshire, and C. E. Lawrence. 2001. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. Nucleic Acids Res **29**:774-782.

McCue, L. A., W. Thompson, C. S. Carmack, and C. E. Lawrence. 2002. Factors influencing the identification of transcription factor binding sites by cross-species comparison. Genome Res **12**:1523-1532.

McInerney, J. O. 1998. GCUA: general codon usage analysis. Bioinformatics. **14**:372-373.

Medigue, C., T. Rouxel, P. Vigier, A. Henaut, and A. Danchin. 1991a. Evidence for horizontal gene transfer in Escherichia coli speciation. J Mol Biol. **222**:851-856.

Medigue, C., T. Rouxel, P. Vigier, A. Henaut, and A. Danchin. 1991b. Evidence for horizontal gene transfer in Escherichia coli speciation. J Mol Biol **222**:851-856.

Medrano-Soto, A., G. Moreno-Hagelsieb, P. Vinuesa, J. A. Christen, and J. Collado-Vides. 2004. Successful Lateral Transfer Requires Codon Usage Compatibility Between Foreign Genes and Recipient Genomes. Mol Biol Evol. **21**:1884-1894.

Mushegian, A. R., and E. V. Koonin. 1996. Gene order is not conserved in bacterial evolution. Trends Genet **12**:289-290.

Nikolaichik, Y. A., and W. D. Donachie. 2000. Conservation of gene order amongst cell wall and cell division genes in Eubacteria, and ribosomal genes in Eubacteria and Eukaryotic organelles. Genetica **108**:1-7.

Norman, E., K. A. De Smet, N. G. Stoker, C. Ratledge, P. R. Wheeler, and J. W. Dale. 1994. Lipid synthesis in mycobacteria: characterization of the biotin carboxyl carrier protein genes from Mycobacterium leprae and M. tuberculosis. J Bacteriol **176**:2525-2531.

Nunn, W. D. 1986. A molecular view of fatty acid catabolism in Escherichia coli. Microbiol Rev **50**:179-192.

Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. **405**:299 - 304.

Ohtsubo, H., K. Nyman, W. Doroszkiewicz, and E. Ohtsubo. 1981. Multiple copies of iso-insertion sequences of IS1 in Shigella dysenteriae chromosome. Nature **292**:640-643.

Olsen, G. J., C. R. Woese, and R. Overbeek. 1994. The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol. **176**:1-6.

Omelchenko, M. V., K. S. Makarova, Y. I. Wolf, I. B. Rogozin, and E. V. Koonin. 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biol **4**:R55.

Peng, W.-T., L. M. Banta, T. C. Charles, and E. W. Nester. 2001. The chvH Locus of Agrobacterium Encodes a Homologue of an Elongation Factor Involved in Protein Synthesis. J Bacteriol. **183**:36-45.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics. **14**:817-818.

Post, L. E., A. E. Arfsten, F. Reusser, and M. Nomura. 1978. DNA sequences of promoter regions for the str and spc ribosomal protein operons in E. coli. Cell **15**:215-229.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2007. Orthologous Transcription Factors in Bacteria Have Different Functions and Regulate Different Genes. PLoS Comput Biol **3**:e175.

Rodriguez, F., J. Oliver, A. Marin, and J. Medina. 1990. The general stochastic model of nucleotide substitution. J Theor Biol. **142**:485-501.

Rogozin, I. B., K. S. Makarova, J. Murvai, E. Czabarka, Y. I. Wolf, R. L. Tatusov, L. A. Szekely, and E. V. Koonin. 2002. Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res **30**:2212-2223.

Rogozin, I. B., K. S. Makarova, Y. I. Wolf, and E. V. Koonin. 2004. Computational approaches for the analysis of gene neighbourhoods in prokaryotic genomes. Brief Bioinform **5**:131-149.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.

Rossello-Mora, R., and R. Amann. 2001. The species concept for prokaryotes. FEMS Microbiol Rev **25**:39-67.

Roth, F. P., J. D. Hughes, P. W. Estep, and G. M. Church. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol **16**:939-945.

Saedler, H., J. Cornelis, B. Cullum, B. Schumacher, and H. Sommer. 1980. IS1 mediated DNA rearrangements. Cold Spring Harbor Symp. Quant. Biol. **45**:93-98.

Saedler, H., H. J. Reif, S. Hu, and N. Davidson. 1974. IS2, a genetic element for turn-off and turn-on of gene activity in E. coli. Mol Gen Genet **132**:265-289.

Salgado, H., S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sanchez-Solano, A. Santos-Zavaleta, I. Martinez-Flores, V. Jimenez-Jacinto, C. Bonavides-Martinez, J. Segura-Salazar, A. Martinez-Antonio, and J. Collado-Vides. 2006. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res **34**:D394-397.

Samonte, R. V., and E. E. Eichler. 2002. Segmental duplications and the evolution of the primate genome. Nat Rev Genet. **3**:65-72.

Sanderson, K. E., and S. L. Liu. 1998. Chromosomal rearrangements in enteric bacteria. Electrophoresis **19**:569-572.

Sankoff, D. 2001. Gene and Genome Duplication. Curr Opin Genet Devel. **11**:681 - 684.

Satoh, M., T. Tanaka, A. Kushiro, T. Hakoshima, and K. Tomita. 1991. Molecular cloning, nucleotide sequence and expression of the tufB gene encoding elongation factor Tu from Thermus thermophilus HB8. FEBS Lett **288**:98-100.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. **18**:502-504.

Sela, S., D. Yogev, S. Razin, and H. Bercovier. 1989. Duplication of the tuf Gene: A New insight into the phylogney of Eubacteria. J Bacteriol. **171**:581-584.

Sharp, P. M., and K. M. Devine. 1989. Codon usage and gene expression level in Dictyostelium discoideum: highly expressed genes do 'prefer' optimal codons. Nucleic Acids Res **17**:5029-5039.

Sharp, P. M., and G. Matassi. 1994. Codon usage and genome evolution. Curr Opin Genet Dev **4**:851-860.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of loglikelihoods with applications to phylogenetic interference. Mol Biol Evol. **16**:1114 - 1116.

Sloane, V., and G. L. Waldrop. 2004. Kinetic characterization of mutations found in propionic acidemia and methylcrotonylglycinuria: evidence for cooperativity in biotin carboxylase. J Biol Chem **279**:15772-15778.

Snyder, L., and W. Champness. 2003. Molecular Genetics of Bacteria Second Edition.

So, M., F. Heffron, and B. J. McCarthy. 1979. The E. coli gene encoding heat stable toxin is a bacterial transposon flanked by inverted repeats of IS1. Nature **277**:453-456.

Sorensen, S. J., M. Bailey, L. H. Hansen, N. Kroer, and S. Wuertz. 2005. Studying plasmid horizontal transfer in situ: a critical review. Nat Rev Microbiol **3**:700-710.

Swaney, S., M. McCroskey, D. Shinabarger, Z. Wang, B. A. Turner, and C. N. Parker. 2006. Characterization of a high-throughput screening assay for inhibitors of elongation factor p and ribosomal peptidyl transferase activity. J Biomol Screen **11**:736-742.

Swofford, D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4.

Tamames, J. 2001. Evolution of gene order conservation in prokaryotes. Genome Biol **2**:1-11.

Tatusova, T. A., and T. L. Madden. 1999. BLAST 2 Sequence, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett. **174**:247-250.

Thompson, J. E., M. T. Hopkins, C. Taylor, and T.-W. Wang. 2004. Regulation of senescence by eukaryotic translation initiation factor 5A: implications for plant growth and development. TRENDS in Plant Science **9**:174-179.

Thorne, J. L. 2000. Models of protein sequence evolution and their applications. Opin. Genet. Dev. **10**:602 - 605.

Valentini, S. R., J. M. Casolari, C. C. Oliveira, P. A. Silver, and A. E. McBride. 2002. Genetic Interactions of Yeast Eukaryotic Translation Initiation Factor 5A (eIF5A) Reveal

Connections to Poly(A)-Binding Protein and Protein Kinase C Signaling. Genetics **160**:393-405.

Weaver, R. F. 2002. Molecular Biology Second Edition. McGraw-Hill Companies, Inc.

Woese, C. R. 1987. Bacterial evolution. Microbiol Rev **51**:221-271.

Woese, C. R. 2000. Interpreting the universal phylogenetic tree. Proc Natl Acad Sci USA. **97**:8392-8396.

Woese, C. R., and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A **74**:5088-5090.

Wolf, D. M., and A. P. Arkin. 2003. Motifs, modules and games in bacteria. Current Opinion in Microbiology **6**:125-134.

Wolf, Y. I., L. Aravind, and E. V. Koonin. 1999. Rickettsiae and Chlamydiae: evidence of horizontal gene transfer and gene exchange. Trends Genet **15**:173-175.

Wolf, Y. I., I. B. Rogozin, A. S. Kondrashov, and E. V. Koonin. 2001. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res **11**:356-372.

Young, F. S., and A. V. Furano. 1981. Regulation of the synthesis of E. coli elongation factor Tu. Cell **24**:695-706.

Zhang, J. 2003. Evolution by gene duplication: an update. TRENDS Ecol Evol. **18**:292 - 298.

Zhang, S. P., G. Zubay, and E. Goldman. 1991. Low-usage codons in Escherichia coli, yeast, fruit fly and primates. Gene **105**:61-72.

Zhang, Z., and M. Gerstein. 2003. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J Biol **2**:11.