Gene Annotation of the Hypothetical Protein-Coding Genes of *Coxiella Burnetii*

A Senior Paper

Presented to

the Department of Biology

of Oral Roberts University

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science

by

Jason Salvato

May 2021

Dr. Celestino Velásquez

Table of Contents

List of Tables

List of Figures

Abstract

Genetic information of organisms and microorganisms has become readily accessible due to advances in genomic sequencing and bioinformatic technology. Despite these advances, there are numerous organisms with genome sequences that have yet to be annotated. Many of these genome sequences require manual annotation, which can uncover hypothetical protein-coding genes. Through the use of publicly available online bioinformatics tools, such as BLAST, T-COFFEE, TMHMM, SignalP, Phobius, and PSORTb, the functions of hypothetical protein-coding genes can be predicted from primary amino acid sequences. Two clusters of properties that aid in determining and predicting the hypothetical genes involve sequence similarity and protein localization. The bioinformatic programs can identify properties such as protein families, conserved domains, signal peptides, and transmembrane regions that belong to the respective clusters. This research project aims to predict the functions of five unannotated hypothetical protein-coding genes in the genome of the bacterium *Coxiella burnetii*. The genes, BMW92_RS10760, BMW92_RS10830, BMW92_RS10835, BMW92_RS10840, BMW92_RS10855, were annotated, analyzed, and had predicted functions of coding proteins. The specific proteins that were predicted to code from the selected genes were uroporphyrinogen-III synthase, pyrroline-5-carboxylate reductase, pyridoxal phosphate-dependent enzyme, phosphoenolpyruvate carboxykinase, and aspartate carbamoyltransferase, respectively. The predicted functions of the hypothetical protein-coding genes provide insight into the proteome of *C. burnetii*. Ultimately, the proposed gene annotations must be validated through molecular cloning and biochemical methods to determine if these proteins expressed by *C. burnetii* are accurate in expression and predicted functions by bioinformatic standards.

Introduction

Whether simple or complex, every living organism is made of a basic unit of life—the cell. With every cell that makes up a living organism, there resides a genetic code that exists in genes contained in chromosomes. The genetic code is a blueprint fundamental to life itself. It contains the genetic information necessary for heredity, development, and phenotype. This genetic code is encoded and organized as Deoxyribonucleic Acid (DNA) in every cell nucleus (Klug et al. 2015). However, it was not until 1944 when Oswald Avery, Colin MacLeod, and Maclyn McCarty, researchers at the Rockefeller Institute in New York, published experiments showing that DNA was the carrier of genetic information in bacteria (Tortora et al. 2018).

This discovery later formed the framework of the structure of DNA. DNA is a long, ladder-like macromolecule that twists to form a double helix. Each linear strand of the helix is made up of subunits called nucleotides (Klug et al. 2015). The nucleotide subunit comprises three components: deoxyribose sugar, phosphate group, and a nitrogenous base. In DNA, there are four different nucleotides, each with a unique nitrogenous base, abbreviated A (adenine), G (guanine), T (thymine), and C (cytosine). These four distinct nitrogenous bases are arranged in various sequence combinations held together by weak chemical bonds, called hydrogen bonds, between each base pair along a phosphate deoxyribose backbone (Klug et al. 2015). The two strands of the double helix of DNA are exact complements of one another; thus, adenine (A) always bonds with its complementary base pair thymine (T), and guanine (G) bonds with its complementary base pair cytosine (C) (Figure 1).

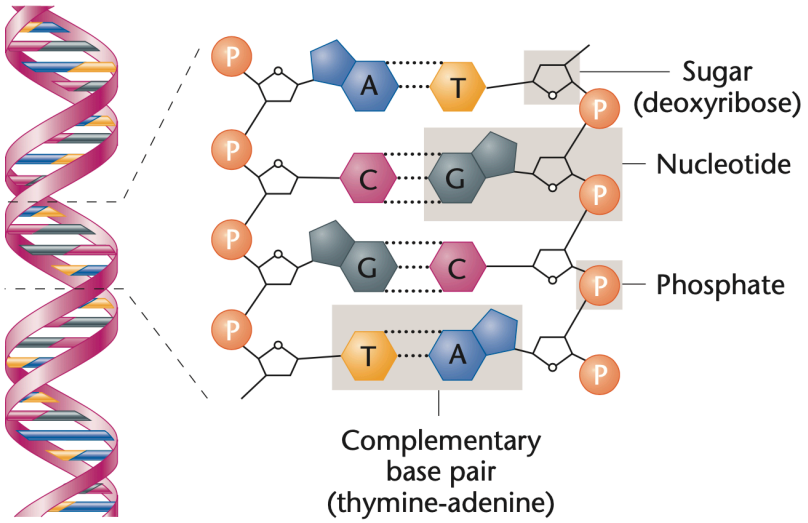


Figure 1: Summary of the structure of DNA, illustrating the arrangement of the double helix (on the left) and the chemical components making up each strand (on the right). The dotted lines on the right represent the hydrogen bonds holding together the two strands of the DNA helix along a phosphate deoxyribose backbone (Klug et al. 2015).

Another type of nucleic acid, ribose nucleic acid (RNA), is chemically similar to DNA but differs in three distinct ways. RNA contains ribose rather than deoxyribose, contains the nitrogenous base uracil (U) in place of thymine (T), and is generally a single-stranded molecule (Klug et al. 2015). As a result of generally being single-stranded, RNA does not have hydrogen bonds between the nucleotides but still contains a phosphate pentose sugar backbone that holds together the single strand (Klug et al. 2015). RNA has an arrangement of various sequence combinations of nucleotides that contain genetic information that can potentially turn into a functional gene product. The functional gene products, most often proteins, have the potential for enormous diversity, which are critical components of all cells and organisms. These genetic codes, organized and encoded in DNA, start gene expression, and serve as the core of all living

organisms. The genetic information carried by the nucleic acids can be transcribed into RNA and then translated into functional gene products, stated by the central dogma of biology (Klug et al. 2015).

Transcription of DNA to RNA

The central dogma of biology begins with the conversion of DNA into RNA. This conversion occurs through the process of transcription, which is a biological process of replication. To protect the genetic code, the cell organizes and contains original copies of the DNA within the cell's nucleus. This serves as biological protection of the genetic code from damage, degradation, or mutations that could highly occur in the cell's cytoplasm (Klug et al. 2015). Without an original genetic code, no template strand would be available to be transcribed into RNA, and consequently, no functional gene product translated. Thus, the nucleus of a cell houses all DNA and provides a library of many genetic codes that can be copied and turned into functional gene products.

There is a sequence of enzymes that must work together for transcription to take place inside the nucleus. The first sequence of events is the initiation steps. Prokaryotic transcription begins when the initial binding is established when the RNA polymerase σ subunit recognizes specific DNA sequences known as promoters. The promoters are located upstream of the initial transcription of a gene. Enzymes move along the DNA until it encounters a promoter region and binds. The promoter regions consist of known consensus sequences, which are homologous in different genes of the same organism or in multiple genes of closely related organisms (Klug et al. 2015). Two consensus sequences found in many bacterial promoters are "TATAAT" and "TTGACA" which are found upstream of the transcriptional start site (Klug et al. 2015). Once the RNA polymerase σ subunit has bound to the promoter, the DNA double helix is converted

into an open structure accessible for the enzyme. Initiation of RNA synthesis begins when 5'-ribonucleoside trisphosphates, which complement the template strand nucleotides, are inserted. Subsequent ribonucleotide complements are inserted and linked together by phosphodiester bonds. This process continues in a 5' to 3' direction whereby DNA and RNA chains run antiparallel to one another.

Upon insertion of the initial eight base-pair ribonucleotides, the RNA polymerase σ subunit dissociates from the holoenzyme attached to the open DNA structure, which commences the elongation step of transcription. The core polymerase enzyme traverses the gene's entirety as the RNA chain grows in size with each newly added base until a termination sequence is read. Termination is dependent on the termination sequence, which causes the newly synthesized RNA transcript to fold in a secondary hairpin structure that stops termination; however, in some cases, termination is dependent on a specific termination factor—rho (ρ) —which physically blocks the growing RNA transcript from elongating (Klug et al. 2015). The RNA transcript, which is precisely complementary to the DNA sequence, is released from the enzyme, and termination of transcription is achieved when the core polymerase enzyme dissociates. The newly synthesized RNA molecule will have matching A, T, G, C nitrogenous bases, except every T base is replaced with U (Klug et al. 2015). This transcriptional process has generated an RNA transcript further capable of turning into a gene product and allows the RNA transcript to begin the next stage—translation (Figure 2).

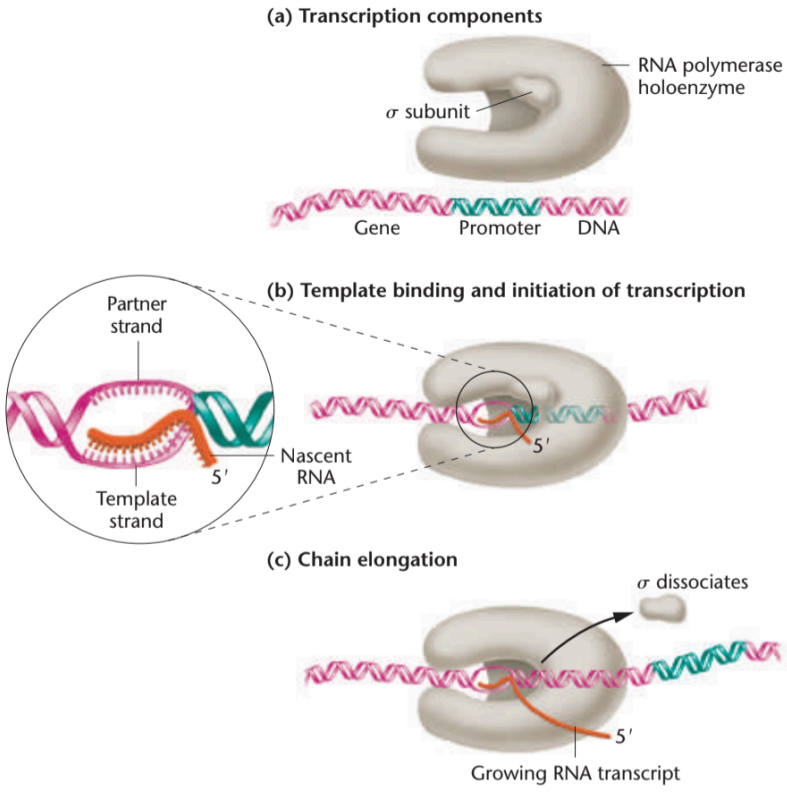


Figure 2: The process of transcription. Step (a) outlines the components assembled for transcription; step (b) outlines the template binding and initiation step; step (c) outlines dissociation of the σ subunit and RNA chain elongation step as the core polymerase enzyme transverse the open DNA structure. The final product of transcription being an RNA transcript capable of entering into the next process of translation (Klug et al. 2015).

Translation of RNA to Protein

Translation is the final step that completes the central dogma of biology. Translation is converting RNA products, generated earlier through transcription, into functional gene products—proteins (Klug et al. 2015). Translation ends with the production of protein; however, the translation process occurs differently across organisms. In prokaryotic cells, the synthesized

mRNA does not require transportation to the cytoplasm to begin (Klug et al. 2015). Translation for prokaryotic cells can occur immediately after the mRNA is generated from transcription in the cytoplasm due to the lack of membrane-bound organelles.

The process of translation follows three designated steps. The first step of initiation commences when the organelles necessary for translation are gathered and assembled (Klug et al. 2015). Ribosomes are composed of two subunits; a small (30S) and large (50S) subunit combined to form a functional and active ribosome (Tortora et al. 2018). The assembled ribosome generates three sites within itself—the aminoacyl (A), peptidyl (P), and the exit (E) sites. This first step requires the use of transfer RNA (tRNA), a small RNA molecule needed to decode and carry amino acids to the ribosomes, to be recruited. The tRNA cannot act alone in this step as this molecule requires a ribosomal binding site (RBS), which signals for the ribosome to bind and facilitate translation (Tortora et al. 2018). This specific ribosomal binding site, the Shine-Dalgarno sequence AGGAGGU, is found only in prokaryotes and originates upstream of the translation initiation codon AUG (Tortora et al. 2018). The initiation step ends with the complete assembly of the ribosomal unit and the recruitment of specialized proteins and RNA molecules needed for successful binding in the subsequent translational steps.

The second step of translation is elongation. During this step, the mRNA will move through the ribosome to be decoded and translated into a polypeptide chain. The ribosome reads the mRNA molecule in base pairs of three, known as codons (Klug et al. 2015). The mRNA molecule binds at the ribosome's A site with a start codon—AUG—present on the mRNA. After being read at the A site, the mRNA shifts to the P site where a tRNA molecule deciphers the codon and carries a corresponding amino acid associated explicitly to that codon to be linked to the amino acid present on the tRNA in the A site by a peptide bond (Klug et al. 2015). After the

amino acid has been attached by the tRNA, the mRNA moves to the E site where both the uncharged tRNA and mRNA are released from the large ribosomal subunit (Klug et al. 2015). Elongation is repeated as the mRNA-tRNA-amino acid chain complex is translocated by a distance of three nucleotides in the direction of the P site. The growing polypeptide chain is elongated with each additional amino acid added as the mRNA is translocated three nucleotides through the ribosome.

This process repeatedly continues with a growing amino acid chain until one of three stop codons is recognized, which begins the termination step. Once the stop codon is recognized, it signals for the action of a GTP-dependent release factor. This action stimulates the hydrolysis of the polypeptide from the peptidyl tRNA, leading to its release from the ribosomal complex and the ribosomal complex's dissociation into two subunits (Figure 3) (Klug et al. 2015).

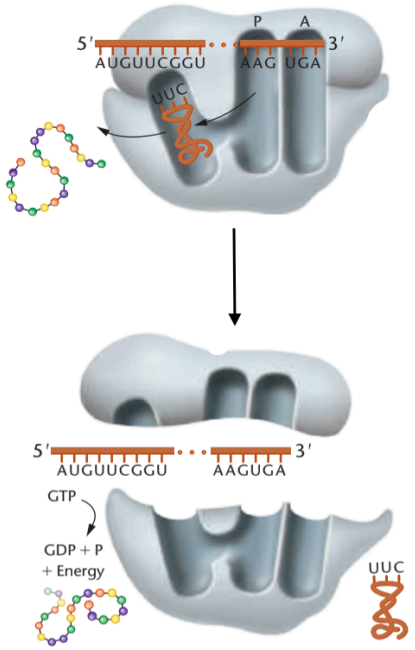


Figure 3: Termination of the process of translation (Klug et al. 2015).

The mRNA transcript is released, charged tRNA is released, and the end of translation results in a polypeptide that folds into native 3-D conformation of protein. This final process of translation of the genetic code completes the central dogma of biology (Figure 4).

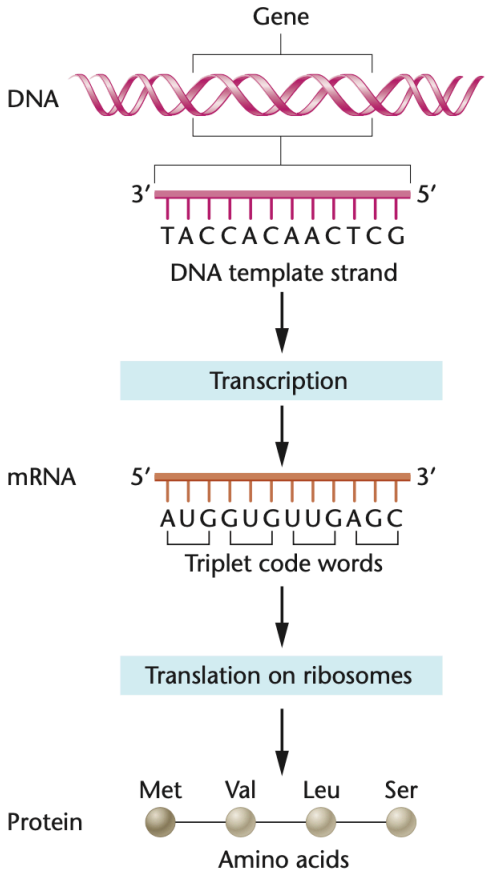


Figure 4: Central dogma of Biology (Klug et al. 2015).

After completing translation, the remaining product is a chain of amino acids—polymer—held together by peptide bonds (Lehninger et al. 2013). The uniqueness of every amino acid sequence is rooted in the genetic code. Despite the genetic code being made up of only four nucleotides—A, G, T, C—there is great diversity in each amino acid sequence attributed to the availability of 20 different amino acids. This diversity can be attributed to

numerous combinations of codons and the genetic code's degeneracy (Lehninger et al. 2013). With 64 possible codon combinations, the genetic code's degeneracy is depicted with a number of those codon combinations coding for the same amino acid. For example, the codons AGC and AGU have a different nucleotide in the third position yet still code for the same amino acid. Every amino acid sequence differs by the amino acids that make up the polypeptide chain. This is dependent on the arrangement of codons and ultimately determined by the sequence of DNA. The redundancy of the genetic code and the numerous arrangements of nucleotides give rise to diverse amino acid sequences and distinct structural protein products essential for living.

Protein Structure

Proteins, composed of varied lengths of amino acids, are essential for many biological processes. The diversity of proteins can be depicted by the different combinations of the possible amino acids. Every amino acid contains a carboxyl (-COOH), hydrogen atom (-H), amine (-NH_2), and variable (R) functional group, specific to each amino acid, bonded to a central carbon atom (Figure 5).

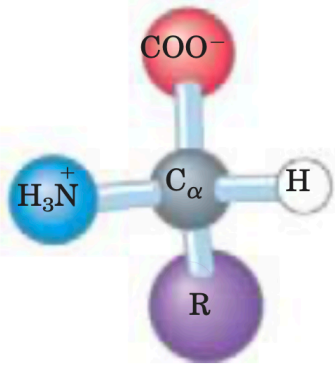


Figure 5: General structure of an amino acid (Lehninger et al. 2013).

There are 20 different amino acids available to make a chain; the combinations can be found with 20^n, where n is equal to the number of amino acids in a chain (Lehninger et al. 2013). The combination of proteins exponentially increases with every addition of an amino acid.

The functionality of proteins can be attributed to the different levels and orders of folding and structure. The versatility of proteins occurs as a result of the three-dimensional arrangement of the amino acid monomers, which are the building blocks of proteins. Proteins differ in the number and the type of amino acids within their chain. The qualities and characteristics—acidity, basicity, hydrophobicity, and hydrophilicity—of amino acids are primarily determined from the (R) functional group and interact highly with other linked amino acids (Lehninger et al. 2013). As a result of these interactions, proteins can be organized into a conceptual hierarchy described by levels of complexity. The four levels of protein structure are commonly defined as primary, secondary, tertiary, and quaternary structures.

The primary structure of a protein is the overall determinate of a protein's conformation and function (Lehninger et al. 2013). The primary structure is the linear amino acid sequence produced after translation. The secondary structure of a protein is synthesized through the interactions between amino acid side chains, which give rise to common patterns that turn, coil, and fold into alpha-helices and B-pleated sheets (Lehninger et al. 2013). The tertiary structure of a protein is the furthering of previous folding, turning, and coiling that involves the R-groups and backbones of the amino acids. This gives rise to distinct three-dimensional globular shapes held together by intermolecular forces such as, ionic interactions, hydrogen bonding, hydrophobic forces, ionic bonds, disulfide bonds, and metallic bonds (Lehninger et al. 2013). The fourth level of protein structure is the quaternary structure. This structure develops when one or more folded proteins interact to form a complex. The components of that complex can involve multiple

identical copies or different polypeptide chains. The various levels of structures of proteins lead to the versatility and functionality within cells (Figure 6).

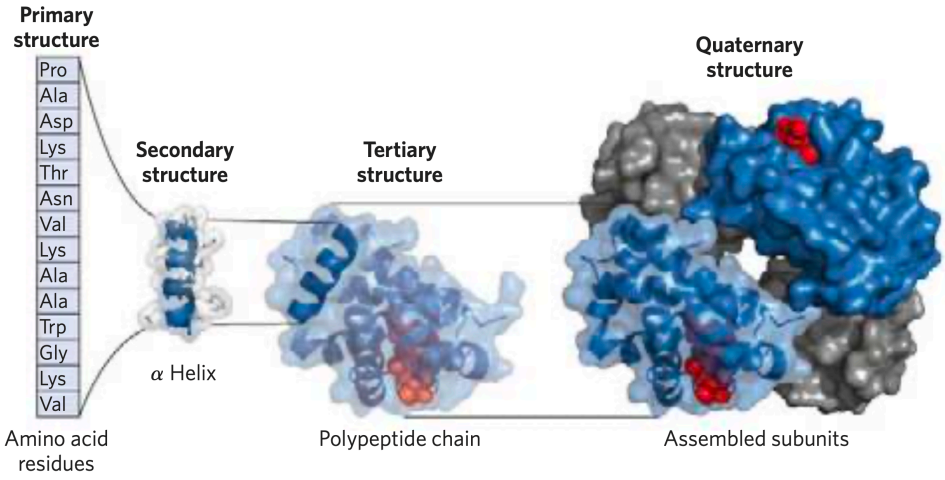


Figure 6: Levels of structure in proteins. (Lehninger et al. 2013).

Gene Regulation

At any point during the production of a functional gene product, there can be numerous regulations regarding that specific gene. Although gene regulation differs in each organism, the commonality is that gene expression is regulated to some degree in every organism. This regulation of gene expression allows certain protein products, whether involved in essential functions of the cell or metabolic pathways, to be produced at a higher or lower frequency dependent on the needs of the cell.

Regulation of gene expression has been studied in cellular activities such as replication, recombination, cell division, and DNA repair (Klug et al. 2015). One of the main benefits of gene regulation is that it allows prokaryotes to adapt to the environment. Specific regulation

allows prokaryotes to produce certain enzymes only when specific chemical substrates are present; these enzymes are known as inducible enzymes. In contrast, some enzymes are produced continuously, regardless of the environment's chemical makeup, which are known as constitutive enzymes (Klug et al. 2015). Furthermore, two systems govern gene regulation. The first system is a repressible system whereby the presence of a specific molecule inhibits gene expression. The second system is an inducible system whereby the presence of a specific molecule stimulates gene expression (Klug et al. 2015). These two systems are governed by either negative or positive control or a combination of both controls. Genetic expression occurs unless it is shut off by some form of a regulator molecule under negative control. In contrast, under positive control, transcription occurs only if a regulator molecule directly stimulates RNA production (Klug et al. 2015).

Notable gene regulators in prokaryotes occur during transcription. For instance, cis-acting elements and trans-acting elements, molecules that bind next to the gene itself or across from the gene, respectively, can promote or repress the binding of RNA polymerases to promoters (Klug et al. 2015). Promoters further display the variability of gene expression that is tightly regulated in prokaryotic cells; strong promoters and weak promoters can regulate the initiation time of transcription from one to two seconds to as little as once every 10 to 20 minutes (Klug et al. 2015). Genes regulation even occurs at the termination process for transcription. When a unique sequence of nucleotides is encountered, the secondary hairpin structure causes the newly formed transcript to fold back on itself; thus, terminating transcription, which can prevent further processing of the transcript and gene expression (Klug et al. 2015).

When examining translation, gene regulation has a major role for successful production gene product. During elongation, the assembled prokaryotic ribosome has two subunits set in

place to regulate gene expression. The small subunit decodes the codons in the mRNA, while the role of the large subunit is peptide-bond synthesis (Klug et al. 2015). This process regulates proper gene expression by minimizing the observable error rate to about 10^{-4}, as a result, incorrect amino acids insertion will occur only once in every 20 polypeptides of an average length of 500 amino acids (Klug et al. 2015).

Gene regulation has been studied in cellular activities such as replication, recombination, cell division, and DNA repair (Klug et al. 2015). As seen in some bacteria, there are clusters of genes that encode proteins with related functions that are regulated and expressed as a unit. These regions of clustered genes are known as operons. The operon region of a bacterial gene comprises the promoter region, operator region, and adjacent structural genes that are transcribed (Figure 7). This region's importance involves the binding of transcription factors at the promoter or operator regions, which can repress or promote transcription of these genes. The operon region invokes a series of molecular interactions between proteins, inducers, and repressors that aids in regulating genes (Klug et al. 2015).

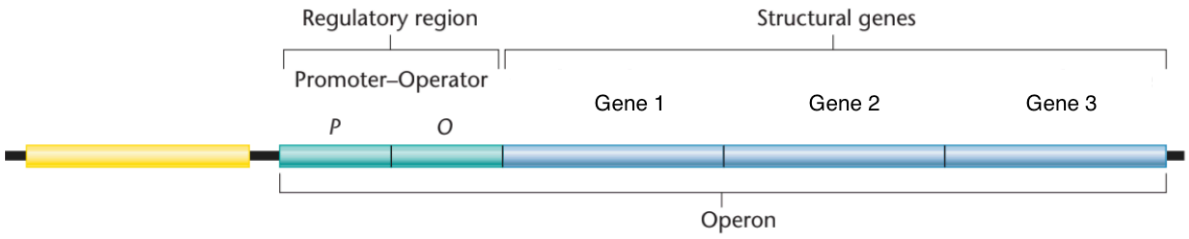


Figure 7: The operon model. Components and general structure of an operon consists of the regulatory region, structural genes, and operon region (Klug et al. 2015).

The frequency of gene expression varies greatly due to the importance and demand of each gene's function. For instance, a number of operons lack gene regulation due to how necessary those genes products are to the cell. Regulation of gene expression is often linked to the metabolic needs of the cell in which the gene products are necessary for the survival of the cell. The frequency of gene expression can be modified by gene regulation of the cell (Klug et al. 2015). As a result, the evolutionary selection of specific genes occurs to promote the frequency of favored gene products that can increase the likelihood of survival of the organism. Thus, certain genes are highly conserved over many generations due to their importance and necessity.

Evolutionary Selection

Conservation of genes over many generations occurs through evolutionary selection and adaptation. Many genes in various organisms look almost identical. The reason for that can be attributed to microbial evolution which keeps genes that are necessary for survival, homeostasis, and functioning while discarding genes that are not as critically needed (Brock et al. 2017). These prioritized genes are highly conserved, or saved, over periods of time to be passed on from one generation to the next. This microbial evolution process is not able to happen spontaneously. It is a long process that is driven by mutations and other heritable changes that accumulate in the genetic code (Tortora et al. 2018). As DNA is replicated, new copies of the gene are created and passed down to each succeeding generation. The offspring of an organism typically receives an exact copy of the gene. Occasionally, there are mistakes, referred to as mutations, that alter the DNA sequence of the gene (Brock et al. 2017). If the mutations that are accumulated are not repaired, then it creates the opportunity for nonidentical genes to be passed to the progeny.

The evolutionary process begins here whereby mutations of the DNA sequences are repeatedly passed from one generation to the next. Since the gene is altered and is not identical, the organisms are considered two different organisms as a result of the mutations accumulated (Tortora et al. 2018). Although mutations can alter the DNA sequence, not all mutations are the same. Neutral mutations have no effect on the function of the gene or the organism carrying it; whereas, beneficial mutations improve the function of the gene or benefits the organism that carries it. In some instances, mutations are detrimental and worsens or removes the function of the gene which can greatly impede the function, fitness, and survival of the organism carrying it (Brock et al. 2017). Regardless of the mutation type, the new mutant organism will reproduce and begin developing a lineage distinguished from the parent with the original gene copy.

Over time, many original and mutant copies of the organism will develop. Evolutionary trees are developed to outline and display the lineage of genes passing from one generation to the next whereby lineages can acquire their own additional mutations that are distinct from the original gene (Brock et al. 2017). The more mutations that are acquired, the farther the two lineages will appear on an evolutionary tree (Brock et al. 2017). Despite the selected gene of interest looking almost identical to the parental gene, each mutant lineage and gene carries a unique mutational signature.

The longer the gene has had the opportunity to accrue change, the more distinguished the resulting genes will be from the ancestral genes. As each mutation is passed on, a new lineage is produced which can be depicted with family and evolutionary trees. As a result, related families of genes share much in common, yet each have distinguishable mutations. As time progresses, changes to the genetic code can eventually develop genes that are related but have entirely different functions from the previous ancestral genes (Brock et al. 2017). Mutations to the genes

that are beneficial, improve fitness, and increases survival of the organism remain conserved over many generations. Many organisms do not survive or pass on their genes which effectively removes those mutations and gene lineages from the population, clearly displaying the process of evolutionary selection (Tortora et al. 2018).

Evolutionary selection can be examined not only with selected genes or the genetic code, but it can also be seen when examining proteins. Proteins function by folding into particular shapes, structures, and conformations with specific chemical properties that allow them to bind to and interact with the things they work on (Lehninger et al. 2013). These shapes are often found in localized regions of the folded protein called domains. These domains play an important role as they are responsible for particular functions or sub-functions of the protein. Domains are modular and have the ability to combine with other domains to produce an entirely unique function from a stand-alone domain (Lehninger et al. 2013). Over time, proteins with similar function will be related and have similar domains to those in the same lineage (Brock et al. 2017). The combination of modular domains allows for new functions to be assimilated, thereby developing some domains to become critical for functioning and survival. These domains remain conserved and used to build other proteins with high priority; thus, displaying the adaptational process (Brock et al. 2017).

Organisms obtain these variant genes and domains that can potentially increase their fitness and likelihood of survival. Just as gene regulation can increase the frequency of a desired gene, evolutionary selection aids organisms by selecting genes that will benefit the organism most (Brock et al. 2017). Particular genes and domains that are beneficial are often highly conserved throughout generations to maintain this adaptational and evolutionary advantage.

Bioinformatics

Bioinformatics is a specialized field of information technology defined as the storage, retrieval, and analysis of biochemical and biological data generated from genomics and proteomics, usually in the form of DNA and amino acid sequences (Klug et al. 2015). This expanding field of bioinformatics began with an international effort to sequence the human genome. This project, known as the Human Genome Project, began in 1990 and was completed by 2003 (Klug et al. 2015). The completion of the Human Genome Project, bioinformatics software, and hardware for processing nucleic acid sequences, protein sequences, and gene-interaction networks has prompted many scientists to derive nucleotide sequences of organism genomes.

Emerging technology has allowed for the genetic code that instructs life in all organisms to be read and understood. Many of the databases and bioinformatics programs focus on manipulating and analyzing DNA and RNA sequences. There is a wide range of bioinformatics tools and databases at the publics' disposal via the Internet. Through the use of proteomics, the study of all proteins expressed in a cell or tissue, information regarding protein-coding genes and non-coding genes can be stored in databases and give insight into the genetic makeup of living organisms (Klug et al. 2015). The use of bioinformatics opens the possibility of understanding how gene regulation and evolutionary selection of genes occurs. For annotated sequences, this data collected can then be stored and gives way to understanding what possible fates are available for specific genetic codes and the normal or variant genes that can arise from those genetic codes (Klug et al. 2015). This information can direct further investigation of unknown genomes.

As research continues, many model organisms and viruses continue to have their genomes sequenced. As of 2013, over 4300 whole genomes have been sequenced (Klug et al. 2015). These studies have demonstrated many similarities between genomes of nearly all species. Each genetic relationship expands the understanding of the evolutionary selection of similar gene sets used by organisms for essential cellular functions, such as DNA replication, transcription, and translation.

Comparative genomics is a subfield that has developed out of the many bioinformatic studies. This field compares the genomes of different organisms to answer questions about genetics and other aspects of biology through gene discovery and the development of model organisms (Klug et al. 2015). These practical applications' importance is the research that can be conducted to study human diseases, genome evolution, and relationships between organisms and environments. Comparative genomics uses a wide range of techniques and resources, revealing genetic differences and similarities between organisms (Klug et al. 2015). These similarities and differences provide data that can be stored while providing insight into how these differences contribute to differences in life cycle, evolutionary selection, gene expression, and survival (Klug et al. 2015).

Early genome projects have focused on prokaryotes since most prokaryotes have small genomes. Many of the prokaryotic genomes that have already been sequence can be traced to causing many human diseases (Klug et al. 2015). Bioinformatics has opened the availability and flood of genomic information regarding prokaryotic genomes. For example, many bacteria have been identified to have a single, circular chromosome; however, there is substantial variation in chromosome organization and number among bacterial species (Klug et al. 2015). Identification of bacterial genomes composed of linear DNA molecules, two circular chromosomes, or more

than two chromosomes has significantly increased as the software and hardware of bioinformatic program tools advances. Despite all the unknown answers, bioinformatics serves as a tool to provide insight regarding prokaryotic genomes, which can detrimentally interact with eukaryotic genomes, as seen with many human diseases such as cholera, tuberculosis, and leprosy (Klug et al. 2015).

There are many unknowns regarding genomes sequences of organisms. Genomics, proteomics, and bioinformatics have already proven to be valuable for identifying members of multigene families. These multigene families are groups of genes that share similar but not identical DNA sequences through duplication and descent from a single ancestral gene (Klug et al. 2015). These fields of study have provided a solid foundation for future research and studies regarding the organization of protein-coding genes in bacteria. Two generalizations, established by previous genome projects, are that gene density is extremely high in bacterial genomes and bacterial genomes contain operons (Klug et al. 2015). This means that many prokaryotic genomes have tightly packed genes; thus, a high proportion of the DNA serves as coding DNA.

Despite the popularity and advancement in biotechnology, there is still a lack of research conducted to study, annotate, and analyze unknown genomic and proteomic data. This can be attributed to the vastness of the genetic codes in every organism and the limited number of trained scientists able to analyze genetic codes successfully. However, each genome classified, sequenced, and analyzed can be stored in genomic libraries and retrieved through databases. Through the use of these database programs, it allows for comparative analysis between organisms' genetic codes, genomes, and proteomes. As the genomic libraries expand, organisms with genomes waiting to be sequenced can be compared and assessed to previously sequenced organisms stored in database programs.

A variety of database programs were used in this research which were organized through Genomics Educational National Initiative – Annotation Collaboration Toolkit, GENI-ACT. Each program investigated a distinct aspect regarding the genetic sequence selected. Aspects that were investigated during the research included protein localization, sequence similarity, and phylogeny. Protein localization involved the use of programs SignalP, LipoP, TMHMM, BOMP, PSORTb, and Phobius (GENIACT n.d.). Sequence similarity involved the use of BLASTp, CDD, MUSCLE, T-COFFEE, and WebLogo (GENI-ACT n.d.).

The expanding field of bioinformatics has provided an understanding of the molecular basis for hundreds of genetic disorders, therapeutic products, and many living organisms' genetic codes. The advancement of technology has accelerated the rate at which a genome can be sequenced and analyzed. Thus, organisms of different species can easily be studied for data regarding harmful genetic sequences and therapeutic products that can be manipulated from these genetic sequences. A microorganism of interest that is being researched for the treatment of similar ricket zoonosis diseases and is the causative pathogen of Query (Q) fever is the bacteria Coxiella burnetii.

Coxiella burnetii

Coxiella burnetii, an obligate gram-negative bacterium, is the etiologic agent of Query (Q) fever (Gürtler et al. 2014). This highly infectious bacterium is a zoonosis recognized by the World Health Organization and the Centers for Disease Control and Prevention as a dangerous pathogenic agent and potential biological weapon. This bacterium has displayed an ability to infect numerous species ranging from arthropods to mammals (Coleman et al. 2004). However, the main reservoirs of *C. burnetii* has been traced to sheep, goats, cattle, and livestock which are typically infected by ticks and the feces of ticks—the sources of infection (Angelakis and Raoult

2010). The primary mode of transmission for *C. burnetti* involves contact or inhalation of aerosolized infectious particles and droplets or consumption of raw milk products from infected organisms (Coleman et al. 2004).

Knowledge of the pathogen has increased significantly due to the distribution worldwide of Q fever. Currently, there are more than 30 genotypes of *C. burnetii* that have been distinguished by genome analysis (Eldin et al. 2016). *C. burnetii* is part of the family of *Coxiellaceae* bacteria and is phylogenetically related to *Legionellae* spp, *Francisella tularensis*, *Rickettsiella* spp, and many Gammaproteobacteria (Eldin et al. 2016). Development of a phylogenetic tree for *C. burnetii* has displayed the diversity of the 30 genotypes which can be regarded as mutations within co-circulating *Coxiella* strains.

C. burnetii possesses a small circular chromosome of approximately 5 million base pairs. Despite a relatively small number of base pairs, sixteen incomplete genomes have yet to be analyzed or studied (Minnick and Raghavan 2011). Dependent on the host system, *C. burnetii* can replicate differently in the cells of the host species because of synthesis of the bacterium's lipopolysaccharide. This lipopolysaccharide is the main component of this bacterium's cell wall, which aids in protection from immune systems and is essential for the bacterium's virulence (Minnick and Raghavan 2011). However, the chemical composition and molecular heterogeneity of the lipopolysaccharides in *C. burnetii* is distinct from the lipopolysaccharides seen in other gram-negative bacteria. Chemical and immunological characterizations of lipopolysaccharides concluded that the lipopolysaccharides from *C. burnetii* have a structural variation similar to the smooth-to-rough mutational variation of saccharide chain length seen in specific gram-negative bacteria species (Amano and Williams 1984).

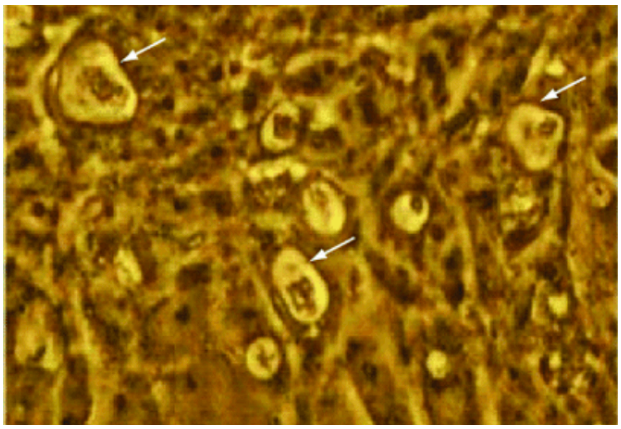


Figure 8: Parasitophorous vacuole of *Coxiella burnetii*. Phase-contrast micrograph showing a synchronously infected Vero cell monolayer at four days post infection of *Coxiella burnetii* (Minnick and Raghavan 2011).

Two important characteristics of *C. burnetii* are the environmental stability and virulence of the bacteria. Firstly, the environmental stability can be attributed to the genome which codes for proteins used in the adaptation of stressful environments. Unlike other bacteria within the *Rickettsia* group, *C. burnetii* is highly resistant to adverse physical and chemical agents which allows it to not be limited in host or geographical distribution (Woldehiwet 2004). Secondly, the virulence of *C. burnetii* corresponds to the biphasic development and spore-like particles. This bacterium will often trigger phagolysosomes vacuoles in the host, but *C. burnetii* has been found to code buffering proteins for acidic environments (Minnick and Raghavan 2011). Some evidence has even suggested the intracellular survival of *C. burnetii* can be attributed to the impairments and undermining abilities of the bacteria on macrophage functions and T-cell responses (Akporiaye et al. 1983). It appears the acidic pH activates *C. burnetii* metabolism and initiates replication of the organism. Furthermore, there has been identification of sodium-proton

exchangers and transporters for osmoprotectants coded by the genome of this bacterium to help relive osmotic pressure and oxidative stress (Coleman et al. 2004).

The biphasic development cycle occurs through the alternation between phase one and phase two (Figure 9).

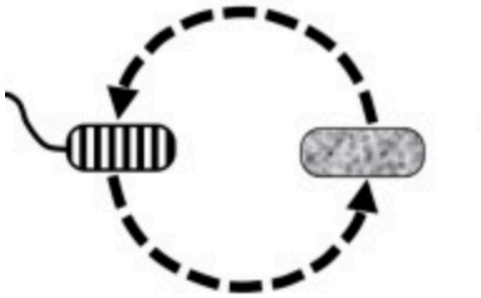


Figure 9: Biphasic developmental cycle. Biphasic cycle can be considered a developmental network in which two forms simply alternate into each other (Robertson et al. 2014).

This biphasic development has revealed the presence of an extreme pleomorphism of *C. burnetii*. During the developmental cycle, two morphological cell types, large cell variant (LCV) and small cell variant (SCV), mature with distinct internal structures. The SCV has an outer membrane that is wider and more prominent than that of the LCV (McCaul and Williams 1981). Complex internal membranous intrusions originate from the cytoplasmic membrane in the SCV while the LCV harbors no extensive membranous system. Furthermore, LCVs contained dense bodies in the periplasmic space which resembles an endospore (McCaul and Williams 1981). Overall, the morphogenesis of *C. burnetii* is comparable to cellular differentiation of endospore formation, which consist of a vegetative and sporogenic differentiation period (McCaul and Williams 1981).

The spore-like particles of *C. burnetii* is morphologically similar to endospores of other bacteria species but has distinct properties. One property being the ability to remain infectious for up to 40 months under very unfavorable external conditions (Woldehiwet 2004). The organism's ability to grow and multiply within phagolysosomes and its propensity to establish persistent infection displays many cellular immunity troubles. Morphogenesis and development of *C. burnetii* is typically reliant on specific environmental conditions that can drive the developmental process (Minnick and Raghavan 2011). With the ability of *C. burnetii* to survive in low pH (~5) of phagocytic vacuoles, the response to develop has been attributed to the depletion of critical metabolites such as amino acids which is known to regulate proper prokaryotic development (Woldehiwet 2004).

Phase one is extremely virulent and contagious, which correlates to *C. burnetii* having a SCV morphological cell type as the genome appears to have homologues of sporulation genes and physical properties for environmental transmission (Cole et al. 2004). In contrast, phase two displays low virulence but can persist in non-immunocompetent host system which correlates with the LCV morphological type (Coleman et al. 2004).

The obligate intracellular nature of *C. burnetii* imposes considerable experimental limitations that can impede the progress in understanding the organism's morphogenesis, genome, and proteome. Numerous researchers have encountered the inability to propagate obligate intracellular pathogens like *C. burnetii* under axenic (host cell-free) culture conditions (Omsland et al. 2009). However, recent studies and advances has led to the understanding of *C. burnetii* replicating exclusively in acidified, lysosome-like vacuoles (Omsland et al. 2009). Successful studies involved axenic cultivation of *C. burnetii* which furthered the understanding

of the organism's pathogenesis, genetics, and development of Q fever preventatives that can be developed from other obligate intracellular pathogens (Omsland et al. 2009).

The cellular and molecular biology of *C. burnetii* remains largely undefined. With a vast majority of the genome unexplored and unsequenced, it leaves room for understanding this bacteria's gene expression and protein interaction. As more of the genome is annotated, sequenced, and analyzed, more information regarding this bacterium can be used for potential therapeutic products or help understand other similar intracellular obligate prokaryotes that can be harmful to other organisms. This bioinformatic research involved the annotation and data collection of five protein-coding genes of *Coxiella burnetii*. The five protein-coding genes had data collected, analyzed, and annotated regarding the gene's sequence similarity, protein localization, structure, conserved domains, and phylogenetic relationships.

Materials and Methods

Basic Information

Basic microbial genomic information of the genes of interest were gathered and recorded. This information was obtained through GENI-ACT databases. The information gathered and recorded included the gene of interest, genome, replicon, locus, old locus tag, products, DNA length, protein length, start and end position of gene, genomic coordinates, nucleotide sequence, amino acid sequence, and isoelectric point of amino acid sequence.

Sequence Similarity

BLAST

The National Center for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (BLAST) server was opened. The Web BLAST section was expanded and the Protein BLAST, “BLASTp”, was selected. The amino acid sequence was converted to FASTA format. FASTA format included the character “>” entered without any spacing prior to the gene of interest followed directly by the entire amino acid sequence underneath on the next line. The amino acid sequence was entered in the “Query Sequence” box in FASTA format and the database “Non-redundant protein sequences (nr)” was selected. The “SwissProt” option was deselected to prevent the limited searches and results regarding genomes manually annotated. The BLAST algorithm parameters were unchanged and set for normal search parameters. The BLAST was then processed, and the results page loaded sequences ranked in similarities of highest similarity to lowest similarity. The result output provided the description of the similar proteins as well as their expected value (e-value). The e-value produced depicted the statistical

significance between the search results outputted. Any e-values greater than e^{-5} were deselected and excluded from the search results. The top ten sequences with significant alignments that were not identical species to the target gene and bacterium were selected. Information of the ten BLAST sequences selected were recorded. The information recorded included the organism name, protein name, percent identity, precent positive, length of alignment match, e-values, and percent gap. A graphic summary was generated from the ten sequences selected which displayed distribution of the top BLAST hits of the ten sequences in relation to the query sequence. Alignment scores were generated in conjunction of the graphic summary detailing the degree of alignment amongst the ten sequences selected.

CDD

The Conserved Domain Database (CDD) of the NCBI was used to identify potential protein domains through the comparison of multiple sequence alignments among domains and full-length proteins. The CDD was automatically ran in parallel with the NCBI BLAST search. The CDD search generated a list of domain hits, which match to protein families such as TigrFam, Cluster of Orthologous Genes (COG), and Pfam. The matches were organized from most conserved to least conserved genes across different organisms and received a corresponding e-value similar to that seen in the BLASTp search. Only domain hits with an e-value less than e^{-5} were selected. Information regarding the list of domain hits, name of matches, accession codes, e-values, conserved domain lengths, query sequences conserved, and descriptions of the domain hits were recorded.

MUSCLE

The MUltiple Sequence Comparison by Log Expectation (MUSCLE) software program was used to generate multiple sequence alignments.

From the BLASTp output, ten sequences were selected by clicking “select all” to uncheck all sequences, and then ten sequences from organisms different than the target organism, *C. burnetii*, with e-values less than e^{-5} were chosen. The selected genes were converted to FASTA format by clicking “download” and then “FASTA (complete sequence)” and used to generate a multiple sequence alignment. The FASTA alignment sequences that were downloaded were edited. The gene that was being investigated was put into the beginning line of FASTA with the corresponding amino acid sequence directly below. The FASTA sequences were edited further by placing the name of the organism after the “>” character. The title was composed of an abbreviated version of the organisms genus and species with no spacing between the bracket character and the organism name such as, “>C.burnetii”. After the FASTA sequences were edited, the program MUSCLE from EMBL-EBI was loaded. The FASTA format of all ten input sequences were entered into the MUSCLE input box and the output format was selected as “ClustalW”, then the task was submitted. The generated multiple sequence alignments were recorded. Additionally, the “Phylogenetic Tree” tab was selected which displayed the neighboring-joining tree without distance corrections. The phylogenetic tree had two options to display branch length. The “Cladogram” and “Real” branch lengths were selected separately and recorded separately.

T-COFFEE

The Tree-based Consistency Objective Function For alignment Evaluation (T-COFFEE) program from EMBL-EBI was loaded and used in conjunction with the MUSCLE software program. The edited FASTA sequences used in the MUSCLE were entered into the T-COFFEE input box. The sequences were submitted, and the generated multiple sequences alignments were recorded. The “Phylogenetic Tree” tab was selected and generated the phylogenetic tree of the

sequences entered from the FASTA sequence. The phylogenetic tree had two options to display branch length. The “Cladogram” and “Real” branch lengths were selected separately and recorded separately.

WEBLOGO

WebLogo was a program used to design and create sequence logos from multiple sequence alignments. The sequence logos provided graphical representation of how often each amino acid or nucleotide was found at a particular position. The WebLogo Version 2.8.2 program, created by the Computational Genomics Research Group from the University of California Berkeley, was loaded. The CLUSTAL multiple sequence alignment generated by MUSCLE and T-COFEE were copied and uploaded into the input box. The parameters and options of the sequence logo were altered under the “Image Format & Size” and “Advanced Logo Options” tabs. The image format was changed from “PNG (bitmap)” to “PDF (vector).” In the “Advanced Logo Options,” the “Sequence Type” was selected for amino acid and the “Multiline Logo (symbols per Line)” was selected for 32. After the parameters were set, the “Create Logo” was selected to produce the sequence logo. The results were recorded and analyzed for sequence similarity of conserved amino acid residues across different organisms.

Protein Localization

SIGNALP

The SignalP 5.0 Server developed by the Center for Biological Sequence Analysis (CBS) was loaded and used to predict the presence of signal peptides and the location of cleavage sites in proteins from Archaea, Gram-positive Bacteria, Gram-negative Bacteria, and Eukarya. The amino acid sequence for the desired gene of interest was inserted into the submission box in FASTA format. The parameters of the search in the software program were changed. The

"Organism group" that was selected was "Gram-negative," which was indicative of Gram-negative Bacteria. The "Output format" was selected as "Long output." After the parameters were selected, the job was submitted. The generated output plot graph and the summary prediction of protein types and cleavage sites were recorded.

LIPOP

The LipoP 1.0 server developed by Denmark Technical University (DTU) Bioinformatics was loaded and used to produce predictions of lipoproteins and discriminates between lipoprotein signal peptides, other signal peptides, and n-terminal membrane helices in Gram-negative bacteria. The amino acid sequence for the desired gene of interest was inserted into the submission box in FASTA format. The "Output format" was then selected for "Extensive, with graphics." The job was submitted after the parameters were set. The result output and the indicated cleavage sites scores, SpI, SpII, TMH, CYT, CleavI, and CleavII values were recorded.

TMHMM

The TMHMM Server v. 2.0 developed by DTU Bioinformatics was loaded and used to predict the presence of transmembrane helices in proteins. The server was loaded and the amino acid sequence for the desired gene of interest was inserted into the submission box in FASTA format. The "Output format" was then selected for "Extensive, with graphics." The parameters were set, and the job was submitted in the server. The outputted graphical image, length of amino acid, number of predicted transmembrane helices, and protein location were all recorded.

BOMP

The beta-barrel outer membrane predicter (BOMP) program from the Computational Biology Unit (CBU) was loaded and used to predict the presence or absence of beta-barrel integral outer membrane proteins. The amino acid sequence for the desired gene of interest was

inserted into the submission box in FASTA format. The parameter regarding the "E-Value" of the search was adjusted and selected as "1^{-5}." After the parameter of the job was set, the job was submitted in the server. The generated output BOMP result was recorded.

PSORTB

The PSORTb v. 3.0.2 software program developed by Simon Fraser University was loaded and used for the prediction of bacterial protein subcellular localization. The search parameters of the program were selected for the specific organism being researched. The "organism type" was selected as "Bacteria." The "Gram stain" was selected as "Negative." The "Output format" was selected as "Normal" and the "Show results" was selected as "Via the web." The amino acid sequence for the desired gene of interest was inserted into the submission box in FASTA format and the job was submitted. The generated output resulted in an outcome table of three parts. Those three parts were the analysis report, localization scores, and final prediction location scores which were all recorded.

PHOBIUS

The Phobius software program developed by the Stockholm Bioinformatics Centre (SBC) was loaded and used as a combined transmembrane topology and single peptide predictor. The amino acid sequence for the desired gene of interest was inserted into the submission box in FASTA format and the job was submitted. The generated graphical result and the prediction results of signal peptide, cytoplasmic region, non-cytoplasmic region, and transmembrane domains located along the amino acid sequence were all recorded.

Results

The hypothetical protein-coding genes of *Coxiella burnetii* BMW92_RS10760, BMW92_RS10830, BMW92_RS10835, BMW92_RS10840, BMW92_RS10855 were individually retrieved and analyzed with the use of bioinformatics software programs, databases, servers, and tools. As stated in the materials and methods section, the information that was retrieved and analyzed was recorded in the following sequence: basic information, sequence similarity, and protein localization.

BMW92_RS10760

The first gene, BMW92_RS10760, was analyzed using bioinformatic technology. Table 1 below contains the provided data regarding basic information. A protein isoelectric point calculator was used to determine the isoelectric point of the protein, protein length, and the number and prevalence of each amino acid that makes up the protein (Figure 10). The BLASTp search tool produced 100 matches ranked from highest sequence similarity to lowest sequence similarity. The top ten sequences with significant alignments that were not identical species to *Coxiella burnetii* were selected. The information recorded included the organism name, protein name, percent identity, precent positive, length of alignment match, e-values, and percent gap. The highest ranked match to the BMW92_RS10760 gene was uroporphyrinogen-III synthase [*Methylomarinum vadi*] (Figure 11). The remaining nine matches to the BMW92_RS10760 gene all had a function as uroporphyrinogen-III synthase (Figure 12-20). The CDD identified four potential protein domains hits conserved (Figure 21). Each of the domain hits conserved and identified by the CDD belong to the HemD superfamily which spans from 8-267 amino acids in reference to the query sequence. The protein classification identified by the CDD was

uroporphyrinogen-III synthase. Each of the four domain hits were recorded of the sequence similarity to the query sequence (Figure 22). The MUSCLE program generated a multiple sequence alignment (MSA); each amino acid in the sequence was assigned a distinct color to distinguish the amino acids being compared (Figure 23). The MUSCLE program generated two phylogenetic trees using the multiple sequence alignment data (Figure 23 and 25). The T-COFEE program generated another multiple sequence alignment to further confirm sequence similarity depicted by MUSCLE MSA (Figure 27). The T-COFFEE program generated two phylogenetic trees using the multiple sequence alignment data (Figure 28 and 29). WebLogo constructed a sequence logo graphical illustration of the amino acid residues of the gene BMW92_RS10760; each of the letter's heights produced correspond to the conservation of the amino acid residue across similar sequences (Figure 26 & 30). Protein localization results included SignalP, LipoP, TMHMM, BOMP, PSORTb, and Phobius. The SignalP graphical illustration identified that there is no presence of a signal peptide (Figure 31). The LipoP resulted in the highest scoring class being the cytoplasmic protein class (Figure 32). The TMHMM test resulted in a graphical illustration, statistics, and a list of the predicted transmembrane helices and the predicted location of the intervening loop regions. The TMHMM test resulted and displayed that the whole sequence contains no transmembrane helices and that the majority of the amino acid residues have a high probability of being located outside of the membrane (Figure 32). The BOMP test result identified there are no integral beta-barrel outer membrane proteins (Figure 34). The PSORTb test resulted in an analysis report that identified no internal helices, motifs, or signal peptides; the localization scores calculated the predictable location of the protein to be unknown (Figure 35). The Phobius resulted in a graphical illustration that identified no transmembrane helices and classified the protein as non-cytoplasmic (Figure 36).

Basic Information

Table 1: Gene BMW92_RS10760 basic information

Genome	Replicon	Locus Tag	Old Locus Tag
Coxiella burnetii	NZ_CP018005	BMW92_RS10760	BMW92_10395
Genomic Coordinates	**Products**	**Length**	**Start and End Position**
1952532..1953338	uroporphyrinogen-III synthase	807 / 268	1952532 - 1953338
Molecular Weight	**Average Isoelectric Point**	**IPC Protein**	**Protein Length**
29687.39284 Da	9.36	8.491	271 amino acids

Nucleotide Sequence	Amino Acid Sequence
atggaaaatgaatccttaaaaaataaaaccatcatgatcactcg ccctgaatggcagggagaattattaaaaaaagccattgaacgt aggggcggcgccgttattttatttccaacccttattataaaaccaa ttaataaatgcaactactcgcccttcgcatccgcaagttttccacc ttcccgcgaaagcgggagtccggatgataaaatgactaaggct ggatttctgaattcaagcgacatcctcatttttctaagtgcaaatg ctgtaaaacattctcctattttaaattttaaagcggaacaaaaatta gttgctattggcacaggtaccgccgcggctttatttcaacgcgg actttctgttgatgccgttcccgaacatttcagtagtgaaggcctt ttagatttgcctttactccatcaggtgactggaaaaacaattgcta ttttttgcggcgaaaactctcgcccttatttagaaaacgaacttat ccatcgcggcgcaaatgtattctccattattacttaccgacgaga aagacctgtcgtcaataaaaaaacgattgacgcactcactcacc aaacgcttcatgcaattgtctccactagcgccgaaagcctccaa aatctctgcaccttattcgaatcacaccaacactggttacaccgt atcccgctcgttgtcattagcaaaagaatggaaaacttagcaaa atctcaagggttccacttggtgcttctagcagacaatcctggag aaaaggctattataaaggttttatccactaaatatccgagttaa	MENESLKNKTIMITRPEWQGELLKKAIER RGGAVILFPTLIIKPINKCNYSPFASASFPP SRESGSPDDKMTKAGFLNSSDILIFLSANA VKHSPILNFKAEQKLVAIGTGTAAALFQR GLSVDAVPEHFSSEGLLDLPLLHQVTGKT IAIFCGENSRPYLENELIHRGANVFSIITYR RERPVVNKKTIDALTHQTLHAIVSTSAES LQNLCTLFESHQHWLHRIPLVVISKRMEN LAKSQGFHLVLLADNPGEKAIIKVLSTKY PS

Figure 10: Protein isoelectric point calculator. The number and prevalence of each amino acid in the protein coded from the BMW92_RS10760 gene of *Coxiella burnetii* (Kozlowski, Biology Direct, <http://isoelectric.org/>).

Sequence Similarity

BLAST

uroporphyrinogen-III synthase [Methylomarinum vadi]

Sequence ID: WP_031434600.1 **Length:** 252 **Number of Matches:** 1

Range 1: 2 to 251 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
145 bits(366)	1e-38	Compositional matrix adjust.	94/269(35%)	138/269(51%)	26/269(9%)

```
Query  3    NESLKNKTIMITRPEWQGELLKKAIERRGGAVILFPTLIIKPINKCNYSPFASASFPPSR  62
            NE L+ K I++TRP  Q   L + IE++GG  + FPTL I+ + +
Sbjct  2    NEQLQAKRILVTRPRHQAGNLCRLIEQQGGVAVRFPTLEIQALER---------------  46

Query  63   ESGSPDDKMTKAGFLNSSDILIFLSANAVK-----HSPILNFKAEQKLVAIGTGTAAALF  117
                P+    +   L   D LIF+SANAV      +S  +N     +L A+G  TA AL
Sbjct  47   ----PETIAARVAALEHVDWLIFISANAVNFVLNSNSGTINRLRRLRLAAVGKATAKALQ  102

Query  118  QRGLSVDAVPEH-FSSEGLLDLPLLHQVTGKTIAIFCGENSRPYLENELIHRGANVFSII  176
             GL+VD +P+H F SE LL  P +  V GK   I  G+  R  L + L  RGA+V  +
Sbjct  103  NNGLTVDLLPQHGFDSESLLRTPAMSAVDGKRCVIVRGQGGREILVDTLRERGADVEYLE  162

Query  177  TYRRERPVV-NKKTIDALTHQTLHAIVSTSAESLQNLCTLFESHQHWLHRIPLVVISKRM  235
             YRR  P   N   ++ L    L AI  TS E+L+NL  +       L  +PLVVIS+R+
Sbjct  163  VYRRVMPQADNSALLERLRENRLDAITITSGEALRNLMEMLGGQACLLLPVPLVVISRRI  222

Query  236  ENLAKSQGFHLVLLADNPGEKAIIKVLST  264
             +A++ GF  ++++D P + +I++ L T
Sbjct  223  GQMAETMGFKRIVVSDGPADTSILQTLIT  251
```

Figure 11: BLAST first match for BMW92_RS10760 sequence from organism *Methylomarinum vadi* with an e-value of 1e-38, 35% identity, 51% positives, 9% gaps (dissimilarity), and an identity of uroporphyrinogen-III synthase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

uroporphyrinogen-III synthase [Methylovulum psychrotolerans]

Sequence ID: WP_103973100.1 **Length:** 256 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 3 to 250 GenPept Graphics Next Match Previous

Score	Expect	Method	Identities	Positives	Gaps
142 bits(359)	2e-37	Compositional matrix adjust.	100/269(37%)	134/269(49%)	31/269(11%)

```
Query  5    SLKNKTIMITRPEWQGELLKKAIERRGGAVILFPTLIIKPINKCNYSPFASASFPPSRES  64
            L    +++TRP  Q E+L + I  +GG  I FPTL I+
Sbjct  3    GLGGAGVLVTRPAHQAEVLCRLIAEQGGTAIRFPTLAIE--------------------  41

Query  65   GSPDDKMTKAGFLNSSDI--LIFLSANAVKHSPILNFKAEQKLVA-----IGTGTAAALF  117
            + D    +    N  +   LIF+SANAV  +   N     KL+A     IG  TA AL
Sbjct  42   ATADTAAVQTALANLGNFQWLIFISANAVNFALKANGGKIPKLIAPRLAAIGQSTAQALA  101

Query  118  QRGLSVDAVP-EHFSSEGLLDLPLLHQVTGKTIAIFCGENSRPYLENELIHRGANVFSII  176
              GL VD VP + F+SE LL  PLL QV G+ I I  GE  R  L  +L HRGA V  I
Sbjct  102  NAGLGVDLVPAQGFNSEALLAEPLLQQVGGQRILIVRGEGGREELAAQLRHRGAEVSYID  161

Query  177  TYRRERPVVNKKTIDA-LTHQTLHAIVSTSAESLQNLCTLFESHQH-WLHRIPLVVISKR  234
            Y+R  P  N   + A LT Q L AI  TS E+LQNL  +     H  L  IP++V+S R
Sbjct  162  VYKRVMPDNNASEVQALLTQQRLQAITITSGEALQNLLMMVAPAYHPLLTAIPVIVVSGR  221

Query  235  MENLAKSQGFHLVLLADNPGEKAIIKVLS  263
            +  +A + GF  V++A+ P + A+IK ++
Sbjct  222  LAQMANNLGFKHVVVAEQPADSAMIKAVT  250
```

Figure 12: BLAST second match for BMW92_RS10760 sequence from organism *Methylomarinum psychrotolerans* with an e-value of 2e-37, 37% identity, 49% positives, 11% gaps (dissimilarity), and an identity of uroporphyrinogen-III synthase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

uroporphyrinogen-III synthase [Methylomonas methanica]

Sequence ID: WP_013820683.1 **Length:** 260 **Number of Matches:** 1

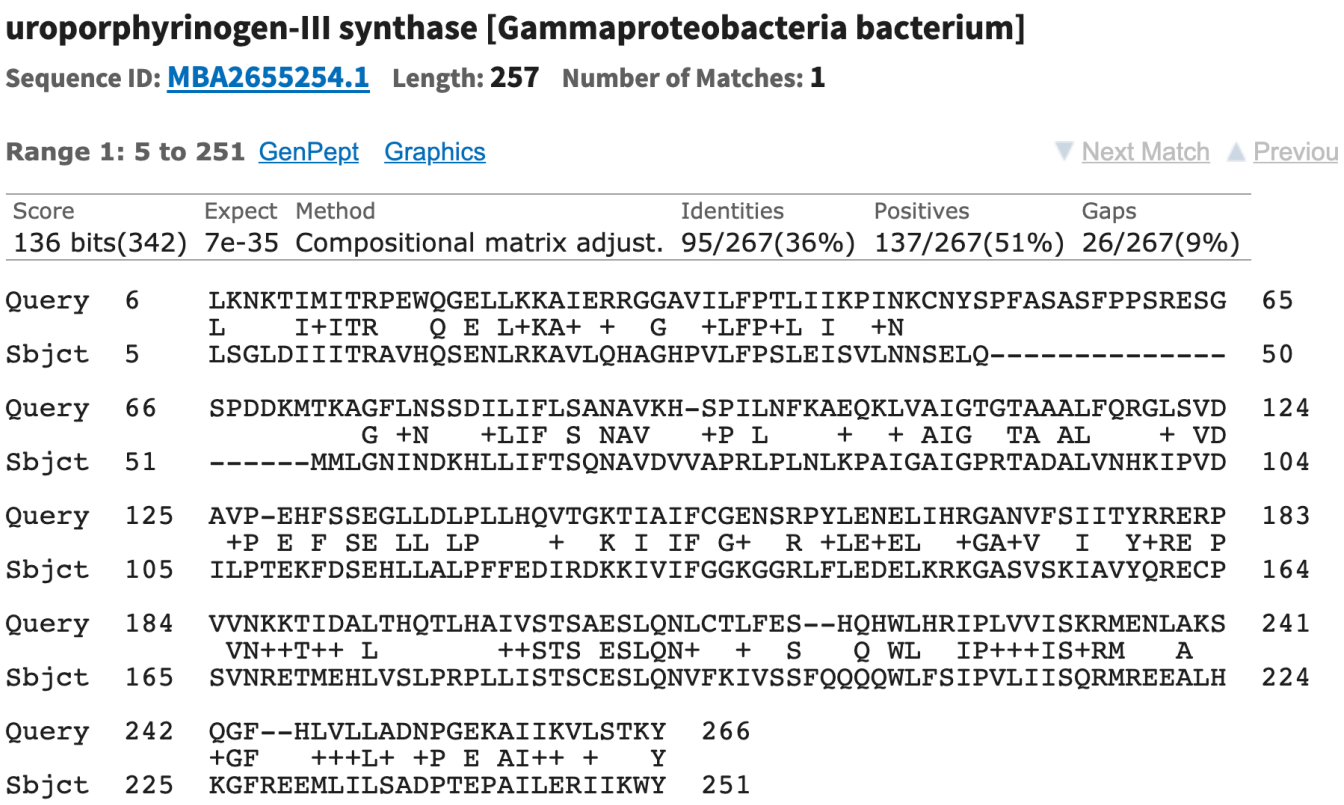
See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 4 to 249 GenPept Graphics ▼ Next Match ▲ Previous

Score	Expect	Method	Identities	Positives	Gaps
139 bits(351)	3e-36	Compositional matrix adjust.	96/269(36%)	144/269(53%)	32/269(11%)

```
Query  5    SLKNKTIMITRPEWQGELLKKAIERRGGAVILFPTLIIKPINKCNYSPFASASFPPSRES  64
            SL+  T+++TRP  Q + L + I +  G  + FPTL I+PI+
Sbjct  4    SLRGATVLVTRPAAQADTLCRLIAQADGRALRFPTLEIQPID------------------  45

Query  65   GSPDDKMTKAGFLNSSDILIFLSANAVKHSPILNFKAEQ------KLVAIGTGTAAALFQ  118
               D+ + +     + + LIF S+NAV  + +  F  +       KL A+G  TA+AL +
Sbjct  46   --VDNALIEKAL--TCNWLIFTSSNAVDFA-LKAFGGKMAGAMAVKLAAVGQATASALQK  100

Query  119  RGLSVDAVPE-HFSSEGLLDLPLLHQVTGKTIAIFCGENSRPYLENELIHRGANVFSIIT  177
             GL V  VP+  FSSEGLL  P + QV+G+ I I  G   R  LE+ L  RGA V  +
Sbjct  101  AGLQVTCVPKTEFSSEGLLAQPAMQQVSGQRIVIVRGMGGREKLEHTLRGRGAEVAYLEV  160

Query  178  YRRERPVVN-KKTIDALTHQTLHAIVSTSAESLQNLCTLFE-SHQHWLHRIPLVVISKRM  235
            YRR RP +   + I +L +Q L+AI  TS E+LQNL T+ + +  + L + PL+V+S R+
Sbjct  161  YRRCRPDIKCDELIQSLRNQQLNAITITSGEALQNLLTMLDPAAANLLRKQPLIVVSDRI  220

Query  236  ENLAKSQGFHLVLLADNPGEKAIIKVLST  264
               LA   GF  V ++  P + AI++ L+T
Sbjct  221  RQLALELGFDQVAVSPQPTDAAILETLTT  249
```

Figure 13: BLAST third match for BMW92_RS10760 sequence from organism *Methylomarinum methanica* with an e-value of 3e-36, 36% identity, 53% positives, 11% gaps (dissimilarity), and an identity of uroporphyrinogen-III synthase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).



uroporphyrinogen-III synthase [Gammaproteobacteria bacterium]

Sequence ID: MBA2655254.1 **Length:** 257 **Number of Matches:** 1

Range 1: 5 to 251 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
136 bits(342)	7e-35	Compositional matrix adjust.	95/267(36%)	137/267(51%)	26/267(9%)

```
Query  6    LKNKTIMITRPEWQGELLKKAIERRGGAVILFPTLIIKPINKCNYSPFASASFPPSRESG  65
            L    I+ITR   Q E L+KA+ +  G  +LFP+L I  +N
Sbjct  5    LSGLDIIITRAVHQSENLRKAVLQHAGHPVLFPSLEISVLNNSELQ--------------  50

Query  66   SPDDKMTKAGFLNSSDILIFLSANAVKH-SPILNFKAEQKLVAIGTGTAAALFQRGLSVD  124
                     G +N   +LIF S NAV   +P L    +  + AIG  TA AL    + VD
Sbjct  51   ------MMLGNINDKHLLIFTSQNAVDVVAPRLPLNLKPAIGAIGPRTADALVNHKIPVD  104

Query  125  AVP-EHFSSEGLLDLPLLHQVTGKTIAIFCGENSRPYLENELIHRGANVFSIITYRRERP  183
             +P E F SE LL LP    +  K I IF G+  R +LE+EL  +GA+V  I  Y+RE P
Sbjct  105  ILPTEKFDSEHLLALPFFEDIRDKKIVIFGGKGGRLFLEDELKRKGASVSKIAVYQRECP  164

Query  184  VVNKKTIDALTHQTLHAIVSTSAESLQNLCTLFES--HQHWLHRIPLVVISKRMENLAKS  241
             VN++T++ L       ++STS ESLQN+  +  S   Q WL  IP+++IS+RM   A
Sbjct  165  SVNRETMEHLVSLPRPLLISTSCESLQNVFKIVSSFQQQQWLFSIPVLIISQRMREEALH  224

Query  242  QGF--HLVLLADNPGEKAIIKVLSTKY  266
            +GF   +++L+ +P E AI++ +   Y
Sbjct  225  KGFREEMLILSADPTEPAILERIIKWY  251
```

Figure 14: BLAST fourth match for BMW92_RS10760 sequence from organism Gammaproteobacteria bacterium with an e-value of 7e-35, 36% identity, 51% positives, 9% gaps (dissimilarity), and an identity of uroporphyrinogen-III synthase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

uroporphyrinogen-III synthase [Methylobacter tundripaludum]

Sequence ID: WP_104427303.1 **Length:** 256 **Number of Matches:** 1

Range 1: 5 to 251 GenPept Graphics ▼ Next Match ▲ Previous

Score	Expect	Method	Identities	Positives	Gaps
133 bits(334)	9e-34	Compositional matrix adjust.	93/269(35%)	134/269(49%)	32/269(11%)

```
Query  6    LKNKTIMITRPEWQGELLKKAIERRGGAVILFPTLIIKPINKCNYSPFASASFPPSRESG  65
            L    +++TRPE Q E L + IE+RGG  + FPTL                     E
Sbjct  5    LNGACVLVTRPEHQAENLSRLIEQRGGVAVRFPTL---------------------EIV  42

Query  66   SPDDKMTKAGF--LNSSDILIFLSANAVK-----HSPILNFKAEQKLVAIGTGTAAALFQ  118
            S DD   K+    L+    ++F+SANAV      +S  +      +  A+G  TA A+
Sbjct  43   SRDDDRIKSTLENLDGFQWVVFISANAVNFALKANSGKIPRTKSVRFAAVGQATAQAMKM  102

Query  119  RGLSVDAVPEH-FSSEGLLDLPLLHQVTGKTIAIFCGENSRPYLENELIHRGANVFSIIT  177
             GL VD VPE+ ++SE LL++P L QV G+   I  GE  R  L   L  RGA V  +
Sbjct  103  AGLPVDLVPEYGYNSEALLEMPQLQQVEGQNCLIVRGEGGREQLATTLRSRGAEVDYLEV  162

Query  178  YRRERPVVNKK-TIDALTHQTLHAIVSTSAESLQNLCTLF-ESHQHWLHRIPLVVISKRM  235
            Y+R  P ++    ++ L    L  I  TSAE+LQNL  +  E +   L  IPLVV+S R+
Sbjct  163  YKRIIPRMDSSPVVELLAQHRLDVITVTSAEALQNLSLMLGEKNNKLLSLIPLVVVSDRI  222

Query  236  ENLAKSQGFHLVLLADNPGEKAIIKVLST  264
              LA   GF+ + + D+P + AI++ + T
Sbjct  223  RCLAADMGFNRITVTDSPIDTAILETVIT  251
```

Figure 15: BLAST fifth match for BMW92_RS10760 sequence from organism *Methylobacter tundripaludum* with an e-value of 9e-34, 35% identity, 49% positives, 11% gaps (dissimilarity), and an identity of uroporphyrinogen-III synthase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

uroporphyrinogen-III synthase [Methylobacter luteus]

Sequence ID: WP_027159695.1 **Length:** 257 **Number of Matches:** 1

Range 1: 4 to 251 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
129 bits(323)	4e-32	Compositional matrix adjust.	86/262(33%)	130/262(49%)	17/262(6%)

```
Query  5    SLKNKTIMITRPEWQGELLKKAIERRGGAVILFPTLIIKPINKCNYSPFASASFPPSRES  64
            L    +++TRP  Q E L + I+ RGG V+ FP L I           A  +    +++
Sbjct  4    GLNGARVLVTRPAHQAENLSRLIQERGGEVVRFPVLDI----------VARDNIEEVQDA  53

Query  65   GSPDDKMTKAGFLNSSDILIFLSANAVKHSPILNFKAEQKLVAIGTGTAAALFQRGLSVD  124
                DK     F++ + +   L AN  K   +       +  A+G  TA AL   GL+VD
Sbjct  54   LKNLDKFQWVVFISPNAVNFALKANNGKIDRLKTV----RFAAVGRATAQALEAAGLTVD  109

Query  125  AVPEH-FSSEGLLDLPLLHQVTGKTIAIFCGENSRPYLENELIHRGANVFSIITYRRERP  183
             VPE  ++SE LL +P + QV G+   I  GE  R  L N L  RGA V  +  Y+R  P
Sbjct  110  VVPEQGYTSEALLAMPQMQQVKGQACLIVRGEGGREELANTLRSRGAVVQYLEVYKRTIP  169

Query  184  VVN-KKTIDALTHQTLHAIVSTSAESLQNLCTLF-ESHQHWLHRIPLVVISKRMENLAKS  241
             ++  + +  L  Q L  I  TS E+LQNL  +  E++   L  IP+VV+S R+  LA
Sbjct  170  SIDSSQVVQLLAQQRLDVITVTSGEALQNLLIMLGENNHQLLLPIPMVVVSDRIRQLAAG  229

Query  242  QGFHLVLLADNPGEKAIIKVLS  263
             GF  + + +NP + AI++ ++
Sbjct  230  MGFKRIAVTENPADTAILETVT  251
```

Figure 16: BLAST sixth match for BMW92_RS10760 sequence from organism *Methylobacter luteus* with an e-value of 4e-32, 33% identity, 49% positives, 6% gaps (dissimilarity), and an identity of uroporphyrinogen-III synthase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

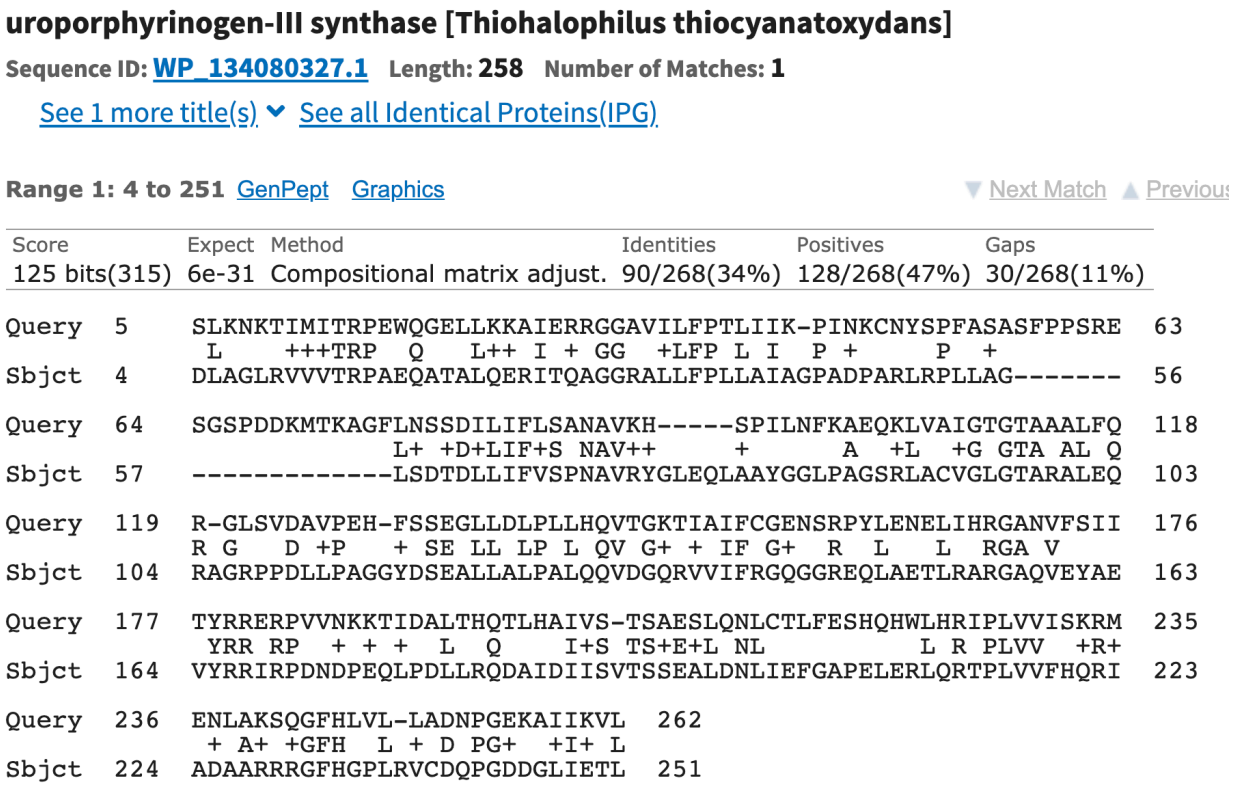


Figure 17: BLAST seventh match for BMW92_RS10760 sequence from organism *Thiohalophilus thiocyanatoxydans* with an e-value of 6e-31, 34% identity, 47% positives, 11% gaps (dissimilarity), and an identity of uroporphyrinogen-III synthase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

uroporphyrinogen-III synthase [Thiothrix nivea]

Sequence ID: WP_002708856.1 **Length:** 259 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 3 to 251 GenPept Graphics Next Match Previou

Score	Expect	Method	Identities	Positives	Gaps
124 bits(310)	3e-30	Compositional matrix adjust.	86/271(32%)	134/271(49%)	33/271(12%)

```
Query  4    ESLKNKTIMITRPEWQGELLKKAIERRGGAVILFPTLIIKPINKCNYSPFASASFPPSRE  63
            E+L+   +++TRP  Q    ++ +E+ G   +LFP ++I P  +
Sbjct  3    ETLRGLNVVVTRPAHQAARFQQMLEQAGANAVLFPVIVIAPPEQ---------------  46

Query  64   SGSPDDKMTKAGFLNSSDILIFLSANAVKHSPILNFKAEQK-------LVAIGTGTAAAL  116
               P    T    L+S D  IF+SANAV+   +      Q+       L A+G  TA  L
Sbjct  47   ---PALAQTMLASLDSYDAAIFISANAVRFG-LEQLDENQRQTLRKLTLGAVGKQTAGVL  102

Query  117  FQRGLSVDAVP-EHFSSEGLLDLPLLHQVTGKTIAIFCGENSRPYLENELIHRGANVFSI  175
             Q G  V  VP   ++SE  L LP + ++ GK I IF G   R +L + L  RGA+V  +
Sbjct  103  QQHGFGVQLVPASGYTSEDFLALPAVQRLVGKRILIFRGAGGREWLADALRSRGASVDYV  162

Query  176  ITYRRERPVVNKKTIDALTH--QTLHAIVSTSAESLQNLCTLFESHQHWLHRIPLVVISK  233
               YRR  P ++   +  L H  Q L  I  TS+E L NL  + + +  W+  +PL+  S+
Sbjct  163  EVYRRICPEIDTSGLK-LRHERQQLDIIAITSSEGLLNLLAMLD-NPDWIKTVPLLAGSQ  220

Query  234  RMENLAKSQGFH-LVLLADNPGEKAIIKVLS  263
            RM   A+  GF   + +ADNPG++A+++ L+
Sbjct  221  RMVEAARQAGFSGTIAIADNPGDEAMLQALT  251
```

Figure 18: BLAST eighth match for BMW92_RS10760 sequence from organism *Thiothrix nivea* with an e-value of 3e-30, 32% identity, 49% positives, 12% gaps (dissimilarity), and an identity of uroporphyrinogen-III synthase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

uroporphyrinogen-III synthase [Candidatus Methylobacter oryzae]

Sequence ID: WP_127027931.1 **Length:** 257 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 2 to 252 GenPept Graphics Next Match Previous

Score	Expect	Method	Identities	Positives	Gaps
124 bits(310)	3e-30	Compositional matrix adjust.	91/273(33%)	135/273(49%)	33/273(12%)

```
Query  3    NESLKNKTIMITRPEWQGELLKKAIERRGGAVILFPTLIIKPINKCNYSPFASASFPPSR  62
            N+ L    I++TRPE Q + L + IE +GG  + FPTL
Sbjct  2    NKLLSGVRILVTRPEHQADNLSRLIEEQGGIAVRFPTL----------------------  39

Query  63   ESGSPDDKMTKAGFLNSSDI---LIFLSANAVKHSPILN-----FKAEQKLVAIGTGTAA  114
            E  + D+ +     L + D+   LIF+SANAV  +   N          +  A+G  TA
Sbjct  40   EIIAKDNALEIKQMLANPDLFQWLIFISANAVNFALKANDGKIACTKSVRFAAVGQSTAQ  99

Query  115  ALFQRGLSVDAVPEH-FSSEGLLDLPLLHQVTGKTIAIFCGENSRPYLENELIHRGANVF  173
            A+   GL+VD VPE  ++SE LL +P L QV G+   I  GE  R  L   L  RGA V
Sbjct  100  AMRMAGLNVDLVPESGYNSEALLAMPELQQVEGQRFLIVRGEGGREQLATALRSRGAEVN  159

Query  174  SIITYRRERPVVNKK-TIDALTHQTLHAIVSTSAESLQNL-CTLFESHQHWLHRIPLVVI  231
             +  YRR  P ++    ++ L   +L  +  TSAE+LQNL   L E +   L  I LVV+
Sbjct  160  YLEVYRRVIPRIDSSPVVELLAQHSLDIVTVTSAEALQNLKLMLDEKNNKLLSLITLVVV  219

Query  232  SKRMENLAKSQGFHLVLLADNPGEKAIIKVLST  264
            S R+  +A   GF+ +++ ++P + AI++ + T
Sbjct  220  SNRIRCIAADMGFNRIIVTNSPIDTAILETVIT  252
```

Figure 19: BLAST ninth match for BMW92_RS10760 sequence from organism *Methylobacter oryzae* with an e-value of 3e-30, 33% identity, 49% positives, 12% gaps (dissimilarity), and an identity of uroporphyrinogen-III synthase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

Figure 20: BLAST tenth match for BMW92_RS10760 sequence from organism *Methylomonas lenta* with an e-value of 8e-30, 36% identity, 49% positives, 11% gaps (dissimilarity), and an identity of uroporphyrinogen-III synthase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

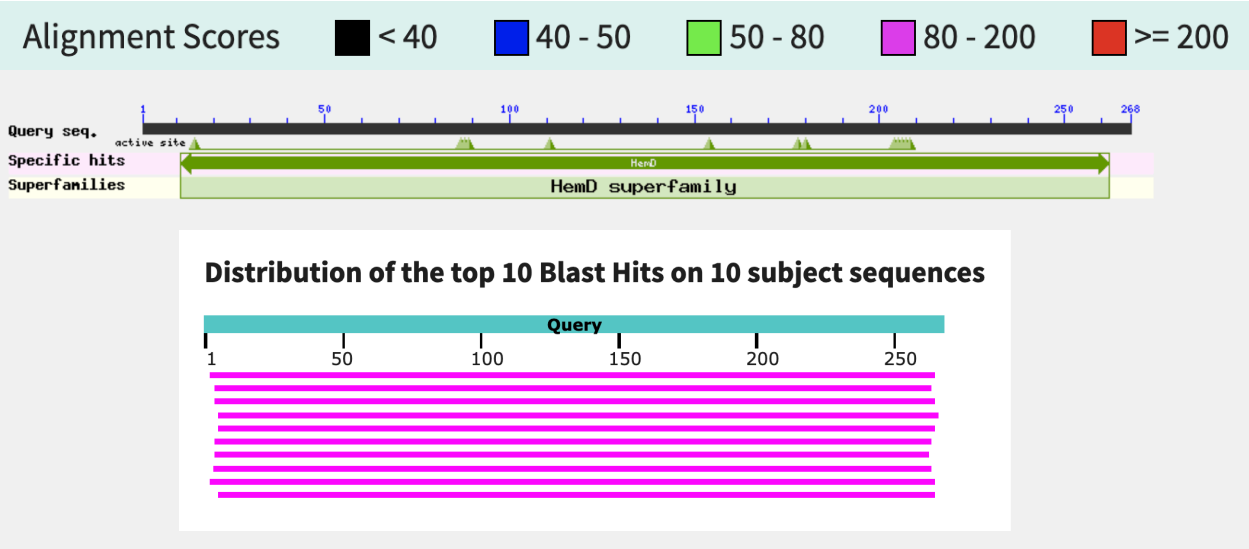


Figure 21: BLAST graphic summary with alignment scores of the top 10 organism sequences similarities selected aligned with *Coxiella burnetii* query sequence of gene BMW92_RS10760. Each of the alignment sequences selected are ordered from highest sequence similarity (top) to lowest sequence similarity (bottom). All organism sequences aligned with the query sequence have an alignment score of 80-200 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

CDD

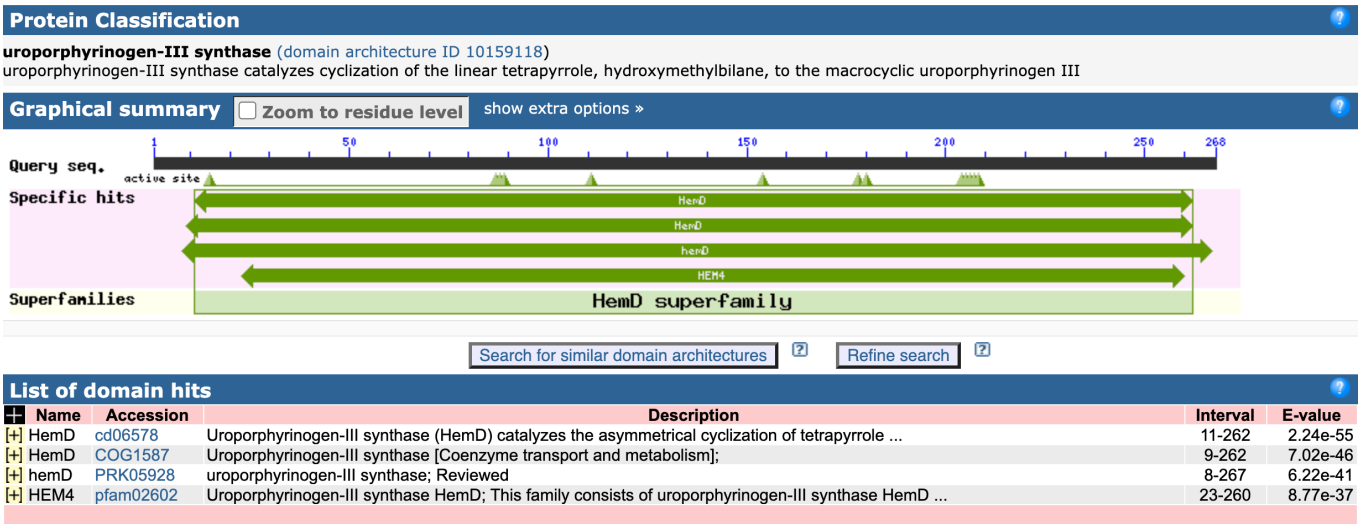


Figure 22: Conserved Domain Database output results for gene BMW92_RS10760. The top domain hit match was HemD: Uroporphyrinogen-III synthase which aligned with the query sequence from amino acid residues 11-262 and had statistically significant e-value of 2.24e-55. The second domain hit match was HemD: Uroporphyrinogen-III synthase which aligned with the query sequence from amino acid residues 9-262 and had a statistically significant e-value of 7.02e-46. The third domain hit match was hemD: uroporphyrinogen-III synthase which aligned with the query sequence from amino acid residues 8-267 and had a statistically significant e-value of 6.22e-41. The last domain hit match was HEM4: Uroporphyrinogen-III synthase which aligned with the query sequence from amino acid residues 23-260 and had a statistically significant e-value of 8.77e-37 ((BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

MUSCLE

```
G.bacterium              -MSALLSGLDIIITRAVHQSENLRKAVLQHAGHPVLFPSLEISVLNN------------
C.burnetii               MENESLKNKTIMITRPEWQGELLKKAIERRGGAVILFPTLIIKPINKCNYSPFASASFPP
M.methanica              -MTLSLRGATVLVTRPAAQADTLCRLIAQADGRALRFPTLEIQPIDV------------
M.lenta                  -MTTGLNGAWVLVTRPVAQAEKLCKLITQQNGQALQFPTLEIQPLKV------------
M.vadi                   -MNEQLQAKRILVTRPRHQAGNLCRLIEQQGGVAVRFPTLEIQALER------------
M.psychrotolerans        --MSGLGGAGVLVTRPAHQAEVLCRLIAEQGGTAIRFPTLAIEATAD------------
M.luteus                 -MIRGLNGARVLVTRPAHQAENLSRLIQERGGEVVRFPVLDIVARDN------------
M.tundripaludum          -MSKVLNGACVLVTRPEHQAENLSRLIEQRGGVAVRFPTLEIVSRDD------------
M.oryzae                 -MNKLLSGVRILVTRPEHQADNLSRLIEEQGGIAVRFPTLEIIAKDN------------
T.nivea                  -MPETLRGLNVVVTRPAHQAARFQQMLEQAGANAVLFPVIVIAPPEQ------------
T.thiocyanatoxydans      -MGCDLAGLRVVVTRPAEQATALQERITQAGGRALLFPLLAIAGPAD------------
                              *    :::**.  *.  :   :    .  : ** : *

G.bacterium              -------SELQMMLGNINDKHLLIFTSQNAVDVVAP------RLPLNLKPAIGAIGPRTA
C.burnetii               SRESGSPDDKMTKAGFLNSSDILIFLSANAVKHSPI-------LNFKAEQKLVAIGTGTA
M.methanica              ------DNALIEKALTCN---WLIFTSSNAVDFALK--AFGGKMAGAMAVKLAAVGQATA
M.lenta                  ------DGELIEKALHCD---WLIFTSTNAVDFALR--ALSGKMTRLHALKLAAVGKATA
M.vadi                   ------PETIAARVAALEHVDWLIFISANAVNFVLN--SNSGTINRLRRLRLAAVGKATA
M.psychrotolerans        ------TAAVQTALANLGNFQWLIFISANAVNFALK--ANGGKIPKLIAPRLAAIGQSTA
M.luteus                 ------IEEVQDALKNLDKFQWVVFISPNAVNFALK--ANNGKIDRLKTVRFAAVGRATA
M.tundripaludum          -------DRIKSTLENLDGFQWVVFISANAVNFALK--ANSGKIPRTKSVRFAAVGQATA
M.oryzae                 ------ALEIKQMLANPDLFQWLIFISANAVNFALK--ANDGKIACTKSVRFAAVGQSTA
T.nivea                  ------PALAQTMLASLDSYDAAIFISANAVRFGLE-QLDENQRQTLRKLTLGAVGKQTA
T.thiocyanatoxydans      ------PARLRPLLAGLSDTDLLIFVSPNAVRYGLEQLAAYGGLP--AGSRLACVGLGTA
                                         :* * ***                    : .:*  **

G.bacterium              DAL-VNHKIPVDILPTEKFDSEHLLALPFFEDIRDKKIVIFGGKGGRLFLEDELKRKGAS
C.burnetii               AAL-FQRGLSVDAVPEH-FSSEGLLDLPLLHQVTGKTIAIFCGENSRPYLENELIHRGAN
M.methanica              SAL-QKAGLQVTCVPKTEFSSEGLLAQPAMQQVSGQRIVIVRGMGGREKLEHTLRGRGAE
M.lenta                  NAL-QEVGLKVACVPETEFSSEGLLAESAMHRVSSQRVMIVRGLGGREKLAQTLHSRGAD
M.vadi                   KAL-QNNGLTVDLLPQHGFDSESLLRTPAMSAVDGKRCVIVRGQGGREILVDTLRERGAD
M.psychrotolerans        QAL-ANAGLGVDLVPAQGFNSEALLAEPLLQQVGGQRILIVRGEGGREELAAQLRHRGAE
M.luteus                 QAL-EAAGLTVDVVPEQGYTSEALLAMPQMQQVKGQACLIVRGEGGREELANTLRSRGAV
M.tundripaludum          QAM-KMAGLPVDLVPEYGYNSEALLEMPQLQQVEGQNCLIVRGEGGREQLATTLRSRGAE
M.oryzae                 QAM-RMAGLNVDLVPESGYNSEALLAMPELQQVEGQRFLIVRGEGGREQLATALRSRGAE
T.nivea                  GVL-QQHGFGVQLVPASGYTSEDFLALPAVQRLVGKRILIFRGAGGREWLADALRSRGAS
T.thiocyanatoxydans      RALEQRAGRPPDLLPAGGYDSEALLALPALQQVDGQRVVIFRGQGGREQLAETLRARGAQ
                          .:          :*   : ** :*  . .  : .:   *. * ..*  *   *  .**
```

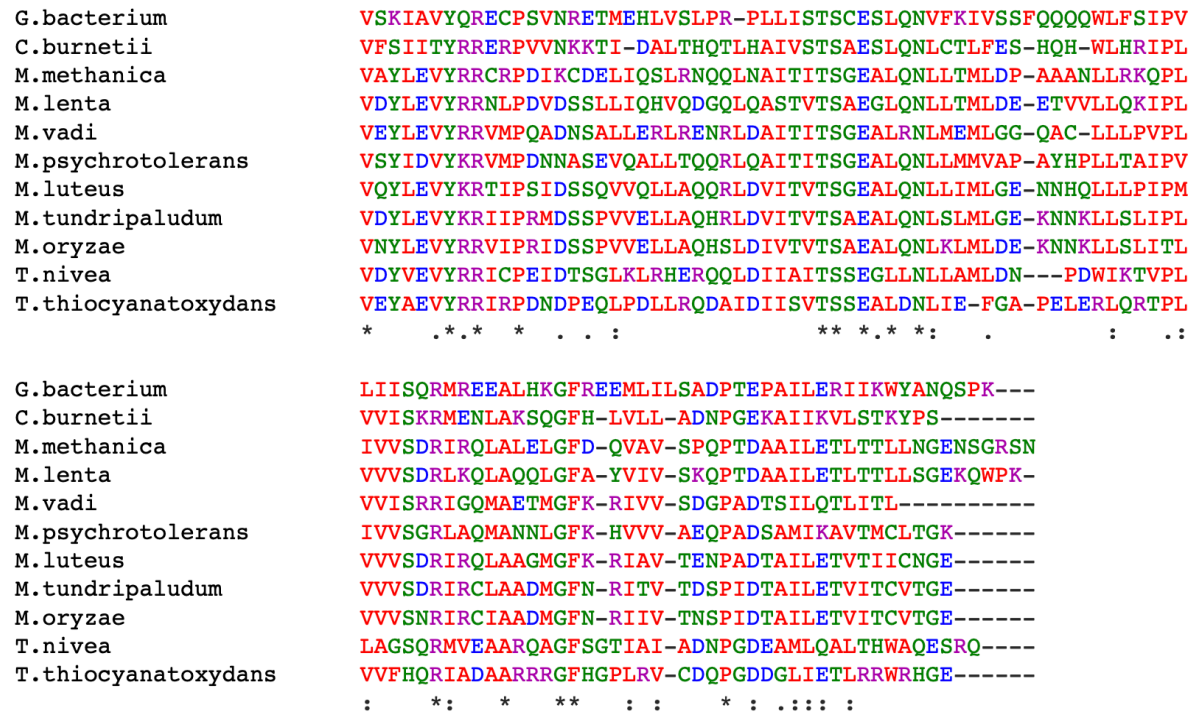


Figure 23: MUSCLE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10760 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered randomly and are listed from top to bottom as followed: *G. bacterium*, *C. burnetii*, *M. methanica*, *M. lenta*, *M. vadi*, *M. psychrotolerans*, *M. luteus*, *M. tundripaludum*, *M. oryzae*, *T. nivea*, *T. thiocyanatoxydans*. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

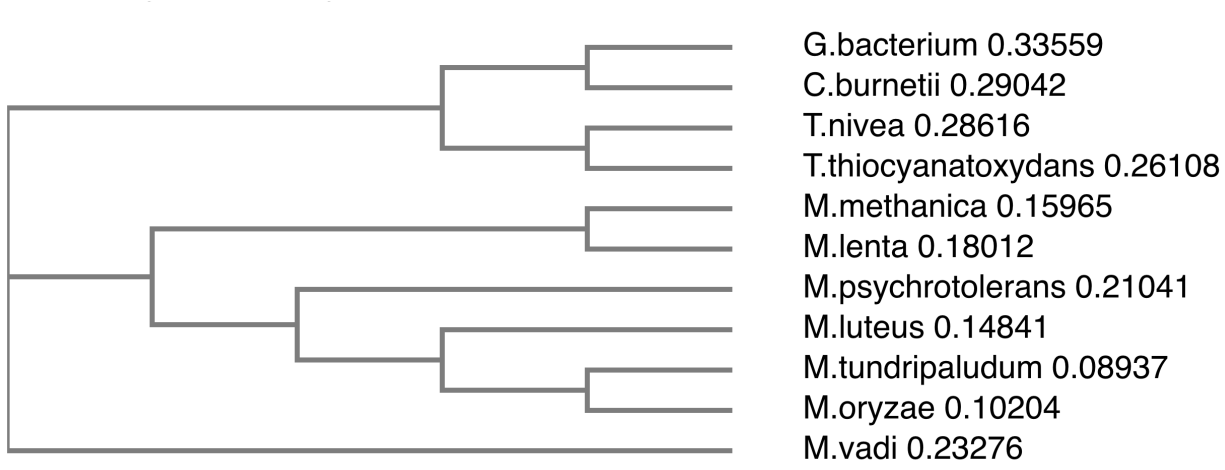


Figure 24: MUSCLE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10760 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered randomly and are listed from top to bottom as followed: *G. bacterium*, *C. burnetii*, *M. methanica*, *M. lenta*, *M. vadi*, *M. psychrotolerans*, *M. luteus*, *M. tundripaludum*, *M. oryzae*, *T. nivea*, *T. thiocyanatoxydans*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

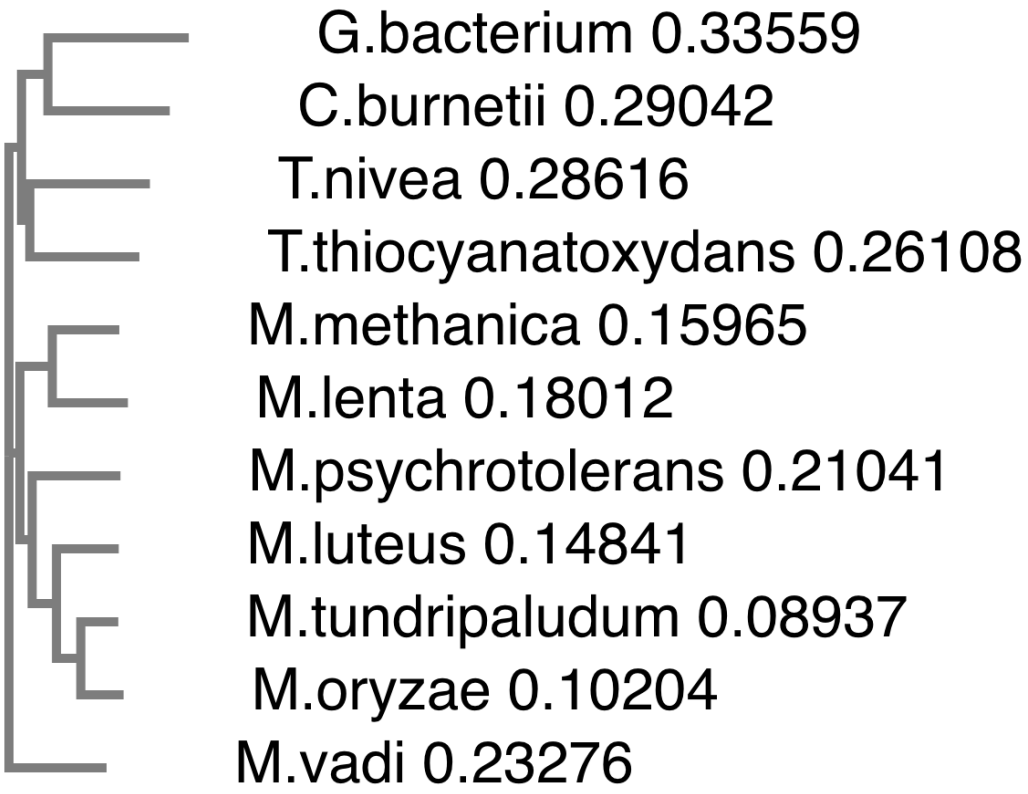
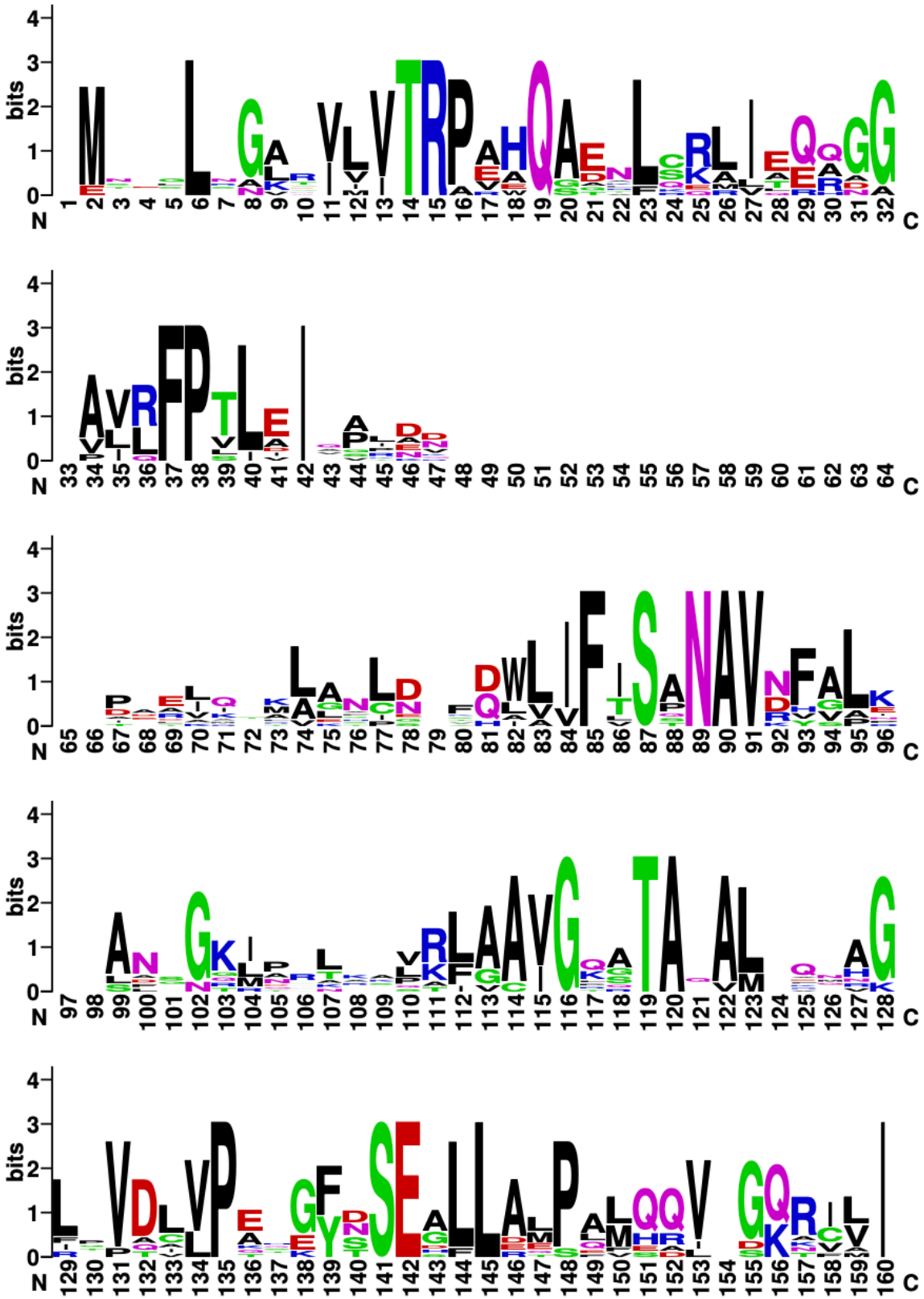


Figure 25: MUSCLE multiple sequence alignment real phylogenetic tree for *Coxiella burnetii* gene BMW92_RS10760 and the top 10 organism sequences similarities selected. Organisms sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered randomly and are listed from top to bottom as followed: *G. bacterium*, *C. burnetii*, *M. methanica*, *M. lenta*, *M. vadi*, *M. psychrotolerans*, *M. luteus*, *M. tundripaludum*, *M. oryzae*, *T. nivea*, *T. thiocyanatoxydans*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

MUSCLE Sequence Logo



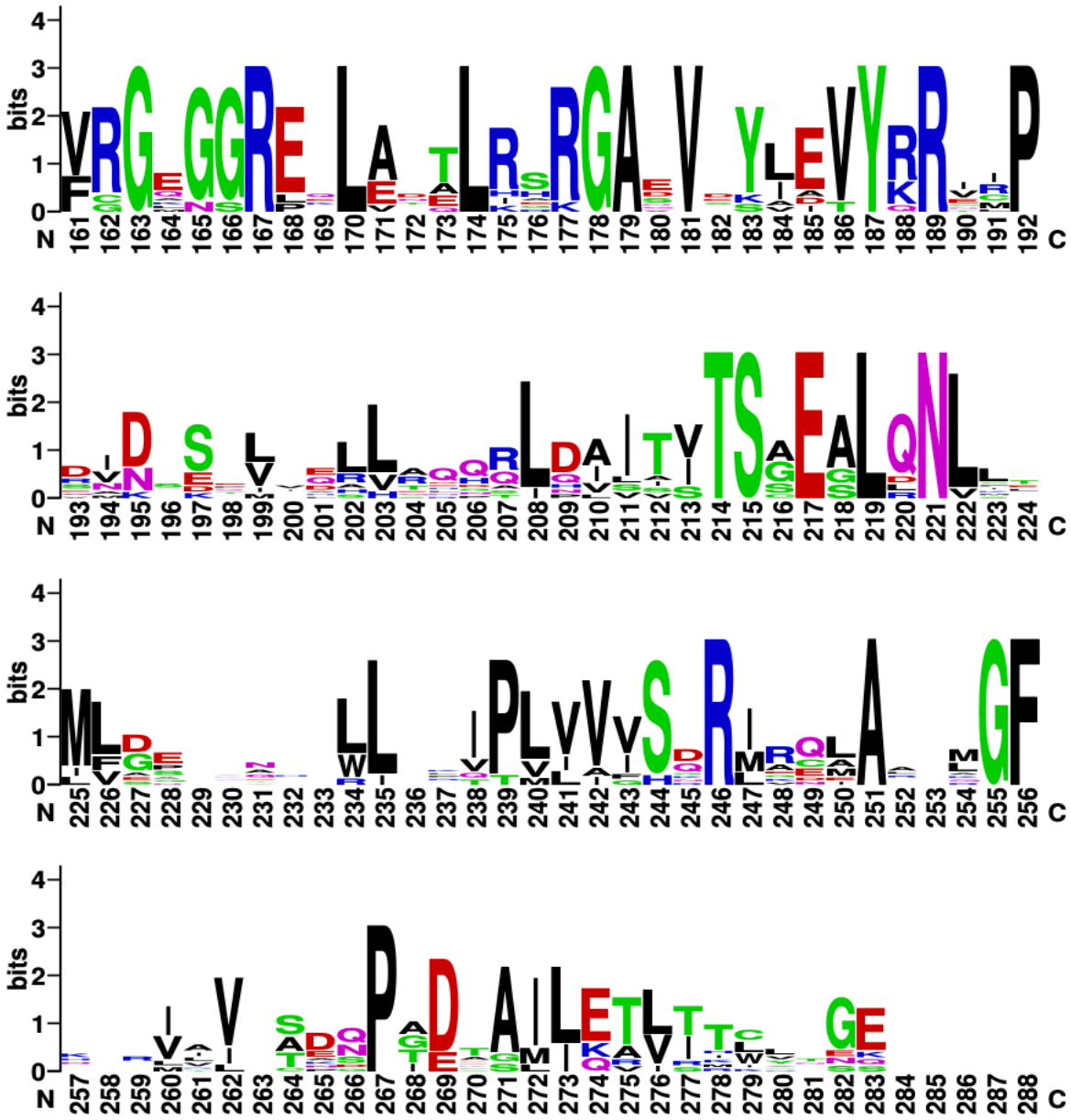


Figure 26: Sequence logo generated from the MUSCLE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10760 and the top 10 organism sequences similarities selected. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. The conservation of each amino acid among residue across similar sequences is represented by the height of each single letter abbreviation of the amino acid (WebLogo, <https://weblogo.berkeley.edu/logo.cgi>).

T-COFEE

```
CLUSTAL W (1.83) multiple sequence alignment

C.burnetii              MENESLKNKTIMITRPEWQGELLKKAIERRGGAVILFPTLIIKPINKCNY
G.bacterium             MS-ALLSGLDIIITRAVHQSENLRKAVLQHAGHPVLFPSLEISVLNNS--
M.lenta                 MT-TGLNGAWVLVTRPVAQAEKLCKLITQQNGQALQFPTLEIQPLKVDG-
M.luteus                MI-RGLNGARVLVTRPAHQAENLSRLIQERGGEVVRFPVLDIVARDNIE-
M.methanica             MT-LSLRGATVLVTRPAAQADTLCRLIAQADGRALRFPTLEIQPIDVDN-
M.oryzae                MN-KLLSGVRILVTRPEHQADNLSRLIEEQGGIAVRFPTLEIIAKDNAL-
M.psychrotolerans       MS--GLGGAGVLVTRPAHQAEVLCRLIAEQGGTAIRFPTLAIEATADTA-
M.tundripaludum         MS-KVLNGACVLVTRPEHQAENLSRLIEQRGGVAVRFPTLEIVSRDDD--
M.vadi                  MN-EQLQAKRILVTRPRHQAGNLCRLIEQQGGVAVRFPTLEIQALERPE-
T.nivea                 MP-ETLRGLNVVVTRPAHQAARFQQMLEQAGANAVLFPVIVIAPPEQPA-
T.thiocyanatoxydans     MG-CDLAGLRVVVTRPAEQATALQERITQAGGRALLFPLLAIAGPADPA-
                        *    *    :::**.  *.  : . : .  .  : ** : *

C.burnetii              SPFASASFPPSRESGSPDDKMTKAGFLNSSDILIFLSANAVKHSPI----
G.bacterium             ------------------ELQMMLGNINDKHLLIFTSQNAVDVVAP----
M.lenta                 ------------------E---LIEKALHCDWLIFTSTNAVDFALRALSG
M.luteus                ------------------EVQDALKNLDKFQWVVFISPNAVNFALKANNG
M.methanica             ------------------A---LIEKALTCNWLIFTSSNAVDFALKAFGG
M.oryzae                ------------------EIKQMLANPDLFQWLIFISANAVNFALKANDG
M.psychrotolerans       ------------------AVQTALANLGNFQWLIFISANAVNFALKANGG
M.tundripaludum         ------------------RIKSTLENLDGFQWVVFISANAVNFALKANSG
M.vadi                  ------------------TIAARVAALEHVDWLIFISANAVNFVLNSNSG
T.nivea                 ------------------LAQTMLASLDSYDAAIFISANAVRFGLEQLDE
T.thiocyanatoxydans     ------------------RLRPLLAGLSDTDLLIFVSPNAVRYGLEQLAA
                                           .  :* * ***

C.burnetii              --LNFKAEQKLVAIGTGTAAALFQ-RGLSVDAVPEH-FSSEGLLDLPLLH
G.bacterium             -RLPLNLKPAIGAIGPRTADALVN-HKIPVDILPTEKFDSEHLLALPFFE
M.lenta                 -KMTRLHALKLAAVGKATANALQE-VGLKVACVPETEFSSEGLLAESAMH
M.luteus                -KIDRLKTVRFAAVGRATAQALEA-AGLTVDVVPEQGYTSEALLAMPQMQ
M.methanica             -KMAGAMAVKLAAVGQATASALQK-AGLQVTCVPKTEFSSEGLLAQPAMQ
M.oryzae                -KIACTKSVRFAAVGQSTAQAMRM-AGLNVDLVPESGYNSEALLAMPELQ
M.psychrotolerans       -KIPKLIAPRLAAIGQSTAQALAN-AGLGVDLVPAQGFNSEALLAEPLLQ
M.tundripaludum         -KIPRTKSVRFAAVGQATAQAMKM-AGLPVDLVPEYGYNSEALLEMPQLQ
M.vadi                  -TINRLRRLRLAAVGKATAKALQN-NGLTVDLLPQHGFDSESLLRTPAMS
T.nivea                 NQRQTLRKLTLGAVGKQTAGVLQQ-HGFGVQLVPASGYTSEDFLALPAVQ
T.thiocyanatoxydans     -YGGLPAGSRLACVGLGTARALEQRAGRPPDLLPAGGYDSEALLALPALQ
                                 : .:*  ** .:          :*   : ** :*  . .
```

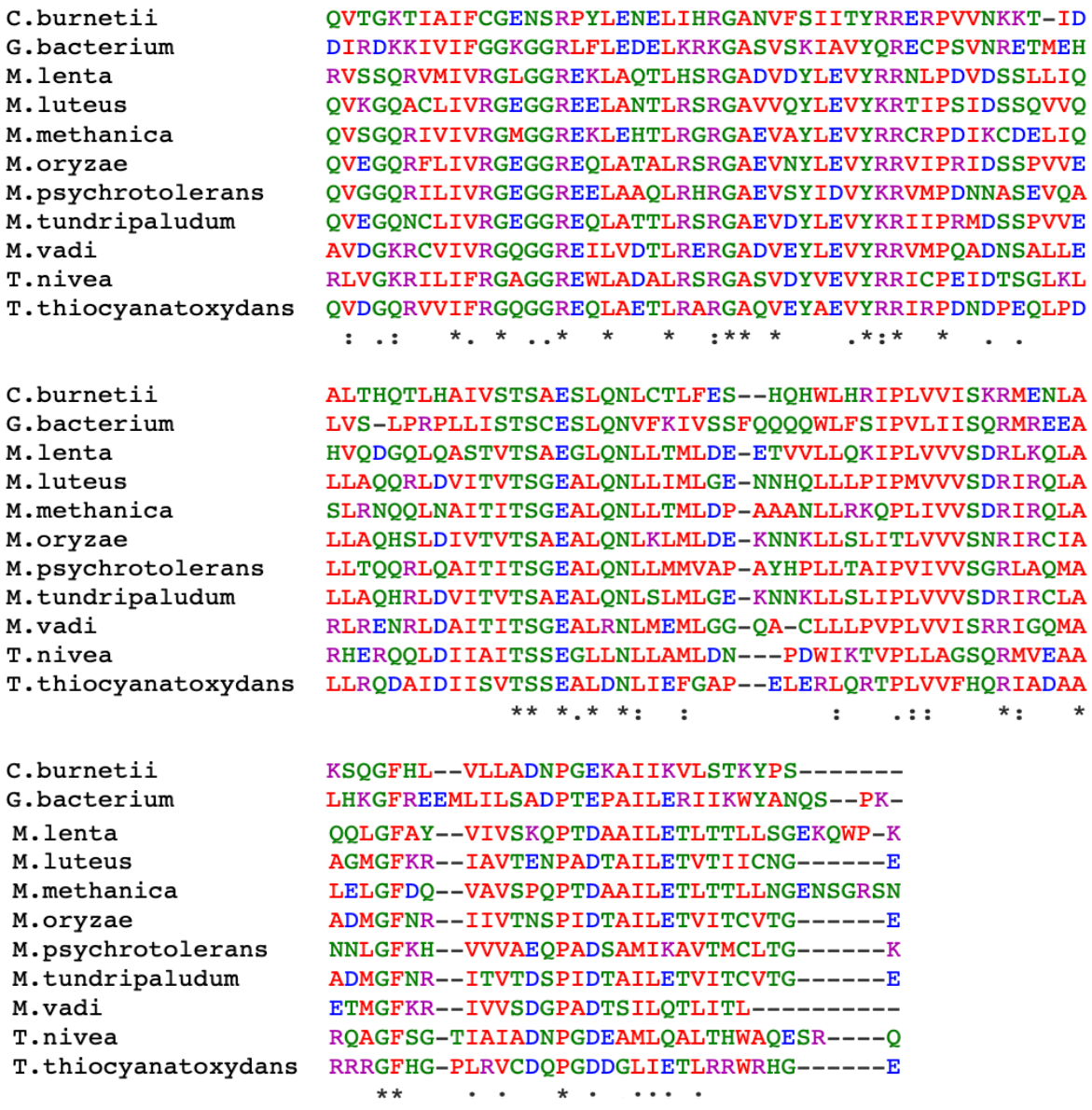


Figure 27: T-COFEE multiple sequence alignment for *Coxiella burnetii* gene BMW92_RS10760 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered randomly and are listed from top to bottom as followed: *C. burnetii*, *G. bacterium*, *M. lenta*, *M. luteus, M. methanica*, *M. oryzae*, *M. psychrotolerans*, *M. tundripaludum, M. vadi*, *T. nivea*, *T. thiocyanatoxydans*. Amino acids are represented by single letter

abbreviations and distinct colors for each respective amino acid (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

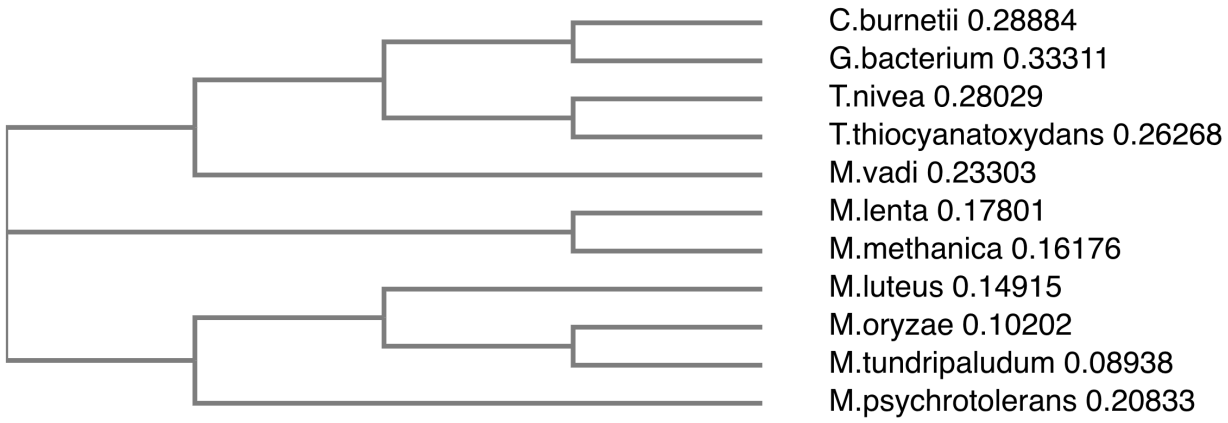


Figure 28: T-COFFEE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10760 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *C. burnetii*, *G. bacterium*, *T. nivea*, *T. thiocyanatoxydans*, *M. vadi, M. lenta, M. methanica, M. luteus, M. oryzae, M. tundripaludum, M. psychrotolerans*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

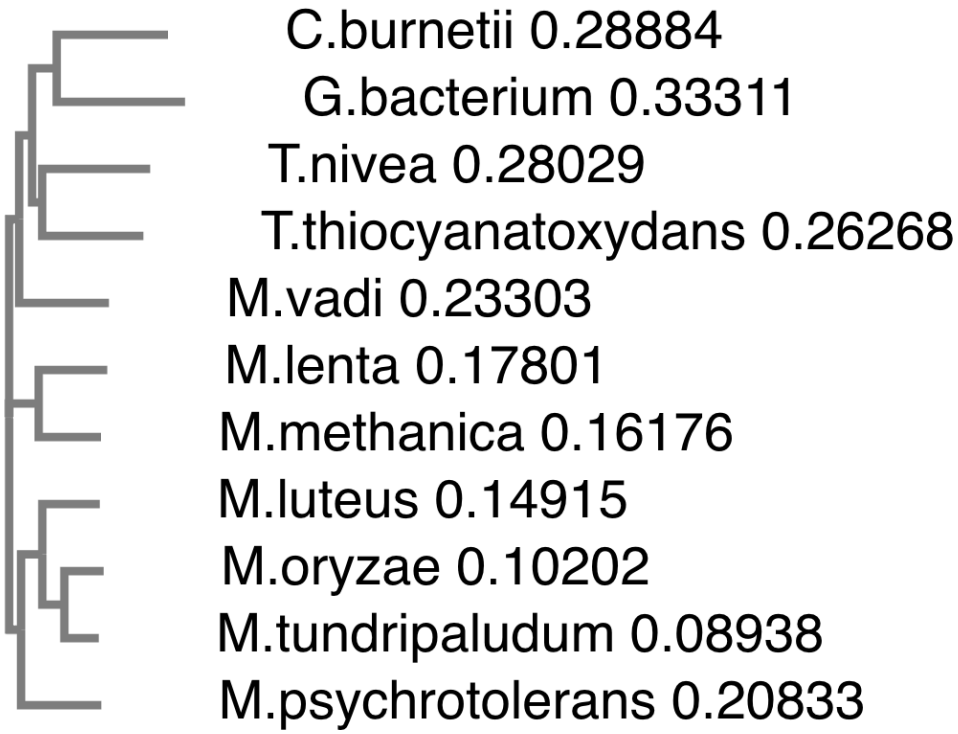
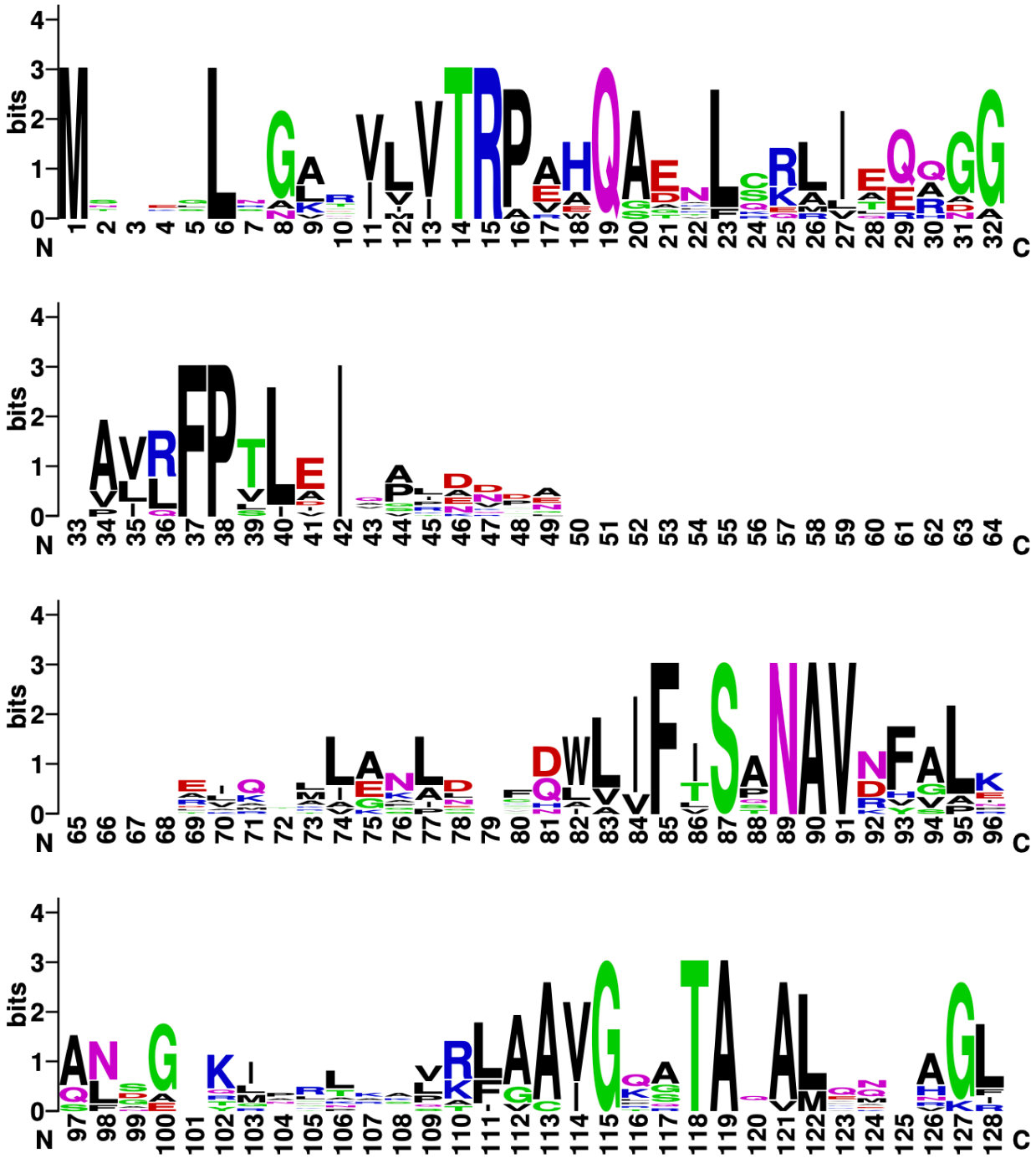


Figure 29: T-COFFEE multiple sequence alignment real phylogenetic tree for *Coxiella burnetii* gene BMW92_RS10760 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *C. burnetii*, *G. bacterium*, *T. nivea*, *T. thiocyanatoxydans*, *M. vadi, M. lenta, M. methanica, M. luteus, M. oryzae, M. tundripaludum, M. psychrotolerans*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

T-COFFEE Sequence Logo



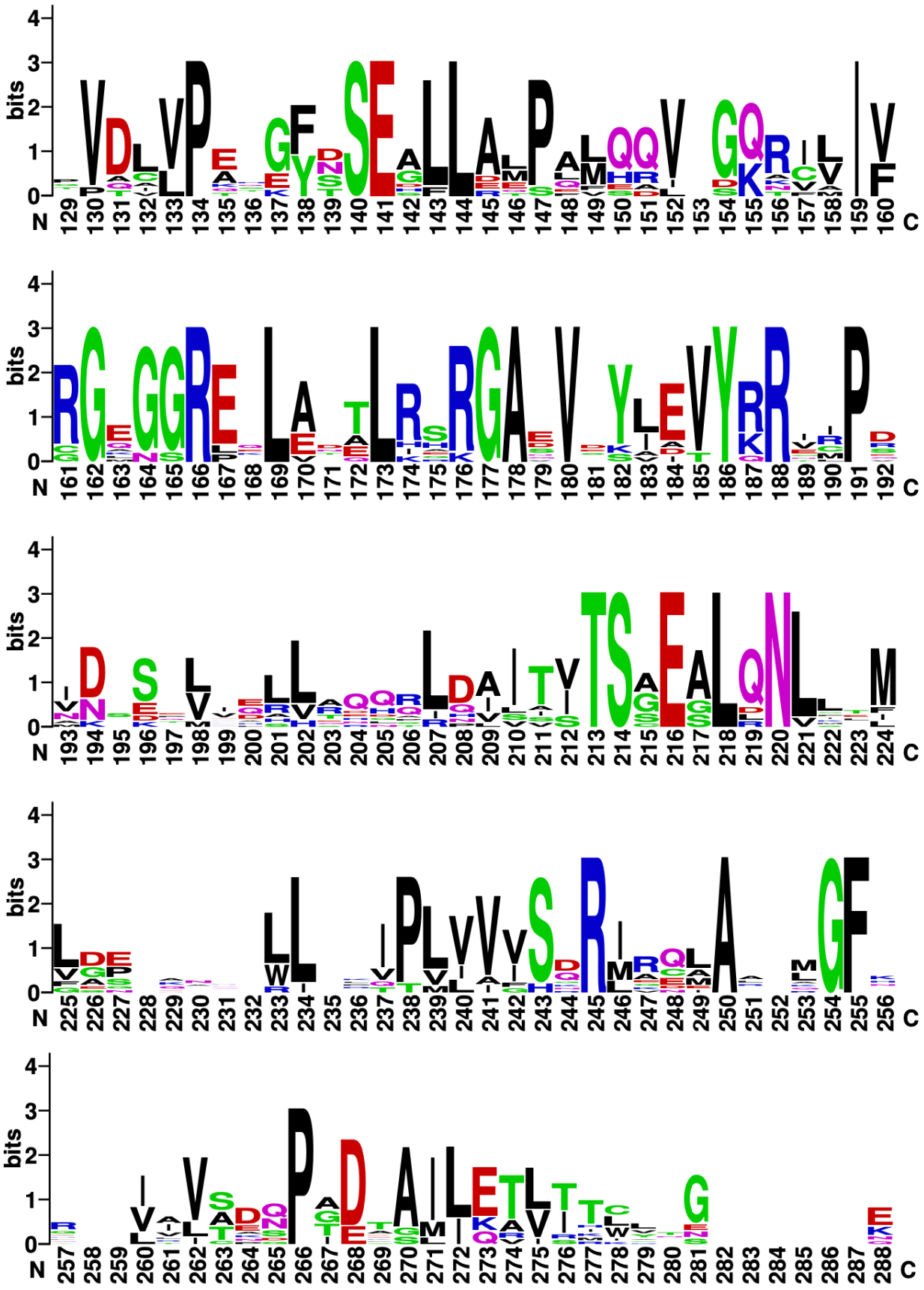
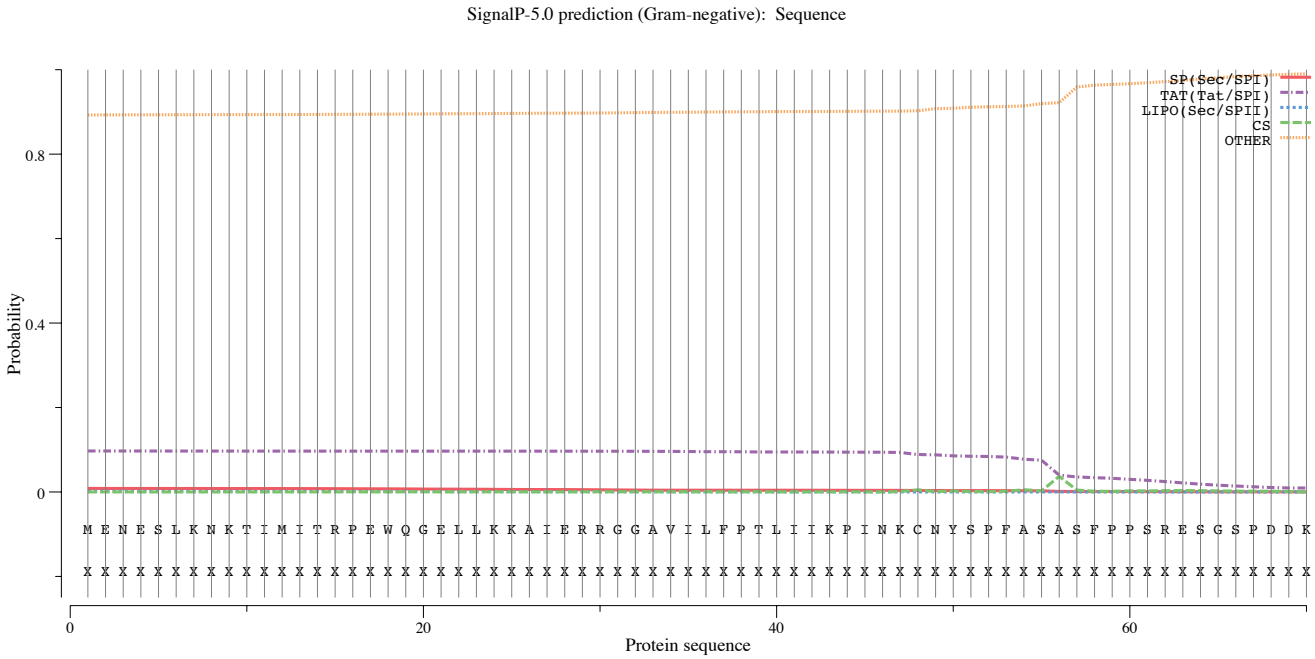

bits
4
3
2
1
0
N
C

Figure 30: Sequence logo generated from the T-COFFEE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10760 and the top 10 organism sequences similarities selected. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. The conservation of each amino acid among residue across similar sequences is represented by the height of each single letter abbreviation of the amino acid (WebLogo, <https://weblogo.berkeley.edu/logo.cgi>).

Protein Localization

SIGNALP



Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.0083	0.0968	0.0017	0.8932

Figure 31: SignalP 5.0 prediction (Gram-negative) for gene BMW92_RS10760 of *Coxiella burnetii*. The SP (Sec/SPI), TAT (Tat/SPI), LIPO (Sec/SPII), and CS probability scores indicate that this protein is not a signal peptide. The program calculated an 89.32% chance that this protein has another protein classification that is not related to similar function and type as signal peptides (SignalP, <http://www.cbs.dtu.dk/services/SignalP/ >).

LIPOP

```
# Sequence CYT score=-0.200913
# Cut-off=-3
Sequence        LipoP1.0:Best   CYT     1       1       -0.200913



# NO PLOT made - less than 4 putative cleavage sites predicted
```

Figure 32: LipoP 1.0 was unable to generate a plot graph due to there being less than four predicted putative cleavage sites. The best localization prediction resulted in the highest scoring class being the cytoplasmic protein class (LipoP, <http://www.cbs.dtu.dk/services/LipoP/>).

TMHMM

```
# WEBSEQUENCE Length: 268
# WEBSEQUENCE Number of predicted TMHs:  0
# WEBSEQUENCE Exp number of AAs in TMHs: 0.11125
# WEBSEQUENCE Exp number, first 60 AAs:  0.05351
# WEBSEQUENCE Total prob of N-in:        0.10637
WEBSEQUENCE     TMHMM2.0        outside      1   268
```

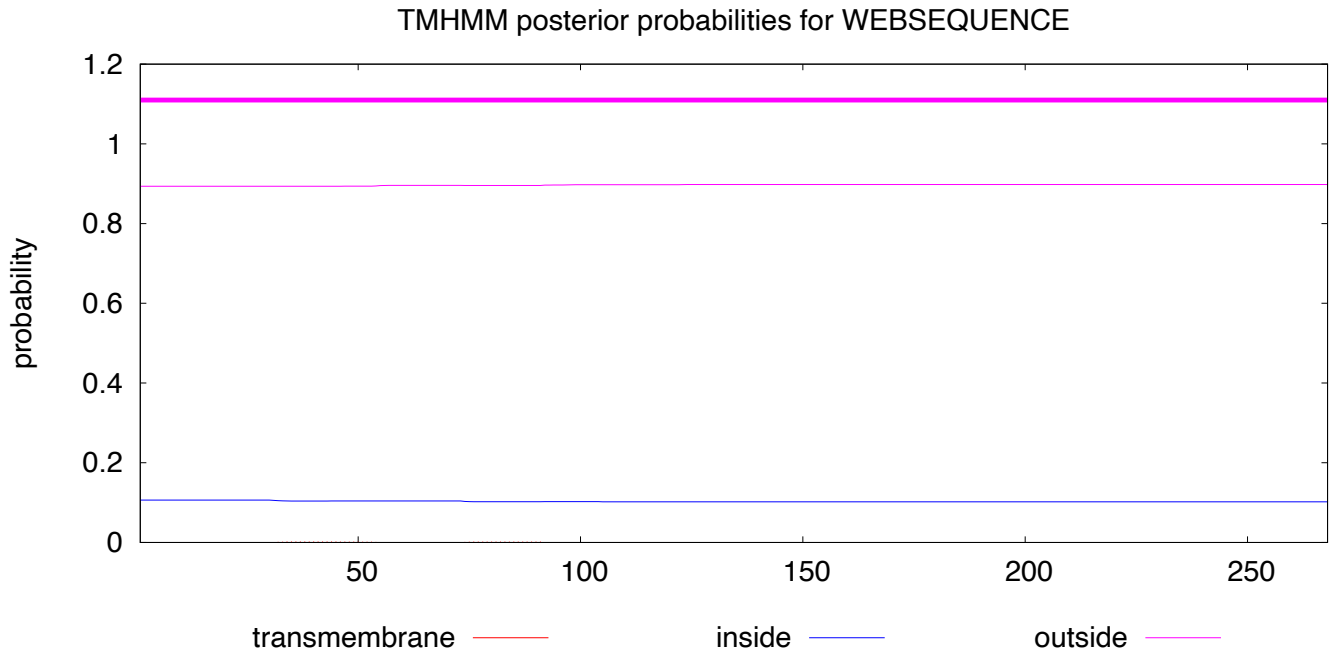


Figure 33: TMHMM posterior probability displayed a line graph that predicts the localization of the protein coded from BMW92_RS10760 as entirely outside the membrane. The red line, representative of the protein being located in the transmembrane, was 0% probability. This is indicative of the protein not being located within the transmembrane. The blue line, representative of the protein being located inside the membrane, was at 0.10 (0.10% probability). This is indicative of the protein being localized inside of the membrane highly unlikely. The

magenta line, representative of the protein being located outside the membrane, was at 0.90 (90% probability). This is indicative of the protein being localized outside of the membrane as highly likely (TMHMM, <http://www.cbs.dtu.dk/services/TMHMM/ >).

BOMP

The total number of valid proteins submitted is: 1
The total number of integral ß-barrel outer membrane proteins predicted is: 0

Sequence name	Category	Best BLAST hit

Figure 34: The BOMP test result identified there are no integral beta-barrel outer membrane proteins for gene BMW92_RS10760 (BOMP, <http://services.cbu.uib.no/tools/bomp>).

PSORTb

```
SeqID: C.burnetii
  Analysis Report:
    CMSVM-              Unknown                         [No details]
    CytoSVM-            Unknown                         [No details]
    ECSVM-              Unknown                         [No details]
    ModHMM-             Unknown                         [No internal helices found]
    Motif-              Unknown                         [No motifs found]
    OMPMotif-           Unknown                         [No motifs found]
    OMSVM-              Unknown                         [No details]
    PPSVM-              Unknown                         [No details]
    Profile-            Unknown                         [No matches to profiles found]
    SCL-BLAST-          Unknown                         [No matches against database]
    SCL-BLASTe-         Unknown                         [No matches against database]
    Signal-             Unknown                         [No signal peptide detected]
  Localization Scores:
    Cytoplasmic             2.00
    CytoplasmicMembrane     2.00
    Periplasmic             2.00
    OuterMembrane           2.00
    Extracellular           2.00
  Final Prediction:
    Unknown
```

Figure 35: The PSORTb test resulted in an analysis report that identified no internal helices, motifs, or signal peptides. The PSORTb localization scores resulted in a 2.0 value for every location (cytoplasmic, cytoplasmic membrane, periplasmic, outer membrane). The calculated localization scores for gene BMW92_RS10760 resulted in the predictable location of the protein to be unknown (PSORTb, <https://www.psort.org/psortb/ >).

Phobius

```
ID   UNNAMED
FT   TOPO_DOM      1    268       NON CYTOPLASMIC.
//
```

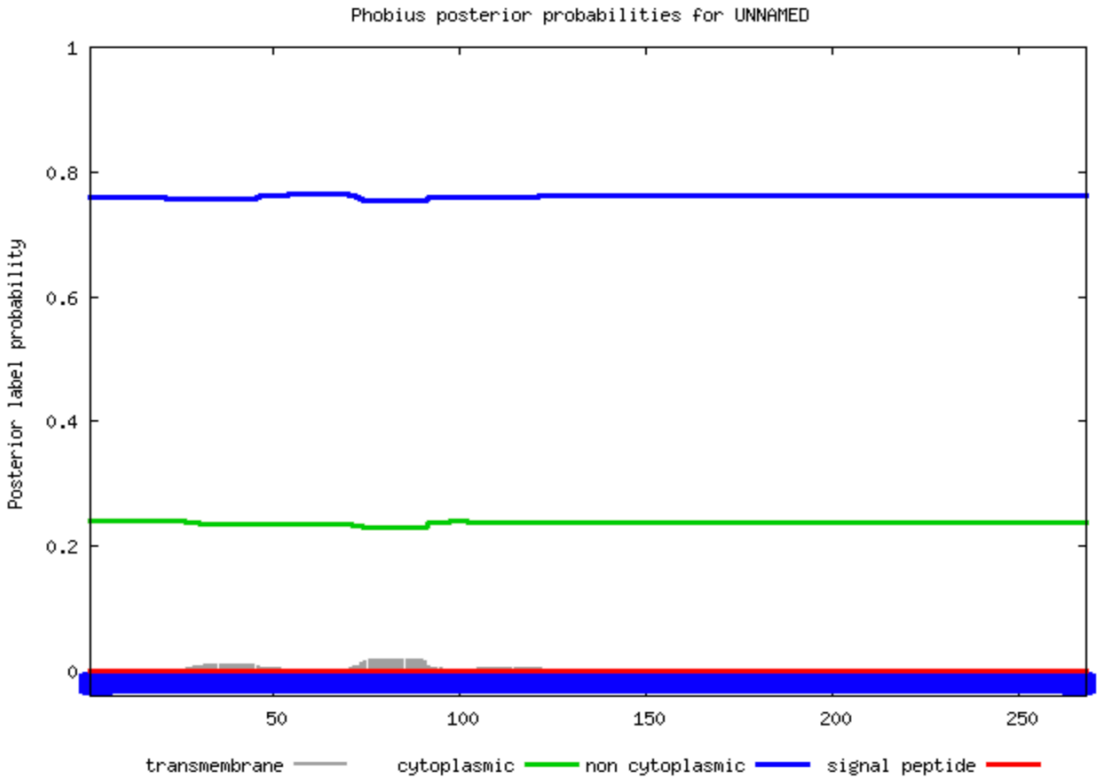


Figure 36: The Phobius posterior probability line graph generated for gene BMW92_RS10760 resulted in a calculated prediction that the whole sequence contains no membrane helices. The grey line, representative of the predicted transmembrane helices location, was less than 0.02 (0.02%) posterior probability. The green line, representative of the predicted cytoplasmic transmembrane helices location, was around 0.25 (25%) posterior probability. The blue line, representative of the predicted non-cytoplasmic transmembrane helices location, was around 0.75 (75%) posterior probability. The red line, representative of the presence or absence of a signal peptide, was 0.00 (0%) posterior probability (Phobius, <http://phobius.sbc.su.se>).

BMW92_RS10830

The second gene, BMW92_RS10830, was analyzed using bioinformatic technology. Table 2 below contains the provided data regarding basic information. A protein isoelectric point calculator was used to determine the isoelectric point of the protein, protein length, and the number and prevalence of each amino acid that makes up the protein (Figure 31). The BLASTp search tool produced 100 matches ranked from highest sequence similarity to lowest sequence similarity. The top ten sequences with significant alignments that were not identical species to *Coxiella burnetii* were selected. The information recorded included the organism name, protein name, percent identity, precent positive, length of alignment match, e-values, and percent gap. The highest ranked match to the BMW92_RS10830 gene was pyrroline-5-carboxylate reductase [*Coxiella mudrowiae*] (Figure 38). The remaining nine matches to the BMW92_RS10830 gene all had a function as pyrroline-5-carboxylate reductase (Figures 39-47) The CDD identified five potential protein domains hits conserved (Figure 48). Four of the domain hits conserved and identified by the CDD belong to the ProC superfamily, PRK11880 superfamily, P5CR dimer superfamily, or proC superfamily. Specific domain hits involved the PRK11880, ProC, and proC superfamilies. One domain hit conserved and identified as a non-specific domain hit was the NADP-binding-Glutamyl-tRNA-reductase, which is part of the NADB Rossman superfamily and further included in the ProC superfamily. The protein classification identified by the CDD was pyrroline-5-carboxylate reductase. Four of the domain hits sequences were aligned with the query sequence based off the amino acids that are highly conserved between both sequences (Figures 50-53). The MUSCLE program generated a multiple sequence alignment (MSA); each amino acid in the sequence was assigned a distinct color to distinguish the amino acids being compared (Figure 54). The MUSCLE program generated two phylogenetic trees using the

multiple sequence alignments to further confirm sequence similarity. The results displayed the numbers followed behind each organism at the end of each leaf node which displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii.* The use of a phylogenetic cladogram (Figure 55) and real phylogenetic tree (Figure 56) provided further understanding of the relatedness of common ancestors and organism sequences that are conserved. Each of the letter's heights produced correspond to the conservation of the amino acid residue across similar sequences. WebLogo produced a sequence logo that was generated from the MUSCLE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10830 and the top 10 organism sequences similarities selected (Figure 57). Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. Each of the single letter amino acid abbreviation's heights correspond to the conservation of the amino acid residue across similar sequences. The T-COFEE program generated another multiple sequence alignment to further confirm sequence similarity depicted with in the MUSCLE MSA (Figure 58). The T-COFFEE program generated two phylogenetic trees, phylogenetic cladogram (Figure 59) and real phylogenetic tree (Figure 60), using the multiple sequence alignment which displayed the genetic proximity and similarity between *Coxiella burnetii* and selected organisms from the BLASTp search. WebLogo constructed a sequence logo from the T-COFFEE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10830 and the top 10 organism sequences similarities selected to further display sequence similarity and conservation of sequences. Each of the single letter amino acid abbreviation's heights correspond to the conservation of the amino acid residue across similar sequences (Figure 61). Protein localization results included SignalP, LipoP, TMHMM, BOMP, PSORTb, and Phobius. The SignalP graphical illustration identified that there

is a small probability of a signal peptide present from amino acids 1-26 and no presence of a signal peptide for the remainder of the protein sequence (Figure 62). The LipoP resulted in the highest scoring class being the cytoplasmic protein class (Figure 63). The TMHMM test resulted in a graphical illustration, statistics, and a list of the predicted transmembrane helices and the predicted location of the intervening loop regions. The TMHMM test resulted and displayed that the whole sequence is highly unlikely to contain any transmembrane helices and that the majority of the protein has a high probability of being located outside of the membrane (Figure 64). The BOMP test result identified there are no integral beta-barrel outer membrane proteins (Figure 65). The PSORTb test resulted in an analysis report that identified no detectable internal helices, motifs, or signal peptides; the localization scores calculated the predictable location of the protein to be cytoplasmic (Figure 66). The Phobius test resulted in a line graphical illustration that identified a low probability of transmembrane helices localized in the cytoplasm and a high probability of cytoplasmic transmembrane helices present; the overall result calculated by Phobius resulted in the entire protein sequence as non-cytoplasmic (Figure 67).

Basic Information

Table 2: Gene BMW92_RS10830 basic information

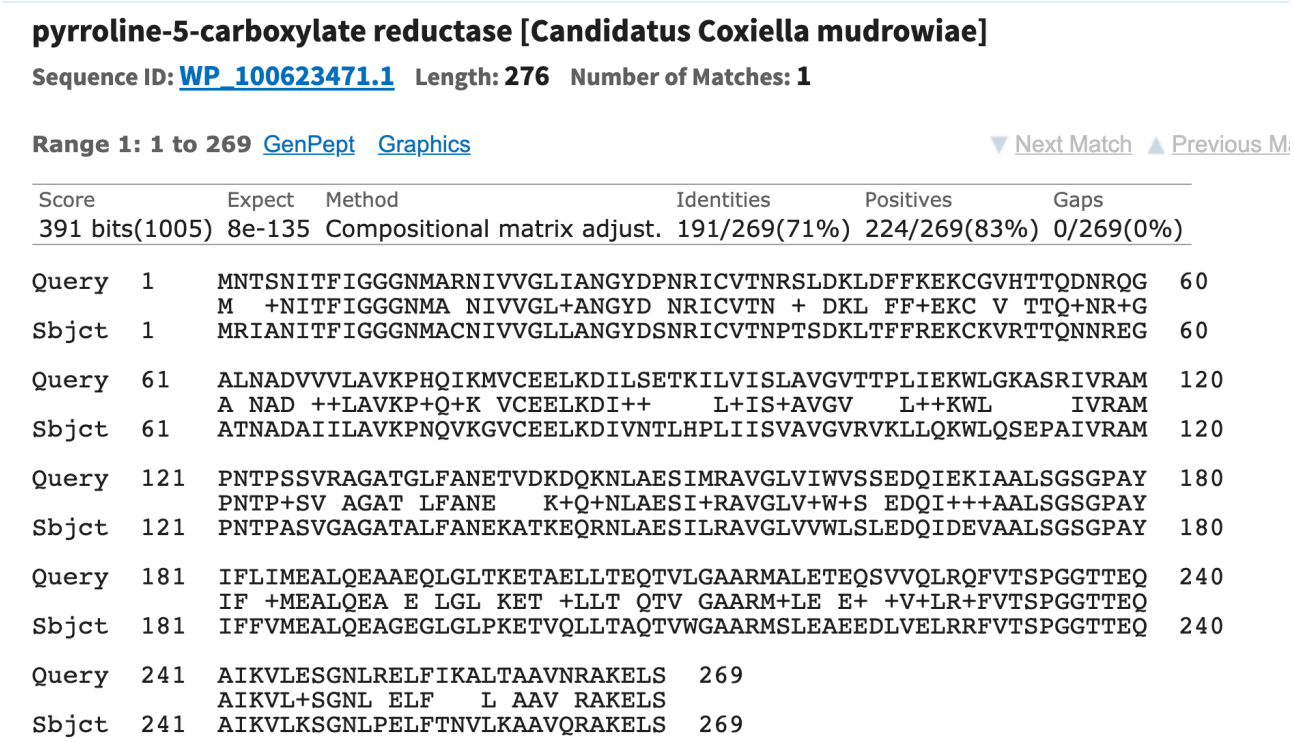
Genome	Replicon	Locus Tag	Old Locus Tag
Coxiella burnetii	NZ_CP018005	BMW92_RS10830	BMW92_10460
Genomic Coordinates	**Products**	**Length**	**Start and End Position**
1964091..1964915	pyrroline-5-carboxylate reductase	825 / 274	1964091 - 1964915
Molecular Weight	**Average Isoelectric Point**	**IPC Protein**	**Protein Length**
29422.01664 Da	6.217	6.04	273 amino acids

Nucleotide Sequence	Amino Acid Sequence
atgaatacttccaatattacttttatcggcggcgggaatatggcg cgcaatatcgtggtaggattaattgccaacggctacgaccctaa ccgtatttgtgttactaatcgaagtttagataaattagatttctttaa ggaaaagtgtggagtccatactactcaagataatcgtcaagga gctttgaacgctgatgtggttgtgttagccgttaaacctcatcaaa ttaaaatggtttgcgaggaattaaaagatattttaagcgaaacga aaattcttgtaatttccttagcagtaggcgttaccacaccgctcat tgaaaaatggttaggcaaggcttcacgtattgtgcgtgctatgcc caatacaccttcctcggtaagagccggtgctacaggtttatttgc aaacgagactgtggataaagaccaaaaaaatctagcggaatc gattatgcgtgcggtggggttggtcatttgggtttcgtctgagga ccaaattgaaaaaatagctgcactttcgggctcgggccctgctt atatttttttaattatggaggcacttcaggaggccgcagagcaatt aggtttaacgaaggaaacagcggaattgcttacggaacaaaca gttttgggcgcggctcgtatggcacttgaaacggaacaaagtgt agtacaattgcgtcaatttgtaacgtcgccaggtggcaccacgg agcaagcgatcaaagtattggaatcaggaaaccttcgtgaatta tttattaaagcgttaacagccgcggttaatcgcgctaaagagtta tcgaaaacggtagaccaatga	MNTSNITFIGGGNMARNIVVGLIANGYDP NRICVTNRSLDKLDFFKEKCGVHTTQDN RQGALNADVVVLAVKPHQIKMVCEELK DILSETKILVISLAVGVTTPLIEKWLGKAS RIVRAMPNTPSSVRAGATGLFANETVKD QKNLAESIMRAVGLVIWVSSEDQIEKIAA LSGSGPAYIFLIMEALQEAAEQLGLTKET AELLTEQTVLGARMALETEQSVVQLRQF VTSPGGTTEQAIKVLESGNLRELFIKALTA AVNRAKELSKTVDQ

Figure 37: Protein isoelectric point calculator. The number and prevalence of each amino acid in the protein coded from the BMW92_RS10830 gene of *Coxiella burnetii* (Kozlowski, Biology Direct, <http://isoelectric.org/>).

Sequence Similarity

BLAST



pyrroline-5-carboxylate reductase [Candidatus Coxiella mudrowiae]

Sequence ID: WP_100623471.1 **Length:** 276 **Number of Matches:** 1

Range 1: 1 to 269 GenPept Graphics ▼ Next Match ▲ Previous M

Score	Expect	Method	Identities	Positives	Gaps
391 bits(1005)	8e-135	Compositional matrix adjust.	191/269(71%)	224/269(83%)	0/269(0%)

```
Query  1    MNTSNITFIGGGNMARNIVVGLIANGYDPNRICVTNRSLDKLDFFKEKCGVHTTQDNRQG  60
            M  +NITFIGGGNMA NIVVGL+ANGYD NRICVTN + DKL FF+EKC V TTQ+NR+G
Sbjct  1    MRIANITFIGGGNMACNIVVGLLANGYDSNRICVTNPTSDKLTFFREKCKVRTTQNNREG  60

Query  61   ALNADVVVLAVKPHQIKMVCEELKDILSETKILVISLAVGVTTPLIEKWLGKASRIVRAM  120
            A NAD ++LAVKP+Q+K VCEELKDI++    L+IS+AVGV   L++KWL     IVRAM
Sbjct  61   ATNADAIILAVKPNQVKGVCEELKDIVNTLHPLIISVAVGVRVKLLQKWLQSEPAIVRAM  120

Query  121  PNTPSSVRAGATGLFANETVDKDQKNLAESIMRAVGLVIWVSSEDQIEKIAALSGSGPAY  180
            PNTP+SV AGAT LFANE   K+Q+NLAESI+RAVGLV+W+S EDQI+++AALSGSGPAY
Sbjct  121  PNTPASVGAGATALFANEKATKEQRNLAESILRAVGLVVWLSLEDQIDEVAALSGSGPAY  180

Query  181  IFLIMEALQEAAEQLGLTKETAELLTEQTVLGAARMALETEQSVVQLRQFVTSPGGTTEQ  240
            IF +MEALQEA E LGL KET +LLT QTV GAARM+LE E+ +V+LR+FVTSPGGTTEQ
Sbjct  181  IFFVMEALQEAGEGLGLPKETVQLLTAQTVWGAARMSLEAEEDLVELRRFVTSPGGTTEQ  240

Query  241  AIKVLESGNLRELFIKALTAAVNRAKELS  269
            AIKVL+SGNL ELF   L AAV RAKELS
Sbjct  241  AIKVLKSGNLPELFTNVLKAAVQRAKELS  269
```

Figure 38: BLAST first match for BMW92_RS10830 sequence from organism *Coxiella mudrowiae* with an e-value of 8e-135, 71% identity, 83% positives, 0% gaps (dissimilarity), and an identity of pyrroline-5-carboxylate reductase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

pyrroline-5-carboxylate reductase [Thioalbus denitrificans]

Sequence ID: WP_114279927.1 **Length:** 277 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 269 GenPept Graphics Next Match Previous

Score	Expect	Method	Identities	Positives	Gaps
286 bits(732)	3e-93	Compositional matrix adjust.	141/269(52%)	191/269(71%)	0/269(0%)

```
Query  1    MNTSNITFIGGGNMARNIVVGLIANGYDPNRICVTNRSLDKLDFFKEKCGVHTTQDNRQG  60
            M    I+FIGGGNM  ++V GLIA+GY P RI V++   + L   + + GVHTT DNR+
Sbjct  1    MEQGIISFIGGGNMCSSLVGGLIADGYAPERIRVSDPGEETLASLRARFGVHTTHDNREA  60

Query  61   ALNADVVVLAVKPHQIKMVCEELKDILSETKILVISLAVGVTTPLIEKWLGKASRIVRAM  120
            A  A VVVLAVKP  +  V  EL  ++ E   LV+S+A G+ T  +++WLG    +VR M
Sbjct  61   AAGAGVVVLAVKPQVLPKVAAELAPVVQEHGTLVVSIAAGIRTTDLQRWLGAGVALVRTM  120

Query  121  PNTPSSVRAGATGLFANETVDKDQKNLAESIMRAVGLVIWVSSEDQIEKIAALSGSGPAY  180
            PNTP+ V++GAT LFA   V   Q++ AES++RAVGL +W+ +E+Q++ + ALSGSGPAY
Sbjct  121  PNTPALVKSGATALFATAAVTAAQRDQAESVLRAVGLTLWLENEEQMDAVTALSGSGPAY  180

Query  181  IFLIMEALQEAAEQLGLTKETAELLTEQTVLGAARMALETEQSVVQLRQFVTSPGGTTEQ  240
             FL+MEA+Q AA+ +GL + TA LLT QT  GAA+MALE+++    LRQ VTSPGGTTE+
Sbjct  181  FFLVMEAMQGAAQAIGLPERTARLLTLQTAFGAAKMALESDEEPSLLRQRVTSPGGTTER  240

Query  241  AIKVLESGNLRELFIKALTAAVNRAKELS  269
            A+ VLE G LRELF  ALT+A +R++EL+
Sbjct  241  ALNVLEEGKLRELFRDALTSARDRSRELA  269
```

Figure 39: BLAST second match for BMW92_RS10830 sequence from organism *Thioalbus denitrificans* with an e-value of 3e-93, 52% identity, 71% positives, 0% gaps (dissimilarity), and an identity of pyrroline-5-carboxylate reductase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

pyrroline-5-carboxylate reductase [Nitrosococcus halophilus]

Sequence ID: WP_013034658.1 **Length:** 277 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 270 GenPept Graphics ▼ Next Match ▲ Previous

Score	Expect	Method	Identities	Positives	Gaps
281 bits(720)	2e-91	Compositional matrix adjust.	138/270(51%)	190/270(70%)	0/270(0%)

```
Query  1    MNTSNITFIGGGNMARNIVVGLIANGYDPNRICVTNRSLDKLDFFKEKCGVHTTQDNRQG  60
            MN   + FIGGGNMA +++ GLIA+G +   I V +    KLD    +  V+TT DN Q
Sbjct  1    MNEKTLAFIGGGNMATSLIGGLIADGRNAQTIWVADPDRSKLDALHHRFSVNTTPDNLQA  60

Query  61   ALNADVVVLAVKPHQIKMVCEELKDILSETKILVISLAVGVTTPLIEKWLGKASRIVRAM  120
            A  A+VVVLAVKP Q++ V   LK +++ ++ L +++A G+  P +E+WLG  + IVRAM
Sbjct  61   AQEAEVVVLAVKPQQLRTVATGLKSVVTSSQPLWLTIAAGIRIPDLERWLGGPAPIVRAM  120

Query  121  PNTPSSVRAGATGLFANETVDKDQKNLAESIMRAVGLVIWVSSEDQIEKIAALSGSGPAY  180
            PNTP+ V+AGAT LFAN   +  Q+ +AES++RAVGL +W+  E+ +E + ALSGSGPAY
Sbjct  121  PNTPALVQAGATALFANAQTNPQQRQMAESVLRAVGLTLWLKDENLMEVVTALSGSGPAY  180

Query  181  IFLIMEALQEAAEQLGLTKETAELLTEQTVLGAARMALETEQSVVQLRQFVTSPGGTTEQ  240
             FL+MEA+++AA  LGL   TA LLT +T  GAA+MALE+E+  ++LRQ VTSPGGTTE+
Sbjct  181  FFLVMEAMEKAAIDLGLDDSTARLLTLETAFGAAKMALESEEDSIRLRQRVTSPGGTTER  240

Query  241  AIKVLESGNLRELFIKALTAAVNRAKELSK  270
            AI  LE  N+RE F  AL AA +R +EL++
Sbjct  241  AITALEEANIREAFAHALRAARDRTRELAE  270
```

Figure 40: BLAST third match for BMW92_RS10830 sequence from organism *Nitrosococcus halophilus* with an e-value of 2e-91, 51% identity, 70% positives, 0% gaps (dissimilarity), and an identity of pyrroline-5-carboxylate reductase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

pyrroline-5-carboxylate reductase [Alkalilimnicola ehrlichii]

Sequence ID: WP_011628091.1 **Length:** 275 **Number of Matches:** 1

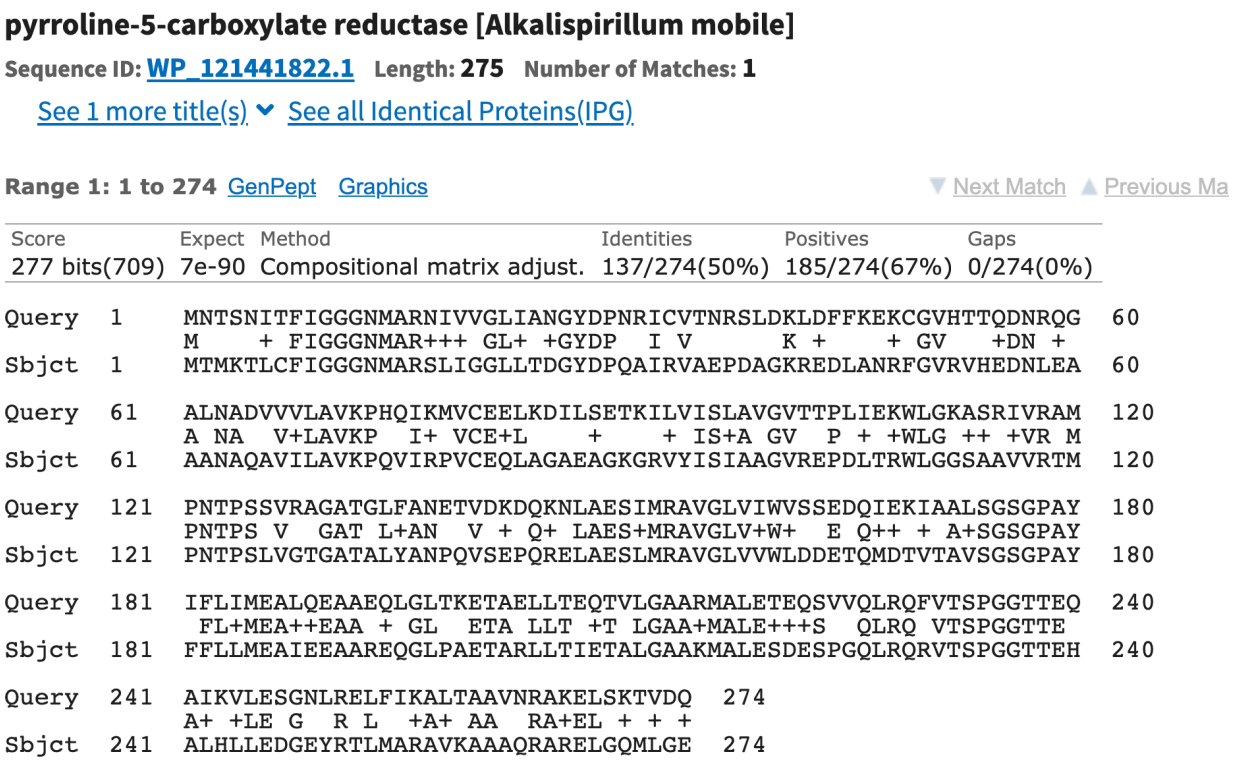
See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 270 GenPept Graphics ▼ Next Match ▲ Previous

Score	Expect	Method	Identities	Positives	Gaps
281 bits(720)	2e-91	Compositional matrix adjust.	137/270(51%)	186/270(68%)	0/270(0%)

```
Query  1    MNTSNITFIGGGNMARNIVVGLIANGYDPNRICVTNRSLDKLDFFKEKCGVHTTQDNRQG  60
            M+ + + FIGGGNMAR+++ GL+A+G+DP  + V +    K D    + GV    DN +
Sbjct  1    MSNNTLCFIGGGNMARSLIGGLLADGFDPQAVRVADPDAGKRDDLANRFGVRVYADNLEA  60

Query  61   ALNADVVVLAVKPHQIKMVCEELKDILSETKILVISLAVGVTTPLIEKWLGKASRIVRAM  120
            A +AD V+LAVKP  ++  CE+L     +   L IS+A GV  P + +WLG  + +VR M
Sbjct  61   AADADTVILAVKPQVVRRTACEQLVAGSGDAGRLFISIAAGVREPDLTRWLGGQAAVVRTM  120

Query  121  PNTPSSVRAGATGLFANETVDKDQKNLAESIMRAVGLVIWVSSEDQIEKIAALSGSGPAY  180
            PNTPS V  GAT L+AN+ V + Q+ LAES+MRAVGLV+W+  E Q++ + A+SGSGPAY
Sbjct  121  PNTPSLVGTGATALYANDRVKERQRELAESLMRAVGLVVWLDDEAQMDTVTAVSGSGPAY  180

Query  181  IFLIMEALQEAAEQLGLTKETAELLTEQTVLGAARMALETEQSVVQLRQFVTSPGGTTEQ  240
             FL+MEA+++AA  LGL  ETA LLT +T LGAA+MALE+++S  QLRQ VTSPGGTTE
Sbjct  181  FFLLMEAIEDAARDLGLPGETARLLTIETALGAAKMALESDESPAQLRQRVTSPGGTTEH  240

Query  241  AIKVLESGNLRELFIKALTAAVNRAKELSK  270
            A+ VLE G  R L  +A+ AA  RA+EL +
Sbjct  241  ALHVLEDGEYRALMTRAVQAAAKRAQELGQ  270
```

Figure 41: BLAST fourth match for BMW92_RS10830 sequence from organism *Alkalilimnicola ehrlichii* with an e-value of 2e-91, 51% identity, 68% positives, 0% gaps (dissimilarity), and an identity of pyrroline-5-carboxylate reductase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).



pyrroline-5-carboxylate reductase [Alkalispirillum mobile]

Sequence ID: WP_121441822.1 **Length:** 275 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 274 GenPept Graphics Next Match Previous Ma

Score	Expect	Method	Identities	Positives	Gaps
277 bits(709)	7e-90	Compositional matrix adjust.	137/274(50%)	185/274(67%)	0/274(0%)

```
Query  1    MNTSNITFIGGGNMARNIVVGLIANGYDPNRICVTNRSLDKLDFFKEKCGVHTTQDNRQG  60
            M    + FIGGGNMAR+++ GL+ +GYDP  I V      K +    + GV   +DN +
Sbjct  1    MTMKTLCFIGGGNMARSLIGGLLTDGYDPQAIRVAEPDAGKREDLANRFGVRVHEDNLEA  60

Query  61   ALNADVVVLAVKPHQIKMVCEELKDILSETKILVISLAVGVTTPLIEKWLGKASRIVRAM  120
            A NA  V+LAVKP  I+ VCE+L    +    + IS+A GV  P + +WLG ++ +VR M
Sbjct  61   AANAQAVILAVKPQVIRPVCEQLAGAEAGKGRVYISIAAGVREPDLTRWLGGSAAVVRTM  120

Query  121  PNTPSSVRAGATGLFANETVDKDQKNLAESIMRAVGLVIWVSSEDQIEKIAALSGSGPAY  180
            PNTPS V  GAT L+AN  V + Q+ LAES+MRAVGLV+W+  E Q++ + A+SGSGPAY
Sbjct  121  PNTPSLVGTGATALYANPQVSEPQRELAESLMRAVGLVVWLDDETQMDTVTAVSGSGPAY  180

Query  181  IFLIMEALQEAAEQLGLTKETAELLTEQTVLGAARMALETEQSVVQLRQFVTSPGGTTEQ  240
             FL+MEA++EAA + GL  ETA LLT +T LGAA+MALE+++S  QLRQ VTSPGGTTE
Sbjct  181  FFLLMEAIEEAAREQGLPAETARLLTIETALGAAKMALESDESPGQLRQRVTSPGGTTEH  240

Query  241  AIKVLESGNLRELFIKALTAAVNRAKELSKTVDQ  274
            A+ +LE G  R L  +A+ AA  RA+EL + + +
Sbjct  241  ALHLLEDGEYRTLMARAVKAAAQRARELGQMLGE  274
```

Figure 42: BLAST fifth match for BMW92_RS10830 sequence from organism *Alkalispirillum mobile* with an e-value of 7e-90, 50% identity, 67% positives, 0% gaps (dissimilarity), and an identity of pyrroline-5-carboxylate reductase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

pyrroline-5-carboxylate reductase [Chromatiales bacterium]

Sequence ID: HDO72906.1 **Length:** 276 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 274 GenPept Graphics ▼ Next Match ▲ Previous M

Score	Expect	Method	Identities	Positives	Gaps
277 bits(709)	9e-90	Compositional matrix adjust.	138/274(50%)	192/274(70%)	1/274(0%)

```
Query  1    MNTSNITFIGGGNMARNIVVGLIANGYDPNRICVTNRSLDKLDFFKEKCGVHTTQDNRQG  60
            M   NI FIGGGNMA +++ GL+A+   P R+CV +R   + +    + GV T++DN
Sbjct  1    MKDVNIAFIGGGNMATSLIGGLLADHVSPARLCVADRDPAQREHLAAQFGVRTSEDNAAC  60

Query  61   ALNADVVVLAVKPHQIKMVCEELKDILSETKILVISLAVGVTTPLIEKWLGKAS-RIVRA  119
            A +ADV+VLAVKP  +  VCE L D +   + LV+S+A GV T  + +WLG     IVRA
Sbjct  61   AEDADVIVLAVKPQVLHEVCEALTDSVQRKQPLVVSVAAGVRTDSLRRWLGGGDVAIVRA  120

Query  120  MPNTPSSVRAGATGLFANETVDKDQKNLAESIMRAVGLVIWVSSEDQIEKIAALSGSGPA  179
            MPNTP+ +++GATGL+A   V ++Q++LAE+I+RA GL +WV  E Q++ + ALSGSGPA
Sbjct  121  MPNTPALLQSGATGLYACTGVSEEQRDLAEAILRATGLTLWVDDEAQMDIVTALSGSGPA  180

Query  180  YIFLIMEALQEAAEQLGLTKETAELLTEQTVLGAARMALETEQSVVQLRQFVTSPGGTTE  239
            Y F +ME L++AA +LGL  +TA LLT QT LGAARMALE+ + V  LR+ VTSPGGTTE
Sbjct  181  YFFRVMEGLEKAATELGLPAQTARLLTLQTALGAARMALESSEPVATLRKRVTSPGGTTE  240

Query  240  QAIKVLESGNLRELFIKALTAAVNRAKELSKTVD  273
            Q +K +E+G++  L  K L AA +R++EL+K +D
Sbjct  241  QGLKAMEAGDIDALLGKVLKAARDRSRELAKLLD  274
```

Figure 43: BLAST sixth match for BMW92_RS10830 sequence from organism *Chromatiales bacterium* with an e-value of 9e-90, 50% identity, 70% positives, 0% gaps (dissimilarity), and an identity of pyrroline-5-carboxylate reductase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

pyrroline-5-carboxylate reductase [Nitrococcus mobilis]

Sequence ID: WP_005003398.1 **Length:** 275 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 270 GenPept Graphics Next Match Previous

Score	Expect	Method	Identities	Positives	Gaps
277 bits(708)	9e-90	Compositional matrix adjust.	145/270(54%)	187/270(69%)	0/270(0%)

```
Query  1    MNTSNITFIGGGNMARNIVVGLIANGYDPNRICVTNRSLDKLDFFKEKCGVHTTQDNRQG  60
            M   +ITFIGGGNMA ++V GLIA+GY   R+ V +    K      +  +H  +DNR+
Sbjct  1    MAEESITFIGGGNMAYSLVGGLIADGYRAERVHVADPDPAKRMDLANRFRIHVHEDNRKA  60

Query  61   ALNADVVVLAVKPHQIKMVCEELKDILSETKILVISLAVGVTTPLIEKWLGKASRIVRAM  120
             L A  VVLAVKP  IK V E L  IL E K LVIS+A GV  P I +WLG    +VR M
Sbjct  61   VLRAAAVVLAVKPQIIKSVLEPLGPILREQKSLVISIAAGVREPDISRWLGGQIAVVRTM  120

Query  121  PNTPSSVRAGATGLFANETVDKDQKNLAESIMRAVGLVIWVSSEDQIEKIAALSGSGPAY  180
            PNTP+ VRAGAT L+ANE V ++Q++LAES++RAVG++ W+  E  ++ + ALSGSGPAY
Sbjct  121  PNTPALVRAGATALYANEYVSQNQRDLAESLLRAVGIIQWLDDETLLDIVTALSGSGPAY  180

Query  181  IFLIMEALQEAAEQLGLTKETAELLTEQTVLGAARMALETEQSVVQLRQFVTSPGGTTEQ  240
             FL+ME L+ AA +LGL ++TA LLT +T LGAARMALE+++   +LR  VTSPGGTTE
Sbjct  181  FFLLMETLEAAAIELGLPEQTARLLTLETALGAARMALESDEDPGRLRLRVTSPGGTTEA  240

Query  241  AIKVLESGNLRELFIKALTAAVNRAKELSK  270
            A +VLESG  ++LF +AL AA  RA EL +
Sbjct  241  ATRVLESGGAQKLFQQALQAATTRAGELGR  270
```

Figure 44: BLAST seventh match for BMW92_RS10830 sequence from organism *Nitrococcus mobilis* with an e-value of 9e-90, 54% identity, 69% positives, 0% gaps (dissimilarity), and an identity of pyrroline-5-carboxylate reductase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

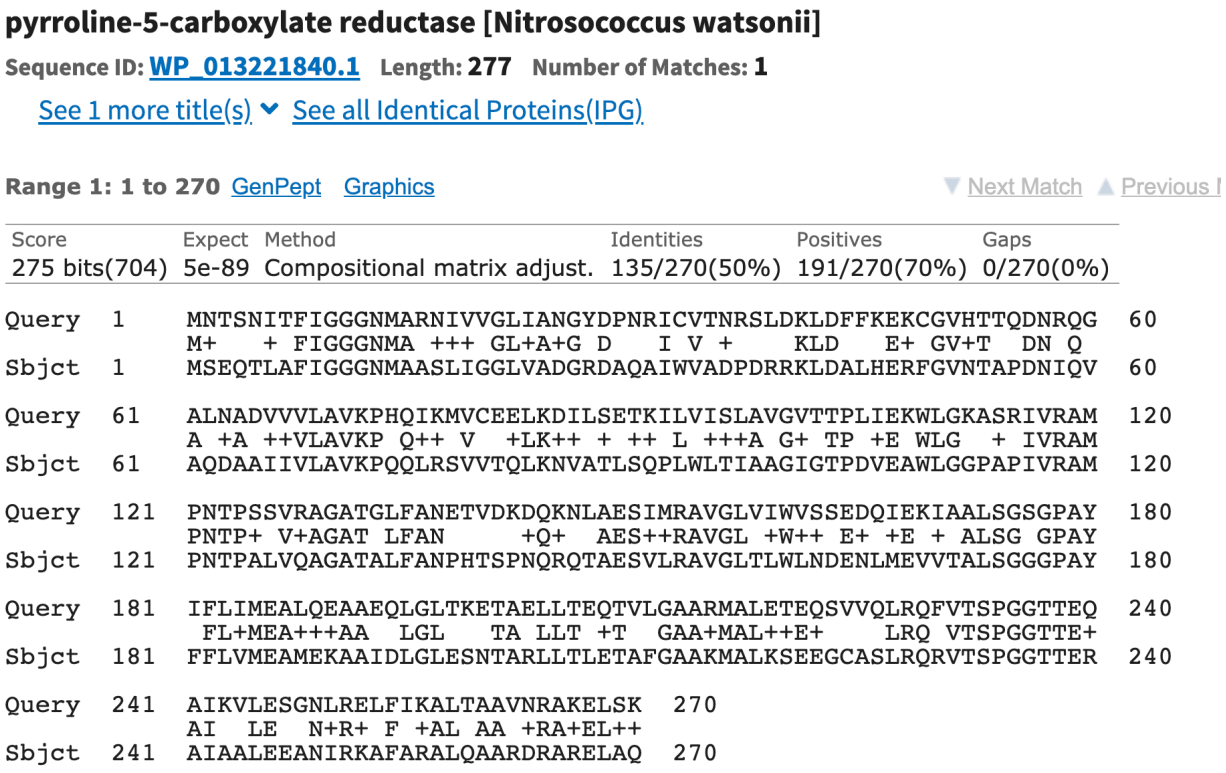
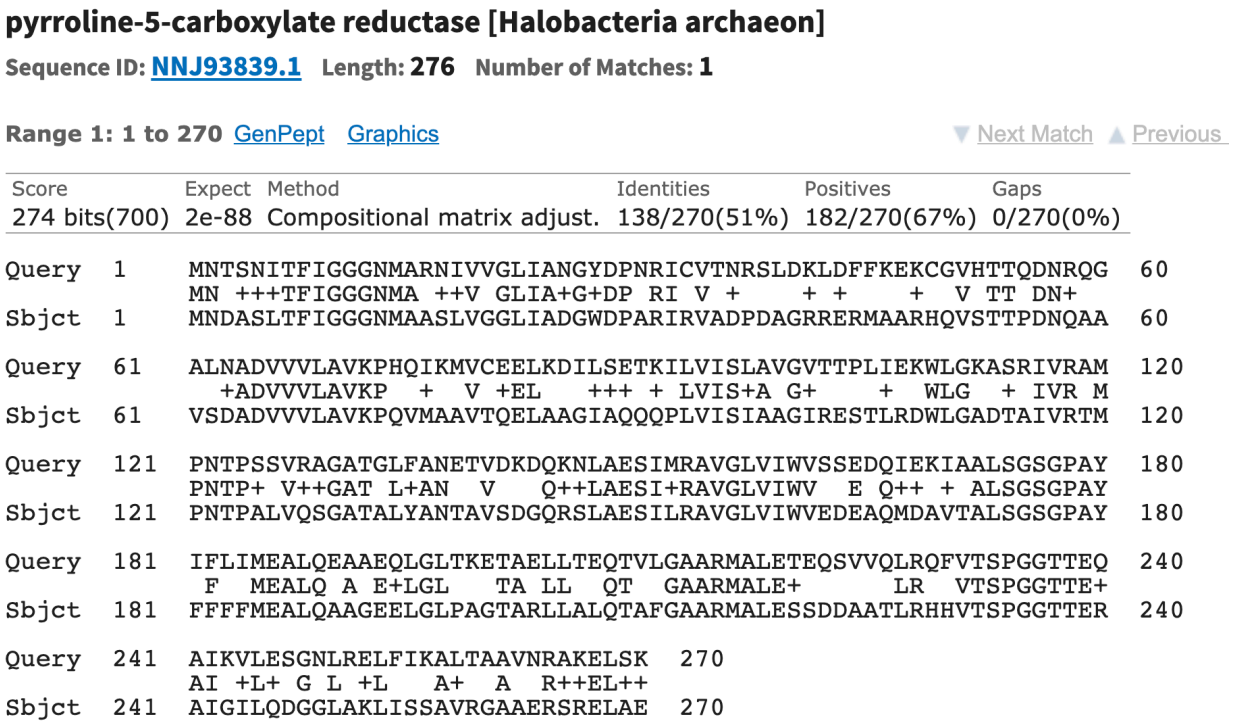


Figure 45: BLAST eighth match for BMW92_RS10830 sequence from organism *Nitrosococcus watsonii* with an e-value of 5e-89, 50% identity, 70% positives, 0% gaps (dissimilarity), and an identity of pyrroline-5-carboxylate reductase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).



pyrroline-5-carboxylate reductase [Halobacteria archaeon]

Sequence ID: NNJ93839.1 **Length:** 276 **Number of Matches:** 1

Range 1: 1 to 270 GenPept Graphics ▼ Next Match ▲ Previous

Score	Expect	Method	Identities	Positives	Gaps
274 bits(700)	2e-88	Compositional matrix adjust.	138/270(51%)	182/270(67%)	0/270(0%)

```
Query  1    MNTSNITFIGGGNMARNIVVGLIANGYDPNRICVTNRSLDKLDFFKEKCGVHTTQDNRQG  60
            MN +++TFIGGGNMA ++V GLIA+G+DP RI V +    + +    +  V TT DN+
Sbjct  1    MNDASLTFIGGGNMAASLVGGLIADGWDPARIRVADPDAGRRERMAARHQVSTTPDNQAA  60

Query  61   ALNADVVVLAVKPHQIKMVCEELKDILSETKILVISLAVGVTTPLIEKWLGKASRIVRAM  120
              +ADVVVLAVKP  +  V +EL   +++ + LVIS+A G+    +  WLG  + IVR M
Sbjct  61   VSDADVVVLAVKPQVMAAVTQELAAGIAQQQPLVISIAAGIRESTLRDWLGADTAIVRTM  120

Query  121  PNTPSSVRAGATGLFANETVDKDQKNLAESIMRAVGLVIWVSSEDQIEKIAALSGSGPAY  180
            PNTP+ V++GAT L+AN  V   Q++LAESI+RAVGLVIWV  E Q++ + ALSGSGPAY
Sbjct  121  PNTPALVQSGATALYANTAVSDGQRSLAESILRAVGLVIWVEDEAQMDAVTALSGSGPAY  180

Query  181  IFLIMEALQEAAEQLGLTKETAELLTEQTVLGAARMALETEQSVVQLRQFVTSPGGTTEQ  240
             F  MEALQ A E+LGL   TA LL  QT  GAARMALE+      LR  VTSPGGTTE+
Sbjct  181  FFFFMEALQAAGEELGLPAGTARLLALQTAFGAARMALESSDDAATLRHHVTSPGGTTER  240

Query  241  AIKVLESGNLRELFIKALTAAVNRAKELSK  270
            AI +L+ G L +L   A+  A  R++EL++
Sbjct  241  AIGILQDGGLAKLISSAVRGAAERSRELAE  270
```

Figure 46: BLAST ninth match for BMW92_RS10830 sequence from organism *Halobacteria archaeon* with an e-value of 2e-88, 51% identity, 67% positives, 0% gaps (dissimilarity), and an identity of pyrroline-5-carboxylate reductase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

pyrroline-5-carboxylate reductase [Aquicella lusitana]

Sequence ID: WP_114834951.1 **Length:** 275 **Number of Matches:** 1

See 2 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 269 GenPept Graphics ▼ Next Match ▲ Previous

Score	Expect	Method	Identities	Positives	Gaps
272 bits(695)	9e-88	Compositional matrix adjust.	139/269(52%)	188/269(69%)	0/269(0%)

```
Query  1    MNTSNITFIGGGNMARNIVVGLIANGYDPNRICVTNRSLDKLDFFKEKCGVHTTQDNRQG  60
            M+T  I+ IG GNM  +++ GLI +G+  +++   + S +KL   K K  ++TT DN Q
Sbjct  1    MHTPVISIIGAGNMGSSLIGGLIKDGHPSDKLWAADPSGEKLTQLKTKFDINTTSDNAQA  60

Query  61   ALNADVVVLAVKPHQIKMVCEELKDILSETKILVISLAVGVTTPLIEKWLGKASRIVRAM  120
               AD ++ AVKP     V   LK +++E K LVIS+A G+    I++WL   + IVRAM
Sbjct  61   IQAADTIIFAVKPQAFAHVALPLKKVIAERKPLVISIAAGIREASIQQWLNGKTPIVRAM  120

Query  121  PNTPSSVRAGATGLFANETVDKDQKNLAESIMRAVGLVIWVSSEDQIEKIAALSGSGPAY  180
            PNTP+ +  GAT L+AN  V + Q+NLAESI+RAVG+V+W++ E  ++ + ALSGSGPAY
Sbjct  121  PNTPALIGCGATALYANPYVTESQRNLAESILRAVGVVVWLNDEKLMDTVTALSGSGPAY  180

Query  181  IFLIMEALQEAAEQLGLTKETAELLTEQTVLGAARMALETEQSVVQLRQFVTSPGGTTEQ  240
             FL+MEALQEAAE LGL  ETA LLT QT LGAARMA+E+  S+ +LR+ VTSPGGTTE+
Sbjct  181  FFLMMEALQEAAEDLGLPTETARLLTLQTALGAARMAIESGTSLAELRRKVTSPGGTTEK  240

Query  241  AIKVLESGNLRELFIKALTAAVNRAKELS  269
            AI VLE  N+R LF +AL AA  R++EL+
Sbjct  241  AISVLEENNIRRLFKQALQAAKLRSEELA  269
```

Figure 47: BLAST tenth match for BMW92_RS10830 sequence from organism *Aquicella lusitana* with an e-value of 9e-88, 52% identity, 69% positives, 0% gaps (dissimilarity), and an identity of pyrroline-5-carboxylate reductase (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

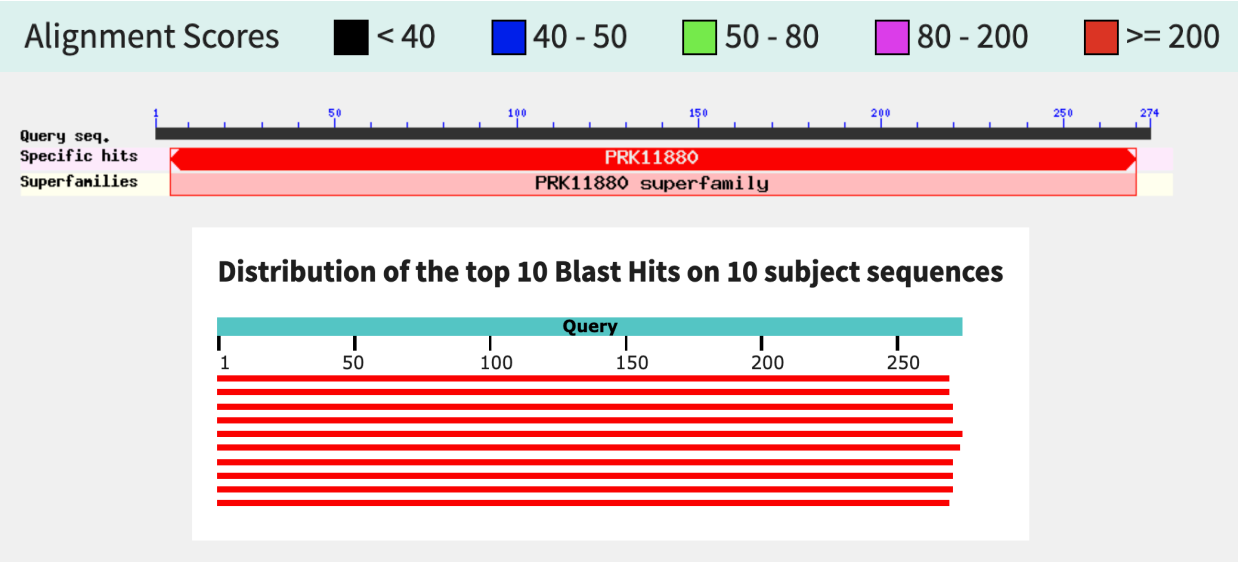
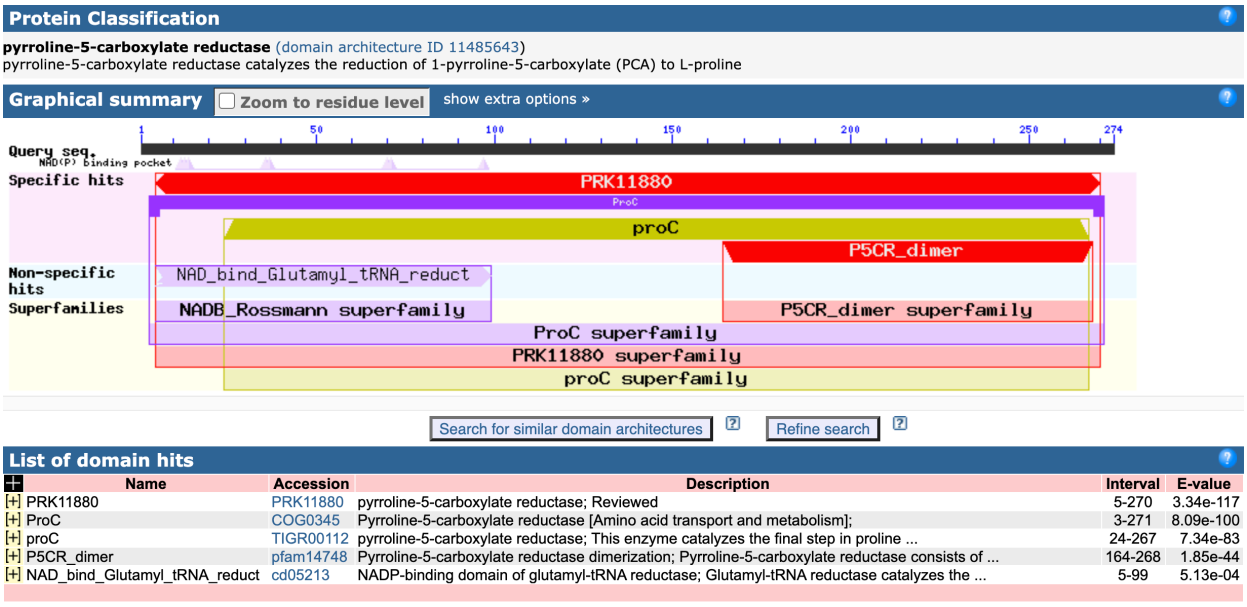


Figure 48: BLAST graphic summary of the top 10 organism sequences similarities selected aligned with *Coxiella burnetii* query sequence of gene BMW92_RS10830. Each of the alignment sequences selected are order from highest sequence similarity (top) to lowest sequence similarity (bottom). All organism sequences aligned with the query sequence have an alignment score of greater than 200 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

CDD



Name	Accession	Description	Interval	E-value
[+] PRK11880	PRK11880	pyrroline-5-carboxylate reductase; Reviewed	5-270	3.34e-117
[+] ProC	COG0345	Pyrroline-5-carboxylate reductase [Amino acid transport and metabolism];	3-271	8.09e-100
[+] proC	TIGR00112	pyrroline-5-carboxylate reductase; This enzyme catalyzes the final step in proline ...	24-267	7.34e-83
[+] P5CR_dimer	pfam14748	Pyrroline-5-carboxylate reductase dimerization; Pyrroline-5-carboxylate reductase consists of ...	164-268	1.85e-44
[+] NAD_bind_Glutamyl_tRNA_reduct	cd05213	NADP-binding domain of glutamyl-tRNA reductase; Glutamyl-tRNA reductase catalyzes the ...	5-99	5.13e-04

Figure 49: Conserved Domain Database output results for gene BMW92_RS10830. The top domain hit match was PRK11880: pyrroline-5-carboxylate reductase which aligned with the query sequence from amino acid residues 5-270 and had statistically significant e-value of 3.34e-117. The second domain hit match was ProC: Pyrroline-5-carboxylate reductase which aligned with the query sequence from amino acid residues 3-271 and had a statistically significant e-value of 8.09e-100. The third domain hit match was proC: pyrroline-5-carboxylate reductase which aligned with the query sequence from amino acid residues 24-267 and had a statistically significant e-value of 7.34e-83. The fourth domain hit match was P5CR dimer: pyrroline-5-carboxylate reductase which aligned with the query sequence from amino acid residues 164-268 and had a statistically significant e-value of 1.85e-44. The last domain hit match was NAD_bind_Glutamyl_tRNA_reduct: NADP-binding domain of glutamyl-tRNA reductase which

aligned with the query sequence from amino acid residues 5-99 and had a statistically non-significant e-value of 5.13e-04 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

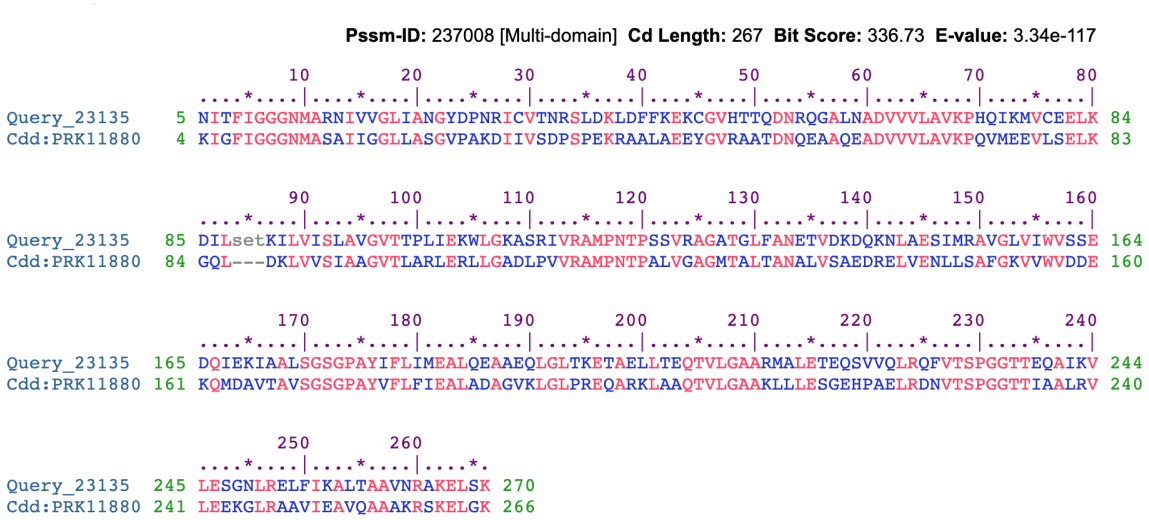


Figure 50: The top domain hit sequence PRK11880: pyrroline-5-carboxylate reductase aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 5-270 and had statistically significant e-value of 3.34e-117 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

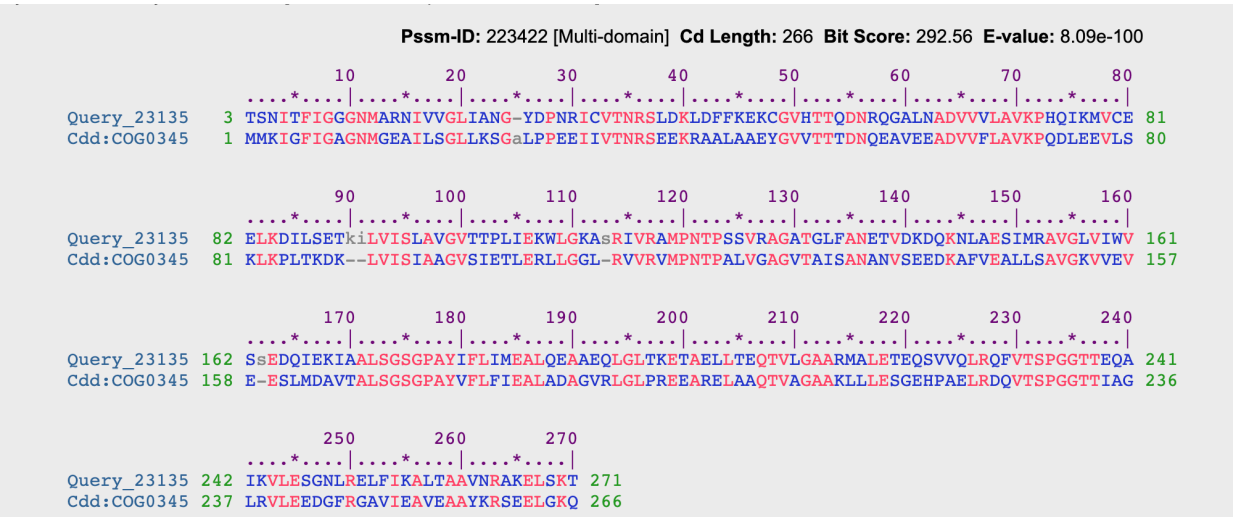


Figure 51: The second domain hit sequence ProC: Pyrroline-5-carboxylate reductase aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 3-271 and had statistically significant e-value of 8.09e-100 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

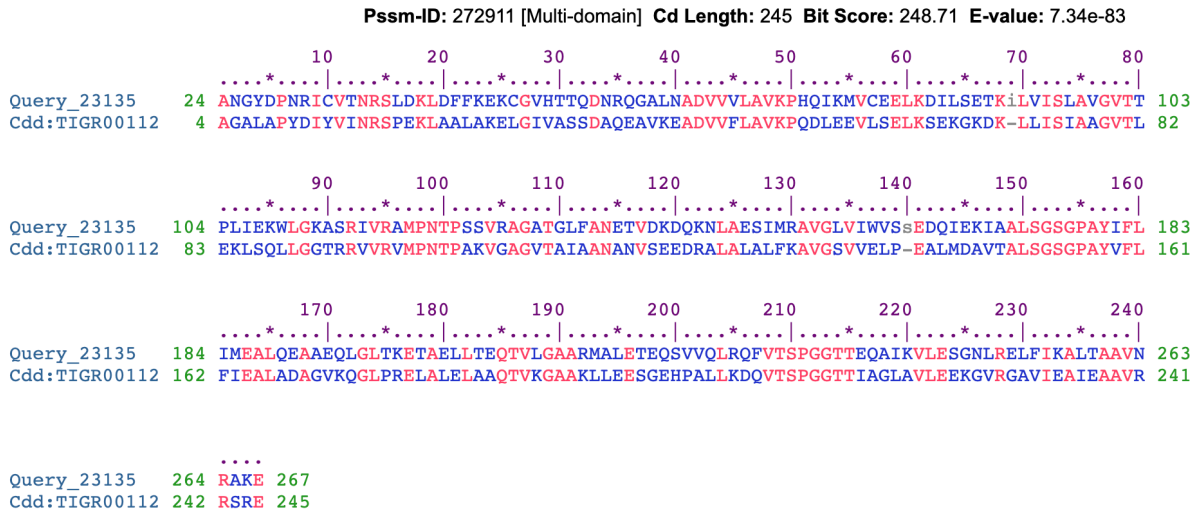


Figure 52: The third domain hit sequence proC: pyrroline-5-carboxylate reductase aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 24-267 and had statistically significant e-value of 7.34e-83 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

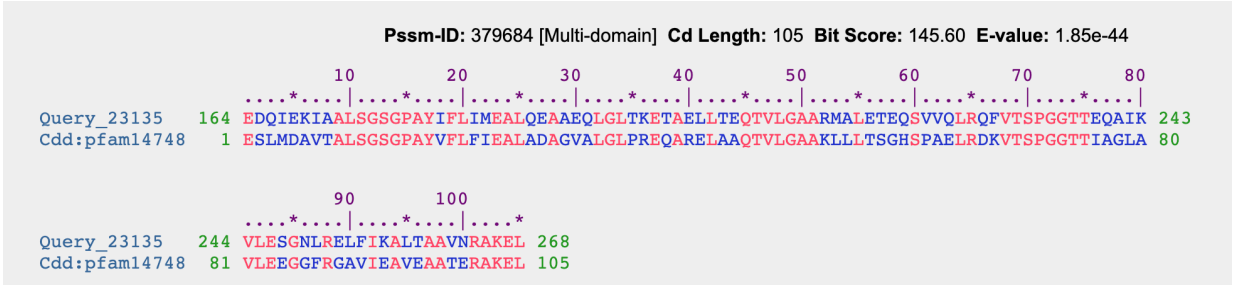


Figure 53: The fourth domain hit sequence P5CR dimer: pyrroline-5-carboxylate reductase aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 164-268 and had statistically significant e-value of 1.85e-44. (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

MUSCLE

```
C.burnetii          MNTSNITFIGGGNMARNIVVGLIANGYDPNRICVTNRSLDKLDFFKEKCGVHTTQDNRQG
C.mudrowiae         MRIANITFIGGGNMACNIVVGLLANGYDSNRICVTNPTSDKLTFFREKCKVRTTQNNREG
A.lusitana          MHTPVISIIGAGNMGSSLIGGLIKDGHPSDKLWAADPSGEKLTQLKTKFDINTTSDNAQA
N.mobilis           MAEESITFIGGGNMAYSLVGGLIADGYRAERVHVADPDPAKRMDLANRFRIHVHEDNRKA
A.ehrlichii         MSNNTLCFIGGGNMARSLIGGLLADGFDPQAVRVADPDAGKRDDLANRFGVRVYADNLEA
A.mobile            MTMKTLCFIGGGNMARSLIGGLLTDGYDPQAIRVAEPDAGKREDLANRFGVRVHEDNLEA
T.denitrificans     MEQGIISFIGGGNMCSSLVGGLIADGYAPERIRVSDPGEETLASLRARFGVHTTHDNREA
N.halophilus        MNEKTLAFIGGGNMATSLIGGLIADGRNAQTIWVADPDRSKLDALHHRFSVNTTPDNLQA
N.watsonii          MSEQTLAFIGGGNMAASLIGGLVADGRDAQAIWVADPDRRKLDALHERFGVNTAPDNIQV
C.bacterium         MKDVNIAFIGGGNMATSLIGGLLADHVSPARLCVADRDPAQREHLAAQFGVRTSEDNAAC
H.archaeon          MNDASLTFIGGGNMAASLVGGLIADGWDPARIRVADPDAGRRERMAARHQVSTTPDNQAA
                    *    : :**.***  .:: **: :   .  : .::        :  .  : .  :*

C.burnetii          ALNADVVVLAVKPHQIKMVCEELKDILSETKILVISLAVGVTTPLIEKWLGKA-SRIVRA
C.mudrowiae         ATNADAIILAVKPNQVKGVCEELKDIVNTLHPLIISVAVGVRVKLLQKWLQSE-PAIVRA
A.lusitana          IQAADTIIFAVKPQAFAHVALPLKKVIAERKPLVISIAAGIREASIQQWLNGK-TPIVRA
N.mobilis           VLRAAAVVLAVKPQIIKSVLEPLGPILREQKSLVISIAAGVREPDISRWLGGQ-IAVVRT
A.ehrlichii         AADADTVILAVKPQVVRTACEQLVAGSGDAGRLFISIAAGVREPDLTRWLGGQ-AAVVRT
A.mobile            AANAQAVILAVKPQVIRPVCEQLAGAEAGKGRVYISIAAGVREPDLTRWLGGS-AAVVRT
T.denitrificans     AAGAGVVVLAVKPQVLPKVAAELAPVVQEHGTLVVSIAAGIRTTDLQRWLGAG-VALVRT
N.halophilus        AQEAEVVVLAVKPQQLRTVATGLKSVVTSSQPLWLTIAAGIRIPDLERWLGGP-APIVRA
N.watsonii          AQDAAIIVLAVKPQQLRSVVTQLKNVATLSQPLWLTIAAGIGTPDVEAWLGGP-APIVRA
C.bacterium         AEDADVIVLAVKPQVLHEVCEALTDSVQRKQPLVVSVAAGVRTDSLRRWLGGGDVAIVRA
H.archaeon          VSDADVVVLAVKPQVMAAVTQELAAGIAQQQPLVISIAAGIRESTLRDWLGAD-TAIVRT
                       *  :::****: .  .   *         : :::*.*:    :  **      :**:

C.burnetii          MPNTPSSVRAGATGLFANETVDKDQKNLAESIMRAVGLVIWVSSEDQIEKIAALSGSGPA
C.mudrowiae         MPNTPASVGAGATALFANEKATKEQRNLAESILRAVGLVVWLSLEDQIDEVAALSGSGPA
A.lusitana          MPNTPALIGCGATALYANPYVTESQRNLAESILRAVGVVVWLNDEKLMDTVTALSGSGPA
N.mobilis           MPNTPALVRAGATALYANEYVSQNQRDLAESLLRAVGIIQWLDDETLLDIVTALSGSGPA
A.ehrlichii         MPNTPSLVGTGATALYANDRVKERQRELAESLMRAVGLVVWLDDEAQMDTVTAVSGSGPA
A.mobile            MPNTPSLVGTGATALYANPQVSEPQRELAESLMRAVGLVVWLDDETQMDTVTAVSGSGPA
T.denitrificans     MPNTPALVKSGATALFATAAVTAAQRDQAESVLRAVGLTLWLENEEQMDAVTALSGSGPA
N.halophilus        MPNTPALVQAGATALFANAQTNPQQRQMAESVLRAVGLTLWLKDENLMEVVTALSGSGPA
N.watsonii          MPNTPALVQAGATALFANPHTSPNQRQTAESVLRAVGLTLWLNDENLMEVVTALSGGGPA
C.bacterium         MPNTPALLQSGATGLYACTGVSEEQRDLAEAILRATGLTLWVDDEAQMDIVTALSGSGPA
H.archaeon          MPNTPALVQSGATALYANTAVSDGQRSLAESILRAVGLVIWVEDEAQMDAVTALSGSGPA
                    *****: :  ***.*:*   .   *.. **:::**.*:  *:. *  :: ::*:**.***

C.burnetii          YIFLIMEALQEAAEQLGLTKETAELLTEQTVLGAARMALETEQSVVQLRQFVTSPGGTTE
C.mudrowiae         YIFFVMEALQEAGEGLGLPKETVQLLTAQTVWGAARMSLEAEEDLVELRRFVTSPGGTTE
A.lusitana          YFFLMMEALQEAAEDLGLPTETARLLTLQTALGAARMAIESGTSLAELRRKVTSPGGTTE
N.mobilis           YFFLLMETLEAAAIELGLPEQTARLLTLETALGAARMALESDEDPGRLRLRVTSPGGTTE
A.ehrlichii         YFFLLMEAIEDAARDLGLPGETARLLTIETALGAAKMALESDESPAQLRQRVTSPGGTTE
A.mobile            YFFLLMEAIEEAAREQGLPAETARLLTIETALGAAKMALESDESPGQLRQRVTSPGGTTE
T.denitrificans     YFFLVMEAMQGAAQAIGLPERTARLLTLQTAFGAAKMALESDEEPSLLRQRVTSPGGTTE
N.halophilus        YFFLVMEAMEKAAIDLGLDDSTARLLTLETAFGAAKMALESEEDSIRLRQRVTSPGGTTE
N.watsonii          YFFLVMEAMEKAAIDLGLESNTARLLTLETAFGAAKMALKSEEGCASLRQRVTSPGGTTE
C.bacterium         YFFRVMEGLEKAATELGLPAQTARLLTLQTALGAARMALESSEPVATLRKRVTSPGGTTE
H.archaeon          YFFFFMEALQAAGEELGLPAGTARLLALQTAFGAARMALESSDDAATLRHHVTSPGGTTE
                    *:* .** :: *.   **   *. **: :*. ***.*::::      **  *********
```

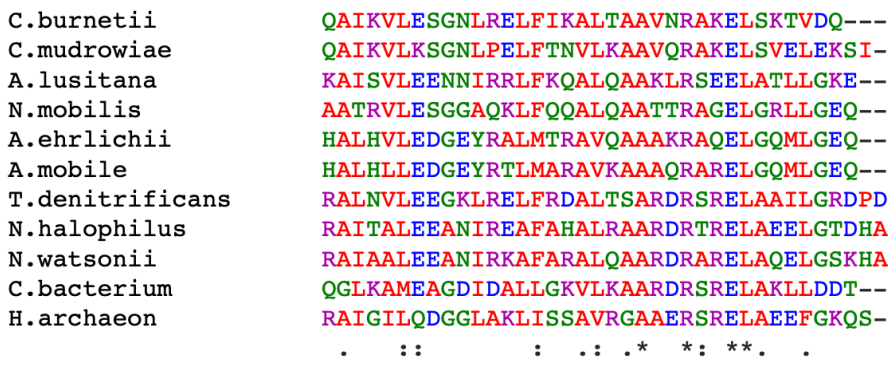


Figure 54: MUSCLE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10830 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered randomly and are listed from top to bottom as followed: *C. burnetii*, *C. mudrowiae*, *A. lusitana*, *N. mobilis*, *A. ehrlichii, A. mobile*, *T. denitrificans*, *N. halophilus*, *N. watsonii*, *C. bacterium*, *H. archaeon*. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

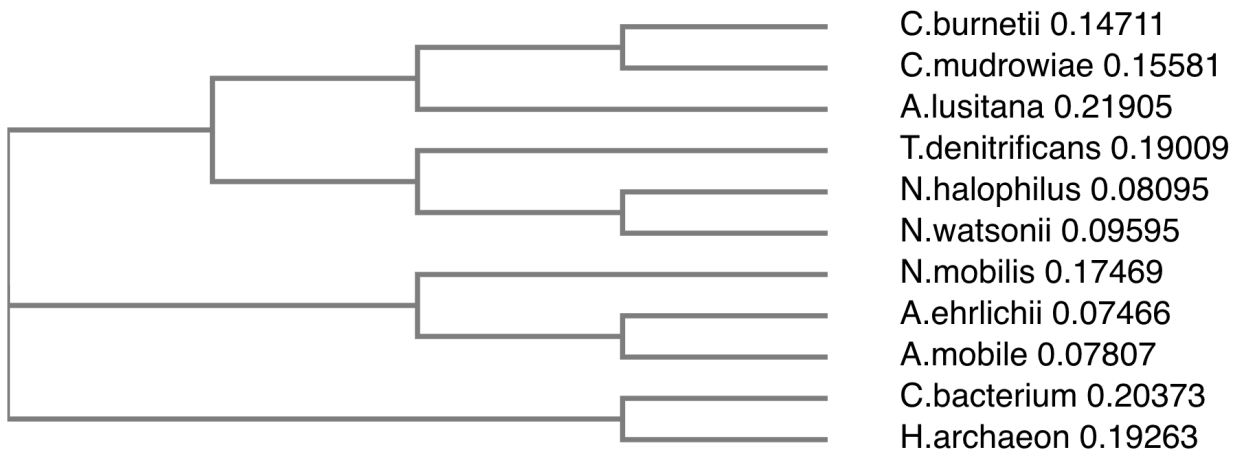


Figure 55: MUSCLE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10830 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *C. burnetii*, *C. mudrowiae*, *A. lusitana*, *N. mobilis*, *A. ehrlichii, A. mobile*, *T. denitrificans*, *N. halophilus*, *N. watsonii*, *C. bacterium*, *H. archaeon*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

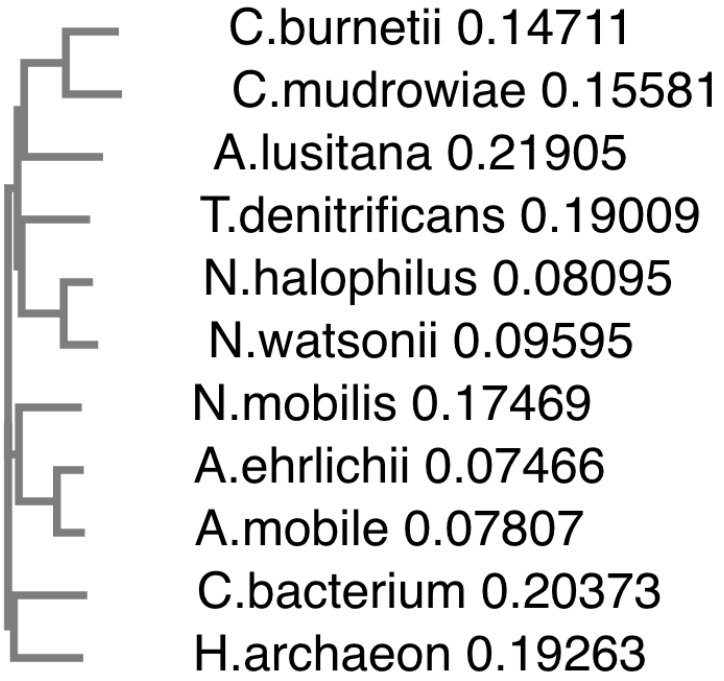
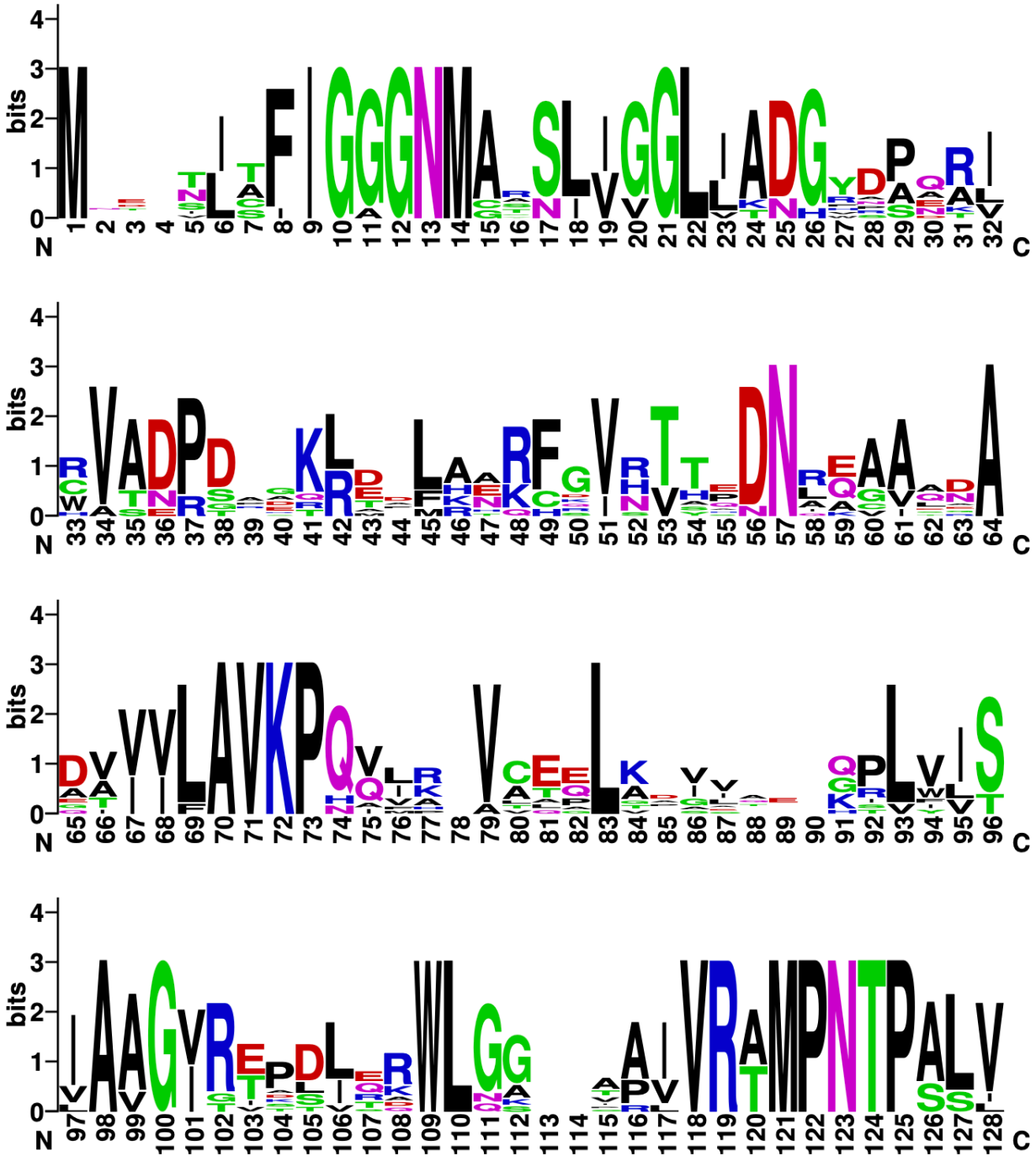


Figure 56: MUSCLE multiple sequence alignment real phylogenetic tree for *Coxiella burnetii* gene BMW92_RS10830 and the top 10 organism sequences similarities selected. Organisms sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *C. burnetii*, *C. mudrowiae*, *A. lusitana*, *N. mobilis*, *A. ehrlichii, A. mobile*, *T. denitrificans*, *N. halophilus*, *N. watsonii*, *C. bacterium*, *H. archaeon*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

MUSCLE Sequence Logo



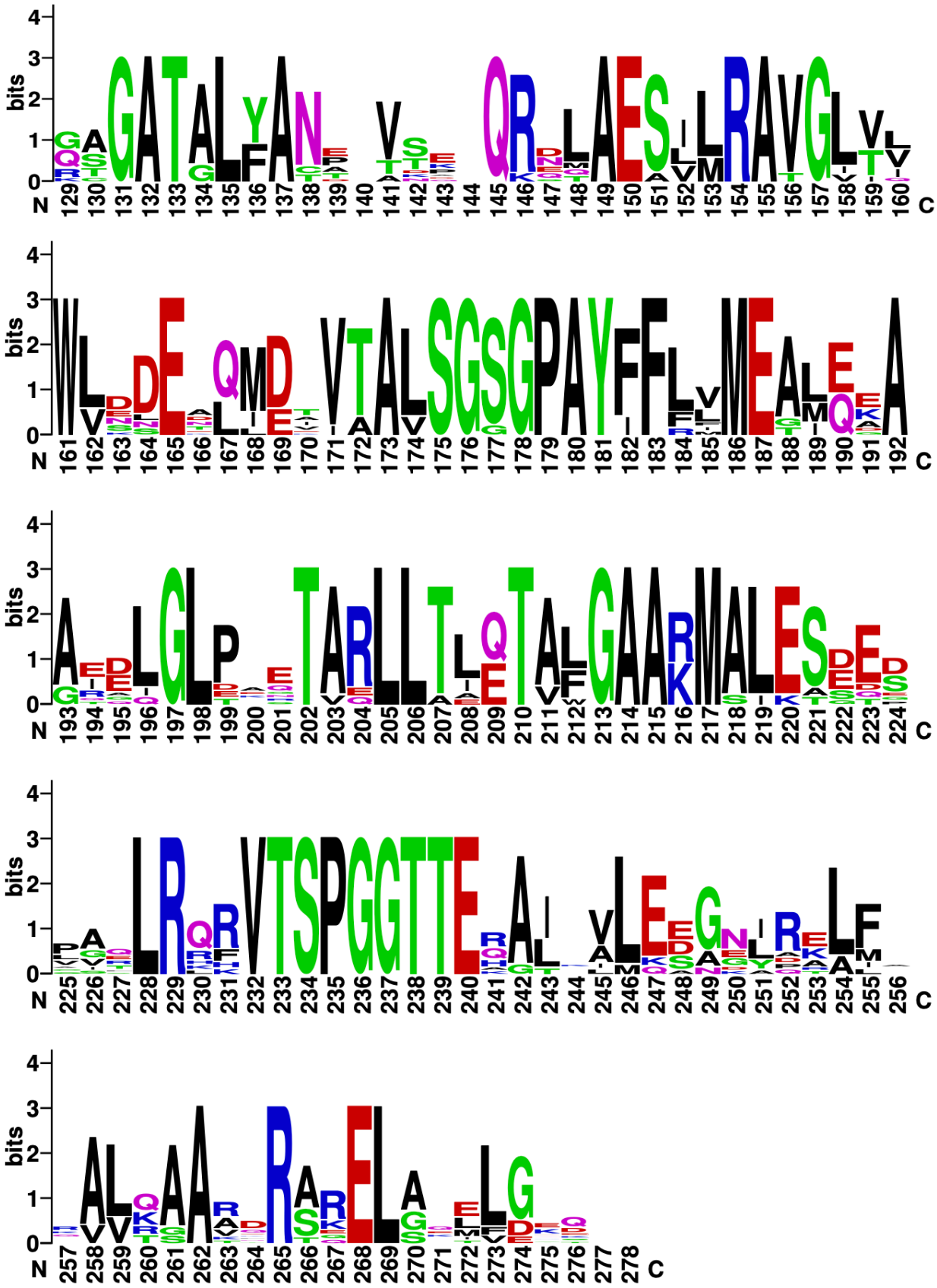

bits
4
3
2
1
0
N
C

Figure 57: Sequence logo generated from the MUSCLE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10830 and the top 10 organism sequences similarities selected. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. The conservation of each amino acid among residue across similar sequences is represented by the height of each single letter abbreviation of the amino acid. (WebLogo, <https://weblogo.berkeley.edu/logo.cgi>).

T-COFFEE

```
A.ehrlichii       MSNNTLCFIGGGNMARSLIGGLLADGFDPQAVRVADPDAGKRDDLANRFG
A.lusitana        MHTPVISIIGAGNMGSSLIGGLIKDGHPSDKLWAADPSGEKLTQLKTKFD
A.mobile          MTMKTLCFIGGGNMARSLIGGLLTDGYDPQAIRVAEPDAGKREDLANRFG
C.bacterium       MKDVNIAFIGGGNMATSLIGGLLADHVSPARLCVADRDPAQREHLAAQFG
C.burnetii        MNTSNITFIGGGNMARNIVVGLIANGYDPNRICVTNRSLDKLDFFKEKCG
C.mudrowiae       MRIANITFIGGGNMACNIVVGLLANGYDSNRICVTNPTSDKLTFFREKCK
H.archaeon        MNDASLTFIGGGNMAASLVGGLIADGWDPARIRVADPDAGRRERMAARHQ
N.halophilus      MNEKTLAFIGGGNMATSLIGGLIADGRNAQTIWVADPDRSKLDALHHRFS
N.mobilis         MAEESITFIGGGNMAYSLVGGLIADGYRAERVHVADPDPAKRMDLANRFR
N.watsonii        MSEQTLAFIGGGNMAASLIGGLVADGRDAQAIWVADPDRRKLDALHERFG
T.denitrificans   MEQGIISFIGGGNMCSSLVGGLIADGYAPERIRVSDPGEETLASLRARFG
                  *    : :**.***  .:: **: :   .  : .::       :  :

A.ehrlichii       VRVYADNLEAAADADTVILAVKPQVVRTACEQLVAGSGDAGRLFISIAAG
A.lusitana        INTTSDNAQAIQAADTIIFAVKPQAFAHVALPLKKVIAERKPLVISIAAG
A.mobile          VRVHEDNLEAAANAQAVILAVKPQVIRPVCEQLAGAEAGKGRVYISIAAG
C.bacterium       VRTSEDNAACAEDADVIVLAVKPQVLHEVCEALTDSVQRKQPLVVSVAAG
C.burnetii        VHTTQDNRQGALNADVVVLAVKPHQIKMVCEELKDILSETKILVISLAVG
C.mudrowiae       VRTTQNNREGATNADAIILAVKPNQVKGVCEELKDIVNTLHPLIISVAVG
H.archaeon        VSTTPDNQAAVSDADVVVLAVKPQVMAAVTQELAAGIAQQQPLVISIAAG
N.halophilus      VNTTPDNLQAAQEAEVVVLAVKPQQLRTVATGLKSVVTSSQPLWLTIAAG
N.mobilis         IHVHEDNRKAVLRAAAVVLAVKPQIIKSVLEPLGPILREQKSLVISIAAG
N.watsonii        VNTAPDNIQVAQDAAIIVLAVKPQQLRSVVTQLKNVATLSQPLWLTIAAG
T.denitrificans   VHTTHDNREAAAGAGVVVLAVKPQVLPKVAAELAPVVQEHGTLVVSIAAG
                  : .  :*      *  :::****: .  .   *         : :::*.*

A.ehrlichii       VREPDLTRWLGG-QAAVVRTMPNTPSLVGTGATALYANDRVKERQRELAE
A.lusitana        IREASIQQWLNG-KTPIVRAMPNTPALIGCGATALYANPYVTESQRNLAE
A.mobile          VREPDLTRWLGG-SAAVVRTMPNTPSLVGTGATALYANPQVSEPQRELAE
C.bacterium       VRTDSLRRWLGGGDVAIVRAMPNTPALLQSGATGLYACTGVSEEQRDLAE
C.burnetii        VTTPLIEKWLGK-ASRIVRAMPNTPSSVRAGATGLFANETVDKDQKNLAE
C.mudrowiae       VRVKLLQKWLQS-EPAIVRAMPNTPASVGAGATALFANEKATKEQRNLAE
H.archaeon        IRESTLRDWLGA-DTAIVRTMPNTPALVQSGATALYANTAVSDGQRSLAE
N.halophilus      IRIPDLERWLGG-PAPIVRAMPNTPALVQAGATALFANAQTNPQQRQMAE
N.mobilis         VREPDISRWLGG-QIAVVRTMPNTPALVRAGATALYANEYVSQNQRDLAE
N.watsonii        IGTPDVEAWLGG-PAPIVRAMPNTPALVQAGATALFANPHTSPNQRQTAE
T.denitrificans   IRTTDLQRWLGA-GVALVRTMPNTPALVKSGATALFATAAVTAAQRDQAE
                  :    :  **      :**:*****: :  ***.*:*   .   *:. **
```

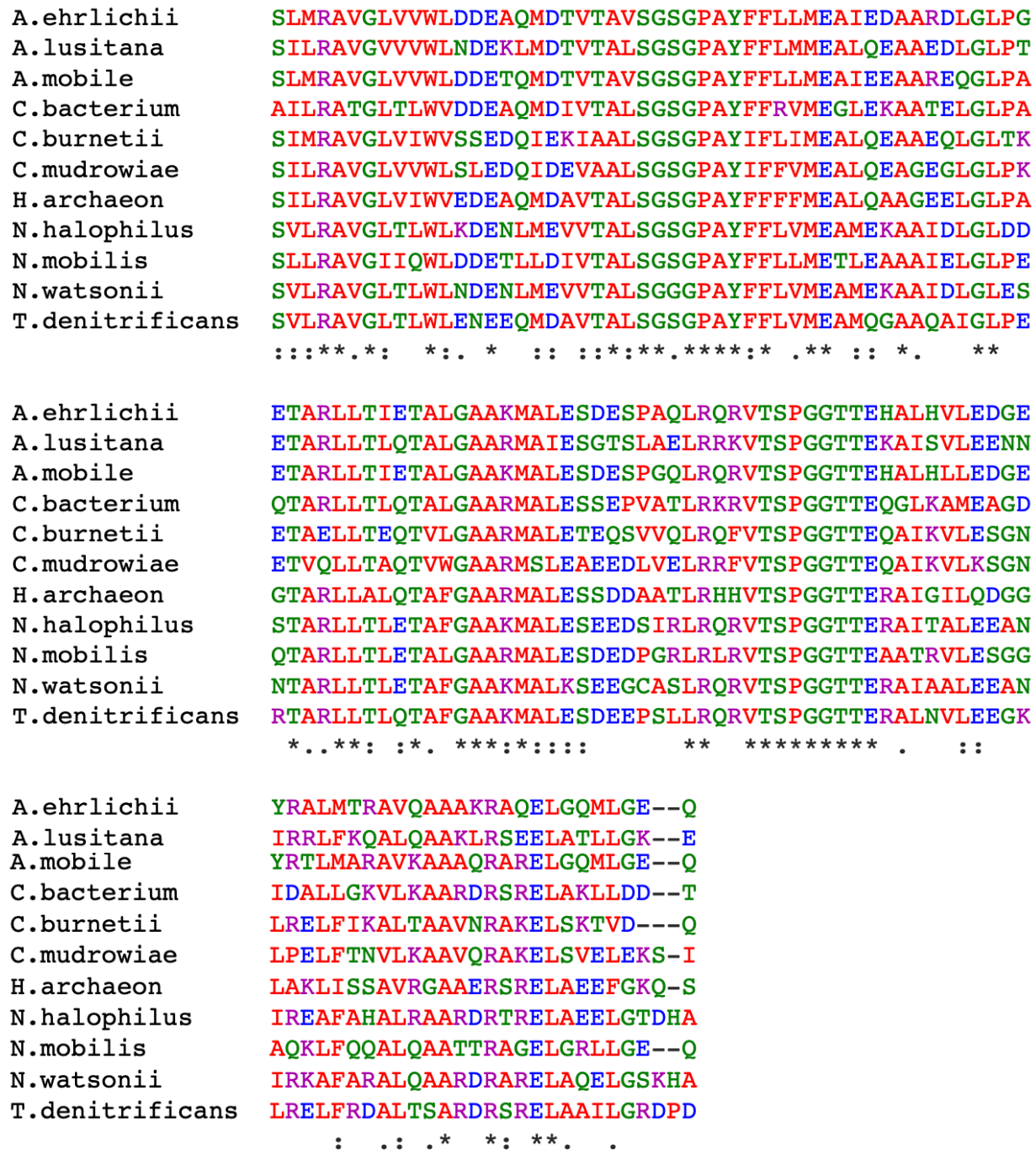


Figure 58: T-COFFEE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10830 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity

originated. Organism sequences were ordered randomly and are listed from top to bottom as followed: *A. ehrlichii, A. lusitana*, *A. mobile*, *C. bacterium*, *C. burnetii*, *C. mudrowiae*, *H. archaeon*, *N. halophilus*, *N. mobilis*, *N. watsonii*, *T. denitrificans*. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

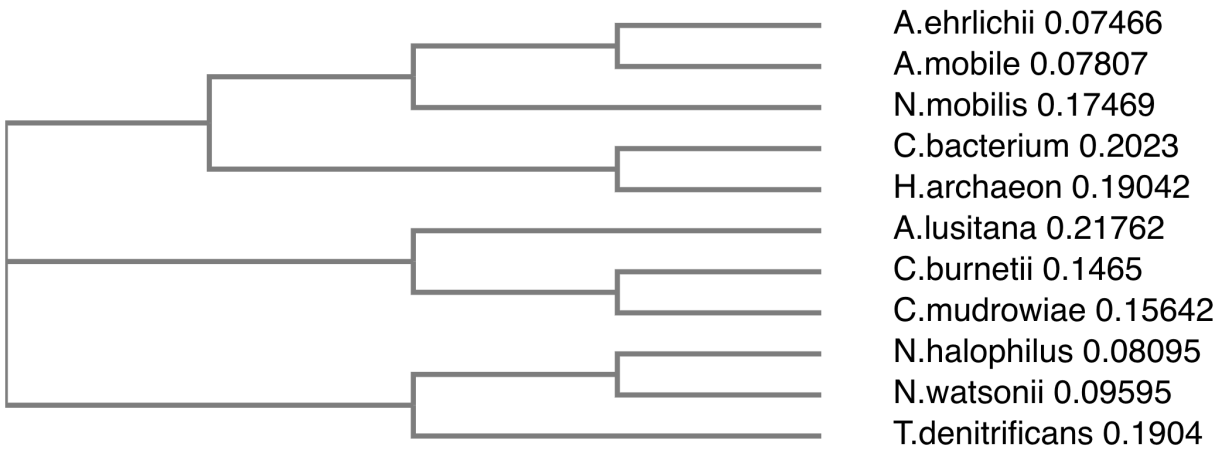


Figure 59: T-COFFEE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10830 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *A. ehrlichii, A. mobile, N. mobilis, C. bacterium, H. archaeon, A. lusitana*, *C. burnetii*, *C. mudrowiae*, *N. halophilus*, *N. watsonii*, *T. denitrificans*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

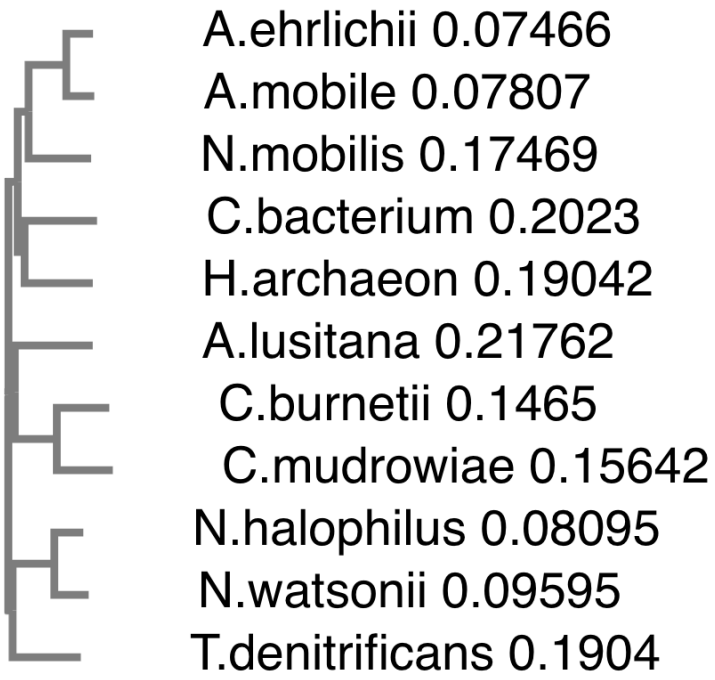
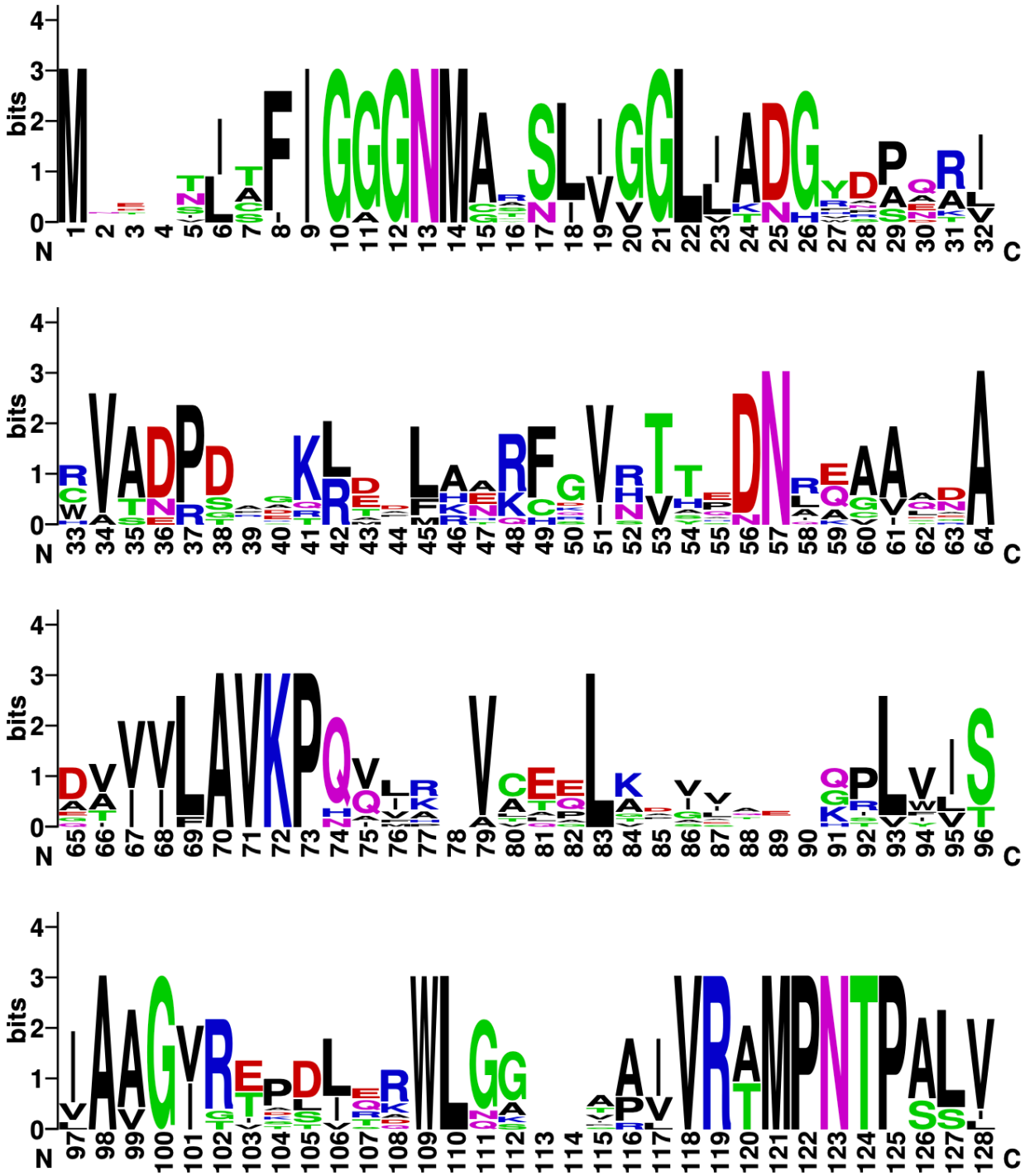


Figure 60: T-COFFEE multiple sequence alignment real phylogenetic tree for *Coxiella burnetii* gene BMW92_RS10830 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *A. ehrlichii, A. mobile, N. mobilis, C. bacterium, H. archaeon, A. lusitana*, *C. burnetii*, *C. mudrowiae*, *N. halophilus*, *N. watsonii*, *T. denitrificans*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

T-COFFEE Sequence Logo



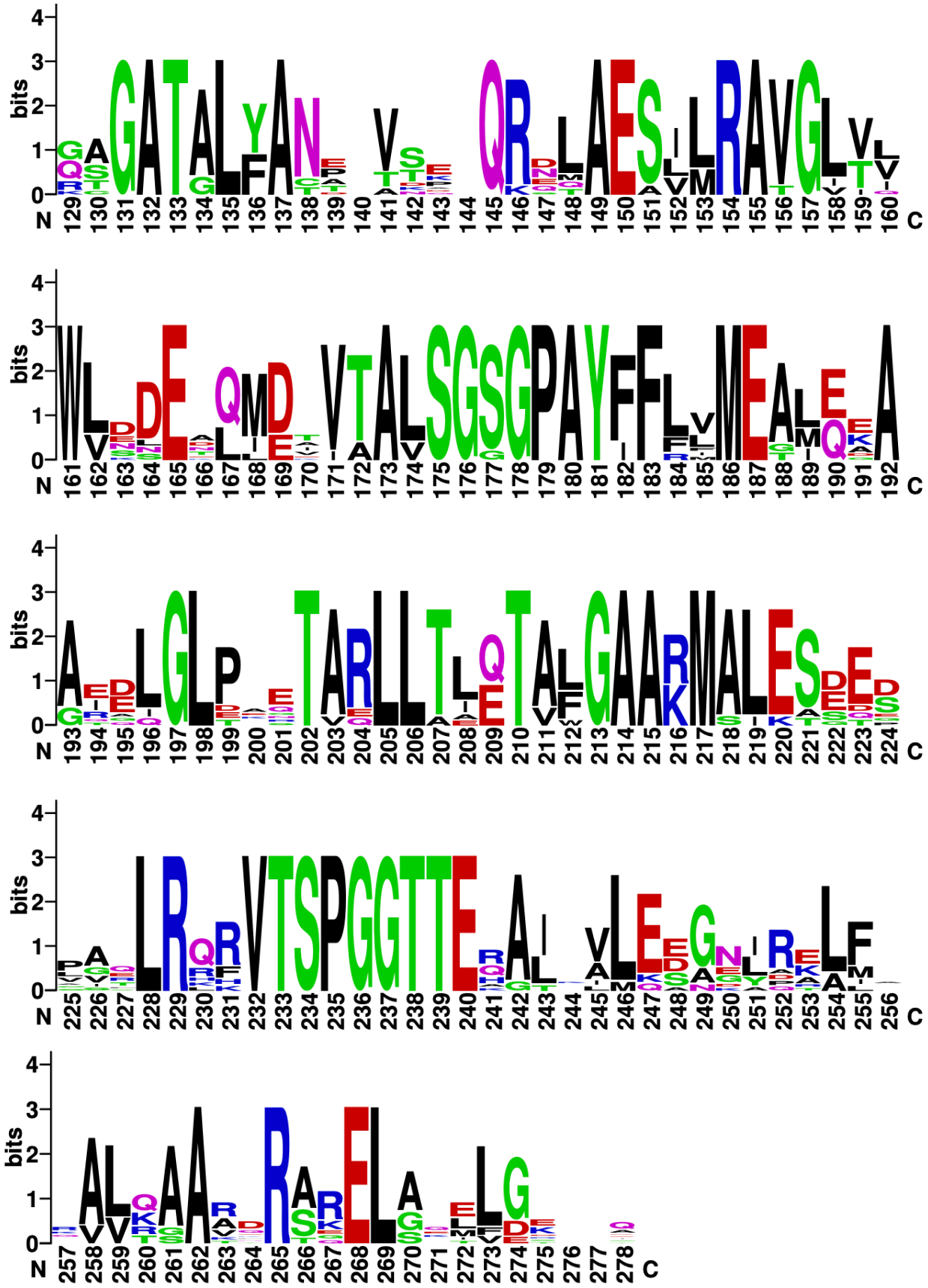
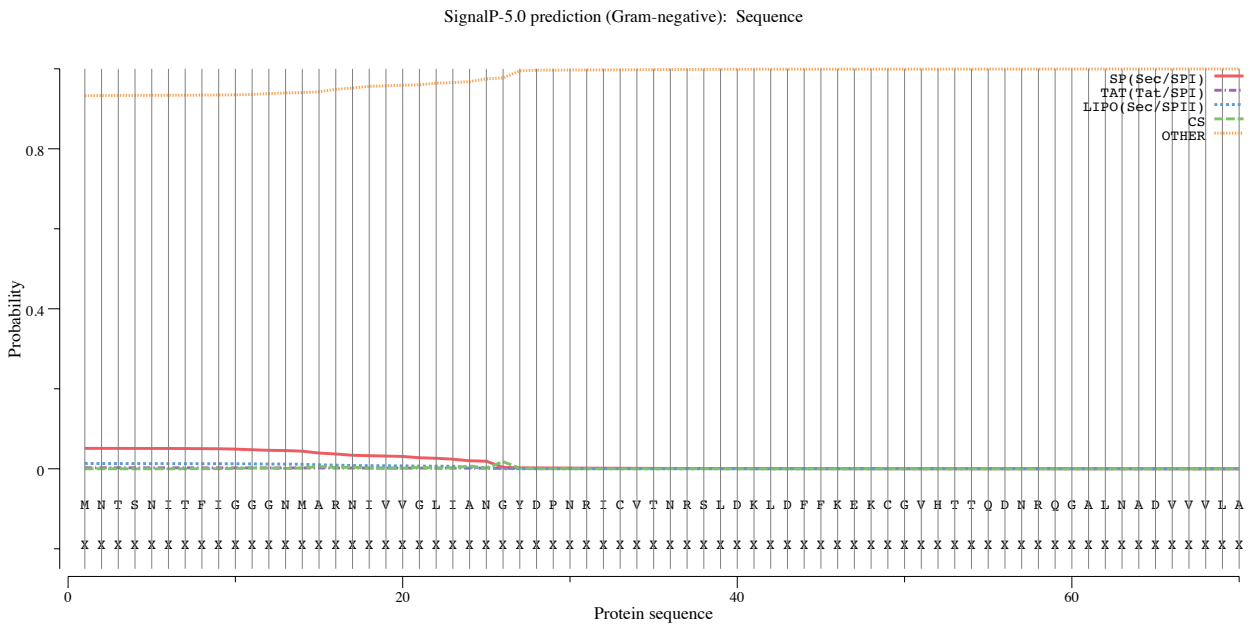

bits
4
3
2
1
0
N
C

Figure 61: Sequence logo generated from the T-COFFEE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10830 and the top 10 organism sequences similarities selected. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. The conservation of each amino acid among residue across similar sequences is represented by the height of each single letter abbreviation of the amino acid. (WebLogo, <https://weblogo.berkeley.edu/logo.cgi>).

Protein Localization

SignalP



Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.0515	0.0022	0.0128	0.9336

Figure 62: SignalP 5.0 prediction (Gram-negative) for gene BMW92_RS10830 of *Coxiella burnetii*. The TAT (Tat/SPI), LIPO (Sec/SPII), and CS probability scores were all below 0.0128 (1.28%) which results in the likelihood of the protein being a signal peptide as highly unlikely and can confirm there is no signal peptide of these protein types. The SP (Sec/SPI) signal peptide probability score was 0.0515 (5.15%). This probability score results in the likelihood of the SP signal peptide as being highly unlikely. The program calculated the probability scores for OTHER as 0.9336 (93.36%). This probability score indicates the protein from gene

BMW92_RS10830 has another protein classification that is not related to similar function or type as a signal peptide (SignalP, <http://www.cbs.dtu.dk/services/SignalP/ >).

LIPOP

```
# Sequence CYT score=-0.200913
# Cut-off=-3
Sequence        LipoP1.0:Best   CYT     1       1       -0.200913



# NO PLOT made - less than 4 putative cleavage sites predicted
```

Figure 63: LipoP 1.0 was unable to generate a plot graph due to there being less than four predicted putative cleavage sites. The best localization prediction resulted in the highest scoring class being the cytoplasmic protein class (LipoP, <http://www.cbs.dtu.dk/services/LipoP/>).

TMHMM

```
# WEBSEQUENCE Length: 274
# WEBSEQUENCE Number of predicted TMHs:  0
# WEBSEQUENCE Exp number of AAs in TMHs: 0.59771
# WEBSEQUENCE Exp number, first 60 AAs:  0.37368
# WEBSEQUENCE Total prob of N-in:        0.05194
WEBSEQUENCE     TMHMM2.0        outside      1   274
```

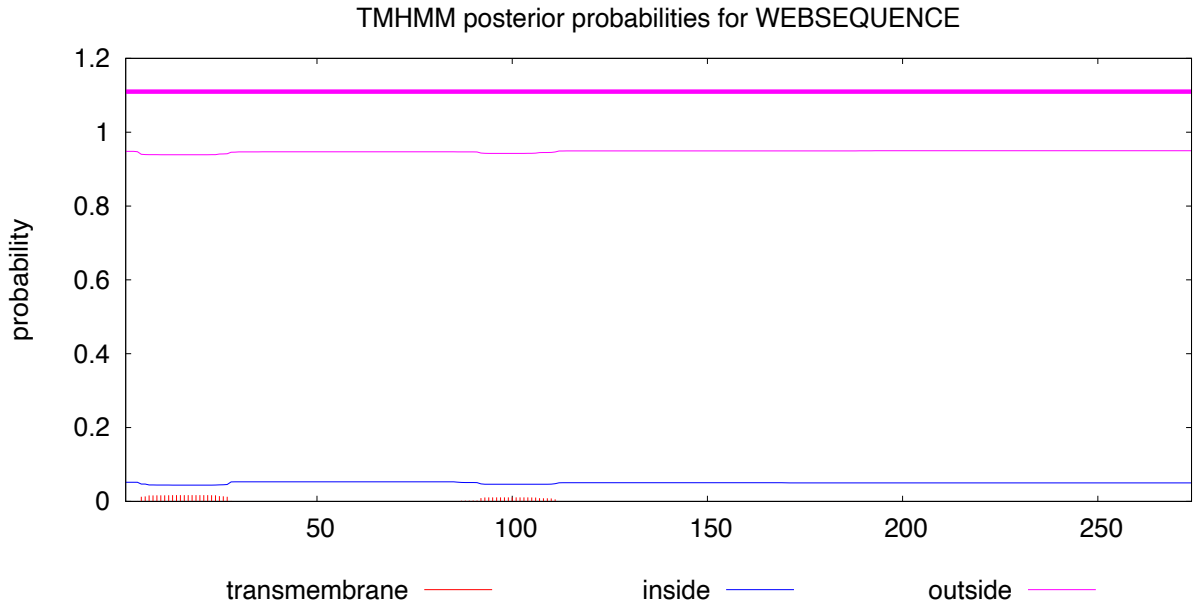


Figure 64: TMHMM posterior probability displayed a line graph that predicts the localization of the protein coded from BMW92_RS10830 as entirely outside the membrane. The red line, representative of the protein being located in the transmembrane, was less than 0.005 (.50% probability) across the entirety of the line graph. This is indicative of the protein being located within the transmembrane as highly unlikely. The blue line, representative of the protein being located inside the membrane, was at 0.05 (5% probability). This is indicative of the protein being located inside of the membrane as highly unlikely. The magenta line, representative of the

protein being located outside the membrane, was at 0.95 (95% probability). This is indicative of the protein being located outside of the membrane as highly likely (TMHMM, <http://www.cbs.dtu.dk/services/TMHMM/ >).

BOMP

The total number of valid proteins submitted is: 1
The total number of integral ß-barrel outer membrane proteins predicted is: 0

Sequence name	Category	Best BLAST hit

Figure 65: The BOMP test result identified there are no integral beta-barrel outer membrane proteins for gene BMW92_RS10830 (BOMP, <http://services.cbu.uib.no/tools/bomp>).

PSORTb

```
SeqID: C.burnetii
  Analysis Report:
    CMSVM-            Unknown                    [No details]
    CytoSVM-          Unknown                    [No details]
    ECSVM-            Unknown                    [No details]
    ModHMM-           Unknown                    [No internal helices found]
    Motif-            Unknown                    [No motifs found]
    OMPMotif-         Unknown                    [No motifs found]
    OMSVM-            Unknown                    [No details]
    PPSVM-            Unknown                    [No details]
    Profile-          Unknown                    [No matches to profiles found]
    SCL-BLAST-        Cytoplasmic                [matched 71159808: Pyrroline-5-carboxylate reductase]
    SCL-BLASTe-       Unknown                    [No matches against database]
    Signal-           Unknown                    [No signal peptide detected]
  Localization Scores:
    Cytoplasmic           9.26
    CytoplasmicMembrane   0.24
    Periplasmic           0.48
    OuterMembrane         0.01
    Extracellular         0.01
  Final Prediction:
    Cytoplasmic           9.26
```

Figure 66: The PSORTb test resulted in an analysis report that identified no detectable internal helices, motifs, or signal peptides. The PSORTb localization scores resulted in a 9.26 value for the cytoplasmic location. The localization score for cytoplasmic membrane was 0.24. The localization score for periplasmic was 0.48. The localization score for the outer membrane location was 0.01. The localization score for the extracellular location was 0.01. The calculated localization scores for gene BMW92_RS10830 resulted in the final predictable location of the protein to be cytoplasmic (PSORTb, <https://www.psort.org/psortb/ >).

PHOBIUS

```
ID   UNNAMED
FT   TOPO_DOM      1     274       NON CYTOPLASMIC.
//
```

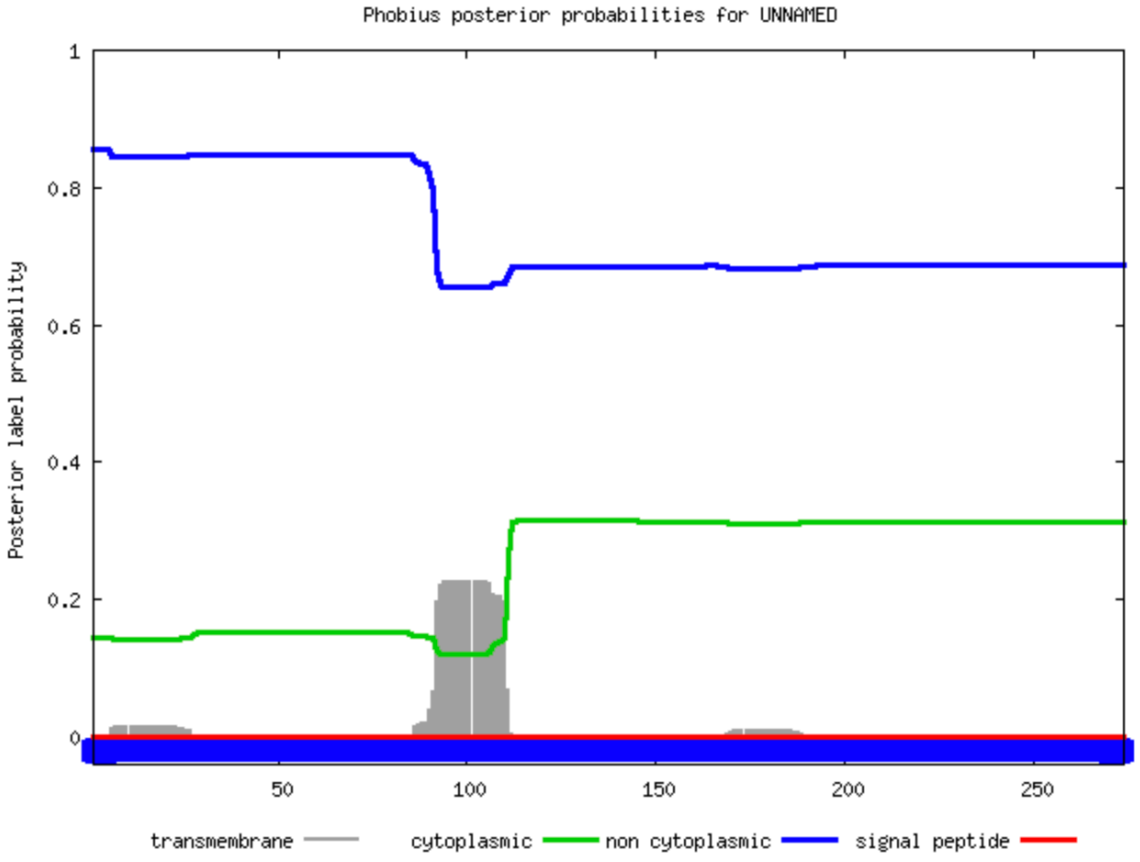


Figure 67: The Phobius posterior probability line graph generated for gene BMW92_RS10830 resulted in a calculated prediction that the whole sequence contains no membrane helices. The grey line, representative of the predicted transmembrane helices location, was around 0.23 (23%) posterior probability from amino acids 91-117. The green line, representative of the predicted cytoplasmic transmembrane helices location, was around 0.15 (15%) posterior probability from amino acids 0-120; this changed to a posterior probability of 0.32 (32%) from amino acids 121-273. The blue line, representative of the predicted non-cytoplasmic transmembrane helices

location, was around 0.83 (83%) posterior probability from amino acids 0-91; the posterior probability changed to 0.68 (68%) from amino acids 121-273. The red line, representative of the presence or absence of a signal peptide, was 0.00 (0%) posterior probability (Phobius, <http://phobius.sbc.su.se>).

BMW92_RS10835

The third gene, BMW92_RS10835, was analyzed using bioinformatic technology. Table 3 below contains the provided data regarding basic information. A protein isoelectric point calculator was used to determine the isoelectric point of the protein, protein length, and the number and prevalence of each amino acid that makes up the protein (Figure 68). The BLASTp search tool produced 100 matches ranked from highest sequence similarity to lowest sequence similarity. The top ten sequences with significant alignments that were not identical species to *Coxiella burnetii* were selected. The information recorded included the organism name, protein name, percent identity, precent positive, length of alignment match, e-values, and percent gap. The highest ranked match to the BMW92_RS10835 gene was pyridoxal phosphate-dependent enzyme [*Rhipicephalus microplus*] (Figure 69). The remaining nine matches to the BMW92_RS10835 gene all had a function as pyridoxal phosphate-dependent enzyme (Figures 70-78). The CDD identified four potential protein domains hits conserved (Figure 79). Four of the domain hits conserved and identified by the CDD belong to the PLPDE_III superfamily (Figure 80). Specific domain hits involved the PLPD_III_Yggs_like, YggS, and Ala_racemase_N. One domain hit conserved and identified as a non-specific domain hit was TIGR00044. The protein classification identified by the CDD was pyridoxal phosphate-dependent enzyme. Four of the domain hits sequences were aligned with the query sequence based off the amino acids that are highly conserved between both sequences (Figures 81-84). The MUSCLE program generated a multiple sequence alignment (MSA); each amino acid in the sequence was assigned a distinct color to distinguish the amino acids being compared (Figure 85). The MUSCLE program generated two phylogenetic trees using the multiple sequence alignments to further confirm sequence similarity. The results displayed the numbers followed

behind each organism at the end of each leaf node which displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii.* The use of a phylogenetic cladogram (Figure 86) and real phylogenetic tree (Figure 87) provided further understanding of the relatedness of common ancestors and organism sequences that are conserved. Each of the letter's heights produced correspond to the conservation of the amino acid residue across similar sequences. WebLogo produced a sequence logo that was generated from the MUSCLE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10835 and the top 10 organism sequences similarities selected (Figure 88). Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. Each of the single letter amino acid abbreviation's heights correspond to the conservation of the amino acid residue across similar sequences. The T-COFEE program generated another multiple sequence alignment to further confirm sequence similarity depicted with in the MUSCLE MSA (Figure 89). The T-COFFEE program generated two phylogenetic trees, phylogenetic cladogram (Figure 90) and real phylogenetic tree (Figure 91), using the multiple sequence alignment which displayed the genetic proximity and similarity between *Coxiella burnetii* and selected organisms from the BLASTp search. WebLogo constructed a sequence logo from the T-COFFEE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10835 and the top 10 organism sequences similarities selected to further display sequence similarity and conservation of sequences. Each of the single letter amino acid abbreviation's heights correspond to the conservation of the amino acid residue across similar sequences (Figure 92). Protein localization results included SignalP, LipoP, TMHMM, BOMP, PSORTb, and Phobius. The SignalP graphical illustration identified that there is no presence of a signal peptide for the entirety of the protein sequence (Figure 93). The LipoP resulted in the

highest scoring class being the cytoplasmic protein class (Figure 94). The TMHMM test resulted in a graphical illustration, statistics, and a list of the predicted transmembrane helices and the predicted location of the intervening loop regions. The TMHMM test resulted and displayed that the whole sequence is highly unlikely to contain any transmembrane helices and that the majority of the protein has a high probability of being located outside of the membrane (Figure 95). The BOMP test result identified there are no integral beta-barrel outer membrane proteins (Figure 96). The PSORTb test resulted in an analysis report that identified no detectable internal helices, motifs, or signal peptides; the localization scores calculated the predictable location of the protein to be cytoplasmic (Figure 97). The Phobius test resulted in a line graphical illustration that identified a low probability of transmembrane helices localized in the cytoplasm and a high probability of non-cytoplasmic transmembrane helices present; the overall result calculated by Phobius resulted in the entire protein sequence as non-cytoplasmic (Figure 98).

Basic Information

Table 3: Gene BMW92_RS10835 basic information

Genome	Replicon	Locus Tag	Old Locus Tag
Coxiella burnetii	NZ_CP018005	BMW92_RS10835	BMW92_10465
Genomic Coordinates	**Products**	**Length**	**Start and End Position**
1964912..1965598	pyridoxal phosphate-dependent enzyme	687 / 228	1964912 - 1965598
Molecular Weight	**Average Isoelectric Point**	**IPC Protein**	**Protein Length**
25645.56854 Da	9.86	8.76	228 amino acids

Nucleotide Sequence	Amino Acid Sequence
atgtccattagcgaaaatattaaaagaattactaccgaaattcgc caagcggaaaaagaatttagccgctcgcctaacgcggtttcgc ttttagctgtgagtaaatcgcaatctcttgataagataaaagaag ctattgcagcaggacaaagacagtttggtgaaaattatttgcaa gaagcgttggtaaaaataaaagcgttgcgcgctcatccattgga atggcattttataggtgtcattcaaactaataagacgcgcctcatt tccacaaattttgattgggtacaaagcgtttctcgtttggaagttg cttcagaattacatcattatcgacctttggaattgccaccattgtct atttgcattcaggtaaacatcagtgaagaaaaaactaaaagcgg tgtagacttaacgaatttatcagaatttgcaaaagcagttagtca gtttgatcgtctgaggttgcgaggtttaatgacgataccggcttat cagaaagattttaatgcgcaaaaggccacatttgaaaagttaaa agaagcgcaacagcaattaataaagaaaggcttaccattggat gtcttgtcattaggaatgacgcacgactttcgcgcggctatcgc agcggggagcactatggttcgtatcggaactggaatttttggtc ctcgggaggatagatga	MSISENIKRITTEIRQAEKEFSRSPNAVSLL AVSKSQSLDKIKEAIAAGQRQFGENYLQE ALVKIKALRAHPLEWHFIGVIQTNKTRLIS TNFDWVQSVSRLEVASELHHYRPLELPPL SICIQVNISEEKTKSGVDLTNLSEFAKAVS QFDRLRLRGLMTIPAYQKDFNAQKATFE KLKEAQQQLIKKGLPLDVLSLGMTHDFR AAIAAGSTMVRIGTGIFGPREDR

Figure 68: Protein isoelectric point calculator. The number and prevalence of each amino acid in the protein coded from the BMW92_RS10835 gene of *Coxiella burnetii* (Kozlowski, Biology Direct, <http://isoelectric.org/>).

Sequence Similarity

BLAST

YggS family pyridoxal phosphate enzyme [Beggiatoa sp. 4572_84]

Sequence ID: OQY55381.1 **Length:** 230 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 9 to 227 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
251 bits(642)	4e-81	Compositional matrix adjust.	122/219(56%)	159/219(72%)	0/219(0%)

```
Query  7    IKRITTEIRQAEKEFSRSPNAVSLLAVSKSQSLDKIKEAIAAGQRQFGENYLQEALVKIK  66
            IK++   I +A ++F+RSP ++ LLAVSK++ ++ I  A  +GQ  FGENYLQEA+ KI
Sbjct  9    IKKVRKRIAEAARQFARSPGSIRLLAVSKTRPVEDIVTAFNSGQTCFGENYLQEAVPKID  68

Query  67   ALRAHPLEWHFIGVIQTNKTRLISTNFDWVQSVSRLEVASELHHYRPLELPPLSICIQVN  126
            ALR +PLEWHFIG +Q+NKTRLI+ NFDWVQS+ +L+ A  L+  RP   PPL++CIQVN
Sbjct  69   ALRDYPLEWHFIGPLQSNKTRLIAENFDWVQSLDKLKHAQRLNAQRPENFPPLNVCIQVN  128

Query  127  ISEEKTKSGVDLTNLSEFAKAVSQFDRLRLRGLMTIPAYQKDFNAQKATFEKLKEAQQQL  186
            ISEE  KSGV LT+L   A+A+++  RLRLRGLM +P   +DF  Q+  F  L+ A  +L
Sbjct  129  ISEETQKSGVHLTDLPTLAQAIAELPRLRLRGLMALPTLCQDFEQQRIPFRALRIAYLKL  188

Query  187  IKKGLPLDVLSLGMTHDFRAAIAAGSTMVRIGTGIFGPR  225
             + GL LD LS+GMT D  AAIA G+T VRIGTGIFG R
Sbjct  189  QESGLALDTLSMGMTGDMVAAIAEGATFVRIGTGIFGER  227
```

Figure 69: BLAST first match for BMW92_RS10835 sequence from organism *Beggiatoa sp.4572_84* with an e-value of 4e-81, 56% identity, 72% positives, 0% gaps (dissimilarity), and an identity of pyridoxal phosphate-dependent enzyme (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

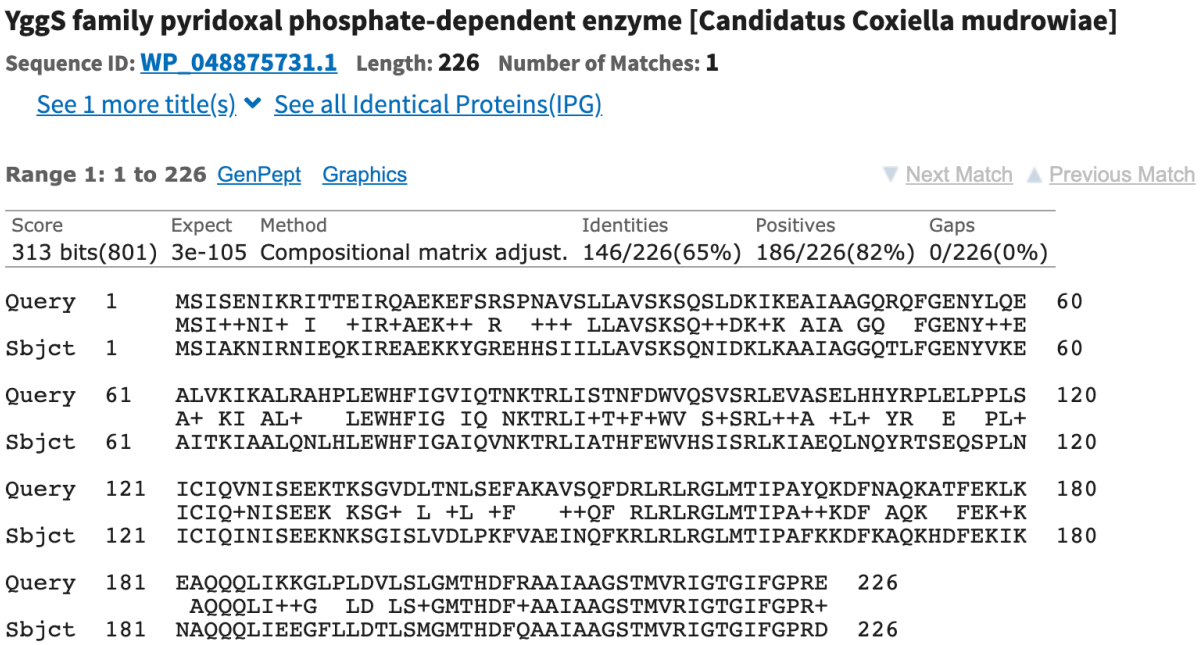


Figure 70: BLAST second match for BMW92_RS10835 sequence from organism *Coxiella mudrowiae* with an e-value of 3e-105, 65% identity, 82% positives, 0% gaps (dissimilarity), and an identity of pyridoxal phosphate-dependent enzyme (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

YggS family pyridoxal phosphate-dependent enzyme [Coxiella endosymbiont of Amblyomma americanum]

Sequence ID: WP_039670151.1 **Length:** 231 **Number of Matches:** 1

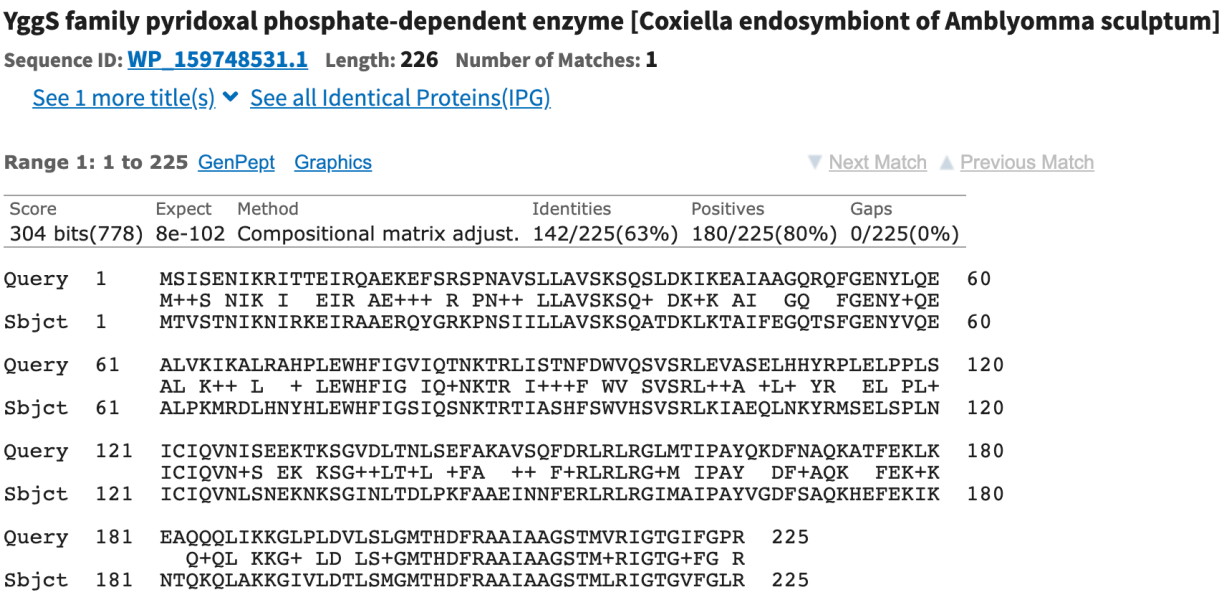
See 2 more title(s) See all Identical Proteins(IPG)

Range 1: 2 to 223 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
304 bits(779)	6e-102	Compositional matrix adjust.	144/222(65%)	178/222(80%)	0/222(0%)

```
Query  6    NIKRITTEIRQAEKEFSRSPNAVSLLAVSKSQSLDKIKEAIAAGQRQFGENYLQEALVKI  65
            NIK I    IR AEK++ R PN+V LLAVSKSQ +DK+K AI+ GQ  FGENY+QEAL+K+
Sbjct  2    NIKNIKKRIRAAEKKYGRKPNSVILLAVSKSQHIDKLKTAISEGQTCFGENYVQEALIKM  61

Query  66   KALRAHPLEWHFIGVIQTNKTRLISTNFDWVQSVSRLEVASELHHYRPLELPPLSICIQV  125
             ALR + LEWHFIG IQTNK  +I+ +F WV SVS+L+ A +L+ YR  ELPPL+ICIQV
Sbjct  62   SALRNYALEWHFIGSIQTNKIPVIAAHFGWVHSVSKLKTAEKLNKYRIPELPPLNICIQV  121

Query  126  NISEEKTKSGVDLTNLSEFAKAVSQFDRLRLRGLMTIPAYQKDFNAQKATFEKLKEAQQQ  185
            N+S E++K+G+ L +LS+FA  +S F RLRLRG+M IPAY  DF AQK  FEK+K AQQ+
Sbjct  122  NVSMEESKNGISLVDLSKFATEISNFKRLRLRGVMAIPAYNLDFCAQKFNFEKIKNAQQK  181

Query  186  LIKKGLPLDVLSLGMTHDFRAAIAAGSTMVRIGTGIFGPRED  227
            LIK+GL LD LS+GMTHDF+AAIAAGSTMVRIGTGIFG R++
Sbjct  182  LIKQGLSLDTLSMGMTHDFQAAIAAGSTMVRIGTGIFGSRDN  223
```

Figure 71: BLAST third match for BMW92_RS10835 sequence from organism *Amblyomma americanum* with an e-value of 6e-102, 65% identity, 80% positives, 0% gaps (dissimilarity), and an identity of pyridoxal phosphate-dependent enzyme (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).



YggS family pyridoxal phosphate-dependent enzyme [Coxiella endosymbiont of Amblyomma sculptum]

Sequence ID: WP_159748531.1 **Length:** 226 **Number of Matches:** 1

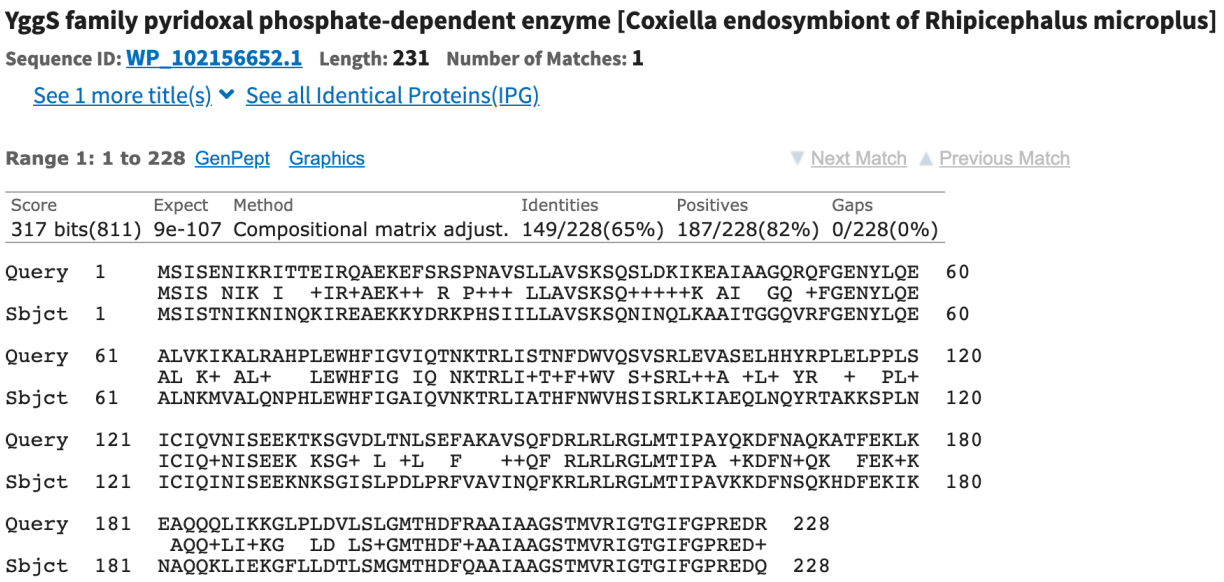
See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 225 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
304 bits(778)	8e-102	Compositional matrix adjust.	142/225(63%)	180/225(80%)	0/225(0%)

```
Query  1    MSISENIKRITTEIRQAEKEFSRSPNAVSLLAVSKSQSLDKIKEAIAAGQRQFGENYLQE  60
            M++S NIK I  EIR AE+++ R PN++ LLAVSKSQ+ DK+K AI  GQ  FGENY+QE
Sbjct  1    MTVSTNIKNIRKEIRAAERQYGRKPNSIILLAVSKSQATDKLKTAIFEGQTSFGENYVQE  60

Query  61   ALVKIKALRAHPLEWHFIGVIQTNKTRLISTNFDWVQSVSRLEVASELHHYRPLELPPLS  120
            AL K++ L  + LEWHFIG IQ+NKTR I+++F WV SVSRL++A +L+ YR  EL PL+
Sbjct  61   ALPKMRDLHNYHLEWHFIGSIQSNKTRTIASHFSWVHSVSRLKIAEQLNKYRMSELSPLN  120

Query  121  ICIQVNISEEKTKSGVDLTNLSEFAKAVSQFDRLRLRGLMTIPAYQKDFNAQKATFEKLK  180
            ICIQVN+S EK KSG++LT+L +FA  ++ F+RLRLRG+M IPAY  DF+AQK  FEK+K
Sbjct  121  ICIQVNLSNEKNKSGINLTDLPKFAAEINNFERLRLRGIMAIPAYVGDFSAQKHEFEKIK  180

Query  181  EAQQQLIKKGLPLDVLSLGMTHDFRAAIAAGSTMVRIGTGIFGPR  225
              Q+QL KKG+ LD LS+GMTHDFRAAIAAGSTM+RIGTG+FG R
Sbjct  181  NTQKQLAKKGIVLDTLSMGMTHDFRAAIAAGSTMLRIGTGVFGLR  225
```

Figure 72: BLAST fourth match for BMW92_RS10835 sequence from organism *Amblyomma sculptum* with an e-value of 8e-102, 63% identity, 80% positives, 0% gaps (dissimilarity), and an identity of pyridoxal phosphate-dependent enzyme (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).



YggS family pyridoxal phosphate-dependent enzyme [Coxiella endosymbiont of Rhipicephalus microplus]

Sequence ID: WP_102156652.1 **Length:** 231 **Number of Matches:** 1

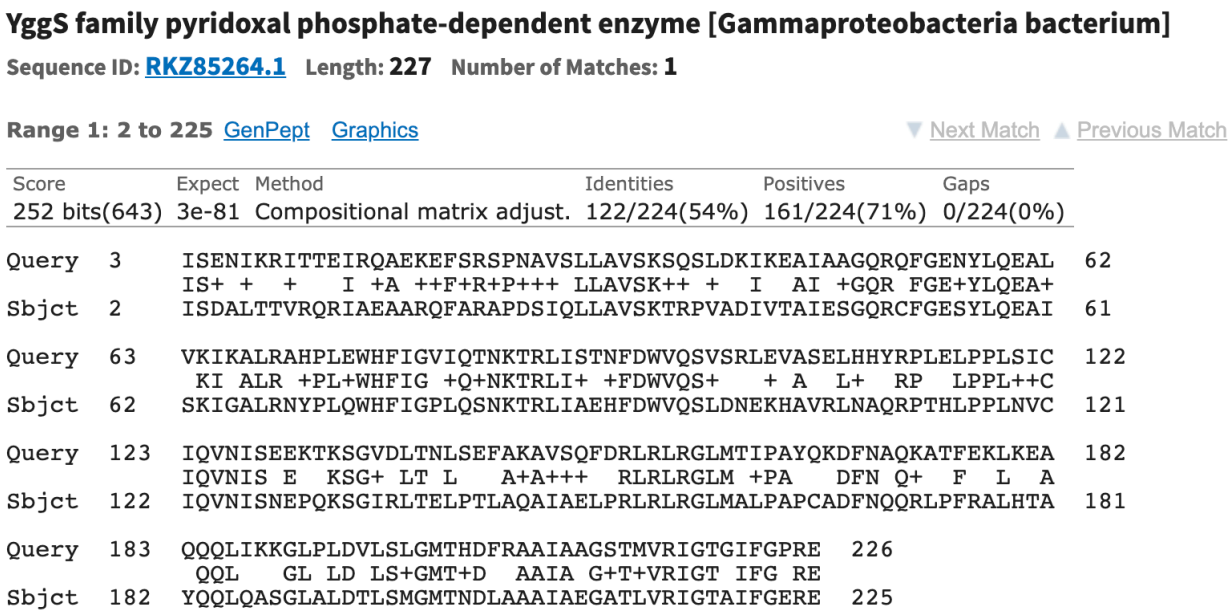
See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 228 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
317 bits(811)	9e-107	Compositional matrix adjust.	149/228(65%)	187/228(82%)	0/228(0%)

```
Query  1    MSISENIKRITTEIRQAEKEFSRSPNAVSLLAVSKSQSLDKIKEAIAAGQRQFGENYLQE  60
            MSIS NIK I  +IR+AEK++ R P+++ LLAVSKSQ+++++K AI  GQ +FGENYLQE
Sbjct  1    MSISTNIKNINQKIREAEKKYDRKPHSIILLAVSKSQNINQLKAAITGGQVRFGENYLQE  60

Query  61   ALVKIKALRAHPLEWHFIGVIQTNKTRLISTNFDWVQSVSRLEVASELHHYRPLELPPLS  120
            AL K+ AL+   LEWHFIG IQ NKTRLI+T+F+WV S+SRL++A +L+ YR  +  PL+
Sbjct  61   ALNKMVALQNPHLEWHFIGAIQVNKTRLIATHFNWVHSISRLKIAEQLNQYRTAKKSPLN  120

Query  121  ICIQVNISEEKTKSGVDLTNLSEFAKAVSQFDRLRLRGLMTIPAYQKDFNAQKATFEKLK  180
            ICIQ+NISEEK KSG+ L +L  F   ++QF RLRLRGLMTIPA +KDFN+QK  FEK+K
Sbjct  121  ICIQINISEEKNKSGISLPDLPRFVAVINQFKRLRLRGLMTIPAVKKDFNSQKHDFEKIK  180

Query  181  EAQQQLIKKGLPLDVLSLGMTHDFRAAIAAGSTMVRIGTGIFGPREDR  228
             AQQ+LI+KG  LD LS+GMTHDF+AAIAAGSTMVRIGTGIFGPRED+
Sbjct  181  NAQQKLIEKGFLLDTLSMGMTHDFQAAIAAGSTMVRIGTGIFGPREDQ  228
```

Figure 73: BLAST fifth match for BMW92_RS10835 sequence from organism *Rhicephalus microplus* with an e-value of 9e-107, 65% identity, 82% positives, 0% gaps (dissimilarity), and an identity of pyridoxal phosphate-dependent enzyme (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).



YggS family pyridoxal phosphate-dependent enzyme [Gammaproteobacteria bacterium]

Sequence ID: RKZ85264.1 Length: 227 Number of Matches: 1

Range 1: 2 to 225 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
252 bits(643)	3e-81	Compositional matrix adjust.	122/224(54%)	161/224(71%)	0/224(0%)

```
Query  3    ISENIKRITTEIRQAEKEFSRSPNAVSLLAVSKSQSLDKIKEAIAAGQRQFGENYLQEAL  62
            IS+ +  +   I +A ++F+R+P+++ LLAVSK++ +  I  AI +GQR FGE+YLQEA+
Sbjct  2    ISDALTTVRQRIAEAARQFARAPDSIQLLAVSKTRPVADIVTAIESGQRCFGESYLQEAI  61

Query  63   VKIKALRAHPLEWHFIGVIQTNKTRLISTNFDWVQSVSRLEVASELHHYRPLELPPLSIC  122
             KI ALR +PL+WHFIG +Q+NKTRLI+ +FDWVQS+   + A  L+  RP  LPPL++C
Sbjct  62   SKIGALRNYPLQWHFIGPLQSNKTRLIAEHFDWVQSLDNEKHAVRLNAQRPTHLPPLNVC  121

Query  123  IQVNISEEKTKSGVDLTNLSEFAKAVSQFDRLRLRGLMTIPAYQKDFNAQKATFEKLKEA  182
            IQVNIS E  KSG+ LT L   A+A+++  RLRLRGLM +PA   DFN Q+  F  L  A
Sbjct  122  IQVNISNEPQKSGIRLTELPTLAQAIAELPRLRLRGLMALPAPCADFNQQRLPFRALHTA  181

Query  183  QQQLIKKGLPLDVLSLGMTHDFRAAIAAGSTMVRIGTGIFGPRE  226
             QQL   GL LD LS+GMT+D  AAIA G+T+VRIGT IFG RE
Sbjct  182  YQQLQASGLALDTLSMGMTNDLAAAIAEGATLVRIGTAIFGERE  225
```

Figure 74: BLAST sixth match for BMW92_RS10835 sequence from organism *Gammaproteobacteria bacterium* with an e-value of 3e-81, 54% identity, 71% positives, 0% gaps (dissimilarity), and an identity of pyridoxal phosphate-dependent enzyme (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

YggS family pyridoxal phosphate-dependent enzyme [Nitrosococcus halophilus]

Sequence ID: WP_013034657.1 **Length:** 231 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 4 to 226 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
249 bits(635)	5e-80	Compositional matrix adjust.	121/223(54%)	157/223(70%)	0/223(0%)

```
Query  3    ISENIKRITTEIRQAEKEFSRSPNAVSLLAVSKSQSLDKIKEAIAAGQRQFGENYLQEAL  62
            I++ + ++ T I +AE+ F R   +V+L+A +K+ S+  I+ AIA GQR FGENYLQEAL
Sbjct  4    IAQQLAQVQTRIAEAEQRFGRPAGSVTLVAATKTCSVSAIRAAIACGQRAFGENYLQEAL  63

Query  63   VKIKALRAHPLEWHFIGVIQTNKTRLISTNFDWVQSVSRLEVASELHHYRPLELPPLSIC  122
             KIK L +  LEWHFIG IQ+NKTR I+ +FDWV SV RL+VA  L+  RP ELPPL++C
Sbjct  64   PKIKELESENLEWHFIGPIQSNKTRDIAAHFDWVHSVDRLKVAQRLNQQRPPELPPLNVC  123

Query  123  IQVNISEEKTKSGVDLTNLSEFAKAVSQFDRLRLRGLMTIPAYQKDFNAQKATFEKLKEA  182
            +QVNIS E +KSG     L+E AKAV++  RL LRGLMT+P    DF AQ+  F  L +
Sbjct  124  LQVNISGEDSKSGTTPEELTELAKAVAEMPRLSLRGLMTLPPLNSDFEAQRQPFRALHQL  183

Query  183  QQQLIKKGLPLDVLSLGMTHDFRAAIAAGSTMVRIGTGIFGPR  225
             Q+L + GL LD LS+GMT D  AAIA G+T+VR+G  IFG R
Sbjct  184  WQELRQGGLKLDTLSIGMTDDLEAAIAEGATLVRVGAAIFGRR  226
```

Figure 75: BLAST seventh match for BMW92_RS10835 sequence from organism *Nitrosococcus halophilus* with an e-value of 5e-80, 54% identity, 70% positives, 0% gaps (dissimilarity), and an identity of pyridoxal phosphate-dependent enzyme (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

YggS family pyridoxal phosphate-dependent enzyme [Nitrosococcus watsonii]

Sequence ID: WP_013221841.1 **Length:** 231 **Number of Matches:** 1

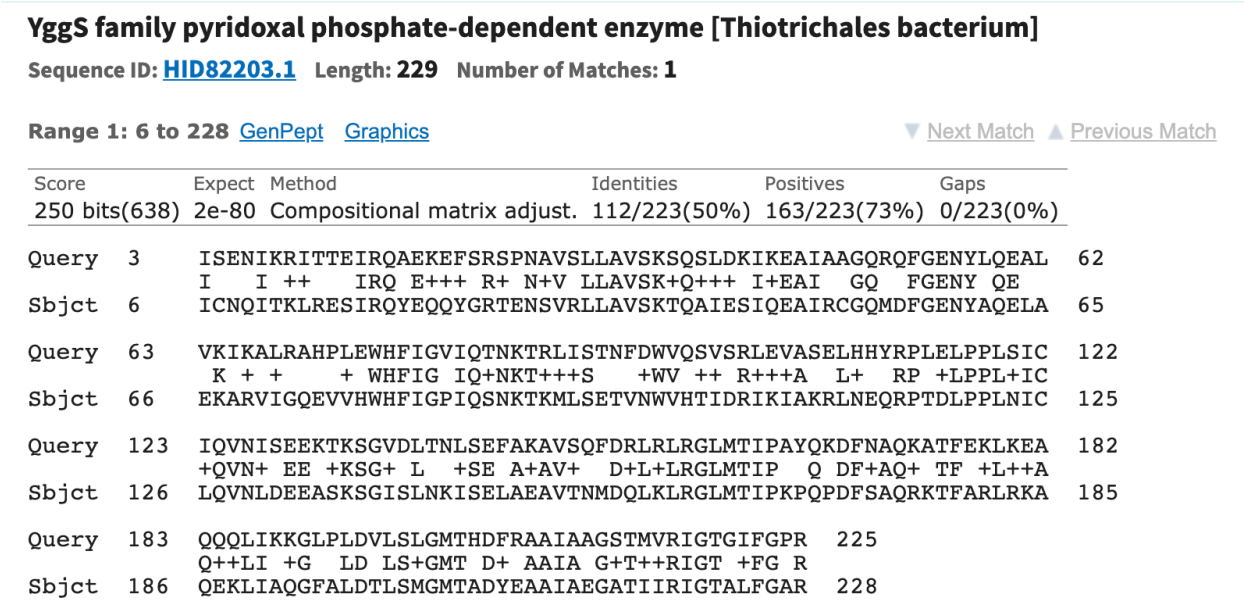
See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 4 to 231 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
249 bits(636)	4e-80	Compositional matrix adjust.	124/228(54%)	158/228(69%)	2/228(0%)

```
Query  3    ISENIKRITTEIRQAEKEFSRSPNAVSLLAVSKSQSLDKIKEAIAAGQRQFGENYLQEAL  62
            I+  +  I T I QAE+ F RS  +VSL+A SK+  +  I+ A+  GQR FGENYLQEAL
Sbjct  4    IALQLAEIYTRIAQAERRFGRSEGSVSLVAASKTCPVSAIRAAVVGGQRAFGENYLQEAL  63

Query  63   VKIKALRAHPLEWHFIGVIQTNKTRLISTNFDWVQSVSRLEVASELHHYRPLELPPLSIC  122
             KIK L    LEWHFIG IQ+NKTR I+T+FDWV SV+RL++A  L   RP EL PL++C
Sbjct  64   PKIKELETEGLEWHFIGPIQSNKTRDIATHFDWVHSVARLKIAQRLSQQRPPELAPLNVC  123

Query  123  IQVNISEEKTKSGVDLTNLSEFAKAVSQFDRLRLRGLMTIPAYQKDFNAQKATFEKLKEA  182
            +QVNIS E +KSG     L+E A AV++  +L LRGLMT+PA   DF AQ+  F  L +
Sbjct  124  LQVNISGESSKSGTTTQELAELAAAVTEMPQLSLRGLMTLPALNSDFEAQRRPFRALHQL  183

Query  183  QQQLIKKGLPLDVLSLGMTHDFRAAIAAGSTMVRIGTGIFG--PREDR  228
             ++L +KG  LD LS+GMT D  AAIA G+T+VR+GT IFG  PR+DR
Sbjct  184  WEELRQKGFALDSLSMGMTDDLEAAIAEGATLVRVGTAIFGSRPRKDR  231
```

Figure 76: BLAST eighth match for BMW92_RS10835 sequence from organism *Nitrosococcus watsonii* with an e-value of 4e-80, 54% identity, 69% positives, 0% gaps (dissimilarity), and an identity of pyridoxal phosphate-dependent enzyme (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).



YggS family pyridoxal phosphate-dependent enzyme [Thiotrichales bacterium]

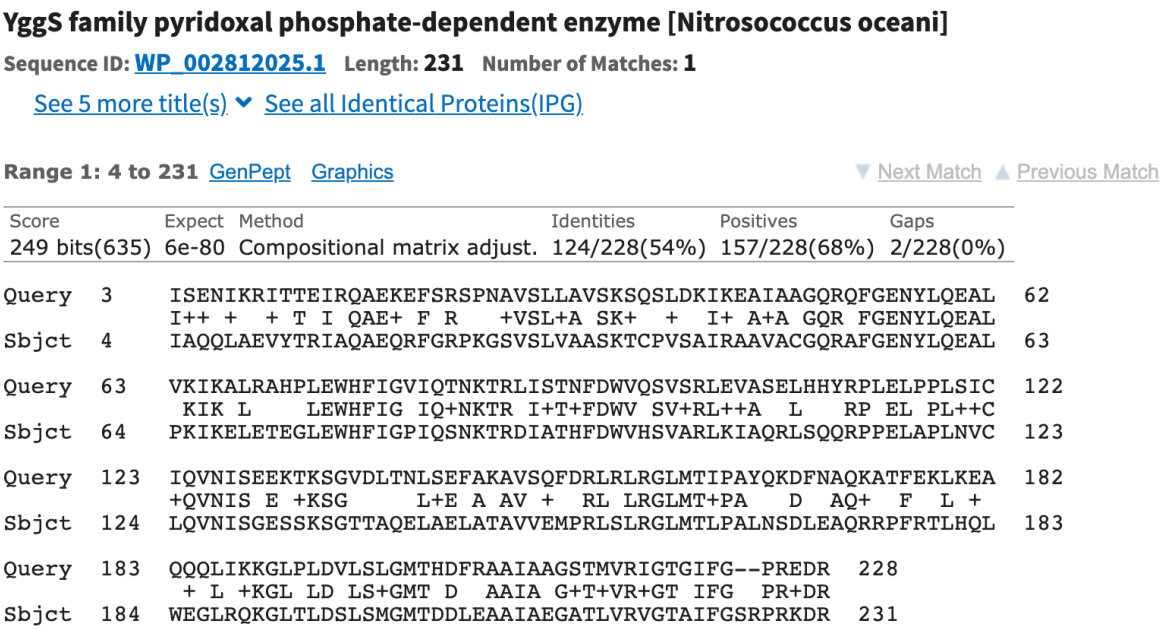
Sequence ID: HID82203.1 **Length:** 229 **Number of Matches:** 1

Range 1: 6 to 228 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
250 bits(638)	2e-80	Compositional matrix adjust.	112/223(50%)	163/223(73%)	0/223(0%)

```
Query  3    ISENIKRITTEIRQAEKEFSRSPNAVSLLAVSKSQSLDKIKEAIAAGQRQFGENYLQEAL  62
            I   I ++   IRQ E+++ R+ N+V LLAVSK+Q+++ I+EAI  GQ  FGENY QE
Sbjct  6    ICNQITKLRESIRQYEQQYGRTENSVRLLAVSKTQAIESIQEAIRCGQMDFGENYAQELA  65

Query  63   VKIKALRAHPLEWHFIGVIQTNKTRLISTNFDWVQSVSRLEVASELHHYRPLELPPLSIC  122
             K + +    + WHFIG IQ+NKT+++S   +WV ++ R+++A  L+  RP +LPPL+IC
Sbjct  66   EKARVIGQEVVHWHFIGPIQSNKTKMLSETVNWVHTIDRIKIAKRLNEQRPTDLPPLNIC  125

Query  123  IQVNISEEKTKSGVDLTNLSEFAKAVSQFDRLRLRGLMTIPAYQKDFNAQKATFEKLKEA  182
            +QVN+ EE +KSG+ L  +SE A+AV+  D+L+LRGLMTIP  Q DF+AQ+ TF +L++A
Sbjct  126  LQVNLDEEASKSGISLNKISELAEAVTNMDQLKLRGLMTIPKPQPDFSAQRKTFARLRKA  185

Query  183  QQQLIKKGLPLDVLSLGMTHDFRAAIAAGSTMVRIGTGIFGPR  225
            Q++LI +G  LD LS+GMT D+ AAIA G+T++RIGT +FG R
Sbjct  186  QEKLIAQGFALDTLSMGMTADYEAAIAEGATIIRIGTALFGAR  228
```

Figure 77: BLAST ninth match for BMW92_RS10835 sequence from organism *Thiotrichales bacterium* with an e-value of 2e-80, 50% identity, 73% positives, 0% gaps (dissimilarity), and an identity of pyridoxal phosphate-dependent enzyme (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).



YggS family pyridoxal phosphate-dependent enzyme [Nitrosococcus oceani]

Sequence ID: WP_002812025.1 **Length:** 231 **Number of Matches:** 1

See 5 more title(s) See all Identical Proteins(IPG)

Range 1: 4 to 231 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
249 bits(635)	6e-80	Compositional matrix adjust.	124/228(54%)	157/228(68%)	2/228(0%)

```
Query  3    ISENIKRITTEIRQAEKEFSRSPNAVSLLAVSKSQSLDKIKEAIAAGQRQFGENYLQEAL  62
            I++ +  + T I QAE+ F R   +VSL+A SK+  +  I+ A+A GQR FGENYLQEAL
Sbjct  4    IAQQLAEVYTRIAQAEQRFGRPKGSVSLVAASKTCPVSAIRAAVACGQRAFGENYLQEAL  63

Query  63   VKIKALRAHPLEWHFIGVIQTNKTRLISTNFDWVQSVSRLEVASELHHYRPLELPPLSIC  122
             KIK L    LEWHFIG IQ+NKTR I+T+FDWV SV+RL++A  L   RP EL PL++C
Sbjct  64   PKIKELETEGLEWHFIGPIQSNKTRDIATHFDWVHSVARLKIAQRLSQQRPPELAPLNVC  123

Query  123  IQVNISEEKTKSGVDLTNLSEFAKAVSQFDRLRLRGLMTIPAYQKDFNAQKATFEKLKEA  182
            +QVNIS E +KSG     L+E A AV +  RL LRGLMT+PA   D  AQ+  F  L +
Sbjct  124  LQVNISGESSKSGTTAQELAELATAVVEMPRLSLRGLMTLPALNSDLEAQRRPFRTLHQL  183

Query  183  QQQLIKKGLPLDVLSLGMTHDFRAAIAAGSTMVRIGTGIFG--PREDR  228
             + L +KGL LD LS+GMT D  AAIA G+T+VR+GT IFG  PR+DR
Sbjct  184  WEGLRQKGLTLDSLSMGMTDDLEAAIAEGATLVRVGTAIFGSRPRKDR  231
```

Figure 78: BLAST tenth match for BMW92_RS10835 sequence from organism *Nitrosococcus oceani* with an e-value of 6e-80, 54% identity, 68% positives, 0% gaps (dissimilarity), and an identity of pyridoxal phosphate-dependent enzyme (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

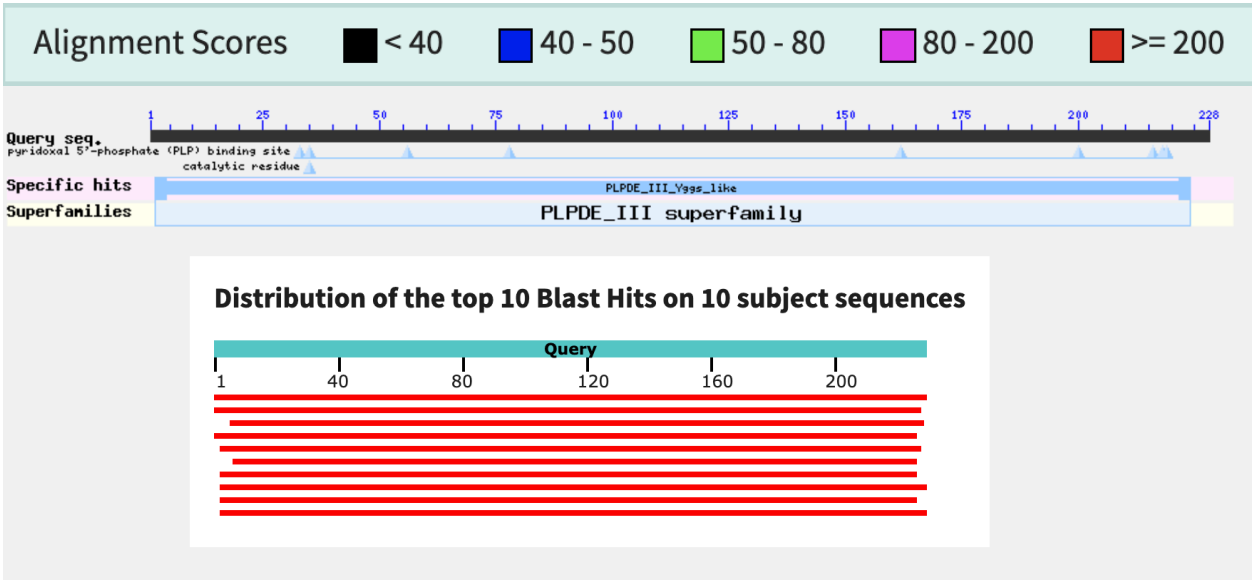


Figure 79: BLAST graphic summary of the top 10 organism sequences similarities selected aligned with *Coxiella burnetii* query sequence of gene BMW92_RS10835. Each of the alignment sequences selected are order from highest sequence similarity (top) to lowest sequence similarity (bottom). All organism sequences aligned with the query sequence have an alignment score of greater than 200 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

CDD

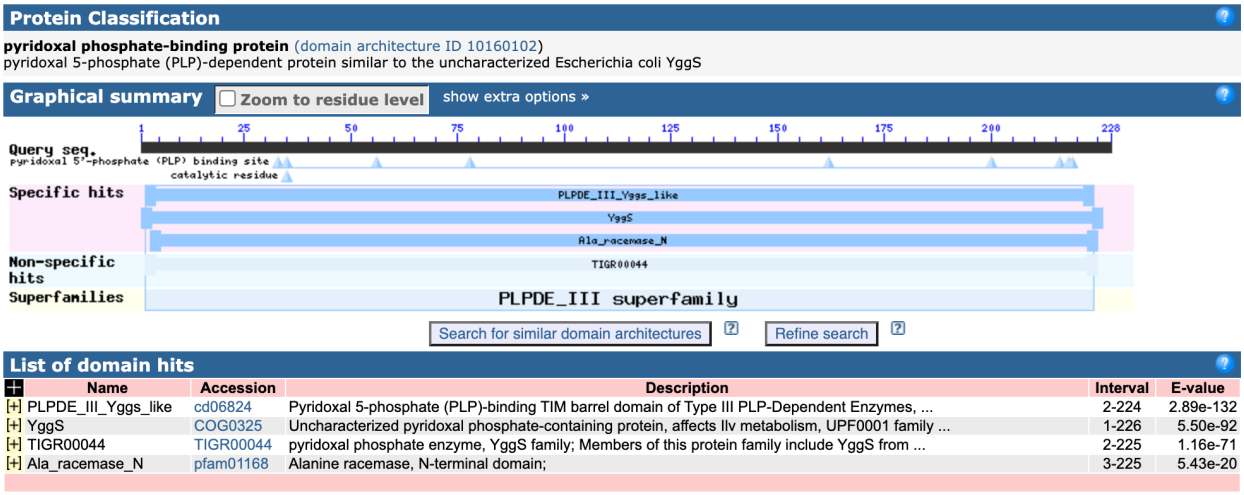


Figure 80: Conserved Domain Database output results for gene BMW92_RS10835. The top domain hit match was PLDE_III_Yggs_like: pyridoxal phosphate-dependent enzyme which aligned with the query sequence from amino acid residues 2-224 and had statistically significant e-value of 2.89e-132. The second domain hit match was YggS: pyridoxal phosphate-dependent enzyme which aligned with the query sequence from amino acid residues 1-226 and had a statistically significant e-value of 5.50e-92. The third domain hit match was TIGR00044: pyridoxal phosphate-dependent enzyme which aligned with the query sequence from amino acid residues 2-225 and had a statistically significant e-value of 1.16e-71. The fourth domain hit match was Ala_racemase_N: pyridoxal phosphate-dependent enzyme which aligned with the query sequence from amino acid residues 3-225 and had a statistically significant e-value of 5.43e-20 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

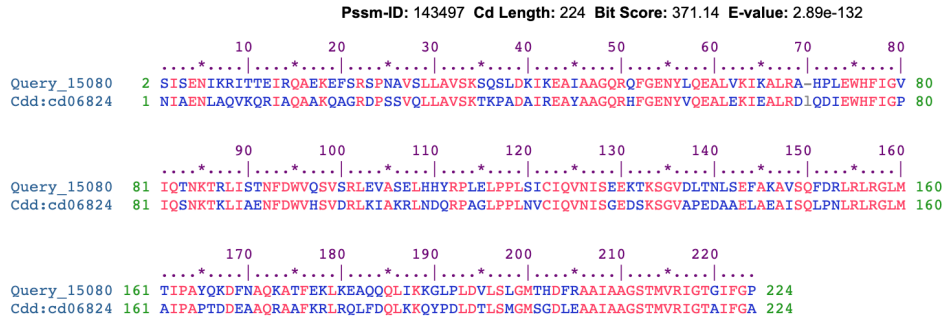


Figure 81: The top domain hit sequence PLDE_III_Yggs_like: Pyridoxal 5-phosphate (PLP)-binding TIM barrel domain of Type III PLD-Dependent Enzyme aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 2-224 and had a statistically significant e-value of 2.89e-132 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

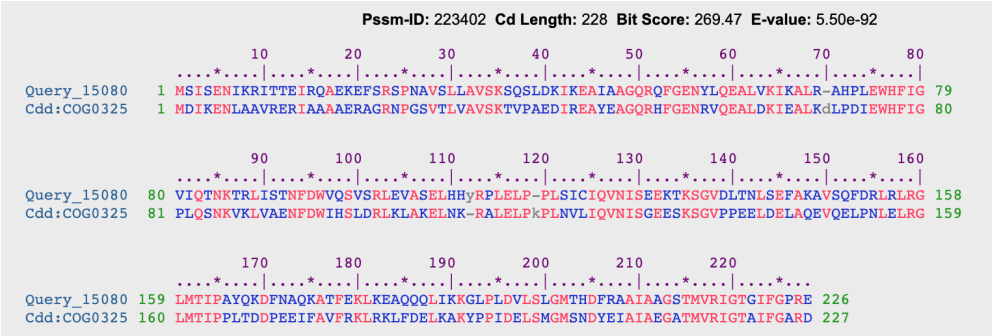


Figure 82: The second domain hit sequence YggS: pyridoxal phosphate-dependent enzyme aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 1-226 and had a statistically significant e-value of 5.50e-92 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

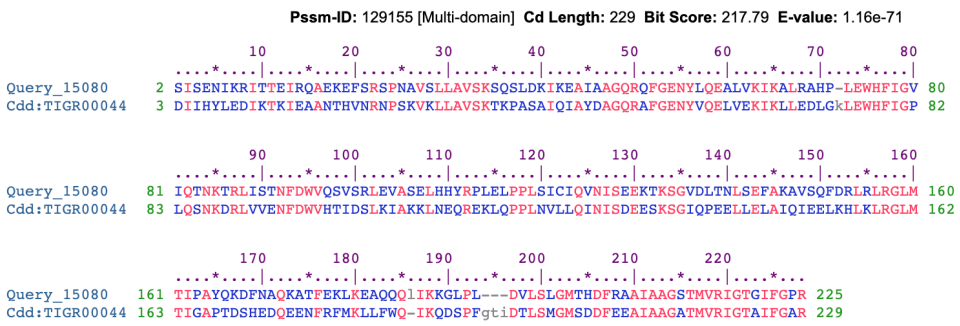


Figure 83: The third domain hit sequence TIGR00044: pyridoxal phosphate-dependent enzyme aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 2-225 and had a statistically significant e-value of 1.16e-71 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

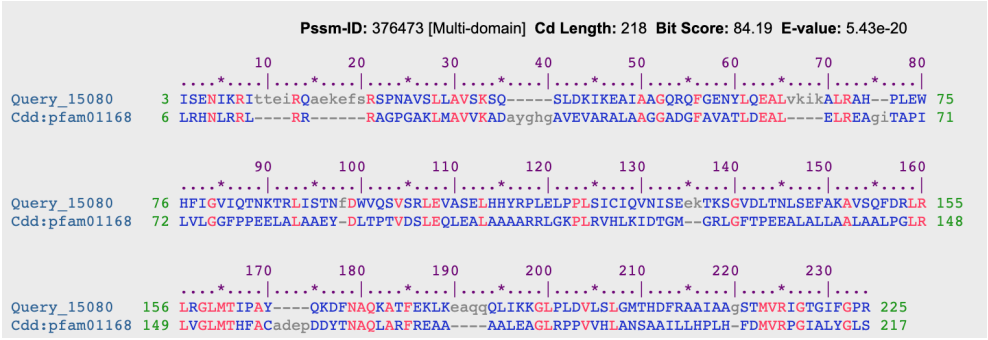


Figure 84: The fourth domain hit sequence Ala_racemase_N: Alanine racemase N-terminal domain aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 3-225 and had a statistically significant e-value of 5.43e-20 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

MUSCLE

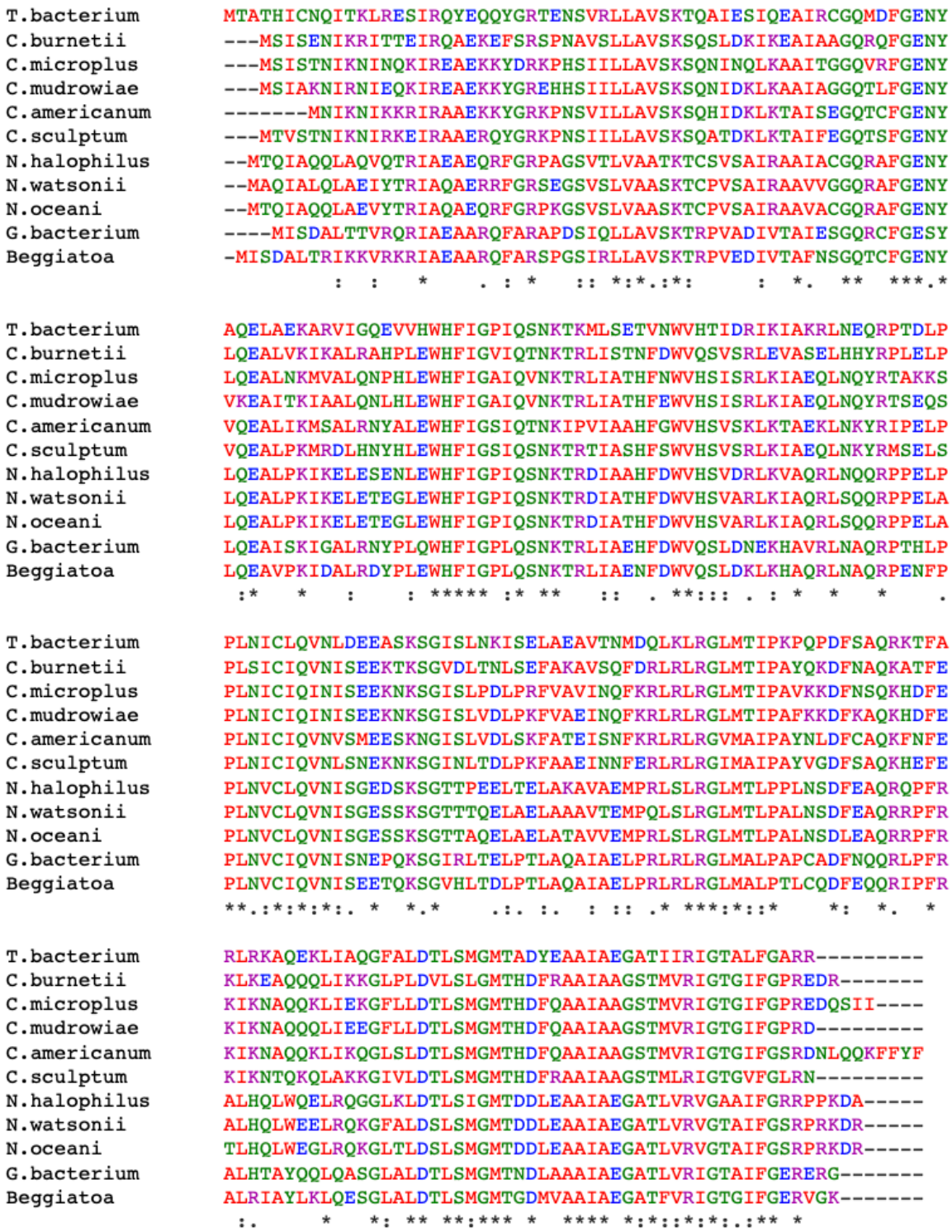

T.bacterium MTATHICNQITKLRESIRQYEQQYGRTENSVRLLAVSKTQAIESIQEAIRCGQMDFGENY
C.burnetii ---MSISENIKRITTEIRQAEKEFSRSPNAVSLLAVSKSQSLDKIKEAIAAGQRQFGENY
C.microplus ---MSISTNIKNINQKIREAEKKYDRKPHSIILLAVSKSQNINQLKAAITGGQVRFGENY
C.mudrowiae ---MSIAKNIRNIEQKIREAEKKYGREHHSIILLAVSKSQNIDKLKAAIAGGQTLFGENY
C.americanum -------MNIKNIKKRIRAAEKKYGRKPNSVILLAVSKSQHIDKLKTAISEGQTCFGENY
C.sculptum ---MTVSTNIKNIRKEIRAAERQYGRKPNSIILLAVSKSQATDKLKTAIFEGQTSFGENY
N.halophilus --MTQIAQQLAQVQTRIAEAEQRFGRPAGSVTLVAATKTCSVSAIRAAIACGQRAFGENY
N.watsonii --MAQIALQLAEIYTRIAQAERRFGRSEGSVSLVAASKTCPVSAIRAAVVGGQRAFGENY
N.oceani --MTQIAQQLAEVYTRIAQAEQRFGRPKGSVSLVAASKTCPVSAIRAAVACGQRAFGENY
G.bacterium ----MISDALTTVRQRIAEAARQFARAPDSIQLLAVSKTRPVADIVTAIESGQRCFGESY
Beggiatoa -MISDALTRIKKVRKRIAEAARQFARSPGSIRLLAVSKTRPVEDIVTAFNSGQTCFGENY
AQELAEKARVIGQEVVHWHFIGPIQSNKTKMLSETVNWVHTIDRIKIAKRLNEQRPTDLP
LQEALVKIKALRAHPLEWHFIGVIQTNKTRLISTNFDWVQSVSRLEVASELHHYRPLELP
LQEALNKMVALQNPHLEWHFIGAIQVNKTRLIATHFNWVHSISRLKIAEQLNQYRTAKKS
VKEAITKIAALQNLHLEWHFIGAIQVNKTRLIATHFEWVHSISRLKIAEQLNQYRTSEQS
VQEALIKMSALRNYALEWHFIGSIQTNKIPVIAAHFGWVHSVSKLKTAEKLNKYRIPELP
VQEALPKMRDLHNYHLEWHFIGSIQSNKTRTIASHFSWVHSVSRLKIAEQLNKYRMSELS
LQEALPKIKELESENLEWHFIGPIQSNKTRDIAAHFDWVHSVDRLKVAQRLNQQRPPELP
LQEALPKIKELETEGLEWHFIGPIQSNKTRDIATHFDWVHSVARLKIAQRLSQQRPPELA
LQEALPKIKELETEGLEWHFIGPIQSNKTRDIATHFDWVHSVARLKIAQRLSQQRPPELA
LQEAISKIGALRNYPLQWHFIGPLQSNKTRLIAEHFDWVQSLDNEKHAVRLNAQRPTHLP
LQEAVPKIDALRDYPLEWHFIGPLQSNKTRLIAENFDWVQSLDKLKHAQRLNAQRPENFP
PLNICLQVNLDEEASKSGISLNKISELAEAVTNMDQLKLRGLMTIPKPQPDFSAQRKTFA
PLSICIQVNISEEKTKSGVDLTNLSEFAKAVSQFDRLRLRGLMTIPAYQKDFNAQKATFE
PLNICIQINISEEKNKSGISLPDLPRFVAVINQFKRLRLRGLMTIPAVKKDFNSQKHDFE
PLNICIQINISEEKNKSGISLVDLPKFVAEINQFKRLRLRGLMTIPAFKKDFKAQKHDFE
PLNICIQVNVSMEESKNGISLVDLSKFATEISNFKRLRLRGVMAIPAYNLDFCAQKFNFE
PLNICIQVNLSNEKNKSGINLTDLPKFAAEINNFERLRLRGIMAIPAYVGDFSAQKHEFE
PLNVCLQVNISGEDSKSGTTPEELTELAKAVAEMPRLSLRGLMTLPPLNSDFEAQRQPFR
PLNVCLQVNISGESSKSGTTTQELAELAAAVTEMPQLSLRGLMTLPALNSDFEAQRRPFR
PLNVCLQVNISGESSKSGTTAQELAELATAVVEMPRLSLRGLMTLPALNSDLEAQRRPFR
PLNVCIQVNISNEPQKSGIRLTELPTLAQAIAELPRLRLRGLMALPAPCADFNQQRLPFR
PLNVCIQVNISEETQKSGVHLTDLPTLAQAIAELPRLRLRGLMALPTLCQDFEQQRIPFR
RLRKAQEKLIAQGFALDTLSMGMTADYEAAIAEGATIIRIGTALFGARR---------
KLKEAQQQLIKKGLPLDVLSLGMTHDFRAAIAAGSTMVRIGTGIFGPREDR-------
KIKNAQQKLIEKGFLLDTLSMGMTHDFQAAIAAGSTMVRIGTGIFGPREDQSII----
KIKNAQQQLIEEGFLLDTLSMGMTHDFQAAIAAGSTMVRIGTGIFGPRD---------
KIKNAQQKLIKQGLSLDTLSMGMTHDFQAAIAAGSTMVRIGTGIFGSRDNLQQKFFYF
KIKNTQKQLAKKGIVLDTLSMGMTHDFRAAIAAGSTMLRIGTGVFGLRN---------
ALHQLWQELRQGGLKLDTLSIGMTDDLEAAIAEGATLVRVGAAIFGRRPPKDA-----
ALHQLWEELRQKGFALDSLSMGMTDDLEAAIAEGATLVRVGTAIFGSRPRKDR-----
TLHQLWEGLRQKGLTLDSLSMGMTDDLEAAIAEGATLVRVGTAIFGSRPRKDR-----
ALHTAYQQLQASGLALDTLSMGMTNDLAAAIAEGATLVRIGTAIFGERERG-------
ALRIAYLKLQESGLALDTLSMGMTGDMVAAIAEGATFVRIGTGIFGERVGK-------

Figure 85: MUSCLE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10835 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered randomly and are listed from top to bottom as followed: *T. bacterium*, *C. burnetii*, *C. microplus*, *C. mudrowiae*, *C. americanum*, *C. sculptum*, *N. halophilus*, *N. watsonii*, *N. oceani*, *G. bacterium*, *Begiatoa.* Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

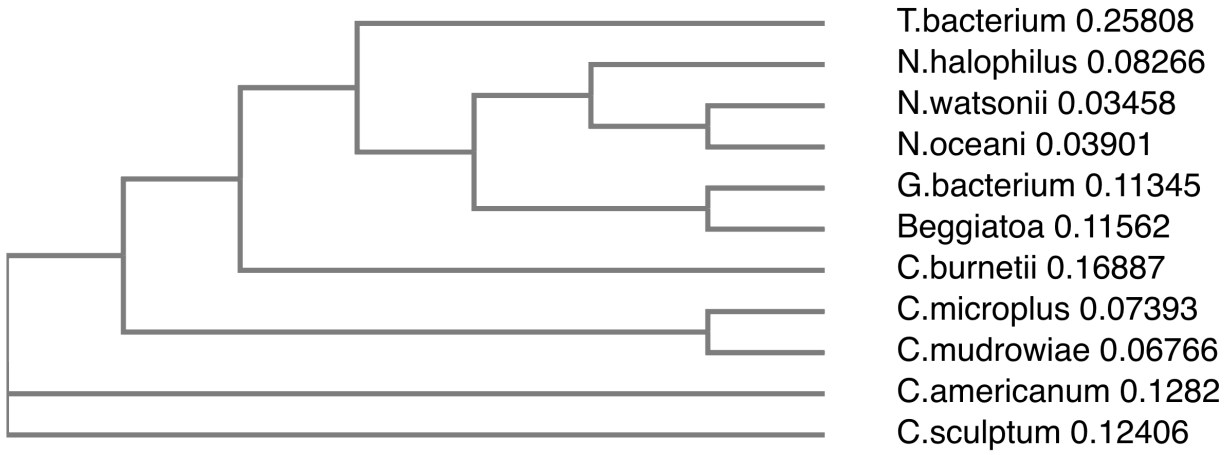


Figure 86: MUSCLE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10835 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *T. bacterium*, *N. halophilus*, *N. watsonii*, *N. oceani*, *G. bacterium*, *Begiatoa, C. burnetii*, *C. microplus*, *C. mudrowiae*, *C. americanum*, *C. sculptum*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

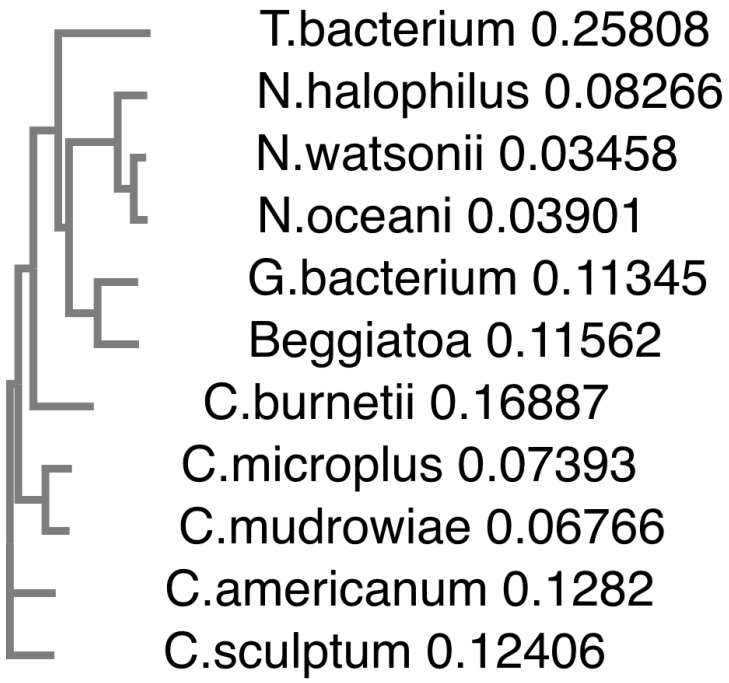
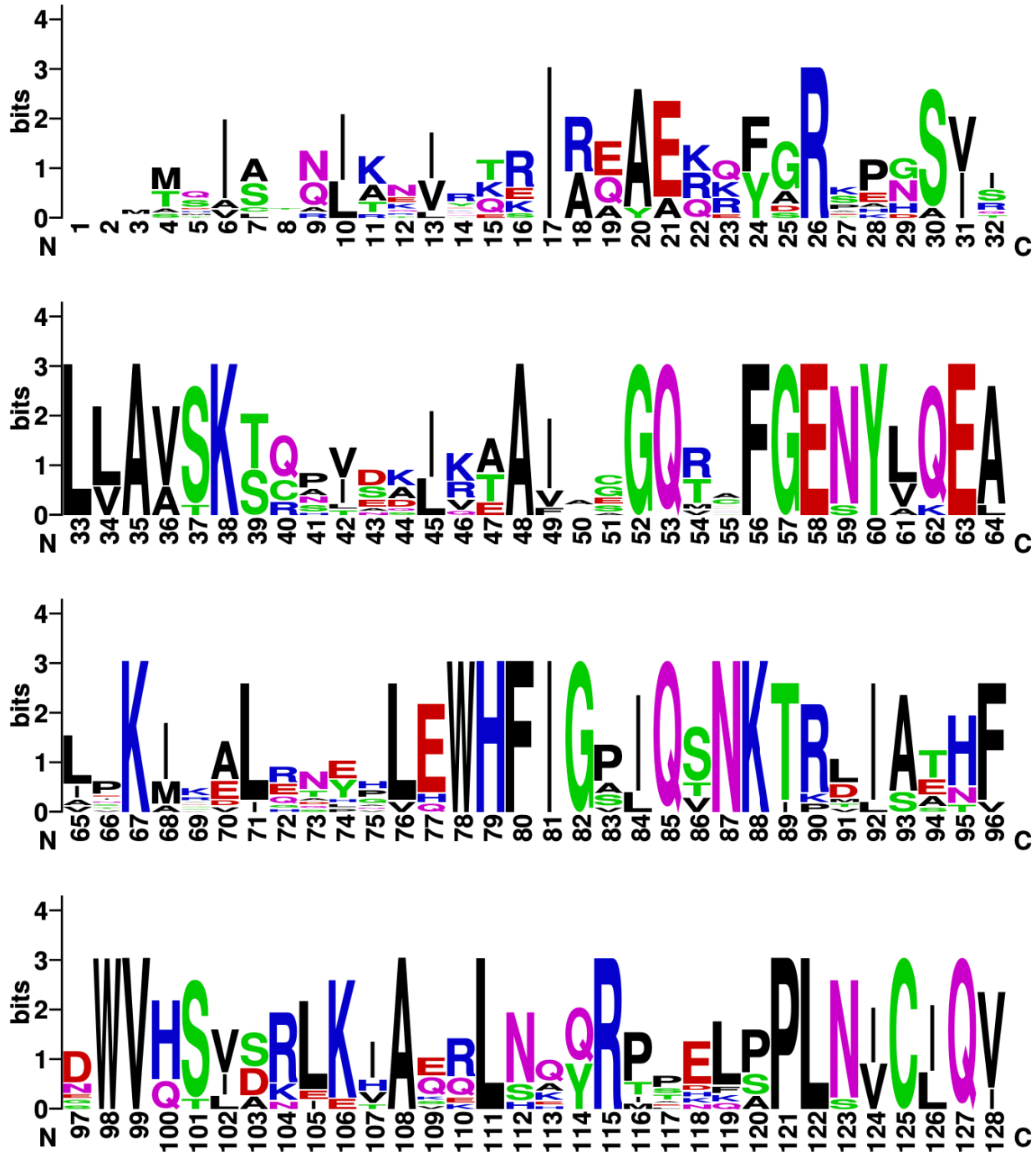


Figure 87: MUSCLE multiple sequence alignment real phylogenetic tree for *Coxiella burnetii* gene BMW92_RS10835 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *T. bacterium*, *N. halophilus*, *N. watsonii*, *N. oceani*, *G. bacterium*, *Begiatoa, C. burnetii*, *C. microplus*, *C. mudrowiae*, *C. americanum*, *C. sculptum*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

MUSCLE Sequence Logo



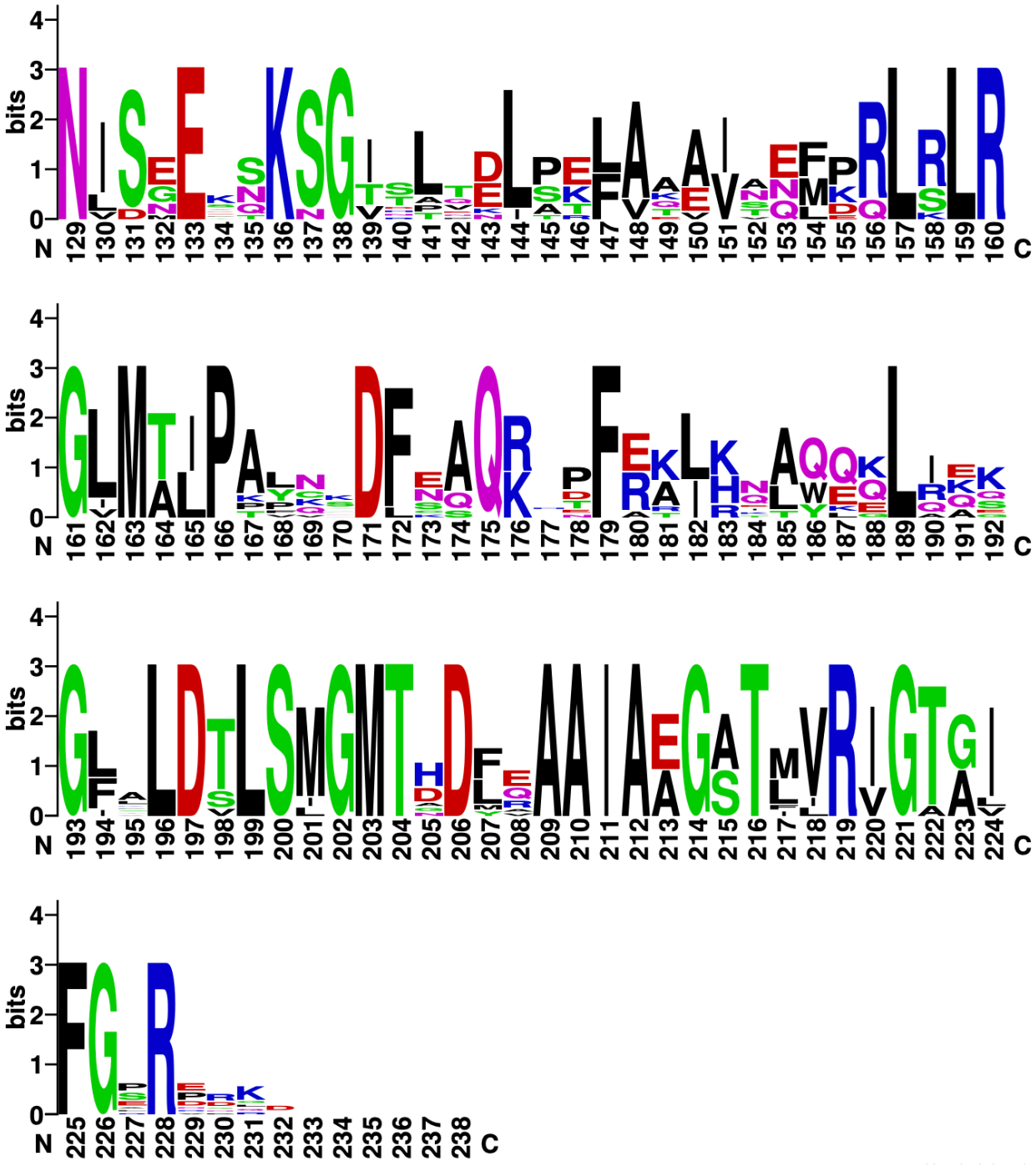


Figure 88: Sequence logo generated from the MUSCLE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10835 and the top 10 organism sequences similarities selected. Amino acids are represented by single letter abbreviations and distinct colors for each

respective amino acid. The conservation of each amino acid among residue across similar sequences is represented by the height of each single letter abbreviation of the amino acid. (WebLogo, <https://weblogo.berkeley.edu/logo.cgi>).

T-COFFEE

```
Beggiatoa          MIS--DALTRIKKVRKRIAEAARQFARSPGSIRLLAVSKTRPVEDIVTAF
C.americanum       M--------NIKNIKKRIRAAEKKYGRKPNSVILLAVSKSQHIDKLKTAI
C.burnetii         MS---I-SENIKRITTEIRQAEKEFSRSPNAVSLLAVSKSQSLDKIKEAI
C.microplus        MS---I-STNIKNINQKIREAEKKYDRKPHSIILLAVSKSQNINQLKAAI
C.mudrowiae        MS---I-AKNIRNIEQKIREAEKKYGREHHSIILLAVSKSQNIDKLKAAI
C.sculptum         MT---V-STNIKNIRKEIRAAERQYGRKPNSIILLAVSKSQATDKLKTAI
G.bacterium        MIS-----DALTTVRQRIAEAARQFARAPDSIQLLAVSKTRPVADIVTAI
N.halophilus       MTQ--I-AQQLAQVQTRIAEAEQRFGRPAGSVTLVAATKTCSVSAIRAAI
N.oceani           MTQ--I-AQQLAEVYTRIAQAEQRFGRPKGSVSLVAASKTCPVSAIRAAV
N.watsonii         MAQ--I-ALQLAEIYTRIAQAERRFGRSEGSVSLVAASKTCPVSAIRAAV
T.bacterium        MTATHI-CNQITKLRESIRQYEQQYGRTENSVRLLAVSKTQAIESIQEAI
                   *         :  :   *    :.: *   :: *:*.:*:     :  *.

Beggiatoa          NSGQTCFGENYLQEAVPKIDALRDYPLEWHFIGPLQSNKTRLIAENFDWV
C.americanum       SEGQTCFGENYVQEALIKMSALRNYALEWHFIGSIQTNKIPVIAAHFGWV
C.burnetii         AAGQRQFGENYLQEALVKIKALRAHPLEWHFIGVIQTNKTRLISTNFDWV
C.microplus        TGGQVRFGENYLQEALNKMVALQNPHLEWHFIGAIQVNKTRLIATHFNWV
C.mudrowiae        AGGQTLFGENYVKEAITKIAALQNLHLEWHFIGAIQVNKTRLIATHFEWV
C.sculptum         FEGQTSFGENYVQEALPKMRDLHNYHLEWHFIGSIQSNKTRTIASHFSWV
G.bacterium        ESGQRCFGESYLQEAISKIGALRNYPLQWHFIGPLQSNKTRLIAEHFDWV
N.halophilus       ACGQRAFGENYLQEALPKIKELESENLEWHFIGPIQSNKTRDIAAHFDWV
N.oceani           ACGQRAFGENYLQEALPKIKELETEGLEWHFIGPIQSNKTRDIATHFDWV
N.watsonii         VGGQRAFGENYLQEALPKIKELETEGLEWHFIGPIQSNKTRDIATHFDWV
T.bacterium        RCGQMDFGENYAQELAEKARVIGQEVVHWHFIGPIQSNKTKMLSETVNWV
                     **  ***.* :*   *   :    :.***** :* **   ::  . **

Beggiatoa          QSLDKLKHAQRLNAQRPENFPPLNVCIQVNISEETQKSGVHLTDLPTLAQ
C.americanum       HSVSKLKTAEKLNKYRIPELPPLNICIQVNVSMEESKNGISLVDLSKFAT
C.burnetii         QSVSRLEVASELHHYRPLELPPLSICIQVNISEEKTKSGVDLTNLSEFAK
C.microplus        HSISRLKIAEQLNQYRTAKKSPLNICIQINISEEKNKSGISLPDLPRFVA
C.mudrowiae        HSISRLKIAEQLNQYRTSEQSPLNICIQINISEEKNKSGISLVDLPKFVA
C.sculptum         HSVSRLKIAEQLNKYRMSELSPLNICIQVNLSNEKNKSGINLTDLPKFAA
G.bacterium        QSLDNEKHAVRLNAQRPTHLPPLNVCIQVNISNEPQKSGIRLTELPTLAQ
N.halophilus       HSVDRLKVAQRLNQQRPPELPPLNVCLQVNISGEDSKSGTTPEELTELAK
N.oceani           HSVARLKIAQRLSQQRPPELAPLNVCLQVNISGESSKSGTTAQELAELAT
N.watsonii         HSVARLKIAQRLSQQRPPELAPLNVCLQVNISGESSKSGTTTQELAELAA
T.bacterium        HTIDRIKIAKRLNEQRPTDLPPLNICLQVNLDEEASKSGISLNKISELAE
                   ::: . : * .*   *  . .**.:*:*:*:. *  *.*    .:. :.
```

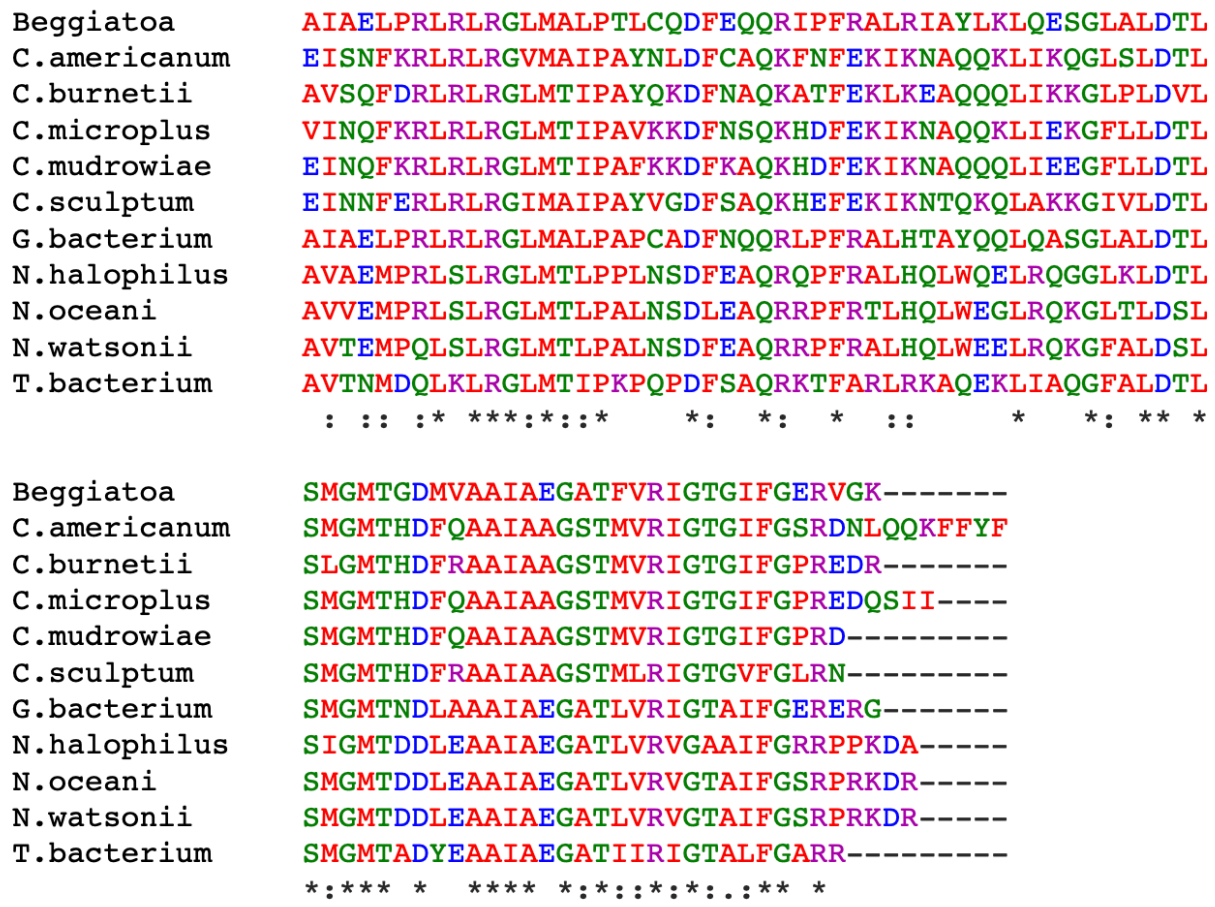


Figure 89: T-COFFEE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10835 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered randomly and are listed from top to bottom as followed: *Begiatoa, C. americanum*, *C. burnetii*, *C. microplus*, *C. mudrowiae*, *C. sculptum*, *G. bacterium*, *N. halophilus*, *N. oceani*, *N. watsonii*, *T. bacterium.* Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

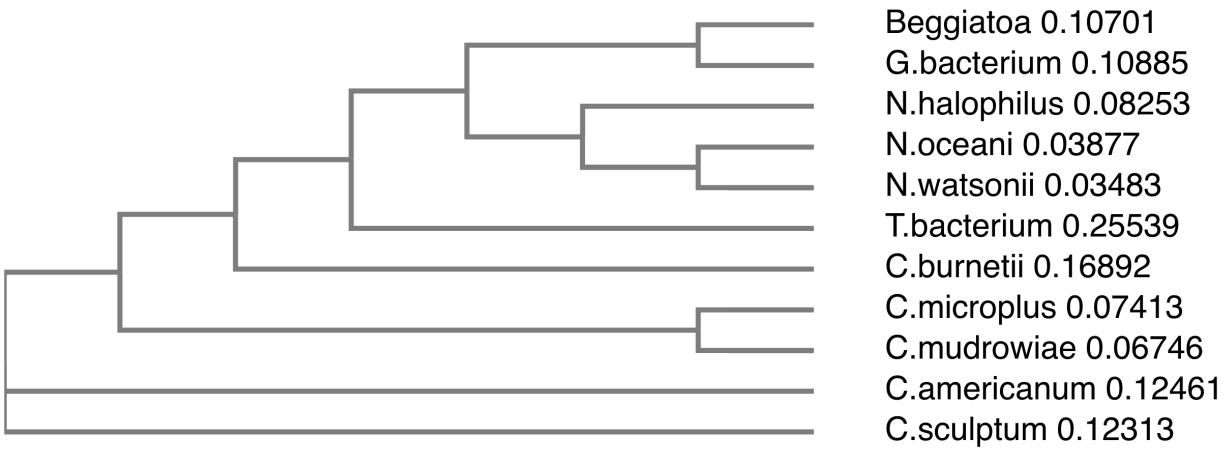


Figure 90: T-COFFEE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10835 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *Beggiatoa, G. bacterium, N. halophilus, N. oceani, N. watsonii, T. bacterium, C. burnetii, C. mudrowiae, C. americanum, C. sculptum*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

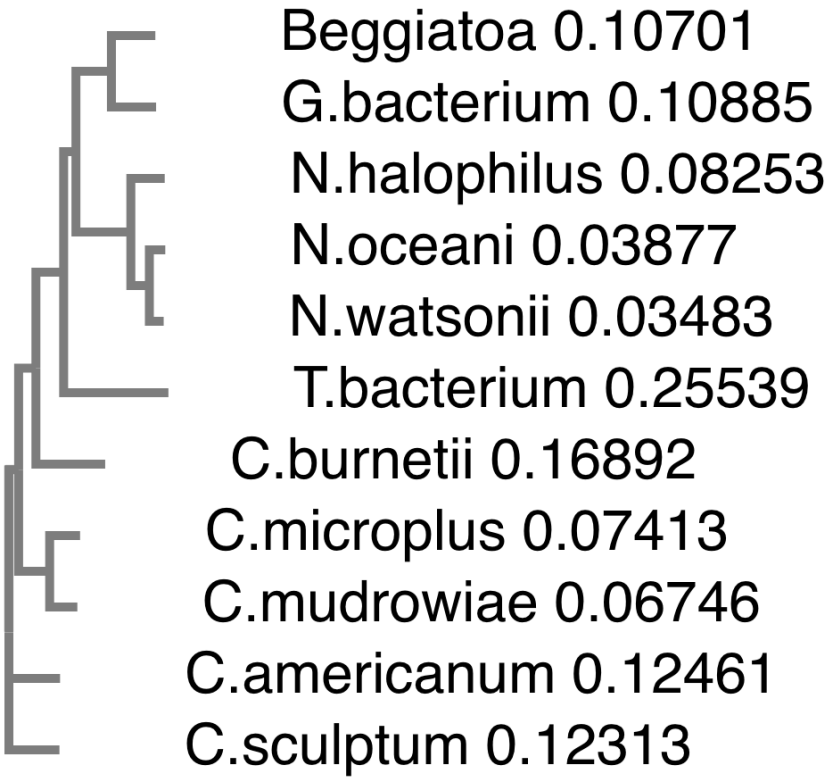
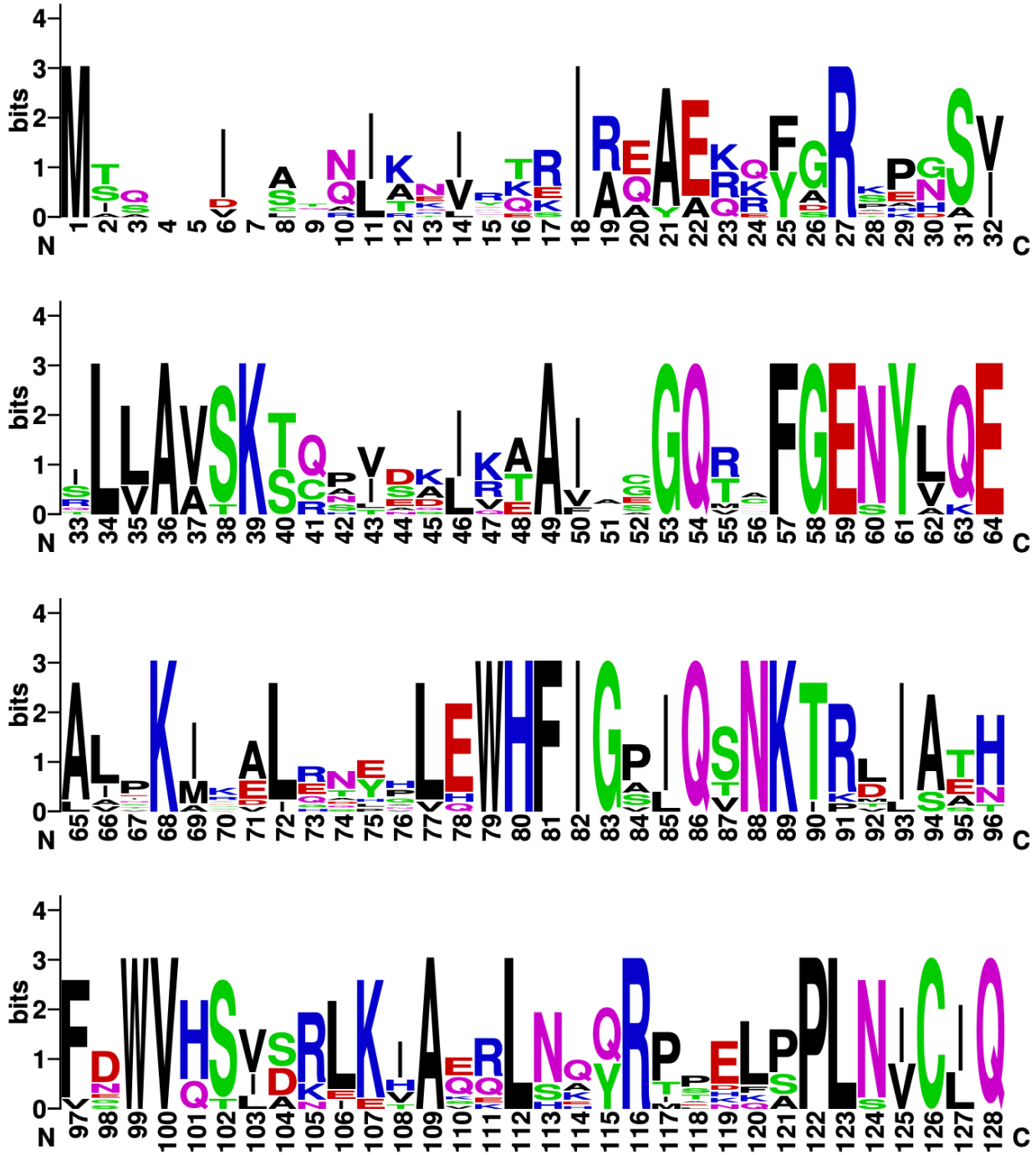


Figure 91: T-COFFEE multiple sequence alignment real phylogenetic tree for *Coxiella burnetii* gene BMW92_RS10835 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *Beggiatoa, G. bacterium, N. halophilus, N. oceani, N. watsonii, T. bacterium, C. burnetii, C. mudrowiae, C. americanum, C. sculptum*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

T-COFFEE Sequence Logo



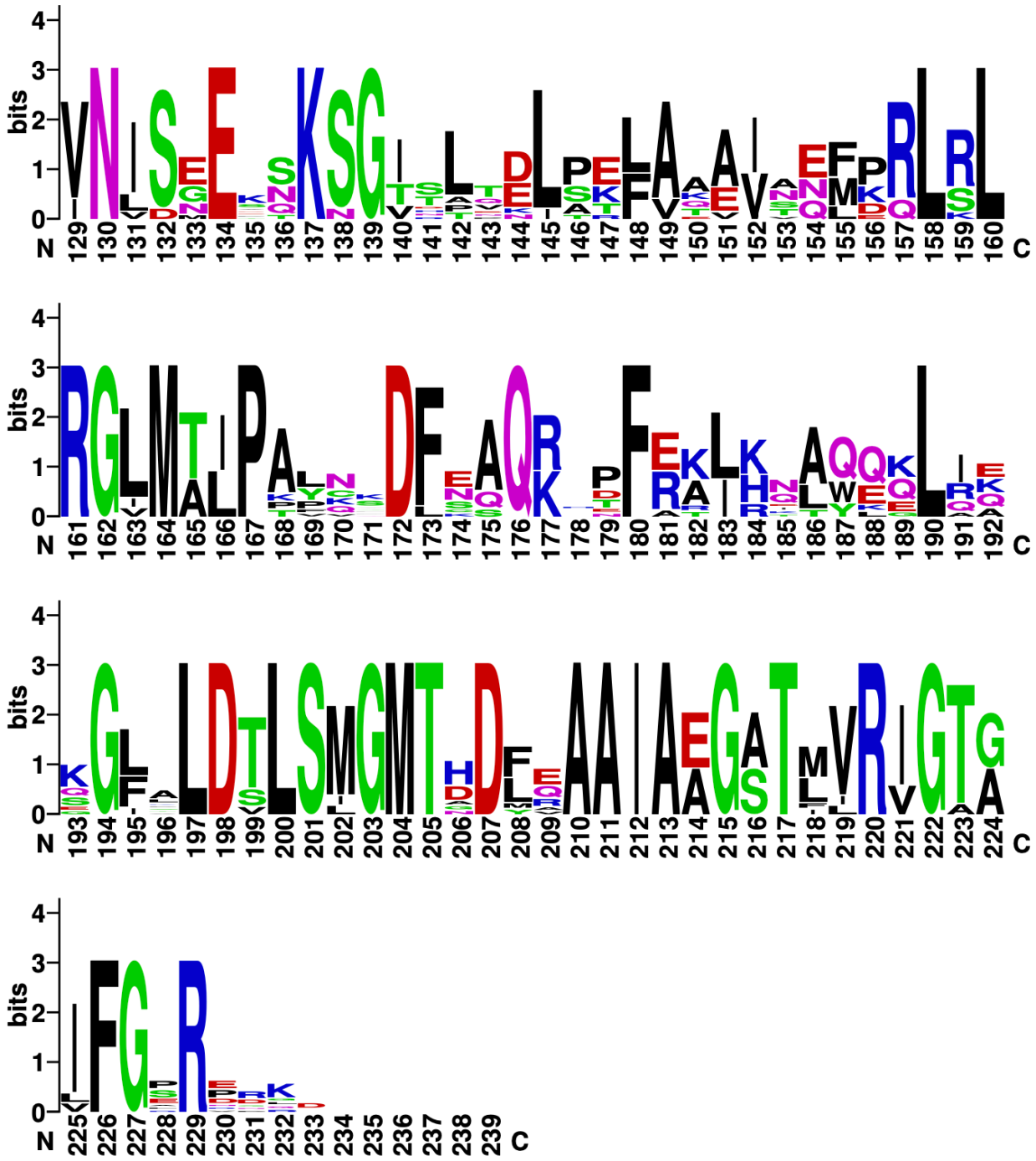
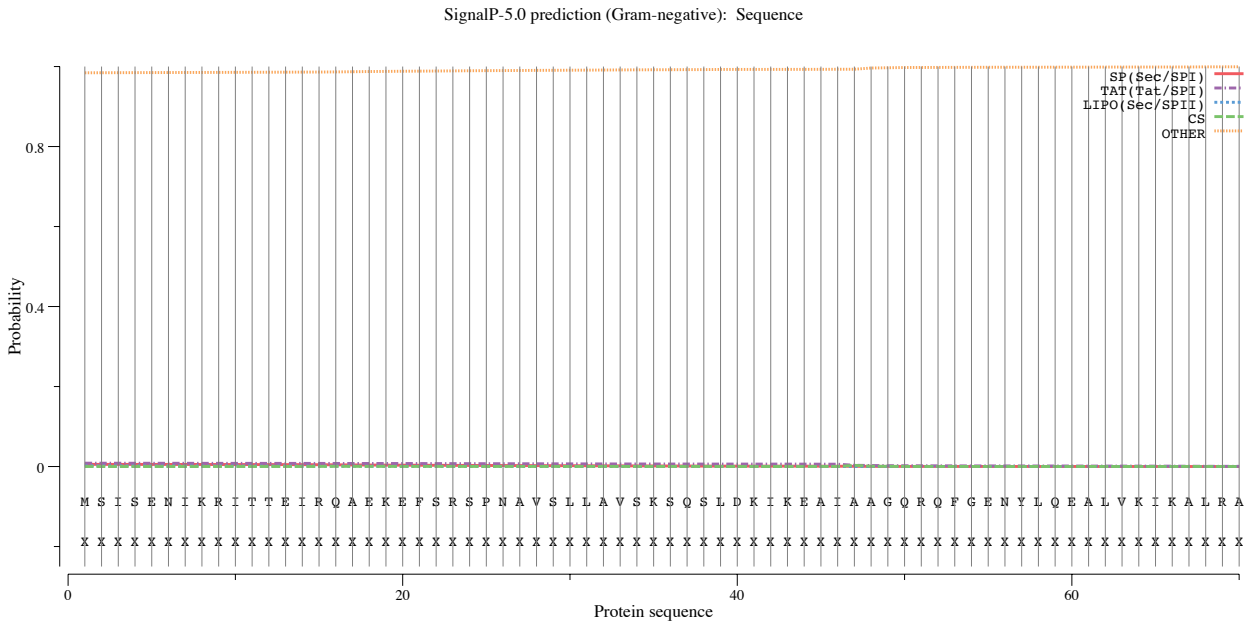


Figure 92: Sequence logo generated from the T-COFFEE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10835 and the top 10 organism sequences similarities

selected. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. The conservation of each amino acid among residue across similar sequences is represented by the height of each single letter abbreviation of the amino acid. (WebLogo, <https://weblogo.berkeley.edu/logo.cgi>).

Protein Localization

SignalP



Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.0052	0.0073	0.0015	0.986

Figure 93: SignalP 5.0 prediction (Gram-negative) for gene BMW92_RS10835 of *Coxiella burnetii*. The SP (Sec/SPI), TAT (Tat/SPI), LIPO (Sec/SPII), and CS probability scores combined were all less than a total 2.0 (2%) which results in the likelihood of the protein being a signal peptide as highly unlikely and can confirm there is no signal peptide of these protein types. The program calculated the probability scores for OTHER as 0.986 (98.6%). This probability score indicates the protein from the gene BMW92_RS10835 has another protein classification that is not related to similar function or type as a signal peptide (SignalP, <http://www.cbs.dtu.dk/services/SignalP/ >).

LipoP

```
# Sequence CYT score=-0.200913
# Cut-off=-3
Sequence        LipoP1.0:Best   CYT     1       1       -0.200913


# NO PLOT made - less than 4 putative cleavage sites predicted
```

Figure 94: LipoP 1.0 was unable to generate a plot graph due to there being less than four predicted putative cleavage sites. The best localization prediction resulted in the highest scoring class being the cytoplasmic protein class (LipoP, <http://www.cbs.dtu.dk/services/LipoP/>).

TMHMM

```
# WEBSEQUENCE Length: 228
# WEBSEQUENCE Number of predicted TMHs:  0
# WEBSEQUENCE Exp number of AAs in TMHs: 0.19731
# WEBSEQUENCE Exp number, first 60 AAs:  0
# WEBSEQUENCE Total prob of N-in:        0.13560
WEBSEQUENCE     TMHMM2.0        outside      1   228
```

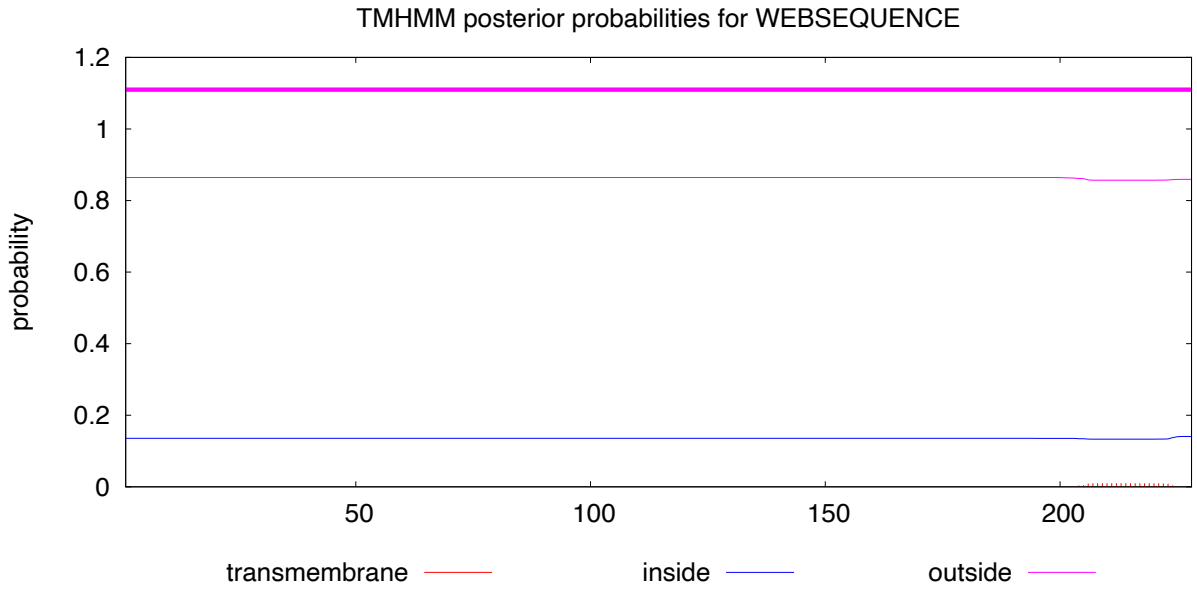


Figure 95: TMHMM posterior probability displayed a line graph that predicts the localization of the protein coded from BMW92_RS10835 as entirely outside the membrane. The red line, representative of the protein being located in the transmembrane, was less than 0.002 (0.20% probability) across the entirety of the line graph. This is indicative of the protein being located within the transmembrane as highly unlikely. The blue line, representative of the protein being located inside the membrane, was at 0.175 (17.55% probability). This is indicative of the protein being located inside of the membrane as unlikely. The magenta line, representative of the protein being located outside the membrane, was at 0.85 (85% probability). This is indicative of the

protein being located outside of the membrane as highly likely (TMHMM, <http://www.cbs.dtu.dk/services/TMHMM/ >).

BOMP

The total number of valid proteins submitted is: 1
The total number of integral ß-barrel outer membrane proteins predicted is: 0

Sequence name	Category	Best BLAST hit

Figure 96: The BOMP test result identified there are no integral beta-barrel outer membrane proteins for gene BMW92_RS10835 (BOMP, <http://services.cbu.uib.no/tools/bomp>).

PSORTb

```
SeqID: C.burnetii
  Analysis Report:
    CMSVM-              Unknown                     [No details]
    CytoSVM-            Cytoplasmic                 [No details]
    ECSVM-              Unknown                     [No details]
    ModHMM-             Unknown                     [No internal helices found]
    Motif-              Unknown                     [No motifs found]
    OMPMotif-           Unknown                     [No motifs found]
    OMSVM-              Unknown                     [No details]
    PPSVM-              Unknown                     [No details]
    Profile-            Unknown                     [No matches to profiles found]
    SCL-BLAST-          Cytoplasmic                 [matched 15595591: Cytoplasmic protein]
    SCL-BLASTe-         Unknown                     [No matches against database]
    Signal-             Unknown                     [No signal peptide detected]
  Localization Scores:
    Cytoplasmic             9.97
    CytoplasmicMembrane     0.01
    Periplasmic             0.01
    OuterMembrane           0.00
    Extracellular           0.00
  Final Prediction:
    Cytoplasmic             9.97
```

Figure 97: The PSORTb test resulted in an analysis report that identified no detectable internal helices, motifs, or signal peptides. The PSORTb localization scores resulted in a 9.97 value for the cytoplasmic location. The localization score for cytoplasmic membrane was 0.01. The localization score for periplasmic was 0.01. The localization score for the outer membrane location was 0.00. The localization score for the extracellular location was 0.00. The calculated localization scores for gene BMW92_RS10835 resulted in the final predictable location of the protein to be cytoplasmic (PSORTb, <https://www.psort.org/psortb/ >).

Phobius

```
ID   UNNAMED
FT   TOPO_DOM      1    228       NON CYTOPLASMIC.
//
```

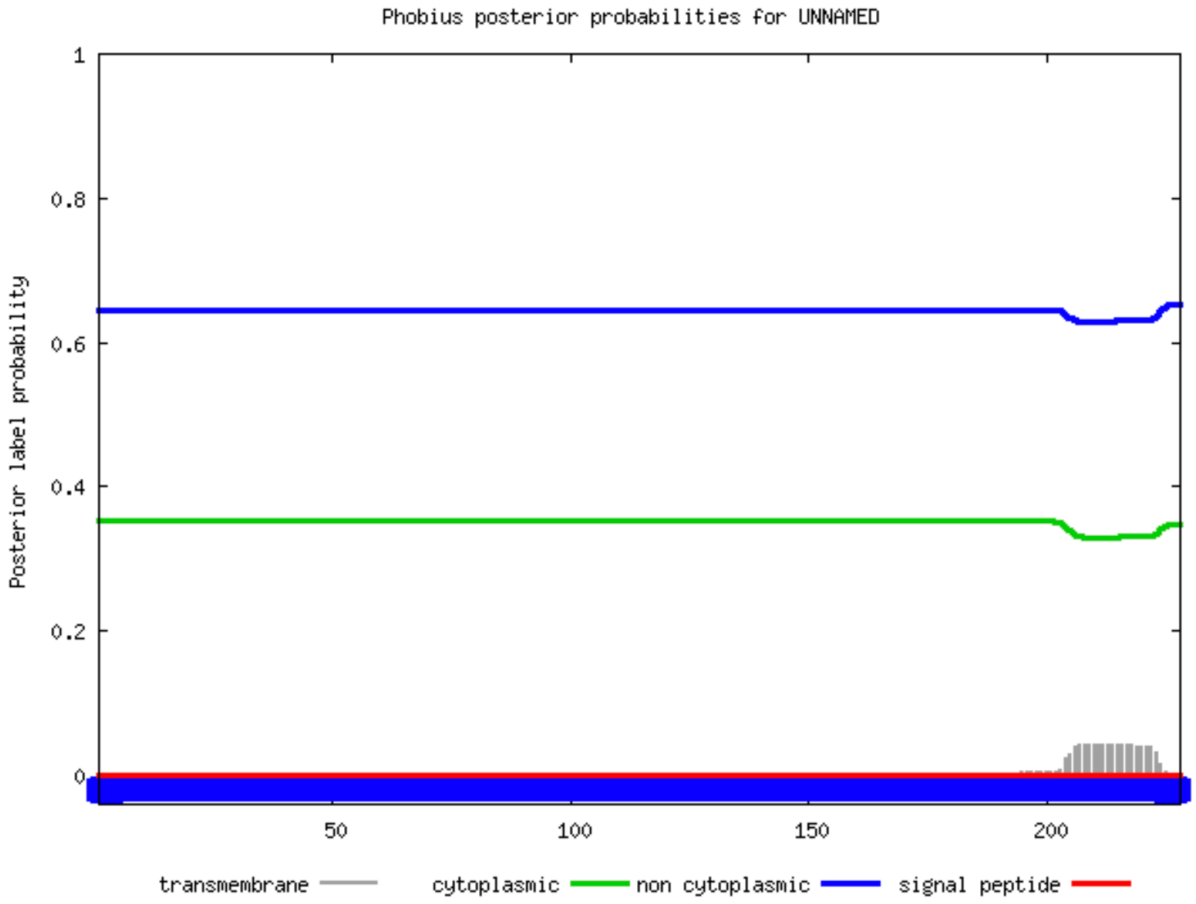


Figure 98: The Phobius posterior probability line graph generated for gene BMW92_RS10835 resulted in a calculated prediction that the whole sequence contains no membrane helices. The grey line, representative of the predicted transmembrane helices location, was around 0.06 (6%) posterior probability from amino acids 190-224. The green line, representative of the predicted cytoplasmic transmembrane helices location, was around 0.35 (35%) posterior probability from amino acids 0-228; this changed to a posterior probability of 0.32 (32%) from amino acids 200-

228. The blue line, representative of the predicted non-cytoplasmic transmembrane helices location, was around 0.64 (64%) posterior probability from amino acids 0-200; the posterior probability changed to 0.62 (62%) from amino acids 201-228. The red line, representative of the presence or absence of a signal peptide, was 0.00 (0%) posterior probability (Phobius, <http://phobius.sbc.su.se>).

BMW92_RS10840

The fourth gene, BMW92_RS10840, was analyzed using bioinformatic technology. Table 4 below contains the provided data regarding basic information. A protein isoelectric point calculator was used to determine the isoelectric point of the protein, protein length, and the number and prevalence of each amino acid that makes up the protein (Figure 99). The BLASTp search tool produced 100 matches ranked from highest sequence similarity to lowest sequence similarity. The top ten sequences with significant alignments that were not identical species to *Coxiella burnetii* were selected. The information recorded included the organism name, protein name, percent identity, precent positive, length of alignment match, e-values, and percent gap. The highest ranked match to the BMW92_RS10840 gene was phosphoenolpyruvate carboxykinase [*Coxiella mudrowiae*] (Figure 100). The remaining nine matches to the BMW92_RS10840 gene all had a function as phosphoenolpyruvate carboxykinase (Figures 101-109). The CDD identified five potential protein domains hits conserved (Figure 110). Five of the domain hits conserved and identified by the CDD belong to the PEPCK_HprK superfamily (Figure 111). Specific domain hits involved the PEPCK_ATP, PRK09344, and PckA. One domain hit conserved and identified as a non-specific domain hit was pckA. The protein classification identified by the CDD was phosphoenolpyruvate carboxykinase. Four of the domain hits sequences were aligned with the query sequence based off the amino acids that are highly conserved between both sequences (Figures 112-115). The MUSCLE program generated a multiple sequence alignment (MSA); each amino acid in the sequence was assigned a distinct color to distinguish the amino acids being compared (Figure 116). The MUSCLE program generated two phylogenetic trees using the multiple sequence alignments to further confirm sequence similarity. The results displayed the numbers followed behind each organism at the end

of each leaf node which displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii.* The use of a phylogenetic cladogram (Figure 117) and real phylogenetic tree (Figure 118) provided further understanding of the relatedness of common ancestors and organism sequences that are conserved. Each of the letter's heights produced correspond to the conservation of the amino acid residue across similar sequences. WebLogo produced a sequence logo that was generated from the MUSCLE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10840 and the top 10 organism sequences similarities selected (Figure 119). Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. Each of the single letter amino acid abbreviation's heights correspond to the conservation of the amino acid residue across similar sequences. The T-COFEE program generated another multiple sequence alignment to further confirm sequence similarity depicted with in the MUSCLE MSA (Figure 120). The T-COFFEE program generated two phylogenetic trees, phylogenetic cladogram (Figure 121) and real phylogenetic tree (Figure 122), using the multiple sequence alignment which displayed the genetic proximity and similarity between *Coxiella burnetii* and selected organisms from the BLASTp search. WebLogo constructed a sequence logo from the T-COFFEE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10840 and the top 10 organism sequences similarities selected to further display sequence similarity and conservation of sequences. Each of the single letter amino acid abbreviation's heights correspond to the conservation of the amino acid residue across similar sequences (Figure 123). Protein localization results included SignalP, LipoP, TMHMM, BOMP, PSORTb, and Phobius. The SignalP graphical illustration identified that there is no presence of a signal peptide for the entirety of the protein sequence (Figure 124). The LipoP resulted in the highest scoring class being the cytoplasmic protein class (Figure 125).

The TMHMM test resulted in a graphical illustration, statistics, and a list of the predicted transmembrane helices and the predicted location of the intervening loop regions. The TMHMM test resulted and displayed that the whole sequence is highly unlikely to contain any transmembrane helices and that the majority of the protein has a high probability of being located outside of the membrane (Figure 126). The BOMP test result identified there are no integral beta-barrel outer membrane proteins (Figure 127). The PSORTb test resulted in an analysis report that identified no detectable internal helices, motifs, or signal peptides; the localization scores calculated the predictable location of the protein to be cytoplasmic (Figure 128). The Phobius test resulted in a line graphical illustration that identified a low probability of transmembrane helices localized in the cytoplasm and a high probability of non-cytoplasmic transmembrane helices present; the overall result calculated by Phobius resulted in the entire protein sequence as non-cytoplasmic, which is contradictory to LipoP, TMHMM and PSORTb results (Figure 129).

Basic Information

Table 4: Gene BMW92_RS10840 basic information

Genome	Replicon	Locus Tag	Old Locus Tag
Coxiella burnetii	NZ_CP018005	BMW92_RS10840	BMW92_10470
Genomic Coordinates	**Products**	**Length**	**Start and End Position**
1965588..1967141	phosphoenolpyruvate carboxykinase	1554 / 517	1965588 - 1967141
Molecular Weight	**Average Isoelectric Point**	**IPC Protein**	**Protein Length**
56807.48644 Da	5.82	5.60	519 amino acids

Nucleotide Sequence	Amino Acid Sequence
atggagcaaattgctgcgcgagttacctatattaacctttctcc tgatgagttgattcaacacgccgtaaaaaatggcgagggcgt attaagttccaccggtgctttagcggttactactgggaaacgc acgggtcgatcgccgaaagatcgttttattgtcaaagatgagc aaaccgccgatcaagtggcgtggggcaatatcaatcagcctg ttgagcaacgcacctttgaccagttgtgggagcgagcgctgcg gtatctttctgaacgtgctgtttatatttcgcatttgcaagtagg ggcggatgataattattttctgccacttaaggtggtcaccgagt ttgcgtggcacaatttatttgcgtgtgatctttttatccgtccttc tggtgatcatgcgaatgggaaaccgtcctgggttattttaagt gccccgggctgaaaactgatcctgagcgagacggcgtgaat agtgatggtgcggtaatgattaatttatcacagcgccgtgtgtt attggtgggcatgccctatgcgggtgaaatgaaaaaagccat gttttccgtgctgaattatcttttgccgccgcacgatgttttacc gatgcattgcgccgctaatgctggccagtcgggcgatgttgca ctatttttcggattatcaggaacgggtaagaccaccttgtcggc tgaccctcatcgatttttaatcggtgacgacgaacacggttgg agcgccacaagcgtttttaattttgagggcgggtgttatgcca agtgcattgatttgtcacaagaacgagagcccatgatttggaa tgcgattcggcacggcgctattatggaaaatgtggttttagat gagaatggcgttcccgattatgcggatgcgcggctaacccaa aattcgcgtgccgcttatccgcgcgagtatattccgttgcgggt ggaaaataatagagggcgcccccccgatgccgtcttatttcta acttgcgatctcgatggtgttttgccgcccgtggcactgctcac gaaagaacaagcggcttattatttttaagcgggtataccgct ttagtgggcagcacggaagtgggcagcgtaaagggcgtcac	MEQIAARVTYINLSPDELIQHAVKNGEGVLSSTGAL AVTTGKRTGRSPKDRFIVKDEQTADQVAWGNINQ PVEQRTFDQLWERALRYLSERAVYISHLQVGADD NYFLPLKVVTEFAWHNLFACDLFIRPSGDHANGKP SWVILSAPGLKTDPERDGVNSDGAVMINLSQRRV LLVGMPYAGEMKKAMFSVLNYLLPPHDVLPMHC AANAGQSGDVALFFGLSGTGKTTLSADPHRFLIGD DEHGWSATSVFNFEGGCYAKCIDLSQEREPMIWN AIRHGAIMENVVLDENGVPDYADARLTQNSRAAY PREYIPLRVENNRGRPPDAVLFLTCDLDGVLPPVAL LTKEQAAYYFLSGYTALVGSTEVGSVKGVTSTFSTC FGAPFFPRPPTVYAELLMKRIEATGCQVYLVNTGW TGGAYGEGGERFSIPTTRAIVNAVLSGKLKEGPTEV LSGFNLTIPKSALGVDDHLLNPRKTWEDVSAYDAR AQRLIQKFRENFEKFKVL AAIREAGPSDVH

ctccaccttcagtacttgctttggcgcacccttttttccacgccc tccgactgtctatgctgaattattaatgaaacgtattgaagca acgggctgtcaagtttacctcgttaatactggctggacagggg gcgcttatggtgaaggaggtgagcgtttttccattcccacgac acgagcgattgttaacgctgttctaagcggaaaactcaaaga gggaccaacagaagtgttgagcggctttaatctcaccattcca aaatcggctttaggtgtggacgatcatttattaaatccccgga agacttgggaagatgttagcgcctacgatgcgcgagcccagc ggttaattcaaaaattccgtgaaaattttgaaaaatttaaagt gcttgctgccattcgggaagccggaccgtctgatgtccattag	

Figure 99: Protein isoelectric point calculator. The number and prevalence of each amino acid in the protein coded from the BMW92_RS10840 gene of *Coxiella burnetii* (Kozlowski, Biology Direct, <http://isoelectric.org/>).

Sequence Similarity

BLAST

phosphoenolpyruvate carboxykinase [Candidatus Coxiella mudrowiae]

Sequence ID: WP_048875732.1 **Length:** 516 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 513 GenPept Graphics Next Match Previou

Score	Expect	Method	Identities	Positives	Gaps
875 bits(2261)	0.0	Compositional matrix adjust.	414/513(81%)	456/513(88%)	0/513(0%)

```
Query  1    MEQIAARVTYINLSPDELIQHAVKNGEGVLSSTGALAVTTGKRTGRSPKDRFIVKDEQTA  60
            M+QIA+R  Y +LS DELIQ A+K GEG LSSTGALAVTTGKRTGRSPKDRFIVKD +TA
Sbjct  1    MDQIASRTVYTDLSVDELIQQALKKGEGKLSSTGALAVTTGKRTGRSPKDRFIVKDAETA  60

Query  61   DQVAWGNINQPVEQRTFDQLWERALRYLSERAVYISHLQVGADDNYFLPLKVVTEFAWHN  120
            DQV WGN+NQ + Q  FDQLW RA  YLS+R +Y+SHLQVGAD+NYFLP++V+TEF WHN
Sbjct  61   DQVQWGNVNQSIVQGVFDQLWNRANAYLSKRPMYVSHLQVGADENYFLPVQVITEFGWHN  120

Query  121  LFACDLFIRPSGDHANGKPSWVILSAPGLKTDPERDGVNSDGAVMINLSQRRVLLVGMPY  180
            LFACDLFIRP GD+A GKP W+ILS PGLKTDPERD VNSD AV+INLSQRRVLLVGM Y
Sbjct  121  LFACDLFIRPDGDYAKGKPEWIILSVPGLKTDPERDKVNSDAAVIINLSQRRVLLVGMAY  180

Query  181  AGEMKKAMFSVLNYLLPPHDVLPMHCAANAGQSGDVALFFGLSGTGKTTLSADPHRFLIG  240
            AGE+KKAMF+VLNYLLPPHDVLPMHCAANAG+SGDVALFFGLSGTGKTTLSADP+RFLIG
Sbjct  181  AGEIKKAMFTVLNYLLPPHDVLPMHCAANAGKSGDVALFFGLSGTGKTTLSADPNRFLIG  240

Query  241  DDEHGWSATSVFNFEGGCYAKCIDLSQEREPMIWNAIRHGAIMENVVLDENGVPDYADAR  300
            DDEHGWS T VFNFEGGCYAKCIDLS EREPMIW AIRHGAIMENVVL  +G PDY +A
Sbjct  241  DDEHGWSRTGVFNFEGGCYAKCIDLSSEREPMIWEAIRHGAIMENVVLQADGQPDYRNAS  300

Query  301  LTQNSRAAYPREYIPLRVENNRGRPPDAVLFLTCDLDGVLPPVALLTKEQAAYYFLSGYT  360
            LTQN+RAAYPRE+I LRV++NRGRPPD+V+FLTCDL GVLPPVALLTKEQAAYYFLSGYT
Sbjct  301  LTQNTRAAYPREHISLRVKDNRGRPPDSVIFLTCDLYGVLPPVALLTKEQAAYYFLSGYT  360

Query  361  ALVGSTEVGSVKGVTSTFSTCFGAPFFPRPPTVYAELLMKRIEATGCQVYLVNTGWTGGA  420
            ALVGSTEVGSVKGVT TFSTCFGAPFFPRPPTVYAELLMKRIE T CQVYLVNTGWTGGA
Sbjct  361  ALVGSTEVGSVKGVTPTFSTCFGAPFFPRPPTVYAELLMKRIEETQCQVYLVNTGWTGGA  420

Query  421  YGEGGERFSIPTTRAIVNAVLSGKLKEGPTEVLSGFNLTIPKSALGVDDHLLNPRKTWED  480
            YGEGG RFSIPTTRAI++A+L+ KL+  PTE L GFNL IPKSA GV+D +LNPR+ W D
Sbjct  421  YGEGGVRFSIPTTRAIIDAILTRKLRNQPTENLKGFNLAIPKSAPGVEDKILNPRQAWTD  480

Query  481  VSAYDARAQRLIQKFRENFEKFKVLAAIREAGP  513
            V AYD +A  LI+KFRENF KF+V  AI++AGP
Sbjct  481  VRAYDIKALTLIEKFRENFVKFQVTDAIQKAGP  513
```

Figure 100: BLAST first match for BMW92_RS10840 sequence from organism *Coxiella mudrowiae* with an e-value of 0.0, 81% identity, 88% positives, 0% gaps, and an identity of phosphoenolpyruvate carboxykinase (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

phosphoenolpyruvate carboxykinase [Coxiella endosymbiont of Rhipicephalus microplus]

Sequence ID: WP_102156648.1 **Length:** 516 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 513 GenPept Graphics ▼ Next Match ▲ Previous Ma

Score	Expect	Method	Identities	Positives	Gaps
824 bits(2129)	0.0	Compositional matrix adjust.	384/513(75%)	438/513(85%)	0/513(0%)

```
Query  1    MEQIAARVTYINLSPDELIQHAVKNGEGVLSSTGALAVTTGKRTGRSPKDRFIVKDEQTA  60
            MEQI +R  Y +L+ DELIQHA+K GEG LS TGALAV TGKRTGRSP+DRFIVKD +T
Sbjct  1    MEQIVSRTVYTDLAIDELIQHALKKGEGTLSVTGALAVRTGKRTGRSPQDRFIVKDSETE  60

Query  61   DQVAWGNINQPVEQRTFDQLWERALRYLSERAVYISHLQVGADDNYFLPLKVVTEFAWHN  120
            DQV WG++NQP+ Q  FDQLW RA  Y+S+R++Y+SHL+VGAD+NY +P++V+TE AWHN
Sbjct  61   DQVQWGDVNQPIVQVVFDQLWNRATAYISKRSMYVSHLKVGADENYSIPVQVITELAWHN  120

Query  121  LFACDLFIRPSGDHANGKPSWVILSAPGLKTDPERDGVNSDGAVMINLSQRRVLLVGMPY  180
            LFAC+LFIRP  D+  GKP W++LS PGL TDP+RD VNSD AV+INLSQRRVLLVGM Y
Sbjct  121  LFACELFIRPDRDYFKGKPKWILLSVPGLTTDPKRDKVNSDAAVIINLSQRRVLLVGMSY  180

Query  181  AGEMKKAMFSVLNYLLPPHDVLPMHCAANAGQSGDVALFFGLSGTGKTTLSADPHRFLIG  240
            AGEMKKAMF+VLNYLLPP DVLPMHCAAN G+SGDVALFFGLSGTGKTTLSADP+RFLIG
Sbjct  181  AGEMKKAMFTVLNYLLPPQDVLPMHCAANTGKSGDVALFFGLSGTGKTTLSADPNRFLIG  240

Query  241  DDEHGWSATSVFNFEGGCYAKCIDLSQEREPMIWNAIRHGAIMENVVLDENGVPDYADAR  300
            DDEHGWS T VFNFEGGCYAKCIDLS EREP+IW +IR+GAIMENVVL  +G P Y DA
Sbjct  241  DDEHGWSRTGVFNFEGGCYAKCIDLSLEREPIIWESIRYGAIMENVVLQADGQPAYNDAS  300

Query  301  LTQNSRAAYPREYIPLRVENNRGRPPDAVLFLTCDLDGVLPPVALLTKEQAAYYFLSGYT  360
            LTQN+RAAYPRE+I  RV+ NRGRPPDAV+FLTCDL GVLPPV+LLTK QAAYYFLSGYT
Sbjct  301  LTQNTRAAYPREHILFRVKENRGRPPDAVIFLTCDLYGVLPPVSLLTKAQAAYYFLSGYT  360

Query  361  ALVGSTEVGSVKGVTSTFSTCFGAPFFPRPPTVYAELLMKRIEATGCQVYLVNTGWTGGA  420
            ALVGSTEVGSVKG+  TFS+CFGAPFFPRPP VYA+LLMKRIE T CQVYLVNTGW GGA
Sbjct  361  ALVGSTEVGSVKGIVPTFSSCFGAPFFPRPPVVYAKLLMKRIEETQCQVYLVNTGWMGGA  420

Query  421  YGEGGERFSIPTTRAIVNAVLSGKLKEGPTEVLSGFNLTIPKSALGVDDHLLNPRKTWED  480
            YGEGG RF IP TR+I++A+L+ KL   PTE L GFNL IP+S  GV+D +LNPRK W D
Sbjct  421  YGEGGVRFRIPITRSIIDAILTRKLINQPTENLKGFNLAIPQSVPGVEDKVLNPRKAWSD  480

Query  481  VSAYDARAQRLIQKFRENFEKFKVLAAIREAGP  513
            + AYD +A  LI+KF+ENF KF+V  AIREAGP
Sbjct  481  LKAYDIKAFSLIEKFQENFVKFQVTEAIREAGP  513
```

Figure 101: BLAST second match for BMW92_RS10840 sequence from organism *Rhipicephalus microplus* with an e-value of 0.0, 75% identity, 85% positives, 0% gaps, and an identity of phosphoenolpyruvate carboxykinase (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

phosphoenolpyruvate carboxykinase (ATP) [Legionellales bacterium]

Sequence ID: MBB71107.1 **Length:** 516 **Number of Matches:** 1

Range 1: 4 to 515 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
733 bits(1893)	0.0	Compositional matrix adjust.	342/512(67%)	410/512(80%)	0/512(0%)

```
Query  2    EQIAARVTYINLSPDELIQHAVKNGEGVLSSTGALAVTTGKRTGRSPKDRFIVKDEQTAD  61
            E  A  +TY +LS ++LI+HA++  EGVLS+  AL+V TG RTGRSP+DRFIV+D+ T +
Sbjct  4    ETKAETMTYTDLSTEQLIKHALERNEGVLSTNQALSVATGTRTGRSPRDRFIVQDDVTTN  63

Query  62   QVAWGNINQPVEQRTFDQLWERALRYLSERAVYISHLQVGADDNYFLPLKVVTEFAWHNL  121
             V WGN+NQP+ Q  FD LW +   YL+++  ++SHL+VGAD  ++LP+KV+ + AWHNL
Sbjct  64   TVDWGNVNQPISQDRFDALWNQIEAYLADKDTFVSHLEVGADSEHYLPVKVINQKAWHNL  123

Query  122  FACDLFIRPSGDHANGKPSWVILSAPGLKTDPERDGVNSDGAVMINLSQRRVLLVGMPYA  181
            F  +LFIRP   +   KP W ILSAP     PERDG NS+ AV++N SQRR+L+ G  YA
Sbjct  124  FTRNLFIRPDTYNRKQKPEWTILSAPDFHASPERDGTNSEAAVILNFSQRRILVCGTHYA  183

Query  182  GEMKKAMFSVLNYLLPPHDVLPMHCAANAGQSGDVALFFGLSGTGKTTLSADPHRFLIGD  241
            GEMKKAMF+V+N+LLP  DVLPMHCA+N G  GDVALFFGLSGTGKTTLSADP RFLIGD
Sbjct  184  GEMKKAMFTVMNFLLPNIDVLPMHCASNIGMEGDVALFFGLSGTGKTTLSADPERFLIGD  243

Query  242  DEHGWSATSVFNFEGGCYAKCIDLSQEREPMIWNAIRHGAIMENVVLDENGVPDYADARL  301
            DEHGW  + VFNFEGGCYAKCIDLS+E+EP+IW+AIRHGAIMENVVLDEN  PDY+D+ L
Sbjct  244  DEHGWGKSGVFNFEGGCYAKCIDLSKEKEPVIWDAIRHGAIMENVVLDENQAPDYSDSTL  303

Query  302  TQNSRAAYPREYIPLRVENNRGRPPDAVLFLTCDLDGVLPPVALLTKEQAAYYFLSGYTA  361
            + NSRAAYPRE+I +R E NRG  PDAVLFLTCDL GVLPPV+LL+KEQAAY+FLSGYTA
Sbjct  304  SMNSRAAYPREHIEMRAEANRGGQPDAVLFLTCDLYGVLPPVSLLSKEQAAYHFLSGYTA  363

Query  362  LVGSTEVGSVKGVTSTFSTCFGAPFFPRPPTVYAELLMKRIEATGCQVYLVNTGWTGGAY  421
            LVGSTEVG  +G+  TFSTCFGAPFFP  P+VYAELL+KRIE TG QVYLVNTGWTGGAY
Sbjct  364  LVGSTEVGQTEGIKPTFSTCFGAPFFPLSPSVYAELLIKRIEETGAQVYLVNTGWTGGAY  423

Query  422  GEGGERFSIPTTRAIVNAVLSGKLKEGPTEVLSGFNLTIPKSALGVDDHLLNPRKTWEDV  481
            G+GGERFSIPTTRAIV A+LSG LK+  T  L GFNL IP++  GVD  LLNP KTW D
Sbjct  424  GQGGERFSIPTTRAIVRAILSGALKDANTITLPGFNLAIPETINGVDSQLLNPVKTWSDS  483

Query  482  SAYDARAQRLIQKFRENFEKFKVLAAIREAGP  513
            +AY+A+   L ++FRENF++F V   I +AGP
Sbjct  484  TAYEAKLMELSEQFRENFKRFDVAPEIVKAGP  515
```

Figure 102: BLAST third match for BMW92_RS10840 sequence from organism *Legionellales bacterium* with an e-value of 0.0, 67% identity, 80% positives, 0% gaps, and an identity of phosphoenolpyruvate carboxykinase (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

phosphoenolpyruvate carboxykinase [Aquicella lusitana]

Sequence ID: WP_114834947.1 **Length:** 524 **Number of Matches:** 1

See 2 more title(s) See all Identical Proteins(IPG)

Range 1: 15 to 522 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
715 bits(1845)	0.0	Compositional matrix adjust.	346/508(68%)	404/508(79%)	3/508(0%)

```
Query  10   YINLSPDELIQHAVKNGEGVLSSTGALAVTTGKRTGRSPKDRFIVKDEQTADQVAWGNIN  69
            +INLS +EL++ A+  GEG L+S  AL V TG RTGRSPKDRFIV+DE T +QV W  IN
Sbjct  15   HINLSAEELVEIALARGEGELASNQALVVKTGARTGRSPKDRFIVRDEITENQVDWNTIN  74

Query  70   QPVEQRTFDQLWERALRYLSER-AVYISHLQVGADDNYFLPLKVVTEFAWHNLFACDLFI  128
            QP+    F+ LW++A  YL  R A +IS L+VGA +   +P+KV+TE AWHNLFA  LFI
Sbjct  75   QPISPEKFNALWQKAQDYLDTRDAHFISFLKVGAHEELGVPVKVITELAWHNLFARVLFI  134

Query  129  RPSGDHANGKPS-WVILSAPGLKTDPERDGVNSDGAVMINLSQRRVLLVGMPYAGEMKKA  187
            RP         P+ W ILS PG KTDP RDGVN D AV++N SQRR+L+ G  YAGEMKKA
Sbjct  135  RPEKPATTVVPNQWTILSVPGFKTDPARDGVNGDAAVILNFSQRRILICGTHYAGEMKKA  194

Query  188  MFSVLNYLLPPHDVLPMHCAANAGQSGDVALFFGLSGTGKTTLSADPHRFLIGDDEHGWS  247
            MFSVLN++LP H++LPMHCAANAG++GD ALFFGLSGTGKTTLSADP RFLIGDDEHGW
Sbjct  195  MFSVLNFILPEHNILPMHCAANAGENGDTALFFGLSGTGKTTLSADPERFLIGDDEHGWG  254

Query  248  ATSVFNFEGGCYAKCIDLSQEREPMIWNAIRHGAIMENVVLDE-NGVPDYADARLTQNSR  306
               VFNFEGGCYAKCIDLS+EREP+IWNAIR+G+++ENVVLD     PDY DA LTQN+R
Sbjct  255  NDGVFNFEGGCYAKCIDLSEEREPLIWNAIRYGSVIENVVLDPVTKNPDYGDASLTQNTR  314

Query  307  AAYPREYIPLRVENNRGRPPDAVLFLTCDLDGVLPPVALLTKEQAAYYFLSGYTALVGST  366
            AAYPRE+IP RVENNRGR P+AVLFLTCDL GVLPPVA LT EQAAYYFLSGYTALVGST
Sbjct  315  AAYPREFIPQRVENNRGRQPNAVLFLTCDLYGVLPPVARLTPEQAAYYFLSGYTALVGST  374

Query  367  EVGSVKGVTSTFSTCFGAPFFPRPPTVYAELLMKRIEATGCQVYLVNTGWTGGAYGEGGE  426
            EVG   G+  TFSTCFGAPFFPRPP VYAELLMKR++    QVYLVNTGW+GGA+GEGG+
Sbjct  375  EVGQGSGIKPTFSTCFGAPFFPRPPRVYAELLMKRLQNFDTQVYLVNTGWSGGAHGEGGK  434

Query  427  RFSIPTTRAIVNAVLSGKLKEGPTEVLSGFNLTIPKSALGVDDHLLNPRKTWEDVSAYDA  486
            RFSIPTTRA+V A+++GKLK+   E L GFN  IPK+  GV+  LLNPRKTW D +A+D
Sbjct  435  RFSIPTTRAVVTAIVNGKLKDAEYEKLPGFNFDIPKAVDGVESKLLNPRKTWNDTAAHDK  494

Query  487  RAQRLIQKFRENFEKFKVLAAIREAGPS  514
             A+ LI++F ENF++F V  AIR AGPS
Sbjct  495  YARILIEQFIENFKRFNVSEAIRNAGPS  522
```

Figure 103: BLAST fourth match for BMW92_RS10840 sequence from organism *Aquicella lusitana* with an e-value of 0.0, 68% identity, 79% positives, 0% gaps, and an identity of phosphoenolpyruvate carboxykinase (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

phosphoenolpyruvate carboxykinase [Coxiellaceae bacterium]

Sequence ID: QLH44014.1 Length: 514 Number of Matches: 1

Range 1: 8 to 514 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
692 bits(1786)	0.0	Compositional matrix adjust.	330/508(65%)	404/508(79%)	3/508(0%)

```
Query  10   YINLSPDELIQHAVKNGEGVLSSTGALAVTTGKRTGRSPKDRFIVKDEQTADQVAWGNIN  69
            +++LS  ELI+ A++  EGVLS+  AL V TGKRTGRSPKDRFIVKDE TAD V WGN+N
Sbjct  8    HVDLSVAELIEMALEREEGVLSANQALVVATGKRTGRSPKDRFIVKDELTADTVDWGNVN  67

Query  70   QPVEQRTFDQLWERALRYLSERAVYISHLQVGADDNYFLPLKVVTEFAWHNLFACDLFIR  129
            QP +   F  LW+RA +Y++++ V++SHL VGAD  +F+P+ V++E+AWHN+F  DLFIR
Sbjct  68   QPFDPAKFTVLWQRAEQYMADQEVFVSHLGVGADIEHFVPVTVISEYAWHNVFVHDLFIR  127

Query  130  PSGDHANGKPSWVILSAPGLKTDPERDGVNSDGAVMINLSQRRVLLVGMPYAGEMKKAMF  189
            P+G + +G+  W IL+A GL TDP RDG NS+  +++N  ++++LL G+ YAGEM KAMF
Sbjct  128  PNGRYPHGRAGWTILNAAGLPTDPARDGTNSEATLILNFKEKKILLCGLRYAGEMXKAMF  187

Query  190  SVLNYLLPPHDVLPMHCAANAGQSGDVALFFGLSGTGKTTLSADPHRFLIGDDEHGWSAT  249
            SVLN++LP  +VLPMHCAAN G+ GDVALFFGLSGTGKTTLSADP R+LIGDDEHGWS
Sbjct  188  SVLNFILPEKNVLPMHCAANVGKQGDVALFFGLSGTGKTTLSADPERYLIGDDEHGWSDH  247

Query  250  SVFNFEGGCYAKCIDLSQEREPMIWNAIRHGAIMENVVLD-ENGVPDYADARLTQNSRAA  308
             VFNFEGGCYAKCI+LS+EREP+IW+AIR+GAIMENVVLD +   P Y DA LT+N+RAA
Sbjct  248  GVFNFEGGCYAKCINLSKEREPVIWDAIRYGAIMENVVLDPKTKEPLYGDASLTENTRAA  307

Query  309  YPREYIPLRVENNRGRPPDAVLFLTCDLDGVLPPVALLTKEQAAYYFLSGYTALVGSTEV  368
            YP E+I +RV  N+   P AV+FLTCDL GVLPPVA+L KEQAAY+FLSGYTALVGSTEV
Sbjct  308  YPLEHIAMRVPENQAGHPQAVIFLTCDLYGVLPPVAILNKEQAAYHFLSGYTALVGSTEV  367

Query  369  GSVKGVTSTFSTCFGAPFFPRPPTVYAELLMKRIEATGCQVYLVNTGWTGGAYGEGGERF  428
            GS  G+ STFSTCFGAPFFPRP  VYA+LL+KR+  TG QVYLVNTGWTGG YGE G+RF
Sbjct  368  GSTAGIKSTFSTCFGAPFFPRPAQVYADLLIKRLTETGAQVYLVNTGWTGGPYGE-GKRF  426

Query  429  SIPTTRAIVNAVLSGKLKEGPTEVLSGFNLTIPKSALGVDDHLLNPRKTWEDVSAYDARA  488
             IPTTRA++ A+L+GKLK  PTEV+ GFNL IPK    V+  LLNP  TW +  AY A
Sbjct  427  DIPTTRAVIRAILTGKLKHVPTEVMPGFNLVIPKEVPDVETRLLNPINTWNNHQAYQASM  486

Query  489  QRLIQKFRENFEKFK-VLAAIREAGPSD  515
            + L+ KF ENF KFK V  AIR+AGP++
Sbjct  487  KELMDKFTENFXKFKHVSEAIRKAGPTE  514
```

Figure 104: BLAST fifth match for BMW92_RS10840 sequence from organism *Coxiellaceae bacterium* with an e-value of 0.0, 65% identity, 79% positives, 0% gaps, and an identity of phosphoenolpyruvate carboxykinase (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

phosphoenolpyruvate carboxykinase [Pseudospirillum japonicum]

Sequence ID: WP_093308432.1 **Length:** 520 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 11 to 516 GenPept Graphics Next Match Previou

Score	Expect	Method	Identities	Positives	Gaps
691 bits(1782)	0.0	Compositional matrix adjust.	327/506(65%)	391/506(77%)	1/506(0%)

```
Query  9    TYINLSPDELIQHAVKNGEGVLSSTGALAVTTGKRTGRSPKDRFIVKDEQTADQVAWGNI  68
            TY NLS  +LI+ A++ GEG L+  GAL V TG+RTGRSP DRFIV +  T+D + WG+I
Sbjct  11   TYTNLSNAQLIELAIQRGEGTLADNGALVVATGQRTGRSPMDRFIVNEPSTSDAIDWGSI  70

Query  69   NQPVEQRTFDQLWERALRYLSERAVYISHLQVGADDNYFLPLKVVTEFAWHNLFACDLFI  128
            N+P     FD LWER   YLS++  +IS L VGAD  ++LP++V TE AWHNLF  +LF+
Sbjct  71   NRPFSAEKFDALWERVEEYLSKQDTFISELHVGADPEHYLPIRVTTETAWHNLFGRNLFV  130

Query  129  RPSGDHANGKPSWVILSAPGLKTDPERDGVNSDGAVMINLSQRRVLLVGMPYAGEMKKAM  188
            RP G +   K  W IL+AP    +P RDG NSDG V++N ++R+VLL GM YAGEMKKAM
Sbjct  131  RPEGYNPKSKGEWQILNAPNFVCEPSRDGTNSDGCVILNFAKRKVLLAGMKYAGEMKKAM  190

Query  189  FSVLNYLLPPHDVLPMHCAANAGQSGDVALFFGLSGTGKTTLSADPHRFLIGDDEHGWSA  248
            FSV N+LLP  DVLPMHC+AN G+ G+  LFFGLSGTGKTTLSADP R+LIGDDEHGW
Sbjct  191  FSVQNFLLPEKDVLPMHCSANVGEDGETTLFFGLSGTGKTTLSADPSRYLIGDDEHGWGK  250

Query  249  TSVFNFEGGCYAKCIDLSQEREPMIWNAIRHGAIMENVVLDENGVPDYADARLTQNSRAA  308
             +VFN EGGCYAKCIDLS E EP+IWNAIR GA++ENV+LDE  VPDY D  LTQNSRAA
Sbjct  251  GTVFNIEGGCYAKCIDLSAENEPVIWNAIRFGAVLENVILDERRVPDYNDDSLTQNSRAA  310

Query  309  YPREYIPLRVENNRGRPPDAVLFLTCDLDGVLPPVALLTKEQAAYYFLSGYTALVGSTEV  368
            YP E+I  RV  NR   P A++FLTCD+ GVLPPV++L+KE AAY+FLSGYTA VGSTE+
Sbjct  311  YPLEHIEKRVLENRAGEPSAIVFLTCDMSGVLPPVSILSKEAAAYHFLSGYTAKVGSTEM  370

Query  369  GSVKGVTSTFSTCFGAPFFPRPPTVYAELLMKRIEATGCQVYLVNTGWTGGAYGEGGERF  428
            GS  G+ +TFSTCFGAPFFPRP  VYA+LL+KRIE  G QVYLVNTGWTGGAYG+GG RF
Sbjct  371  GSSSGLEATFSTCFGAPFFPRPAHVYADLLIKRIEEFGSQVYLVNTGWTGGAYGQGGNRF  430

Query  429  SIPTTRAIVNAVLSGKLKEGPTEVLSGFNLTIPKSALGVDDHLLNPRKTWEDVSAYDARA  488
            SIPTTRAI+NAV +G LK+   E L G NL++PK   GV+D LLNPR TWED +AYDA+A
Sbjct  431  SIPTTRAIINAVQTGVLKDAEIEQLPGLNLSVPKHIPGVEDRLLNPRNTWEDTAAYDAQA  490

Query  489  QRLIQKFRENFEKFK-VLAAIREAGP  513
             RL+ +F ENF+KF+ V  AI EAGP
Sbjct  491  ARLVAQFVENFKKFQGVDEAIIEAGP  516
```

Figure 105: BLAST sixth match for BMW92_RS10840 sequence from organism *Pseudospirillum japonicum* with an e-value of 0.0, 65% identity, 77% positives, 0% gaps, and an identity of phosphoenolpyruvate carboxykinase (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

phosphoenolpyruvate carboxykinase [Aquicella siphonis]

Sequence ID: WP_148337466.1 **Length:** 524 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 15 to 522 GenPept Graphics Next Match Previou

Score	Expect	Method	Identities	Positives	Gaps
689 bits(1778)	0.0	Compositional matrix adjust.	334/508(66%)	393/508(77%)	3/508(0%)

```
Query  10   YINLSPDELIQHAVKNGEGVLSSTGALAVTTGKRTGRSPKDRFIVKDEQTADQVAWGNIN  69
            ++NLS  EL++ A+  GEG L+S  AL V TG RTGRSPKDRFIV+ + T  QV W  IN
Sbjct  15   HLNLSAKELVELALARGEGELASNQALVVKTGSRTGRSPKDRFIVRGQATETQVDWNQIN  74

Query  70   QPVEQRTFDQLWERALRYL-SERAVYISHLQVGADDNYFLPLKVVTEFAWHNLFACDLFI  128
            QP+    F+ LWE+AL YL S+ A + S+L+VGA +   + +KV+ E AWH LFA  LFI
Sbjct  75   QPISADKFEALWEKALHYLNSKDARFTSYLKVGAHETLGVSVKVMAELAWHTLFAHVLFI  134

Query  129  RPSGDHANGKPS-WVILSAPGLKTDPERDGVNSDGAVMINLSQRRVLLVGMPYAGEMKKA  187
            RP     + +P+ W ILS PG KTDP RDGVNSD AV+++  + R+L+ G  YAGEMKKA
Sbjct  135  RPVTPPTSDQPNQWTILSTPGFKTDPARDGVNSDAAVILDFEKHRILICGTYYAGEMKKA  194

Query  188  MFSVLNYLLPPHDVLPMHCAANAGQSGDVALFFGLSGTGKTTLSADPHRFLIGDDEHGWS  247
            MFSVLN++LP HD+LPMHCAANA + GD ALFFGLSGTGKTTLSADP R LIGDDEHGW
Sbjct  195  MFSVLNFVLPQHDILPMHCAANASKEGDTALFFGLSGTGKTTLSADPKRLLIGDDEHGWG  254

Query  248  ATSVFNFEGGCYAKCIDLSQEREPMIWNAIRHGAIMENVVLD-ENGVPDYADARLTQNSR  306
               +FNFEGGCYAKCIDLS EREP+IWNAIR G ++ENVVL+ +   PDYADA LTQN+R
Sbjct  255  EDGIFNFEGGCYAKCIDLSPEREPLIWNAIRFGTVIENVVLNPQTREPDYADASLTQNTR  314

Query  307  AAYPREYIPLRVENNRGRPPDAVLFLTCDLDGVLPPVALLTKEQAAYYFLSGYTALVGST  366
            AAYPRE+IP RVENNRGR P AVLFLTCDL GVLPPVA LT EQAAYYFLSGYTALVGST
Sbjct  315  AAYPREFIPERVENNRGRQPHAVLFLTCDLYGVLPPVARLTPEQAAYYFLSGYTALVGST  374

Query  367  EVGSVKGVTSTFSTCFGAPFFPRPPTVYAELLMKRIEATGCQVYLVNTGWTGGAYGEGGE  426
            EVG   G+  TFSTCFGAPFFPRPP VYAELLMKR+     QVYLVNTGWTGG++GEGG+
Sbjct  375  EVGQGSGIKPTFSTCFGAPFFPRPPGVYAELLMKRLRNFDTQVYLVNTGWTGGSHGEGGK  434

Query  427  RFSIPTTRAIVNAVLSGKLKEGPTEVLSGFNLTIPKSALGVDDHLLNPRKTWEDVSAYDA  486
            RFSIPTTR++V A++ G LK    E L GFN+ IPK   GVD  LLNPRKTW++ +A+DA
Sbjct  435  RFSIPTTRSVVTAIVEGTLKNAEFETLPGFNIEIPKDVPGVDTRLLNPRKTWDNQAAHDA  494

Query  487  RAQRLIQKFRENFEKFKVLAAIREAGPS  514
              A+ LI +F ENF++F V  AIR AGP+
Sbjct  495  NARTLISQFIENFKRFNVSDAIRNAGPT  522
```

Figure 106: BLAST seventh match for BMW92_RS10840 sequence from organism *Aquicella siphonis* with an e-value of 0.0, 66% identity, 77% positives, 0% gaps, and an identity of phosphoenolpyruvate carboxykinase (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

phosphoenolpyruvate carboxykinase [Candidatus Rickettsiella isopodorum]

Sequence ID: WP_071662850.1 **Length:** 524 **Number of Matches:** 1

See 1 more title(s) ❯ See all Identical Proteins(IPG)

Range 1: 13 to 519 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
686 bits(1770)	0.0	Compositional matrix adjust.	322/507(64%)	402/507(79%)	3/507(0%)

```
Query  10   YINLSPDELIQHAVKNGEGVLSSTGALAVTTGKRTGRSPKDRFIVKDEQTADQVAWGNIN  69
            +++LS +EL+  AV+  EGV+++ GAL+V+TGKRTGRSPKD+FIV + ++   + W +IN
Sbjct  13   FVDLSVEELLNFAVERKEGVIAANGALSVSTGKRTGRSPKDKFIVAEPKSEKDIDWDSIN  72

Query  70   QPVEQRTFDQLWERALRYLSERAVYISHLQVGADDNYFLPLKVVTEFAWHNLFACDLFIR  129
            Q + +  F  LW+RA +Y+ +  ++IS+LQVGAD  Y+LP+KV+TE+AWHNLFA  LFIR
Sbjct  73   QALSEERFHALWQRAEQYVKDADLFISNLQVGADPTYYLPVKVITEYAWHNLFARQLFIR  132

Query  130  PSGDHAN-GKPSWVILSAPGLKTDPERDGVNSDGAVMINLSQRRVLLVGMPYAGEMKKAM  188
            P   +    KP W ILS PGLKTDP+RDGVNSD  ++I+L++R+VLL G  YAGE+KKAM
Sbjct  133  PDDFYGKVSKPEWTILSVPGLKTDPQRDGVNSDATLVIHLTERKVLLCGHRYAGEIKKAM  192

Query  189  FSVLNYLLPPHDVLPMHCAANAGQSGDVALFFGLSGTGKTTLSADPHRFLIGDDEHGWSA  248
            FSV+NYLLP  DVLPMHC+AN G+ GDVALFFGLSGTGKTTLSADP RFLIGDDEH WS
Sbjct  193  FSVMNYLLPAVDVLPMHCSANVGKEGDVALFFGLSGTGKTTLSADPDRFLIGDDEHAWSE  252

Query  249  TSVFNFEGGCYAKCIDLSQEREPMIWNAIRHGAIMENVVLD-ENGVPDYADARLTQNSRA  307
            T VFNFEGGCYAKCIDLS+EREP+IWNAIRHGA+MENVVLD E   P+Y DARLTQN+R
Sbjct  253  TGVFNFEGGCYAKCIDLSKEREPLIWNAIRHGAVMENVVLDPETLDPNYKDARLTQNTRV  312

Query  308  AYPREYIPLRVENNR-GRPPDAVLFLTCDLDGVLPPVALLTKEQAAYYFLSGYTALVGST  366
            AYP  +I  R   NR  R PDAV+FL CDL GVLPP+A L  EQAAYYFLSGYTALVGST
Sbjct  313  AYPLNFIESRFRANRVDRLPDAVIFLCCDLYGVLPPIACLNHEQAAYYFLSGYTALVGST  372

Query  367  EVGSVKGVTSTFSTCFGAPFFPRPPTVYAELLMKRIEATGCQVYLVNTGWTGGAYGEGGE  426
            EVG  + + +TFSTCFGAPFFPRP  VYAELL+KR++ +  +VYLVNTGWTGGAYG+GG+
Sbjct  373  EVGQTEPIKTTFSTCFGAPFFPRPAKVYAELLIKRLKNSHAKVYLVNTGWTGGAYGDGGQ  432

Query  427  RFSIPTTRAIVNAVLSGKLKEGPTEVLSGFNLTIPKSALGVDDHLLNPRKTWEDVSAYDA  486
            RFSIP TRA++ A+L+ ++ +  +E+L GFN +IPK    +++HLLNP+KTW++   YD
Sbjct  433  RFSIPATRAVIKAILNDEVGKAESELLLGFNFSIPKQLPNIENHLLNPKKTWKNPKDYDV  492

Query  487  RAQRLIQKFRENFEKFKVLAAIREAGP  513
            +A  LI KF  NF++F V   IR+AGP
Sbjct  493  KAHELINKFINNFKQFDVNPVIRDAGP  519
```

Figure 107: BLAST eighth match for BMW92_RS10840 sequence from organism *Rickettsiella isopodorum* with an e-value of 0.0, 64% identity, 79% positives, 0% gaps, and an identity of phosphoenolpyruvate carboxykinase (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

phosphoenolpyruvate carboxykinase [Candidatus Rickettsiella viridis]

Sequence ID: WP_126322187.1 **Length:** 523 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 15 to 521 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
684 bits(1766)	0.0	Compositional matrix adjust.	324/508(64%)	402/508(79%)	4/508(0%)

```
Query  9    TYINLSPDELIQHAVKNGEGVLSSTGALAVTTGKRTGRSPKDRFIVKDEQTADQVAWGNI  68
            +Y++L+ ++LI  A++  EGV+++ GAL+V+TG+RTGRSPKD+FIV++ +T   + WG +
Sbjct  15   SYVDLTVEQLINFAIERKEGVIAANGALSVSTGERTGRSPKDKFIVQEAKTEKDIDWGPV  74

Query  69   NQPVEQRTFDQLWERALRYLSERAVYISHLQVGADDNYFLPLKVVTEFAWHNLFACDLFI  128
            NQP+ +  F  LW+RA  Y  E  ++IS+LQVGAD +Y+LP+KV+T++AWHNLFA  LFI
Sbjct  75   NQPIAEEHFHALWQRAESYAKEVDLFISNLQVGADPDYYLPVKVITQYAWHNLFARQLFI  134

Query  129  RPSGDHANG-KPSWVILSAPGLKTDPERDGVNSDGAVMINLSQRRVLLVGMPYAGEMKKA  187
            RP   H    K  W ILS PGLKTDP  DGV+SD  +M++LS+R+VLL G  YAGE+KKA
Sbjct  135  RPENFHGKANKAEWTILSLPGLKTDPRCDGVHSDATLMLHLSERKVLLCGHRYAGEIKKA  194

Query  188  MFSVLNYLLPPHDVLPMHCAANAGQSGDVALFFGLSGTGKTTLSADPHRFLIGDDEHGWS  247
            MFSVLNYLLP  DVLPMHC+AN G+ GDVALFFGLSGTGKTTLSADP R+LIGDDEHGWS
Sbjct  195  MFSVLNYLLPASDVLPMHCSANVGKQGDVALFFGLSGTGKTTLSADPERYLIGDDEHGWS  254

Query  248  ATSVFNFEGGCYAKCIDLSQEREPMIWNAIRHGAIMENVVLDENGV-PDYADARLTQNSR  306
              SVFNFEGGCYAKCIDLS+EREP+IWNAIRHGA+MENVVLD + + PDY DA LTQN+R
Sbjct  255  ENSVFNFEGGCYAKCIDLSKEREPVIWNAIRHGAVMENVVLDPHTLEPDYKDASLTQNTR  314

Query  307  AAYPREYIPLRVENNR-GRPPDAVLFLTCDLDGVLPPVALLTKEQAAYYFLSGYTALVGS  365
             AYP ++I LRV  NR  + P AV+FLTCDL GVLPPVA L+ EQAAYYFLSGYTALVGS
Sbjct  315  VAYPLDFISLRVPENRVEQLPSAVIFLTCDLYGVLPPVARLSHEQAAYYFLSGYTALVGS  374

Query  366  TEVGSVKGVTSTFSTCFGAPFFPRPPTVYAELLMKRIEATGCQVYLVNTGWTGGAYGEGG  425
            TEVG  + + +TFSTCFGAPFFPRP  VYAELL+KR++ +   VYLVNTGWTGGAYG+ G
Sbjct  375  TEVGQTEAIKTTFSTCFGAPFFPRPAKVYAELLIKRLKNSDANVYLVNTGWTGGAYGQ-G  433

Query  426  ERFSIPTTRAIVNAVLSGKLKEGPTEVLSGFNLTIPKSALGVDDHLLNPRKTWEDVSAYD  485
             RF IP TRAI+ A+LS ++K      L GFN  IPK+   +D  LL+PR+TW+D++AYD
Sbjct  434  RRFPIPVTRAIIQAILSDEMKTAEYTTLPGFNFAIPKNLKDIDACLLDPRQTWDDIAAYD  493

Query  486  ARAQRLIQKFRENFEKFKVLAAIREAGP  513
             + + LI KF +NF+KF+V   IR+AGP
Sbjct  494  YKTKELIAKFIDNFKKFEVSKEIRDAGP  521
```

Figure 108: BLAST ninth match for BMW92_RS10840 sequence from organism *Rickettsiella viridis* with an e-value of 0.0, 64% identity, 79% positives, 0% gaps, and an identity of phosphoenolpyruvate carboxykinase (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

phosphoenolpyruvate carboxykinase [Modicisalibacter sp. 'Wilcox']

Sequence ID: WP_163647577.1 **Length:** 525 **Number of Matches:** 1

Range 1: 18 to 523 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
671 bits(1732)	0.0	Compositional matrix adjust.	321/506(63%)	388/506(76%)	1/506(0%)

```
Query  10   YINLSPDELIQHAVKNGEGVLSSTGALAVTTGKRTGRSPKDRFIVKDEQTADQVAWGNIN  69
            + NL   ELI+ AV +GEG L++ GAL V TG+RTGRSP DRFIV +  TAD + WG++N
Sbjct  18   HTNLCAAELIERAVADGEGRLATNGALVVNTGERTGRSPADRFIVDEPSTADLIDWGSVN  77

Query  70   QPVEQRTFDQLWERALRYLSERAVYISHLQVGADDNYFLPLKVVTEFAWHNLFACDLFIR  129
            +P +   FD LWER   YL+E + Y++ L VGAD  ++LP++V TE AWHNLFA +LF+R
Sbjct  78   RPFDAERFDALWERVEDYLAEGSSYVAELHVGADPEHYLPIRVTTETAWHNLFARNLFVR  137

Query  130  PSGDHANGKPSWVILSAPGLKTDPERDGVNSDGAVMINLSQRRVLLVGMPYAGEMKKAMF  189
            P   +  GK  W IL+AP    DP RDG NSDGAV+IN ++RRVLL GM YAGEMKKAMF
Sbjct  138  PEAFNPKGKNEWTILNAPHFTCDPSRDGTNSDGAVVINFARRRVLLAGMRYAGEMKKAMF  197

Query  190  SVLNYLLPPHDVLPMHCAANAGQSGDVALFFGLSGTGKTTLSADPHRFLIGDDEHGWSAT  249
            SV N+LLP  DVLPMHC+AN G+ G+  LFFGLSGTGKTTLSADP R+LIGDDEHGW
Sbjct  198  SVQNFLLPEKDVLPMHCSANVGEDGETTLFFGLSGTGKTTLSADPARYLIGDDEHGWGEG  257

Query  250  SVFNFEGGCYAKCIDLSQEREPMIWNAIRHGAIMENVVLDENGVPDYADARLTQNSRAAY  309
            +VFN EGGCYAKCIDLS++ EP+IW AIR GA++ENVVLD+   PDYAD  LTQNSRAAY
Sbjct  258  TVFNIEGGCYAKCIDLSEKNEPVIWQAIRFGAVLENVVLDDRRAPDYADDSLTQNSRAAY  317

Query  310  PREYIPLRVENNRGRPPDAVLFLTCDLDGVLPPVALLTKEQAAYYFLSGYTALVGSTEVG  369
            P E+I  RVE NR   P A++FLTCD+ GVLPPV++L+KE AAY+FLSGYTA VGSTE+G
Sbjct  318  PLEHIDKRVEENRAGEPSAIIFLTCDMSGVLPPVSVLSKEAAAYHFLSGYTAKVGSTEMG  377

Query  370  SVKGVTSTFSTCFGAPFFPRPPTVYAELLMKRIEATGCQVYLVNTGWTGGAYGEGGERFS  429
            S  G+ +TFSTCFGAPFFPRP   YA+LL+KRIEA G +VYLVNTGWTGG+YG+GG RFS
Sbjct  378  SSAGLEATFSTCFGAPFFPRPAREYADLLIKRIEAFGSRVYLVNTGWTGGSYGQGGSRFS  437

Query  430  IPTTRAIVNAVLSGKLKEGPTEVLSGFNLTIPKSALGVDDHLLNPRKTWEDVSAYDARAQ  489
            IPTTR I++AV SG LK+  T  + G NL +P +  GVD  LL+PR+TW D +AYD + Q
Sbjct  438  IPTTRGIISAVQSGALKDVETRRVDGLNLDVPVAVPGVDSRLLDPRETWGDPAAYDRQRQ  497

Query  490  RLIQKFRENFEKFK-VLAAIREAGPS  514
             L+ KF ENF+KF  V  AI  AGPS
Sbjct  498  ELVAKFVENFKKFAGVDEAIIAAGPS  523
```

Figure 109: BLAST tenth match for BMW92_RS10840 sequence from organism *Modicisalibacter wilcox* with an e-value of 0.0, 63% identity, 76% positives, 0% gaps, and an identity of phosphoenolpyruvate carboxykinase (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

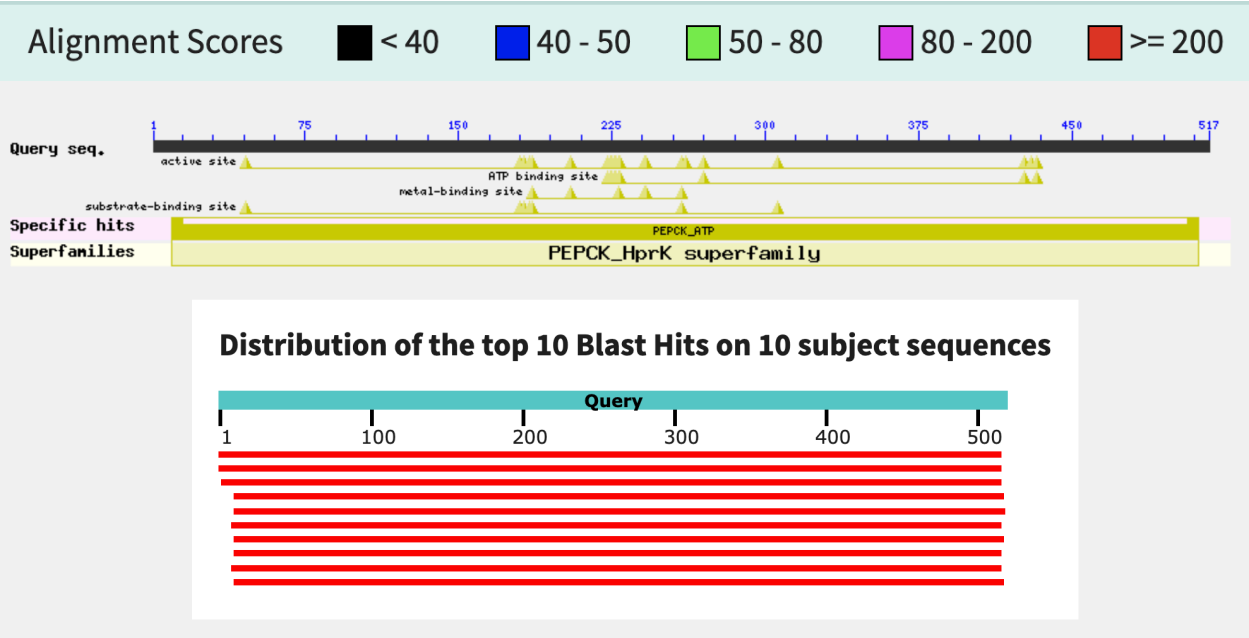


Figure 110: BLAST graphic summary of the top 10 organism sequences similarities selected aligned with *Coxiella burnetii* query sequence of gene BMW92_RS10840. Each of the alignment sequences selected are order from highest sequence similarity (top) to lowest sequence similarity (bottom). All organism sequences aligned with the query sequence have an alignment score of greater than 200 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

CDD

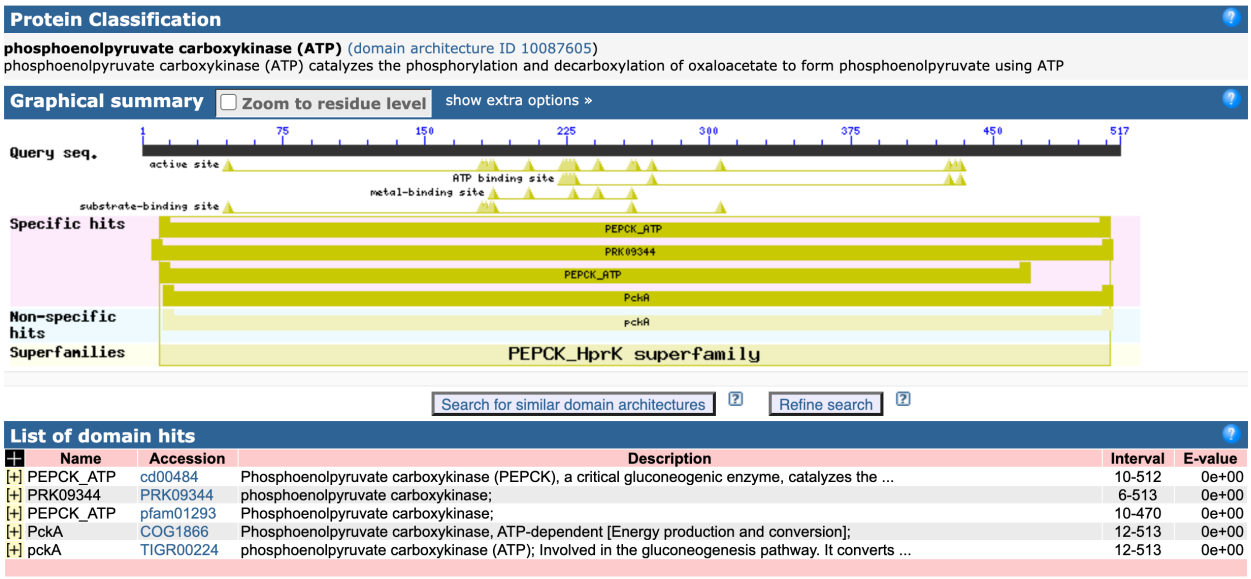


Figure 111: Conserved Domain Database output results for gene BMW92_RS10840. The top domain hit match was PEPCK_ATP: Phosphoenolpyruvate carboxykinase (PEPCK) which aligned with the query sequence from amino acid residues 10-512 and had statistically significant e-value of 0e+00. The second domain hit match was PRK09344: Phosphoenolpyruvate carboxykinase (PEPCK) which aligned with the query sequence from amino acid residues 6-513 and had a statistically significant e-value of 0e+00. The third domain hit match was PEPCK_ATP: Phosphoenolpyruvate carboxykinase (PEPCK) which aligned with the query sequence from amino acid residues 10-470 and had a statistically significant e-value of 0e+00. The fourth domain hit match was PckA: Phosphoenolpyruvate carboxykinase (PEPCK) which aligned with the query sequence from amino acid residues 12-513 and had a statistically significant e-value of 0e+00. The fifth domain hit match was pckA: Phosphoenolpyruvate carboxykinase (PEPCK) which aligned with the query sequence from amino acid residues 12-513 and had a statistically significant e-value of 0e+00 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

Figure 112: The top domain hit sequence PEPCK_ATP: Phosphoenolpyruvate carboxykinase (PEPCK) aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 10-512 and had a statistically significant e-value of 0e+00 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

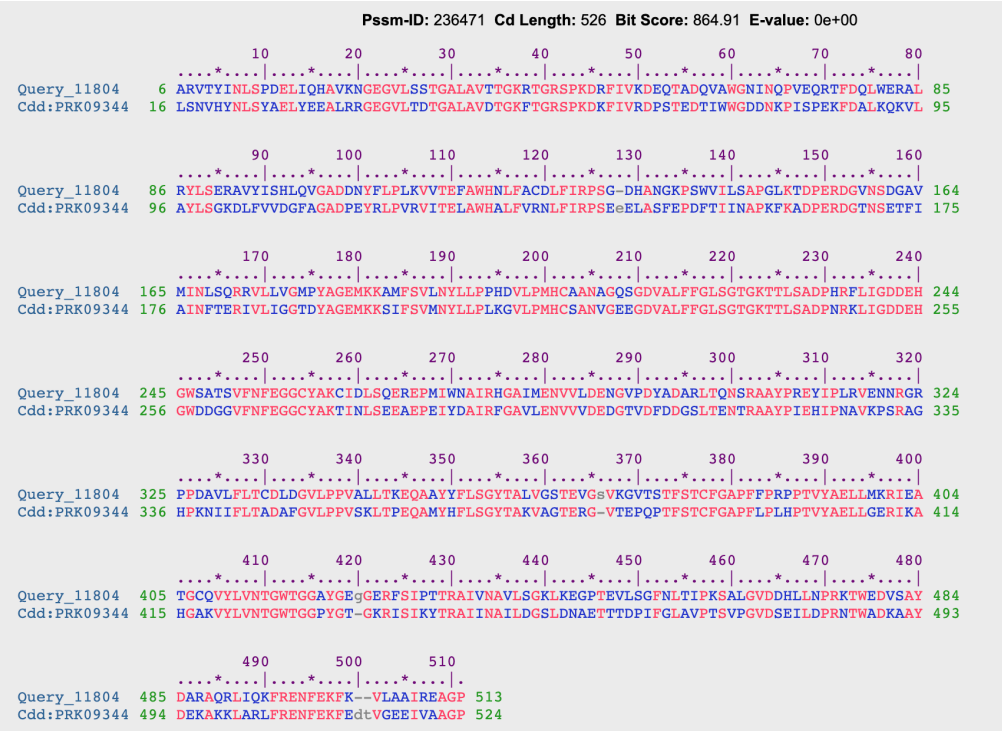


Figure 113: The second domain hit sequence PRK09344: Phosphoenolpyruvate carboxykinase (PEPCK) aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 6-513 and had a statistically significant e-value of 0e+00 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

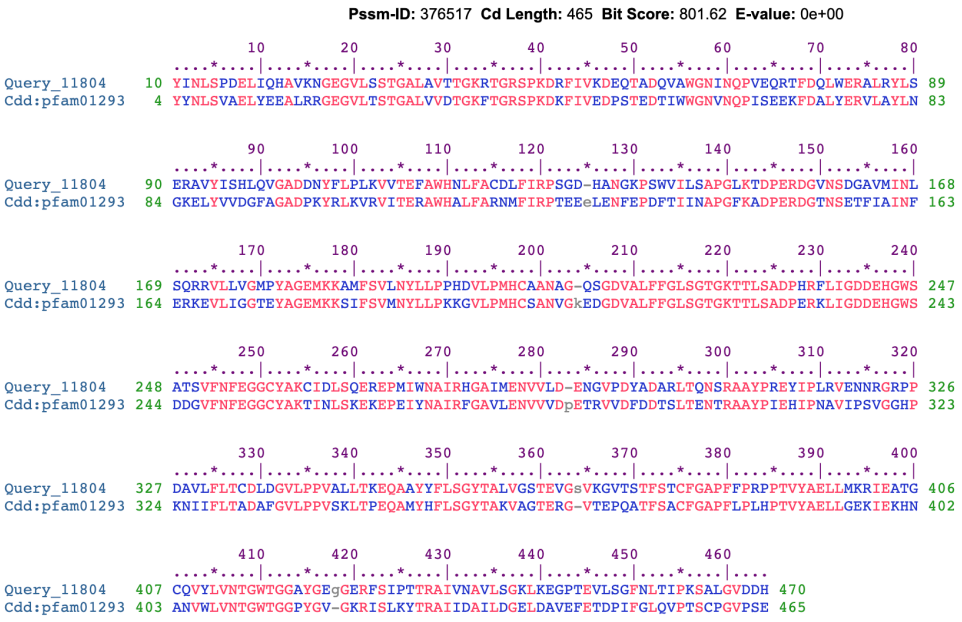


Figure 114: The third domain hit sequence PEPCK_ATP: Phosphoenolpyruvate carboxykinase (PEPCK) aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 10-470 and had a statistically significant e-value of 0e+00 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

Figure 115: The fourth domain hit sequence PckA: Phosphoenolpyruvate carboxykinase (PEPCK) aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 12-513 and had a statistically significant e-value of 0e+00 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

MUSCLE

```
P.japonicum       --MTTNNESAVN----TYTNLSNAQLIELAIQRGEGTLADNGALVVATGQRTGRSPMDRF
M.wilcox          MTTNSAAAQPTARRASVHTNLCAAELIERAVADGEGRLATNGALVVNTGERTGRSPADRF
R.isopodorum      ---MSAVQSKNSSK--IFVDLSVEELLNFAVERKEGVIAANGALSVSTGKRTGRSPKDKF
R.viridis         MSSSLDAGTLVSGK--SYVDLTVEQLINFAIERKEGVIAANGALSVSTGERTGRSPKDKF
C.bacterium       --------MATENQ--RHVDLSVAELIEMALEREEGVLSANQALVVATGKRTGRSPKDRF
L.bacterium       ----MHAETKAETM--TYTDLSTEQLIKHALERNEGVLSTNQALSVATGTRTGRSPRDRF
A.lusitana        -MVESNEVVTMNTK--AHINLSAEELVEIALARGEGELASNQALVVKTGARTGRSPKDRF
A.siphonis        -MVQSTNEVFIQSK--NHLNLSAKELVELALARGEGELASNQALVVKTGSRTGRSPKDRF
C.burnetii        ------MEQIAARV--TYINLSPDELIQHAVKNGEGVLSSTGALAVTTGKRTGRSPKDRF
C.mudrowiae       ------MDQIASRT--VYTDLSVDELIQQALKKGEGKLSSTGALAVTTGKRTGRSPKDRF
R.microplus       ------MEQIVSRT--VYTDLAIDELIQHALKKGEGTLSVTGALAVRTGKRTGRSPQDRF
                                  . :*   :*:: *:   ** :: . ** * ** ****** *.*

P.japonicum       IVNEPSTSDAIDWGSINRPFSAEKFDALWERVEEYLSKQDT-FISELHVGADPEHYLPIR
M.wilcox          IVDEPSTADLIDWGSVNRPFDAERFDALWERVEDYLAEGSS-YVAELHVGADPEHYLPIR
R.isopodorum      IVAEPKSEKDIDWDSINQALSEERFHALWQRAEQYVKDADL-FISNLQVGADPTYYLPVK
R.viridis         IVQEAKTEKDIDWGPVNQPIAEEHFHALWQRAESYAKEVDL-FISNLQVGADPDYYLPVK
C.bacterium       IVKDELTADTVDWGNVNQPFDPAKFTVLWQRAEQYMADQEV-FVSHLGVGADIEHFVPVT
L.bacterium       IVQDDVTTNTVDWGNVNQPISQDRFDALWNQIEAYLADKDT-FVSHLEVGADSEHYLPVK
A.lusitana        IVRDEITENQVDWNTINQPISPEKFNALWQKAQDYLDTRDAHFISFLKVGAHEELGVPVK
A.siphonis        IVRGQATETQVDWNQINQPISADKFEALWEKALHYLNSKDARFTSYLKVGAHETLGVSVK
C.burnetii        IVKDEQTADQVAWGNINQPVEQRTFDQLWERALRYLSERAV-YISHLQVGADDNYFLPLK
C.mudrowiae       IVKDAETADQVQWGNVNQSIVQGVFDQLWNRANAYLSKRPM-YVSHLQVGADENYFLPVQ
R.microplus       IVKDSETEDQVQWGDVNQPIVQVVFDQLWNRATAYISKRSM-YVSHLKVGADENYSIPVQ
                  **    :   : *. :*...    *  **:.   *       : : * ***     :.:

P.japonicum       VTTETAWHNLFGRNLFVRPEGYNPKSKG-EWQILNAPNFVCEPSRDGTNSDGCVILNFAK
M.wilcox          VTTETAWHNLFARNLFVRPEAFNPKGKN-EWTILNAPHFTCDPSRDGTNSDGAVVINFAR
R.isopodorum      VITEYAWHNLFARQLFIRPDDFYGKVSKPEWTILSVPGLKTDPQRDGVNSDATLVIHLTE
R.viridis         VITQYAWHNLFARQLFIRPENFHGKANKAEWTILSLPGLKTDPRCDGVHSDATLMLHLSE
C.bacterium       VISEYAWHNVFVHDLFIRPNGRYPHGRA-GWTILNAAGLPTDPARDGTNSEATLILNFKE
L.bacterium       VINQKAWHNLFTRNLFIRPDTYNRKQKP-EWTILSAPDFHASPERDGTNSEAAVILNFSQ
A.lusitana        VITELAWHNLFARVLFIRPEKPATTVVPNQWTILSVPGFKTDPARDGVNGDAAVILNFSQ
A.siphonis        VMAELAWHTLFAHVLFIRPVTPPTSDQPNQWTILSTPGFKTDPARDGVNSDAAVILDFEK
C.burnetii        VVTEFAWHNLFACDLFIRPSGDHANGKP-SWVILSAPGLKTDPERDGVNSDGAVMINLSQ
C.mudrowiae       VITEFGWHNLFACDLFIRPDGDYAKGKP-EWIILSVPGLKTDPERDKVNSDAAVIINLSQ
R.microplus       VITELAWHNLFACELFIRPDRDYFKGKP-KWILLSVPGLTTDPKRDKVNSDAAVIINLSQ
                  *  : .**.:*   **:**           * :*. . :  .*  * .:.:. ::: :

P.japonicum       RKVLLAGMKYAGEMKKAMFSVQNFLLPEKDVLPMHCSANVGEDGETTLFFGLSGTGKTTL
M.wilcox          RRVLLAGMRYAGEMKKAMFSVQNFLLPEKDVLPMHCSANVGEDGETTLFFGLSGTGKTTL
R.isopodorum      RKVLLCGHRYAGEIKKAMFSVMNYLLPAVDVLPMHCSANVGKEGDVALFFGLSGTGKTTL
R.viridis         RKVLLCGHRYAGEIKKAMFSVLNYLLPASDVLPMHCSANVGKQGDVALFFGLSGTGKTTL
C.bacterium       KKILLCGLRYAGEMXKAMFSVLNFILPEKNVLPMHCAANVGKQGDVALFFGLSGTGKTTL
L.bacterium       RRILVCGTHYAGEMKKAMFTVMNFLLPNIDVLPMHCASNIGMEGDVALFFGLSGTGKTTL
A.lusitana        RRILICGTHYAGEMKKAMFSVLNFILPEHNILPMHCAANAGENGDTALFFGLSGTGKTTL
A.siphonis        HRILICGTYYAGEMKKAMFSVLNFVLPQHDILPMHCAANASKEGDTALFFGLSGTGKTTL
C.burnetii        RRVLLVGMPYAGEMKKAMFSVLNYLLPPHDVLPMHCAANAGQSGDVALFFGLSGTGKTTL
C.mudrowiae       RRVLLVGMAYAGEIKKAMFTVLNYLLPPHDVLPMHCAANAGKSGDVALFFGLSGTGKTTL
R.microplus       RRVLLVGMSYAGEMKKAMFTVLNYLLPPQDVLPMHCAANTGKSGDVALFFGLSGTGKTTL
                  ..:*: *  ****: ****:* *::**  ::******::* . .*:.:**************
```

```
P.japonicum       SADPSRYLIGDDEHGWGKGTVFNIEGGCYAKCIDLSAENEPVIWNAIRFGAVLENVILD-
M.wilcox          SADPARYLIGDDEHGWGEGTVFNIEGGCYAKCIDLSEKNEPVIWQAIRFGAVLENVVLD-
R.isopodorum      SADPDRFLIGDDEHAWSETGVFNFEGGCYAKCIDLSKEREPLIWNAIRHGAVMENVVLDP
R.viridis         SADPERYLIGDDEHGWSENSVFNFEGGCYAKCIDLSKEREPVIWNAIRHGAVMENVVLDP
C.bacterium       SADPERYLIGDDEHGWSDHGVFNFEGGCYAKCINLSKEREPVIWDAIRYGAIMENVVLDP
L.bacterium       SADPERFLIGDDEHGWGKSGVFNFEGGCYAKCIDLSKEKEPVIWDAIRHGAIMENVVLD-
A.lusitana        SADPERFLIGDDEHGWGNDGVFNFEGGCYAKCIDLSEEREPLIWNAIRYGSVIENVVLDP
A.siphonis        SADPKRLLIGDDEHGWGEDGIFNFEGGCYAKCIDLSPEREPLIWNAIRFGTVIENVVLNP
C.burnetii        SADPHRFLIGDDEHGWSATSVFNFEGGCYAKCIDLSQEREPMIWNAIRHGAIMENVVLD-
C.mudrowiae       SADPNRFLIGDDEHGWSRTGVFNFEGGCYAKCIDLSSEREPMIWEAIRHGAIMENVVLQ-
R.microplus       SADPNRFLIGDDEHGWSRTGVFNFEGGCYAKCIDLSLEREPIIWESIRYGAIMENVVLQ-
                  **** * *******.*.   :**:**********:** :.**:**::**.*:::***:*:

P.japonicum       ERRVPDYNDDSLTQNSRAAYPLEHIEKRVLENR-AGEPSAIVFLTCDMSGVLPPVSILSK
M.wilcox          DRRAPDYADDSLTQNSRAAYPLEHIDKRVEENR-AGEPSAIIFLTCDMSGVLPPVSVLSK
R.isopodorum      ETLDPNYKDARLTQNTRVAYPLNFIESRFRANRVDRLPDAVIFLCCDLYGVLPPIACLNH
R.viridis         HTLEPDYKDASLTQNTRVAYPLDFISLRVPENRVEQLPSAVIFLTCDLYGVLPPVARLSH
C.bacterium       KTKEPLYGDASLTENTRAAYPLEHIAMRVPENQ-AGHPQAVIFLTCDLYGVLPPVAILNK
L.bacterium       ENQAPDYSDSTLSMNSRAAYPREHIEMRAEANR-GGQPDAVLFLTCDLYGVLPPVSLLSK
A.lusitana        VTKNPDYGDASLTQNTRAAYPREFIPQRVENNR-GRQPNAVLFLTCDLYGVLPPVARLTP
A.siphonis        QTREPDYADASLTQNTRAAYPREFIPERVENNR-GRQPHAVLFLTCDLYGVLPPVARLTP
C.burnetii        ENGVPDYADARLTQNSRAAYPREYIPLRVENNR-GRPPDAVLFLTCDLDGVLPPVALLTK
C.mudrowiae       ADGQPDYRNASLTQNTRAAYPREHISLRVKDNR-GRPPDSVIFLTCDLYGVLPPVALLTK
R.microplus       ADGQPAYNDASLTQNTRAAYPREHILFRVKENR-GRPPDAVIFLTCDLYGVLPPVSLLTK
                      * * :  *: *:*.*** :.*  *   *.    * :::** **: *****:: *.

P.japonicum       EAAAYHFLSGYTAKVGSTEMGSSSGLEATFSTCFGAPFFPRPAHVYADLLIKRIEEFGSQ
M.wilcox          EAAAYHFLSGYTAKVGSTEMGSSAGLEATFSTCFGAPFFPRPAREYADLLIKRIEAFGSR
R.isopodorum      EQAAYYFLSGYTALVGSTEVGQTEPIKTTFSTCFGAPFFPRPAKVYAELLIKRLKNSHAK
R.viridis         EQAAYYFLSGYTALVGSTEVGQTEAIKTTFSTCFGAPFFPRPAKVYAELLIKRLKNSDAN
C.bacterium       EQAAYHFLSGYTALVGSTEVGSTAGIKSTFSTCFGAPFFPRPAQVYADLLIKRLTETGAQ
L.bacterium       EQAAYHFLSGYTALVGSTEVGQTEGIKPTFSTCFGAPFFPLSPSVYAELLIKRIEETGAQ
A.lusitana        EQAAYYFLSGYTALVGSTEVGQGSGIKPTFSTCFGAPFFPRPPRVYAELLMKRLQNFDTQ
A.siphonis        EQAAYYFLSGYTALVGSTEVGQGSGIKPTFSTCFGAPFFPRPPGVYAELLMKRLRNFDTQ
C.burnetii        EQAAYYFLSGYTALVGSTEVGSVKGVTSTFSTCFGAPFFPRPPTVYAELLMKRIEATGCQ
C.mudrowiae       EQAAYYFLSGYTALVGSTEVGSVKGVTPTFSTCFGAPFFPRPPTVYAELLMKRIEETQCQ
R.microplus       AQAAYYFLSGYTALVGSTEVGSVKGIVPTFSSCFGAPFFPRPPVVYAKLLMKRIEETQCQ
                    ***:******* *****:*.   : .***:******** ..  **.**:**:     .

P.japonicum       VYLVNTGWTGGAYGQGGNRFSIPTTRAIINAVQTGVLKDAEIEQLPGLNLSVPKHIPGVE
M.wilcox          VYLVNTGWTGGSYGQGGSRFSIPTTRGIISAVQSGALKDVETRRVDGLNLDVPVAVPGVD
R.isopodorum      VYLVNTGWTGGAYGDGGQRFSIPATRAVIKAILNDEVGKAESELLLGFNFSIPKQLPNIE
R.viridis         VYLVNTGWTGGAYGQ-GRRFPIPVTRAIIQAILSDEMKTAEYTTLPGFNFAIPKNLKDID
C.bacterium       VYLVNTGWTGGPYGE-GKRFDIPTTRAVIRAILTGKLKHVPTEVMPGFNLVIPKEVPDVE
L.bacterium       VYLVNTGWTGGAYGQGGERFSIPTTRAIVRAILSGALKDANTITLPGFNLAIPETINGVD
A.lusitana        VYLVNTGWSGGAHGEGGKRFSIPTTRAVVTAIVNGKLKDAEYEKLPGFNFDIPKAVDGVE
A.siphonis        VYLVNTGWTGGSHGEGGKRFSIPTTRSVVTAIVEGTLKNAEFETLPGFNIEIPKDVPGVD
C.burnetii        VYLVNTGWTGGAYGEGGERFSIPTTRAIVNAVLSGKLKEGPTEVLSGFNLTIPKSALGVD
C.mudrowiae       VYLVNTGWTGGAYGEGGVRFSIPTTRAIIDAILTRKLRNQPTENLKGFNLAIPKSAPGVE
R.microplus       VYLVNTGWMGGAYGEGGVRFRIPITRSIIDAILTRKLINQPTENLKGFNLAIPQSVPGVE
                  ******** **.:*: * ** ** **.:: *:    :       : *:*: :*    .::
```

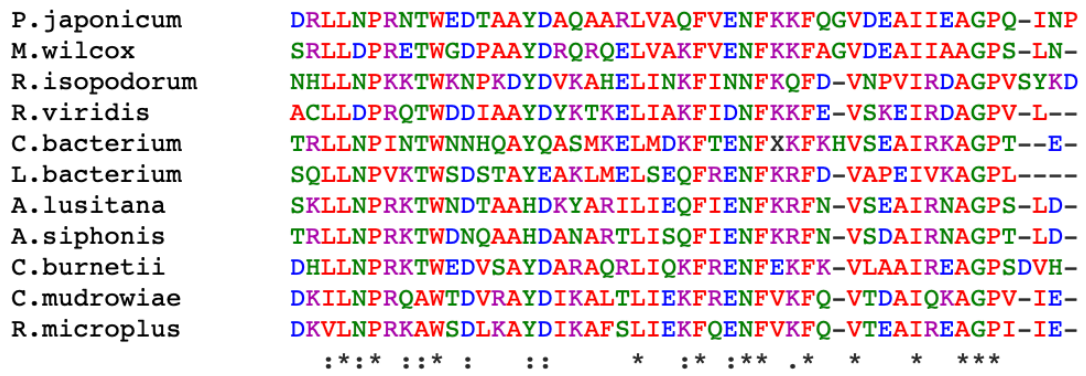


Figure 116: MUSCLE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10840 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered randomly and are listed from top to bottom as followed: *P. japonicum*, *M. wilcox*, *R. isopodorum*, *R. viridis*, *C. bacterium*, *L. lusitana*, *A. siphonis*, *C. burnetii*, *C. mudrowiae*, *R. microplus*. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

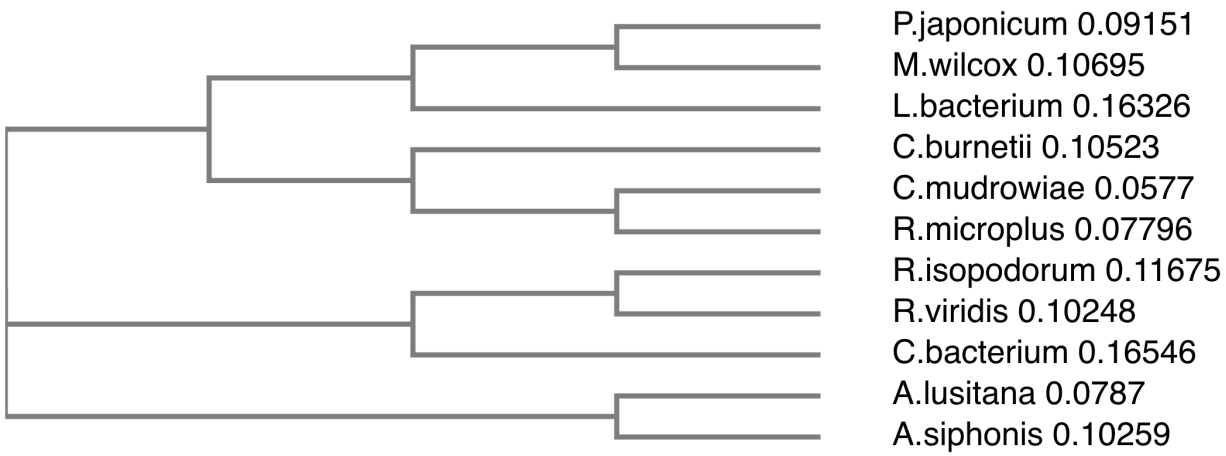


Figure 117: MUSCLE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10840 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *P. japonicum*, *M. wilcox*, *L. bacterium*, *C. burnetii*, *C. mudrowiae*, *R. microplus*, *R. isopodorum*, *R. viridis*, *C. bacterium*, *A. lusitana*, *A. siphonis*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

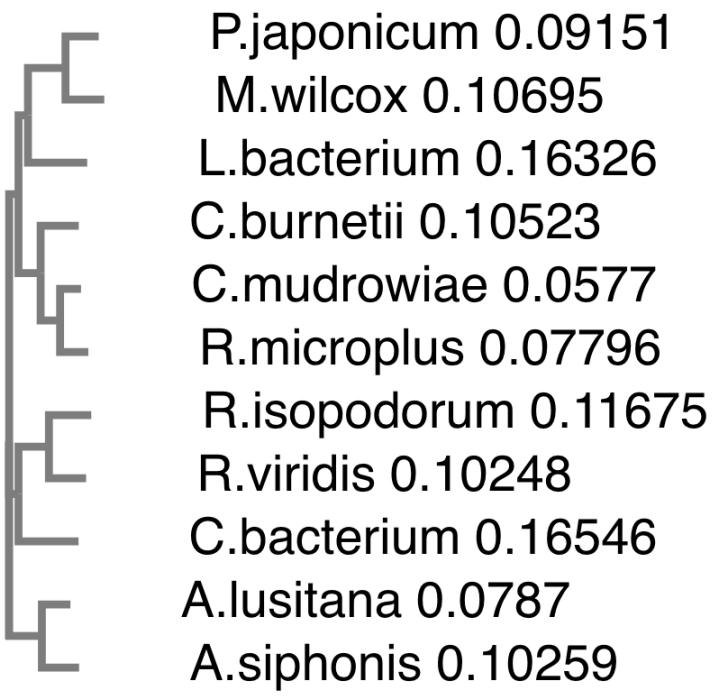
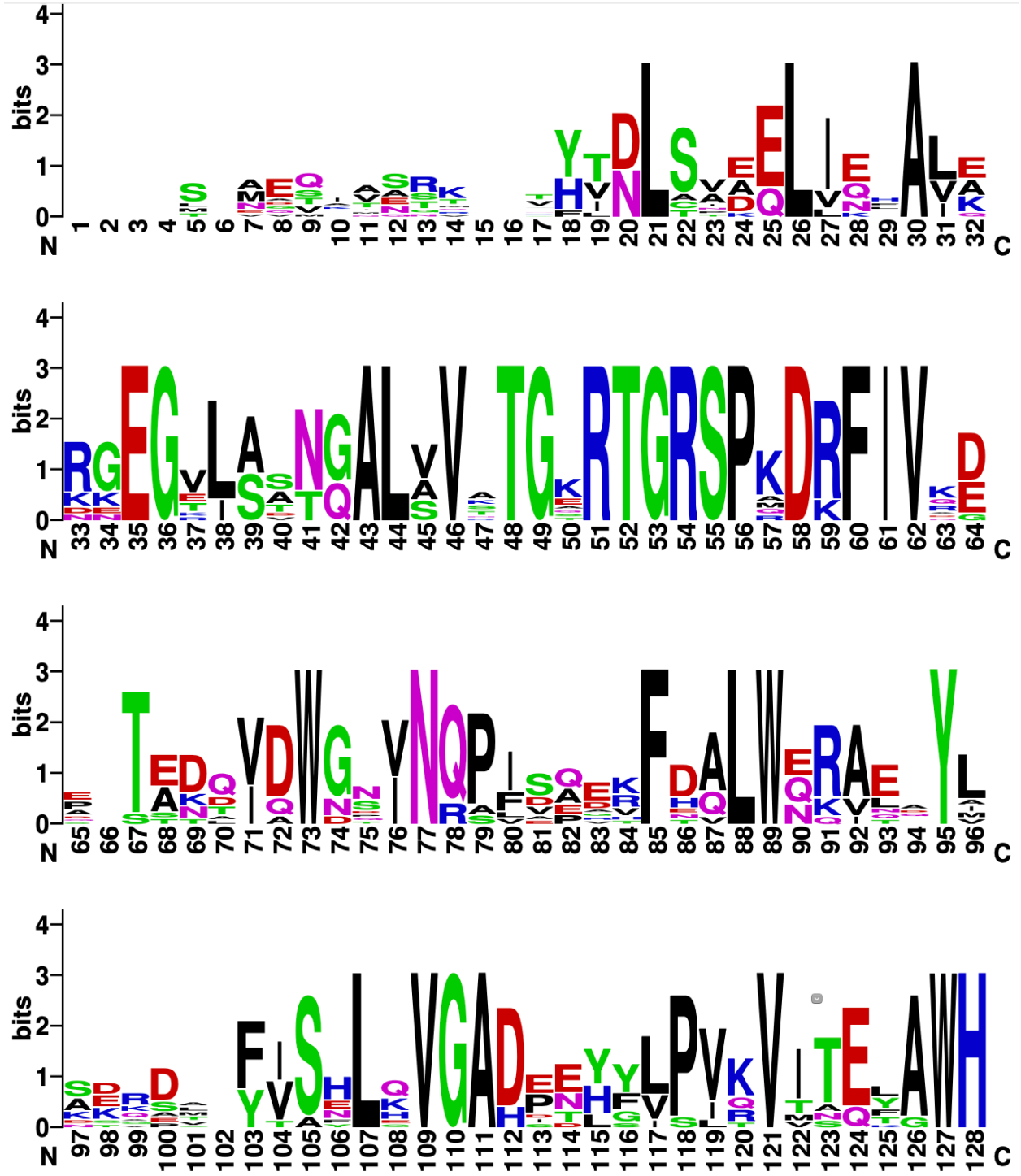
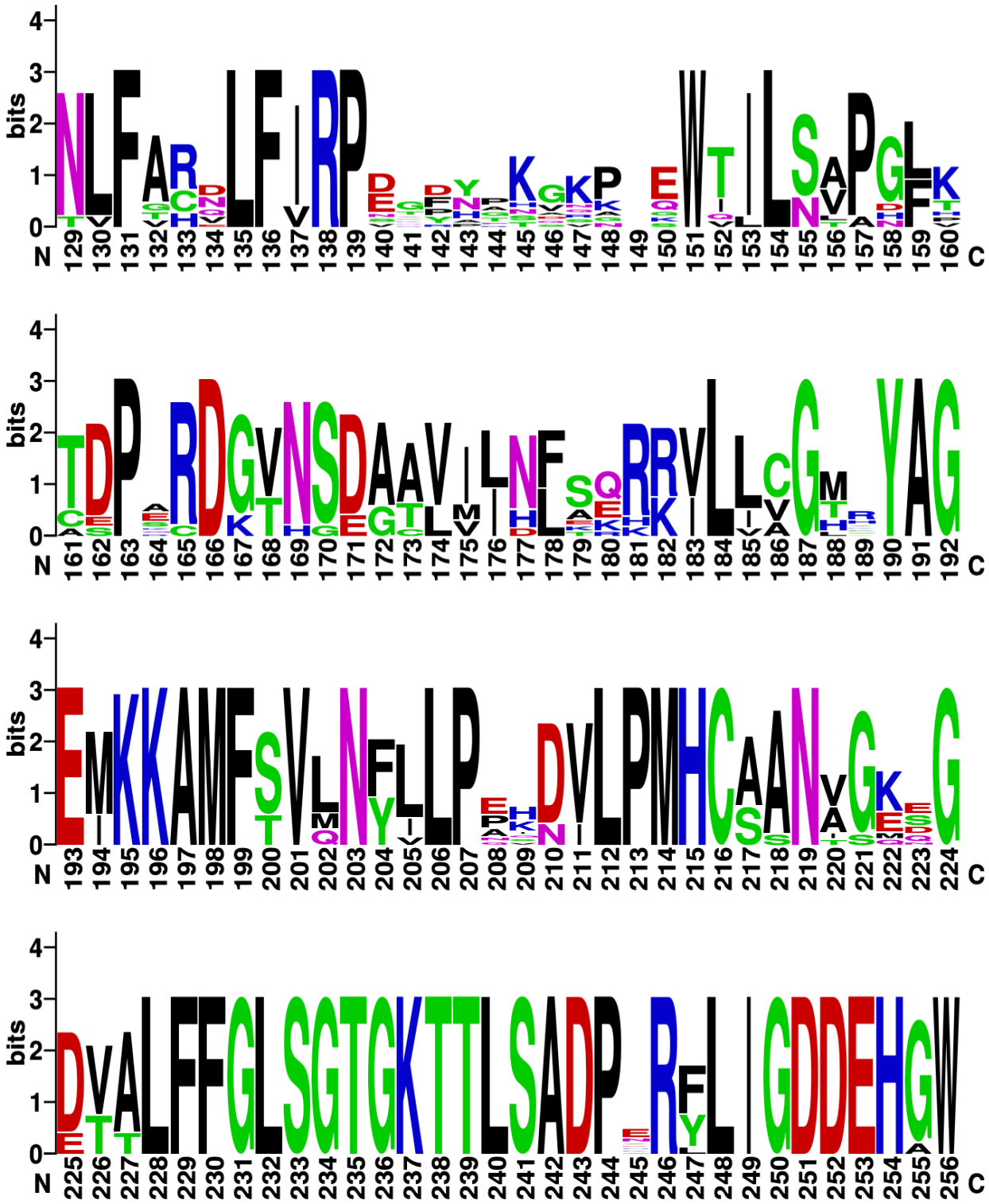
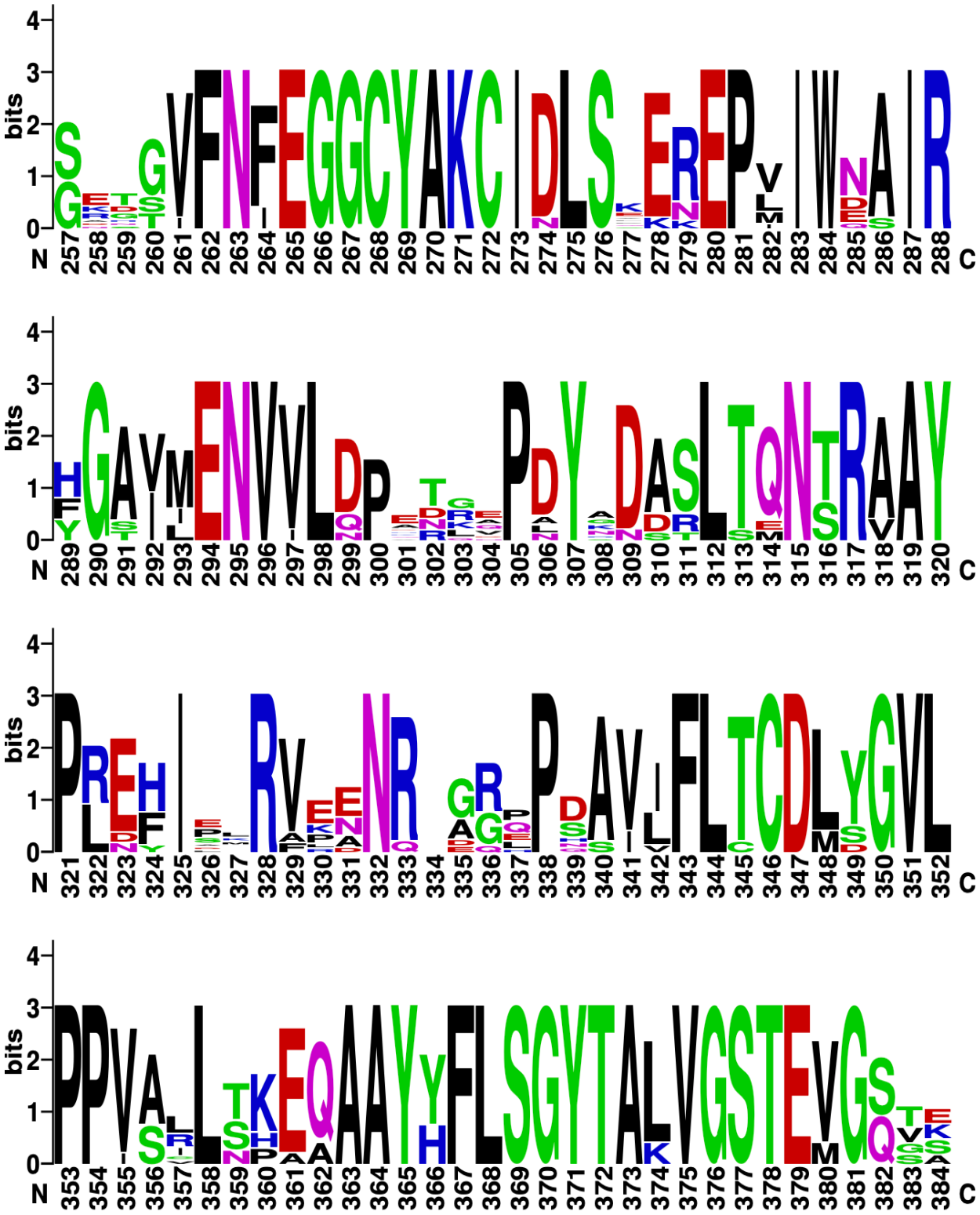


Figure 118: MUSCLE multiple sequence alignment real phylogenetic tree for *Coxiella burnetii* gene BMW92_RS10840 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *P. japonicum*, *M. wilcox*, *L. bacterium*, *C. burnetii*, *C. mudrowiae*, *R. microplus*, *R. isopodorum*, *R. viridis*, *C. bacterium*, *A. lusitana*, *A. siphonis*. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

MUSCLE Sequence Logo




bits
N
C


bits
4
3
2
1
0
N
C

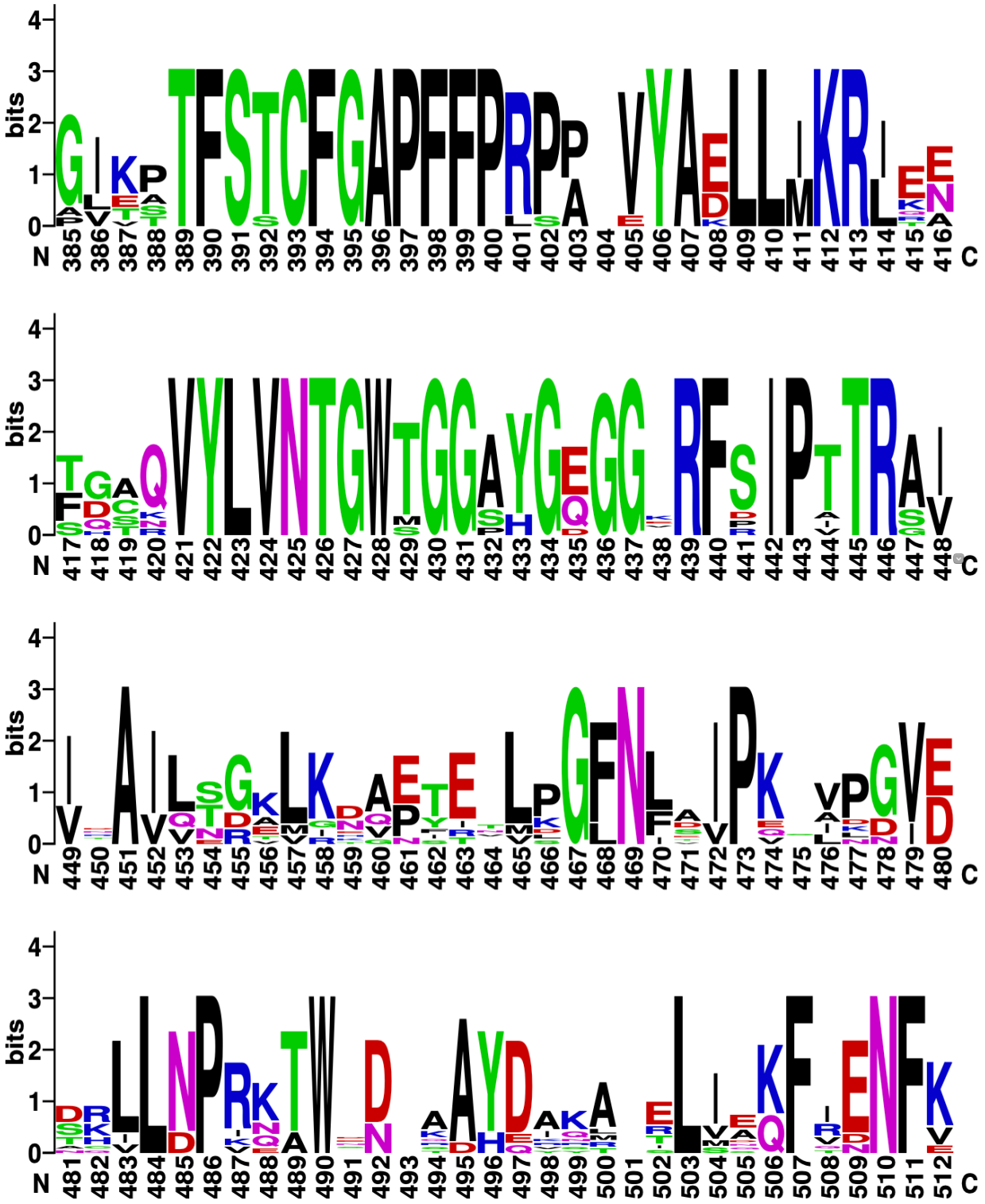

bits
N
C

Figure 119: Sequence logo generated from the MUSCLE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10840 and the top 10 organism sequences similarities selected. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. The conservation of each amino acid among residue across similar sequences is represented by the height of each single letter abbreviation of the amino acid. (WebLogo, <https://weblogo.berkeley.edu/logo.cgi>).

T-COFFEE

```
A.lusitana      MVESNEV---VTMNTKAHINLSAEELVEIALARGEGELASNQALVVKTGA
A.siphonis      MVQSTNE---VFIQSKNHLNLSAKELVELALARGEGELASNQALVVKTGS
C.bacterium     MATE----------NQRHVDLSVAELIEMALEREEGVLSANQALVVATGK
C.burnetii      MEQI--------AARVTYINLSPDELIQHAVKNGEGVLSSTGALAVTTGK
C.mudrowiae     MDQI--------ASRTVYTDLSVDELIQQALKKGEGKLSSTGALAVTTGK
L.bacterium     MHAETK------AETMTYTDLSTEQLIKHALERNEGVLSTNQALSVATGT
M.wilcox        MTTNSAAAQPTARRASVHTNLCAAELIERAVADGEGRLATNGALVVNTGE
P.japonicum     MTTNNES------AVNTYTNLSNAQLIELAIQRGEGTLADNGALVVATGQ
R.isopodorum    MSAVQS-----KNSSKIFVDLSVEELLNFAVERKEGVIAANGALSVSTGK
R.microplus     MEQI--------VSRTVYTDLAIDELIQHALKKGEGTLSVTGALAVRTGK
R.viridis       MSSSLDAG--TLVSGKSYVDLTVEQLINFAIERKEGVIAANGALSVSTGE
                *                . :*   :*:: *:   ** :: . ** * **

A.lusitana      RTGRSPKDRFIVRDEITENQVDWNTINQPISPEKFNALWQKAQDYLDTRD
A.siphonis      RTGRSPKDRFIVRGQATETQVDWNQINQPISADKFEALWEKALHYLNSKD
C.bacterium     RTGRSPKDRFIVKDELTADTVDWGNVNQPFDPAKFTVLWQRAEQYMADQE
C.burnetii      RTGRSPKDRFIVKDEQTADQVAWGNINQPVEQRTFDQLWERALRYLSERA
C.mudrowiae     RTGRSPKDRFIVKDAETADQVQWGNVNQSIVQGVFDQLWNRANAYLSKRP
L.bacterium     RTGRSPRDRFIVQDDVTTNTVDWGNVNQPISQDRFDALWNQIEAYLADKD
M.wilcox        RTGRSPADRFIVDEPSTADLIDWGSVNRPFDAERFDALWERVEDYLAEGS
P.japonicum     RTGRSPMDRFIVNEPSTSDAIDWGSINRPFSAEKFDALWERVEEYLSKQD
R.isopodorum    RTGRSPKDKFIVAEPKSEKDIDWDSINQALSEERFHALWQRAEQYVKDAD
R.microplus     RTGRSPQDRFIVKDSETEDQVQWGDVNQPIVQVVFDQLWNRATAYISKRS
R.viridis       RTGRSPKDKFIVQEAKTEKDIDWGPVNQPIAEEHFHALWQRAESYAKEVD
                ****** *:***    :   : *. :*:..    *  **::   *

A.lusitana      AHFISFLKVGAHEELGVPVKVITELAWHNLFARVLFIRPEKPATTVVPNQ
A.siphonis      ARFTSYLKVGAHETLGVSVKVMAELAWHTLFAHVLFIRPVTPPTSDQPNQ
C.bacterium     -VFVSHLGVGADIEHFVPVTVISEYAWHNVFVHDLFIRPNGRYPH-GRAG
C.burnetii      -VYISHLQVGADDNYFLPLKVVTEFAWHNLFACDLFIRPSGDHAN-GKPS
C.mudrowiae     -MYVSHLQVGADENYFLPVQVITEFGWHNLFACDLFIRPDGDYAK-GKPE
L.bacterium     -TFVSHLEVGADSEHYLPVKVINQKAWHNLFTRNLFIRPDTYNRK-QKPE
M.wilcox        -SYVAELHVGADPEHYLPIRVTTETAWHNLFARNLFVRPEAFNPK-GKNE
P.japonicum     -TFISELHVGADPEHYLPIRVTTETAWHNLFGRNLFVRPEGYNPK-SKGE
R.isopodorum    -LFISNLQVGADPTYYLPVKVITEYAWHNLFARQLFIRPDDFYGKVSKPE
R.microplus     -MYVSHLKVGADENYSIPVQVITELAWHNLFACELFIRPDRDYFK-GKPK
R.viridis       -LFISNLQVGADPDYYLPVKVITQYAWHNLFARQLFIRPENFHGKANKAE
                  : : * ***.    :.: *  : .**.:*   **:**

A.lusitana      WTILSVPGFKTDPARDGVNGDAAVILNFSQRRILICGTHYAGEMKKAMFS
A.siphonis      WTILSTPGFKTDPARDGVNSDAAVILDFEKHRILICGTYYAGEMKKAMFS
C.bacterium     WTILNAAGLPTDPARDGTNSEATLILNFKEKKILLCGLRYAGEMXKAMFS
C.burnetii      WVILSAPGLKTDPERDGVNSDGAVMINLSQRRVLLVGMPYAGEMKKAMFS
C.mudrowiae     WIILSVPGLKTDPERDKVNSDAAVIINLSQRRVLLVGMAYAGEIKKAMFT
L.bacterium     WTILSAPDFHASPERDGTNSEAAVILNFSQRRILVCGTHYAGEMKKAMFT
M.wilcox        WTILNAPHFTCDPSRDGTNSDGAVVINFARRRVLLAGMRYAGEMKKAMFS
P.japonicum     WQILNAPNFVCEPSRDGTNSDGCVILNFAKRKVLLAGMKYAGEMKKAMFS
R.isopodorum    WTILSVPGLKTDPQRDGVNSDATLVIHLTERKVLLCGHRYAGEIKKAMFS
R.microplus     WILLSVPGLTTDPKRDKVNSDAAVIINLSQRRVLLVGMSYAGEMKKAMFT
R.viridis       WTILSLPGLKTDPRCDGVHSDATLMLHLSERKVLLCGHRYAGEIKKAMFS
                * :*. . :  .*  * .:.:. ::::.: .:::*: *  ****: ****:
```

```
A.lusitana        VLNFILPEHNILPMHCAANAGENGDTALFFGLSGTGKTTLSADPERFLIG
A.siphonis        VLNFVLPQHDILPMHCAANASKEGDTALFFGLSGTGKTTLSADPKRLLIG
C.bacterium       VLNFILPEKNVLPMHCAANVGKQGDVALFFGLSGTGKTTLSADPERYLIG
C.burnetii        VLNYLLPPHDVLPMHCAANAGQSGDVALFFGLSGTGKTTLSADPHRFLIG
C.mudrowiae       VLNYLLPPHDVLPMHCAANAGKSGDVALFFGLSGTGKTTLSADPNRFLIG
L.bacterium       VMNFLLPNIDVLPMHCASNIGMEGDVALFFGLSGTGKTTLSADPERFLIG
M.wilcox          VQNFLLPEKDVLPMHCSANVGEDGETTLFFGLSGTGKTTLSADPARYLIG
P.japonicum       VQNFLLPEKDVLPMHCSANVGEDGETTLFFGLSGTGKTTLSADPSRYLIG
R.isopodorum      VMNYLLPAVDVLPMHCSANVGKEGDVALFFGLSGTGKTTLSADPDRFLIG
R.microplus       VLNYLLPPQDVLPMHCAANTGKSGDVALFFGLSGTGKTTLSADPNRFLIG
R.viridis         VLNYLLPASDVLPMHCSANVGKQGDVALFFGLSGTGKTTLSADPERYLIG
                  * *:::**  ::******::* . .*:.:***************** * ***

A.lusitana        DDEHGWGNDGVFNFEGGCYAKCIDLSEEREPLIWNAIRYGSVIENVVLDP
A.siphonis        DDEHGWGEDGIFNFEGGCYAKCIDLSPEREPLIWNAIRFGTVIENVVLNP
C.bacterium       DDEHGWSDHGVFNFEGGCYAKCINLSKEREPVIWDAIRYGAIMENVVLDP
C.burnetii        DDEHGWSATSVFNFEGGCYAKCIDLSQEREPMIWNAIRHGAIMENVVLDE
C.mudrowiae       DDEHGWSRTGVFNFEGGCYAKCIDLSSEREPMIWEAIRHGAIMENVVLQA
L.bacterium       DDEHGWGKSGVFNFEGGCYAKCIDLSKEKEPVIWDAIRHGAIMENVVLDE
M.wilcox          DDEHGWGEGTVFNIEGGCYAKCIDLSEKNEPVIWQAIRFGAVLENVVLDD
P.japonicum       DDEHGWGKGTVFNIEGGCYAKCIDLSAENEPVIWNAIRFGAVLENVILDE
R.isopodorum      DDEHAWSETGVFNFEGGCYAKCIDLSKEREPLIWNAIRHGAVMENVVLDP
R.microplus       DDEHGWSRTGVFNFEGGCYAKCIDLSLEREPIIWESIRYGAIMENVVLQA
R.viridis         DDEHGWSENSVFNFEGGCYAKCIDLSKEREPVIWNAIRHGAVMENVVLDP
                  ****.*.   :**:**********:** :.**:**::**.*::::***:*:

A.lusitana        VTKNPDYGDASLTQNTRAAYPREFIPQRVENNRG-RQPNAVLFLTCDLYG
A.siphonis        QTREPDYADASLTQNTRAAYPREFIPERVENNRG-RQPHAVLFLTCDLYG
C.bacterium       KTKEPLYGDASLTENTRAAYPLEHIAMRVPENQA-GHPQAVIFLTCDLYG
C.burnetii        N-GVPDYADARLTQNSRAAYPREYIPLRVENNRG-RPPDAVLFLTCDLDG
C.mudrowiae       D-GQPDYRNASLTQNTRAAYPREHISLRVKDNRG-RPPDSVIFLTCDLYG
L.bacterium       N-QAPDYSDSTLSMNSRAAYPREHIEMRAEANRG-GQPDAVLFLTCDLYG
M.wilcox          R-RAPDYADDSLTQNSRAAYPLEHIDKRVEENRA-GEPSAIIFLTCDMSG
P.japonicum       R-RVPDYNDDSLTQNSRAAYPLEHIEKRVLENRA-GEPSAIVFLTCDMSG
R.isopodorum      ETLDPNYKDARLTQNTRVAYPLNFIESRFRANRVDRLPDAVIFLCCDLYG
R.microplus       D-GQPAYNDASLTQNTRAAYPREHILFRVKENRG-RPPDAVIFLTCDLYG
R.viridis         HTLEPDYKDASLTQNTRVAYPLDFISLRVPENRVEQLPSAVIFLTCDLYG
                      * * :  *: *:*.*** :.*  *   *:    * :::** **: *

A.lusitana        VLPPVARLTPEQAAYYFLSGYTALVGSTEVGQGSGIKPTFSTCFGAPFFP
A.siphonis        VLPPVARLTPEQAAYYFLSGYTALVGSTEVGQGSGIKPTFSTCFGAPFFP
C.bacterium       VLPPVAILNKEQAAYHFLSGYTALVGSTEVGSTAGIKSTFSTCFGAPFFP
C.burnetii        VLPPVALLTKEQAAYYFLSGYTALVGSTEVGSVKGVTSTFSTCFGAPFFP
C.mudrowiae       VLPPVALLTKEQAAYYFLSGYTALVGSTEVGSVKGVTPTFSTCFGAPFFP
L.bacterium       VLPPVSLLSKEQAAYHFLSGYTALVGSTEVGQTEGIKPTFSTCFGAPFFP
M.wilcox          VLPPVSVLSKEAAAYHFLSGYTAKVGSTEMGSSAGLEATFSTCFGAPFFP
P.japonicum       VLPPVSILSKEAAAYHFLSGYTAKVGSTEMGSSSGLEATFSTCFGAPFFP
R.isopodorum      VLPPIACLNHEQAAYYFLSGYTALVGSTEVGQTEPIKTTFSTCFGAPFFP
R.microplus       VLPPVSLLTKAQAAYYFLSGYTALVGSTEVGSVKGIVPTFSSCFGAPFFP
R.viridis         VLPPVARLSHEQAAYYFLSGYTALVGSTEVGQTEAIKTTFSTCFGAPFFP
                  ****:: *.   ***:******* ******:*.   : .***:********
```

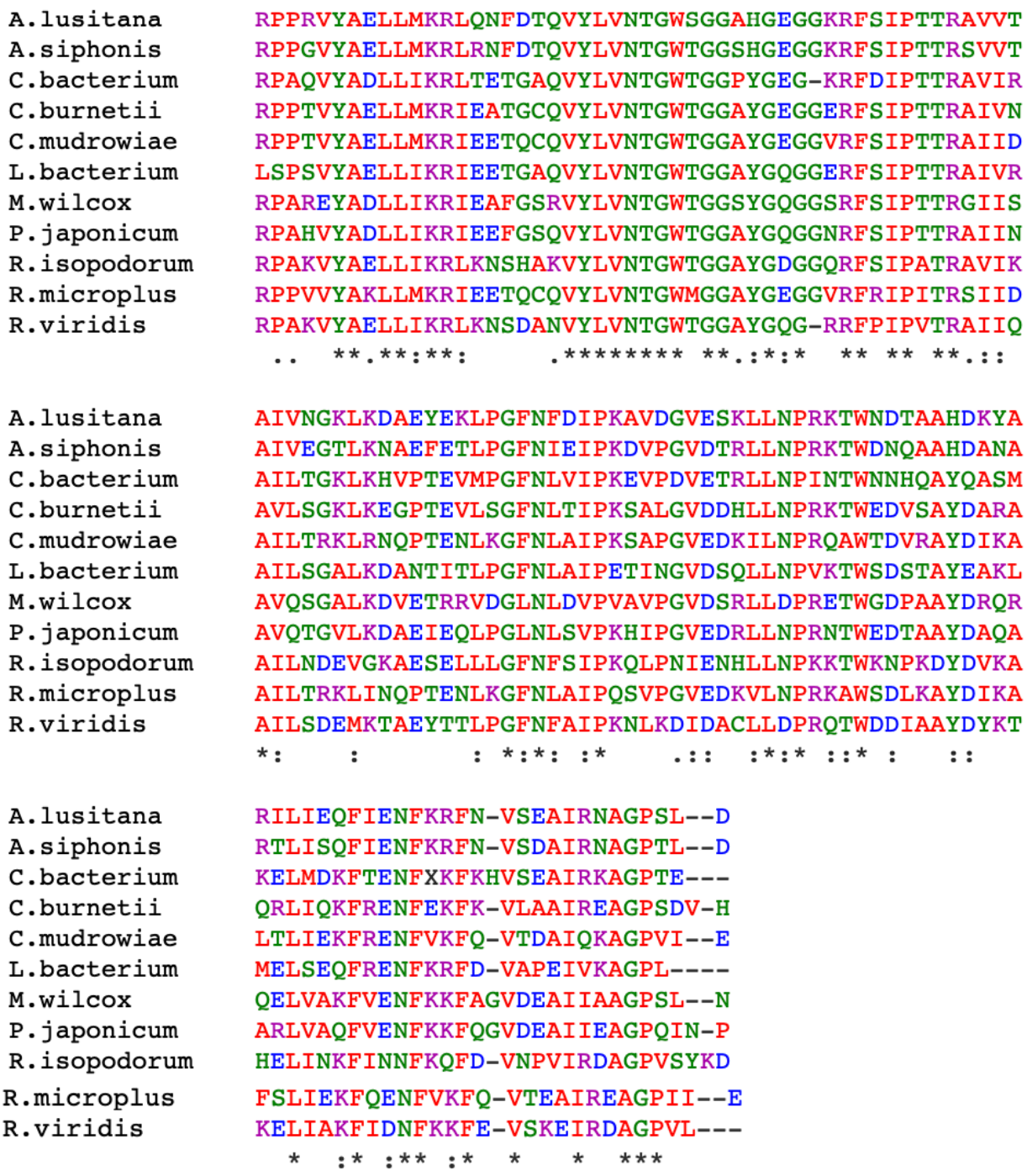


Figure 120: T-COFFEE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10840 and the top 10 organism sequences similarities selected.

Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered randomly and are listed from top to bottom as followed: *A. Lusitana*, *A. siphonis*, *C. bacterium*, *C. burnetii*, *C. mudrowiae*, *L. bacterium*, *M. wilcox*, *P. japonicum*, *R. isopodorum*, *R. microplus*, *R. viridis*. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

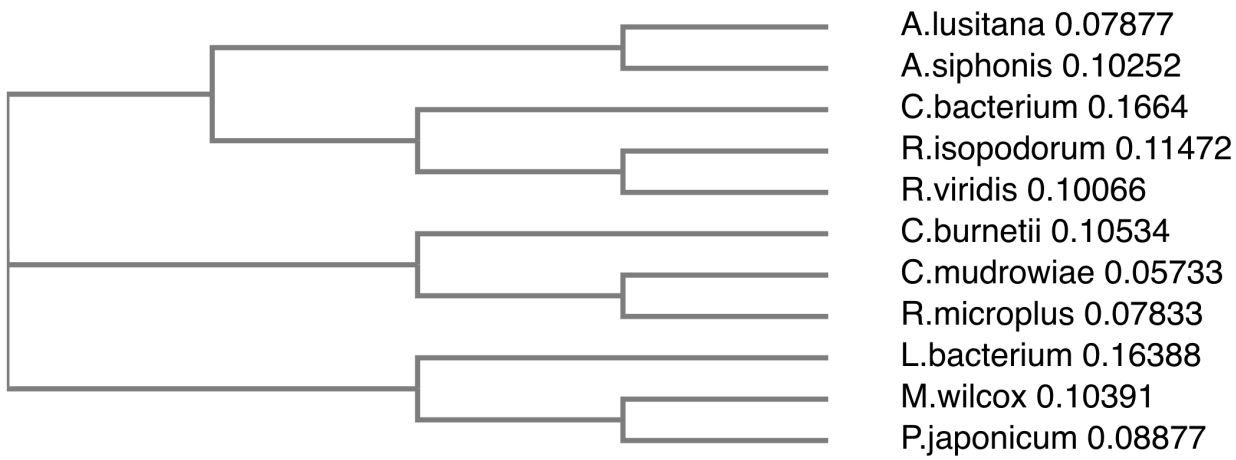


Figure 121: T-COFFEE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10840 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *A. lusitana*, *A. siphonis*, *C. bacterium*, *R. isopodorum*, *R. viridis*, *C. burnetii*, *C. mudrowiae*, *R. microplus*, *L. bacterium*, *M. wilcox*, *P. japonicum.* The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

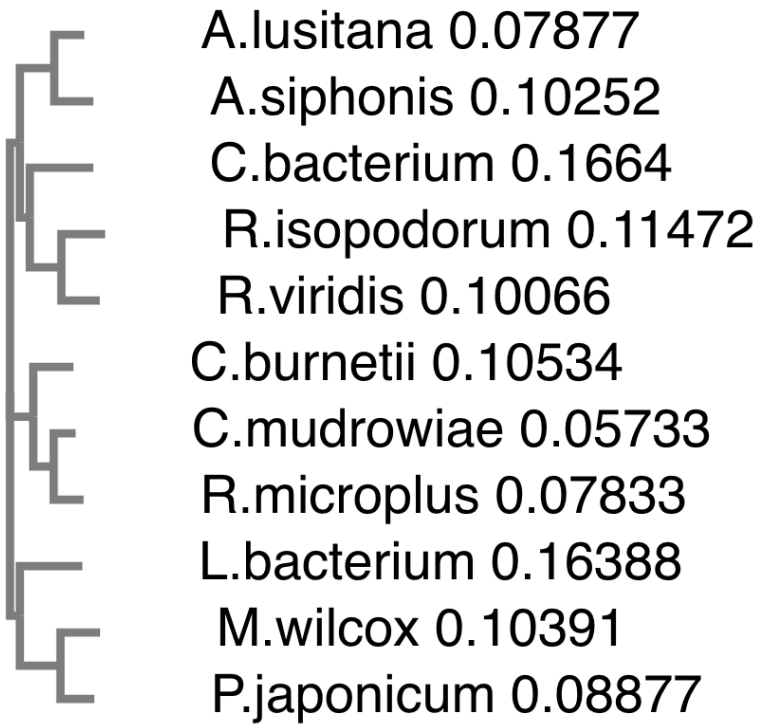
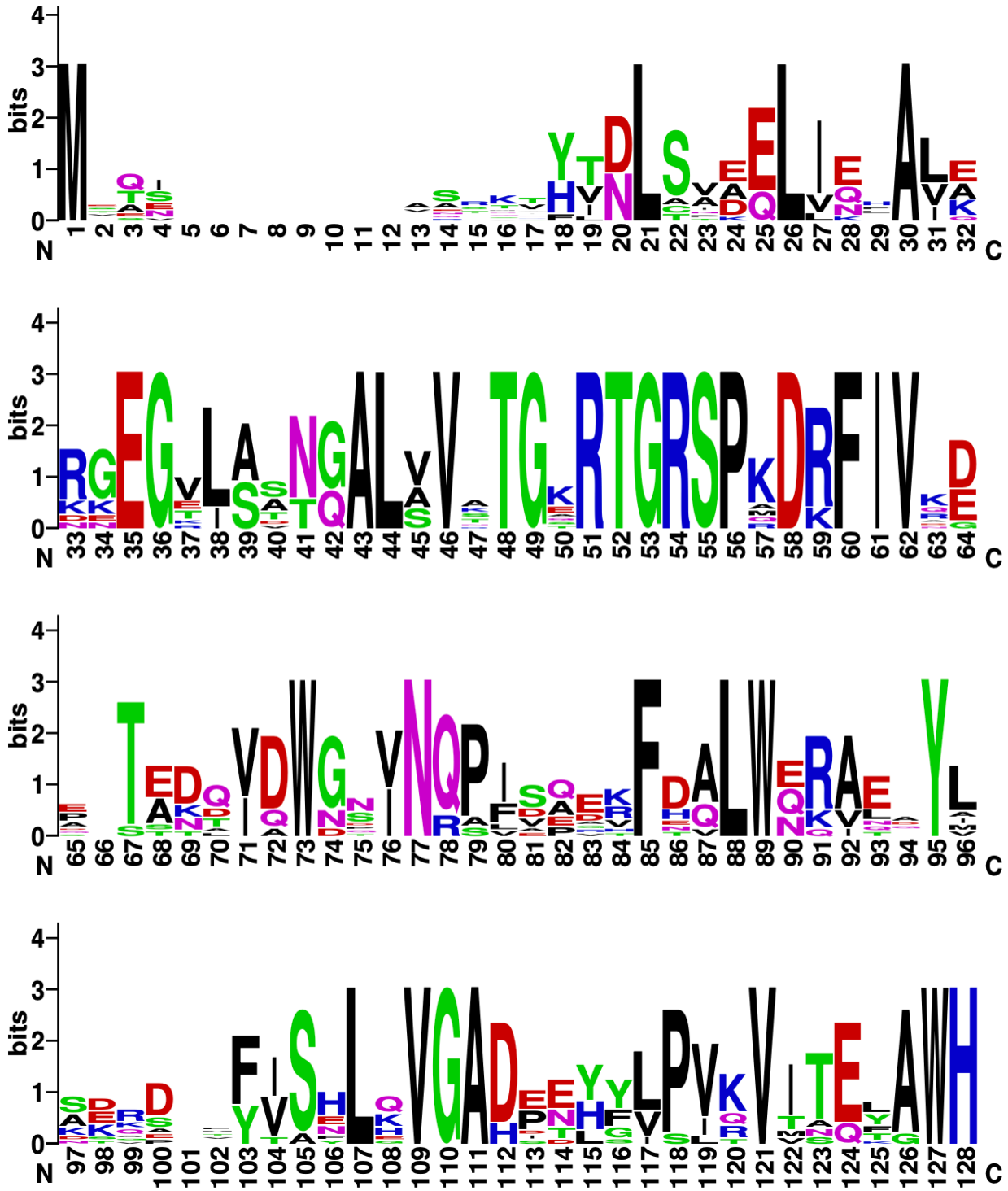
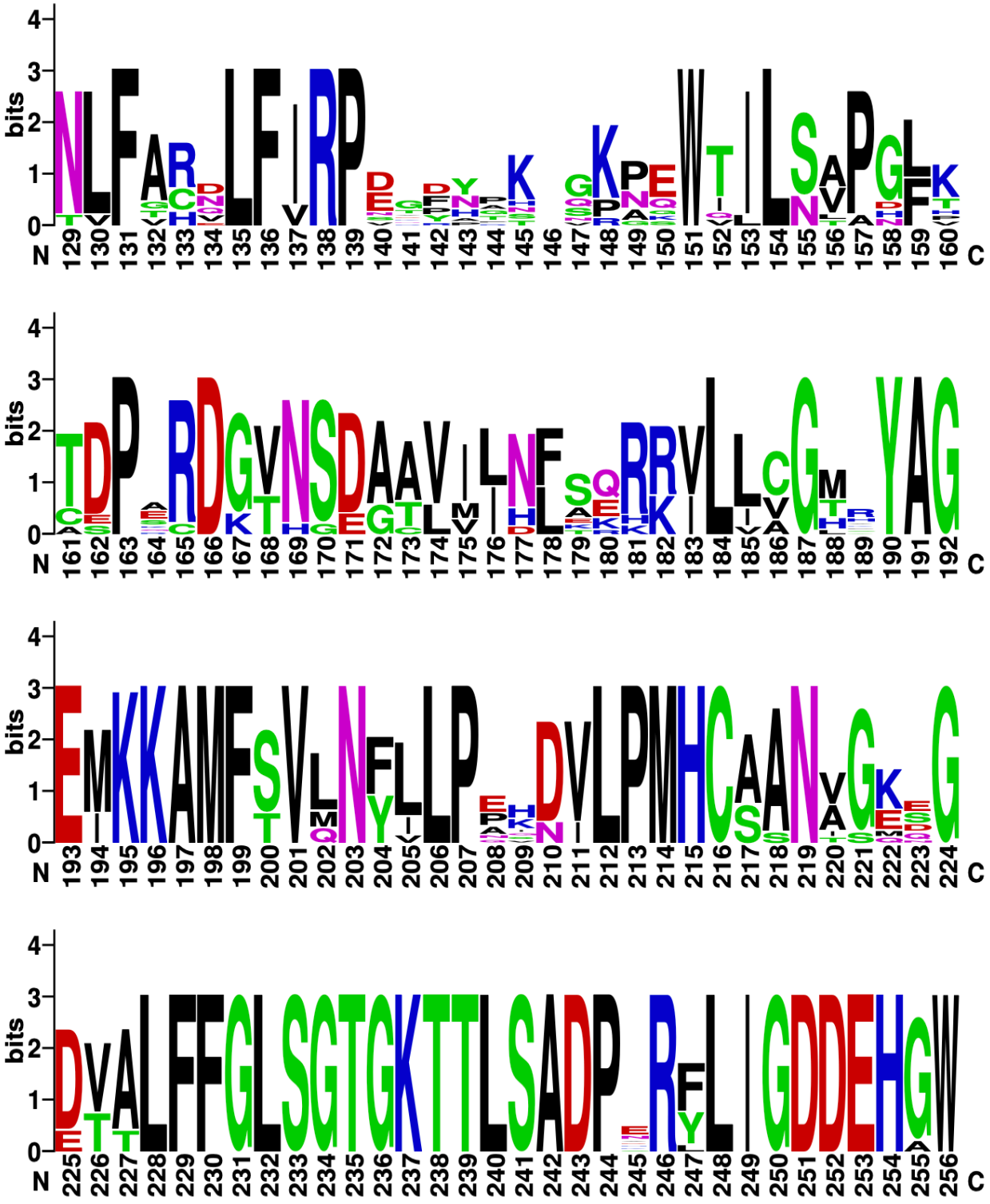
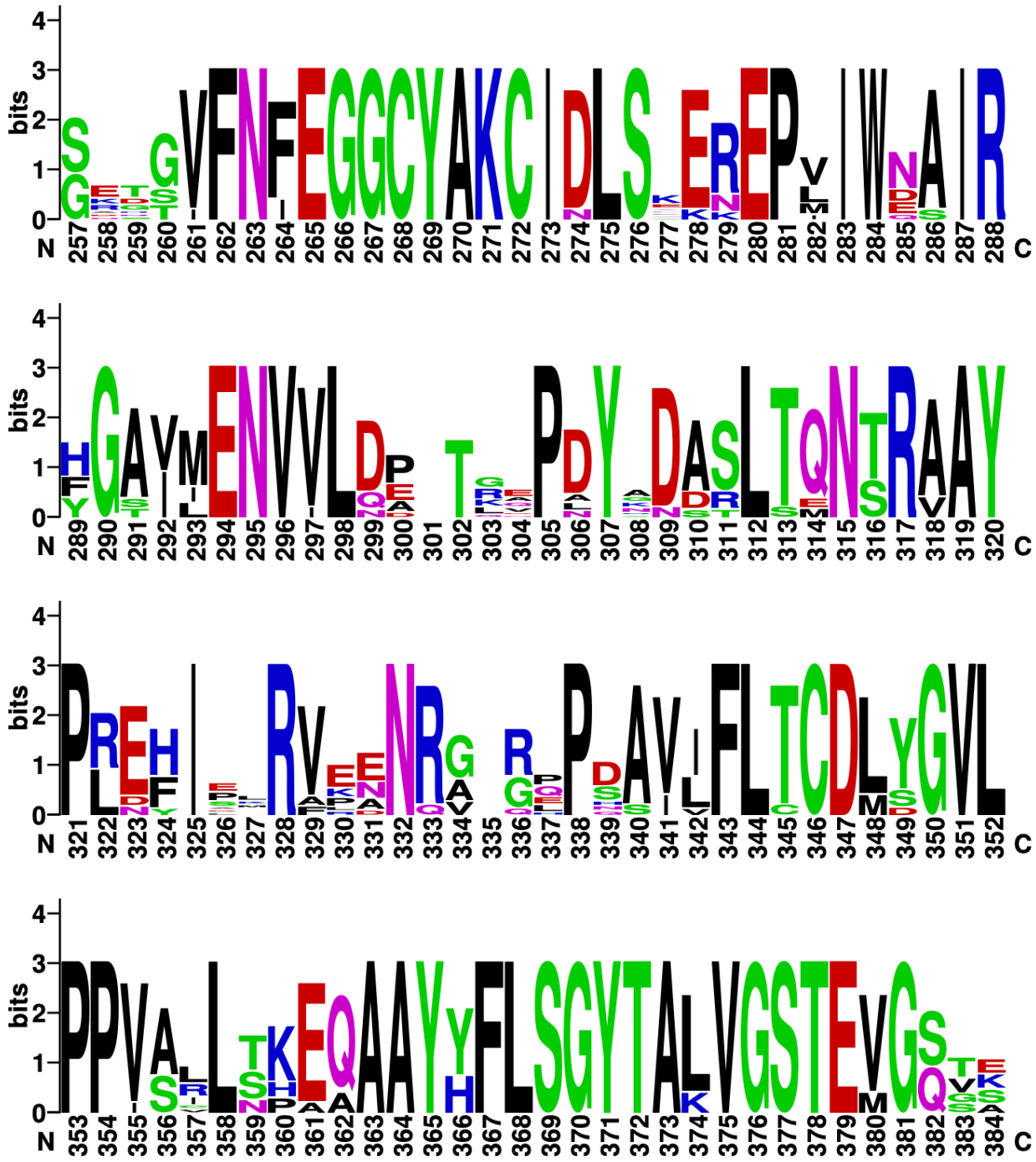


Figure 122: T-COFFEE multiple sequence alignment real phylogenetic tree for *Coxiella burnetii* gene BMW92_RS10840 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *A. lusitana*, *A. siphonis*, *C. bacterium*, *R. isopodorum*, *R. viridis*, *C. burnetii*, *C. mudrowiae*, *R. microplus*, *L. bacterium*, *M. wilcox*, *P. japonicum.* The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

T-COFFEE Sequence Logo




bits
4
3
2
1
0
N
C


bits
4
3
2
1
0
N
C

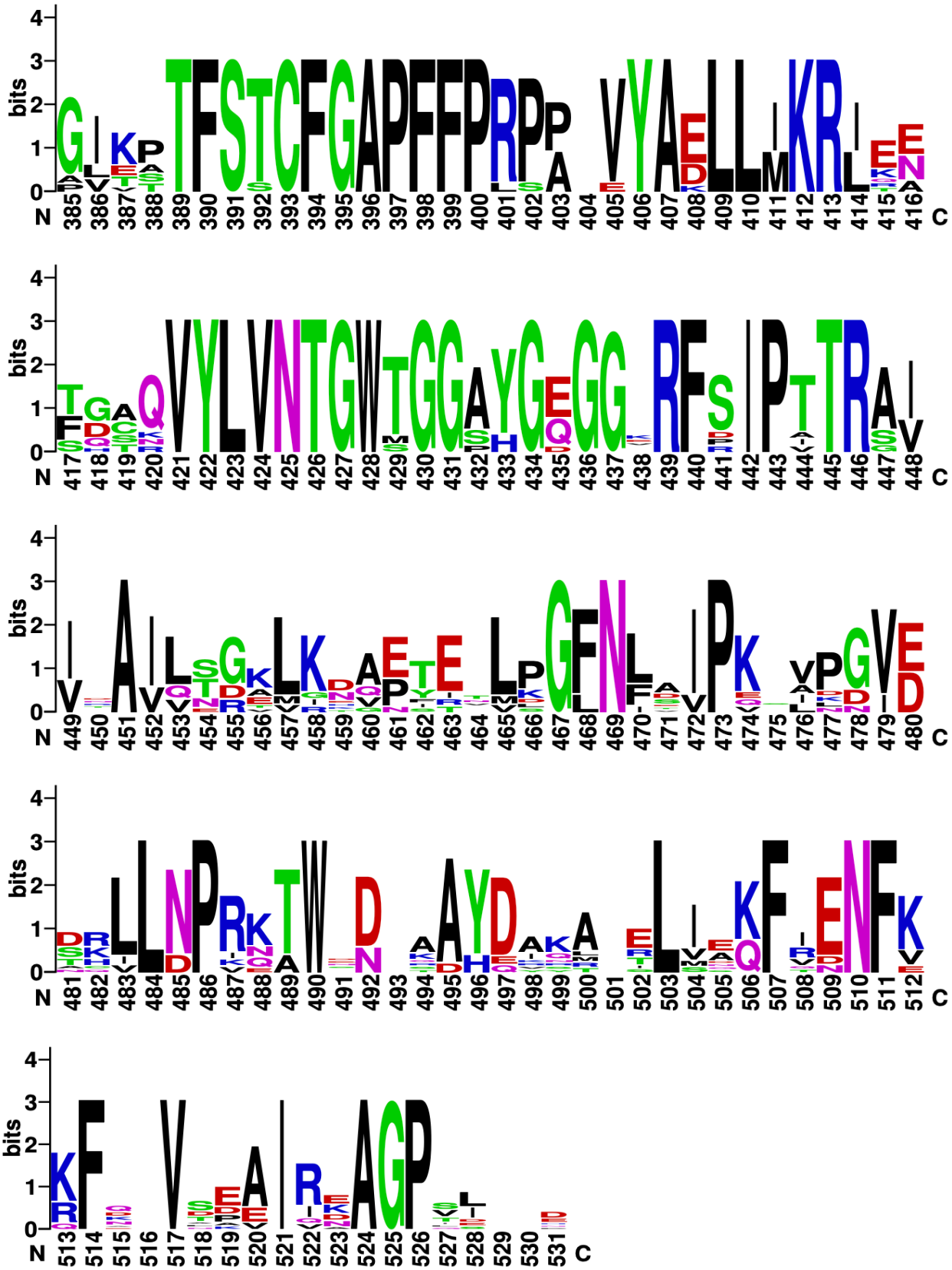
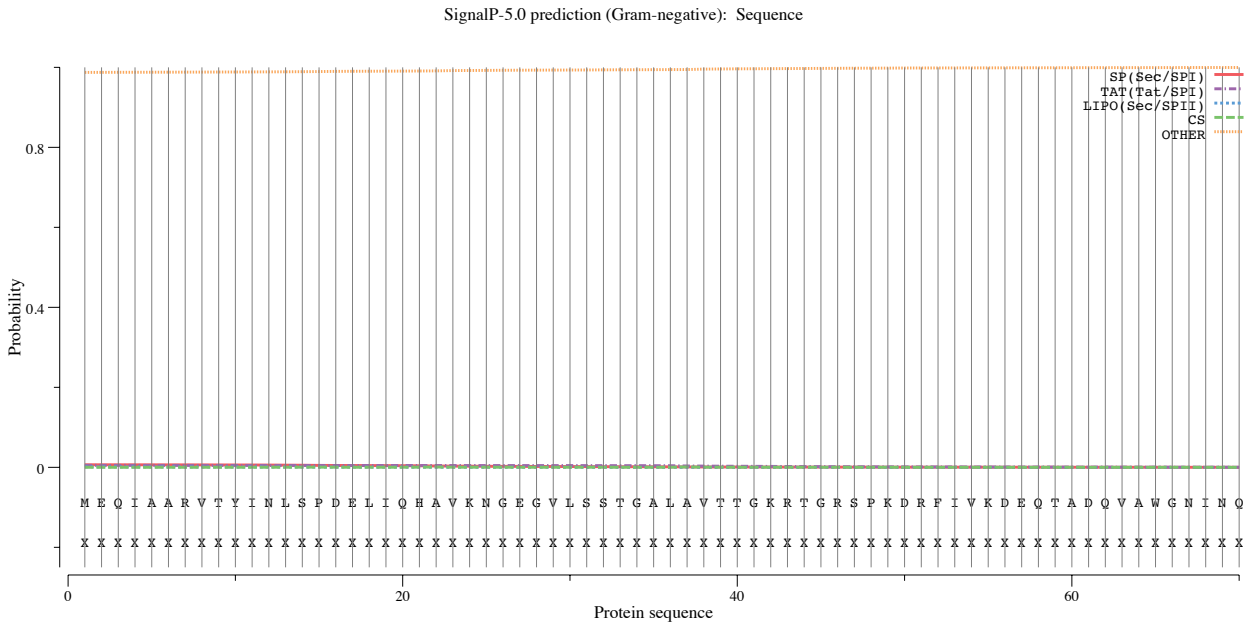


Figure 123: Sequence logo generated from the T-COFFEE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10840 and the top 10 organism sequences similarities selected. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. The conservation of each amino acid among residue across similar sequences is represented by the height of each single letter abbreviation of the amino acid. (WebLogo, <https://weblogo.berkeley.edu/logo.cgi>).

Protein Localization

SignalP



Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.0063	0.0047	0.0011	0.9879

Figure 124: SignalP 5.0 prediction (Gram-negative) for gene BMW92_RS10840 of *Coxiella burnetii*. The SP (Sec/SPI), TAT (Tat/SPI), LIPO (Sec/SPII), and CS probability scores combined were all less than a total 0.0184 (1.84%) which results in the likelihood of the protein being a signal peptide as highly unlikely and can confirm there is no signal peptide of these protein types. The program calculated the probability scores for OTHER as 0.9879 (98.79%). This probability score indicates the protein from the gene BMW92_RS10840 has another protein

classification that is not related to similar function or type as a signal peptide (SignalP, <http://www.cbs.dtu.dk/services/SignalP/ >).

LipoP

```
# Sequence CYT score=-0.200913
# Cut-off=-3
Sequence        LipoP1.0:Best   CYT     1       1       -0.200913



# NO PLOT made - less than 4 putative cleavage sites predicted
```

Figure 125: LipoP 1.0 was unable to generate a plot graph due to there being less than four predicted putative cleavage sites. The best localization prediction resulted in the highest scoring class being the cytoplasmic protein class (LipoP, <http://www.cbs.dtu.dk/services/LipoP/>).

TMHMM

```
# WEBSEQUENCE Length: 517
# WEBSEQUENCE Number of predicted TMHs:  0
# WEBSEQUENCE Exp number of AAs in TMHs: 0.05524
# WEBSEQUENCE Exp number, first 60 AAs:  0.00167
# WEBSEQUENCE Total prob of N-in:        0.00762
WEBSEQUENCE     TMHMM2.0        outside      1   517
```

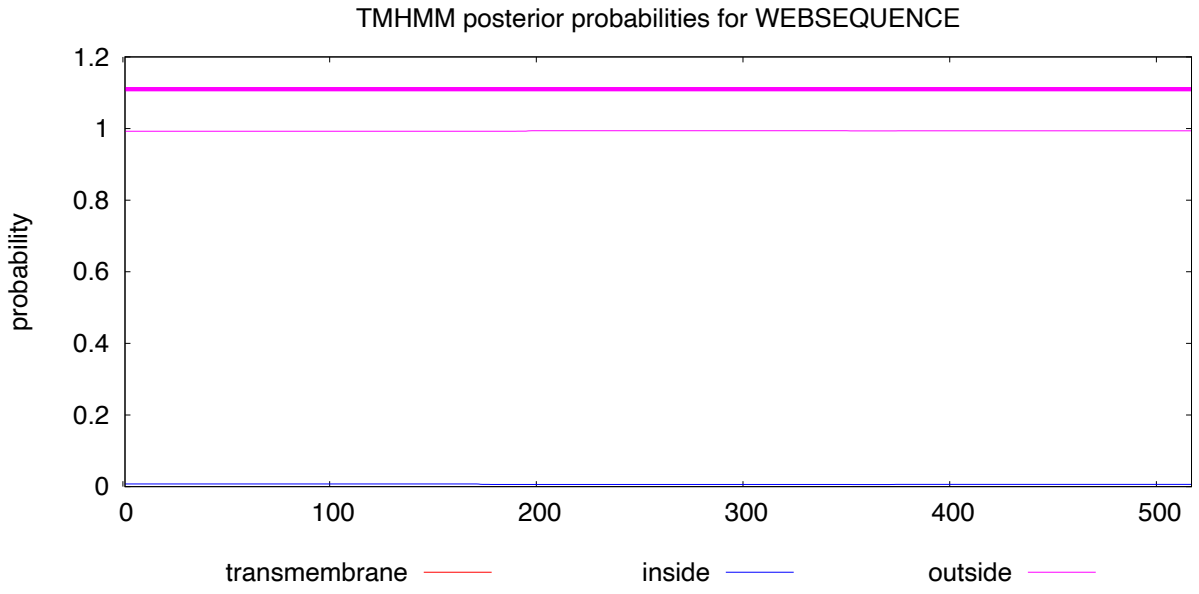


Figure 126: TMHMM posterior probability displayed a line graph that predicts the localization of the protein coded from BMW92_RS10840 as entirely outside the membrane. The red line, representative of the protein being located in the transmembrane, was less than 0.001 (0.10% probability) across the entirety of the line graph. This is indicative of the protein being located within the transmembrane as highly unlikely. The blue line, representative of the protein being located inside the membrane, was at 0.001 (0.10% probability). This is indicative of the protein being located inside of the membrane as highly unlikely. The magenta line, representative of the protein being located outside the membrane, was at 0.99 (99% probability). This is indicative of

the protein being located outside of the membrane as highly likely (TMHMM, <http://www.cbs.dtu.dk/services/TMHMM/ >).

BOMP

The total number of valid proteins submitted is: 1
The total number of integral ß-barrel outer membrane proteins predicted is: 0

Sequence name	Category	Best BLAST hit

Figure 127: The BOMP test result identified there are no integral beta-barrel outer membrane proteins for gene BMW92_RS10840 (BOMP, <http://services.cbu.uib.no/tools/bomp>).

PSORTb

```
SeqID: C.burnetii
  Analysis Report:
    CMSVM-              Unknown                     [No details]
    CytoSVM-            Unknown                     [No details]
    ECSVM-              Unknown                     [No details]
    ModHMM-             Unknown                     [No internal helices found]
    Motif-              Unknown                     [No motifs found]
    OMPMotif-           Unknown                     [No motifs found]
    OMSVM-              Unknown                     [No details]
    PPSVM-              Unknown                     [No details]
    Profile-            Unknown                     [No matches to profiles found]
    SCL-BLAST-          Cytoplasmic                 [matched 1172572: Phosphoenolpyruvate carboxykinase [ATP]]
    SCL-BLASTe-         Unknown                     [No matches against database]
    Signal-             Unknown                     [No signal peptide detected]
  Localization Scores:
    Cytoplasmic             9.26
    CytoplasmicMembrane     0.24
    Periplasmic             0.48
    OuterMembrane           0.01
    Extracellular           0.01
  Final Prediction:
    Cytoplasmic             9.26
```

Figure 128: The PSORTb test resulted in an analysis report that identified no detectable internal helices, motifs, or signal peptides. The PSORTb localization scores resulted in a 9.26 value for the cytoplasmic location. The localization score for cytoplasmic membrane was 0.24. The localization score for periplasmic was 0.48. The localization score for the outer membrane location was 0.01. The localization score for the extracellular location was 0.01. The calculated localization scores for gene BMW92_RS10840 resulted in the final predictable location of the protein to be cytoplasmic (PSORTb, <https://www.psort.org/psortb/ >).

Phobius

```
ID   UNNAMED
FT   TOPO_DOM      1     517       NON CYTOPLASMIC.
//
```

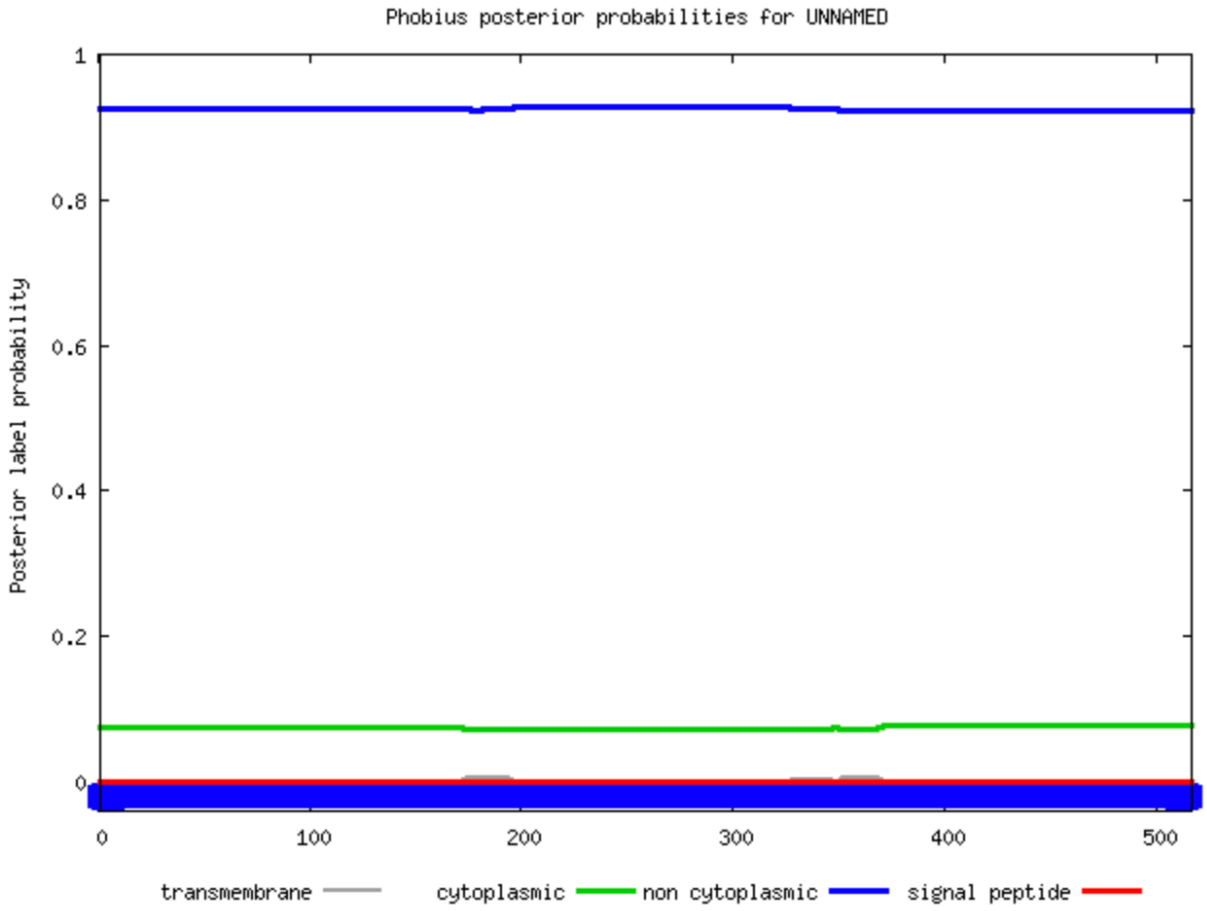


Figure 129: The Phobius posterior probability line graph generated for gene BMW92_RS10840 resulted in a calculated prediction that the whole sequence contains no membrane helices. The grey line, representative of the predicted transmembrane helices location, was around 0.001 (0.1%) posterior probability from amino acids 180-195. The green line, representative of the predicted cytoplasmic transmembrane helices location, was around 0.09 (9%) posterior

probability from amino acids 0-517. The blue line, representative of the predicted non-cytoplasmic transmembrane helices location, was around 0.93 (93%) posterior probability from amino acids 0-517. The red line, representative of the presence or absence of a signal peptide, was 0.00 (0%) posterior probability (Phobius, <http://phobius.sbc.su.se>).

BMW92_RS10855

The fifth gene, BMW92_RS10855, was analyzed using bioinformatic technology. Table 5 below contains the provided data regarding basic information. A protein isoelectric point calculator was used to determine the isoelectric point of the protein, protein length, and the number and prevalence of each amino acid that makes up the protein (Figure 130). The BLASTp search tool produced 100 matches ranked from highest sequence similarity to lowest sequence similarity. The top ten sequences with significant alignments that were not identical species to *Coxiella burnetii* were selected. The information recorded included the organism name, protein name, percent identity, precent positive, length of alignment match, e-values, and percent gap. The highest ranked match to the BMW92_RS10855 gene was aspartate cabamoyltransferase catalytic subunit [*Coxiella mudrowiae*] (Figure 131). The remaining nine matches to the BMW92_RS10855 gene all had a function as aspartate cabamoyltransferase catalytic subunit (Figures 132-140). The CDD identified five potential protein domains hits conserved (Figure 141). Four of the domain hits conserved and identified by the CDD belong to the pyrB, PyrB, asp_carb_tr, and OTCace_N superfamilies (Figure 142) Specific domain hits involved PyrB and OTCace_N. Two domain hits conserved and identified as a non-specific domain hit were pyrB and asp_carb_tr. The protein classification identified by the CDD was aspartate cabamoyltransferase catalytic subunit. Four of the domain hits sequences were aligned with the query sequence based off the amino acids that are highly conserved between both sequences (Figures 143-146). The MUSCLE program generated a multiple sequence alignment (MSA); each amino acid in the sequence was assigned a distinct color to distinguish the amino acids being compared (Figure 147). The MUSCLE program generated two phylogenetic trees using the multiple sequence alignments to further confirm sequence similarity. The results displayed

the numbers followed behind each organism at the end of each leaf node which displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii.* The use of a phylogenetic cladogram (Figure 148) and real phylogenetic tree (Figure 149) provided further understanding of the relatedness of common ancestors and organism sequences that are conserved. Each of the letter's heights produced correspond to the conservation of the amino acid residue across similar sequences. WebLogo produced a sequence logo that was generated from the MUSCLE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10855 and the top 10 organism sequences similarities selected (Figure 150). Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. Each of the single letter amino acid abbreviation's heights correspond to the conservation of the amino acid residue across similar sequences. The T-COFEE program generated another multiple sequence alignment to further confirm sequence similarity depicted with in the MUSCLE MSA (Figure 151). The T-COFFEE program generated two phylogenetic trees, phylogenetic cladogram (Figure 152) and real phylogenetic tree (Figure 153), using the multiple sequence alignment which displayed the genetic proximity and similarity between *Coxiella burnetii* and selected organisms from the BLASTp search. WebLogo constructed a sequence logo from the T-COFFEE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10855 and the top 10 organism sequences similarities selected to further display sequence similarity and conservation of sequences. Each of the single letter amino acid abbreviation's heights correspond to the conservation of the amino acid residue across similar sequences (Figure 154). Protein localization results included SignalP, LipoP, TMHMM, BOMP, PSORTb, and Phobius. The SignalP graphical illustration identified that there is no presence of a signal peptide for the entirety of the protein sequence (Figure 155). The

LipoP resulted in the highest scoring class being the cytoplasmic protein class (Figure 156). The TMHMM test resulted in a graphical illustration, statistics, and a list of the predicted transmembrane helices and the predicted location of the intervening loop regions. The TMHMM test resulted and displayed that the whole sequence is highly unlikely to contain any transmembrane helices and that the majority of the protein has a high probability of being located outside of the membrane (Figure 157). The BOMP test result identified there are no integral beta-barrel outer membrane proteins (Figure 158). The PSORTb test resulted in an analysis report that identified no detectable internal helices, motifs, or signal peptides; the localization scores calculated the predictable location of the protein to be cytoplasmic (Figure 159). The Phobius test resulted in a line graphical illustration that identified a low probability of transmembrane helices localized in the cytoplasm and a high probability of non-cytoplasmic transmembrane helices present; the overall result calculated by Phobius resulted in the entire protein sequence as non-cytoplasmic, which is contradictory to LipoP and PSORTb results (Figure 160).

Basic Information

Table 3: Gene BMW92_RS10835 basic information

Genome	Replicon	Locus Tag	Old Locus Tag
Coxiella burnetii	NZ_CP018005	BMW92_RS10855	BMW92_10485
Genomic Coordinates	**Products**	**Length**	**Start and End Position**
1968231..1969163	Aspartate carbamoyltransferase	933 / 310	1968231 - 1969163
Molecular Weight	**Average Isoelectric Point**	**IPC Protein**	**Protein Length**
34944.39434 Da	8.913	8.174	312 amino acids

Nucleotide Sequence	Amino Acid Sequence
atgaatgaacttcctttacatttattgaatatgcgctcactcac gcgcgaccatattgaaaaactcatccaacgggcgaattatttt ttaactcagggcatggaaaaaaattcggtctttgaaacattaa aagggcacgtagtcgccaacttatttttttgaacccagcacacg aacgcgcaactccttttgaaattgcggcaaaacgtttgggcgcc atggttcttaaccctaatcttaaaatttcggcaataagtaaag gtgaaactctttttgatacgattaaaactttggaagcgatgggt gtttatttttttcattgtacgccattctgaaaatgaaaccccgga gcagatagcaaaacaattatcctcaggcgtcgtcatcaacgc gggtgacggtaatcatcaacatccctcacaagctttaattgatt taatgacaataaagcaacacaaaccccattggaataaattgt gcgtcacgattattggcgatattcgtcattctcgcgtggcaaac tcattaatggacggcttagtcacgatgggcgttccggaaattc gattggtaggcccatcgtcattattgccggacaaggtcgggaa cgactcgattaaaaaattcaccgaattaaaaccaagtctcctt aacagcgacgttattgtcacccttcgtttgcaaaaggaacgcc atgataattctgtcgatatcgatgcttttcgcggatcttttcgat tgacacctgaaaaattatattccgcaaaacccgatgccattgt tatgcatccgggtcccgtcaaccgcgaagtcgaaattaattct gatgtcgcagataaccaacaatctgtcatccttcaacaagtac gtaacggcgtcgctatgcggatggctgtgttggaattatttttg ttgcgagattttcgatttttttga	MNELPLHLLNMRSLTRDHIEKLIQRANYFLTQGME KNSVFETLKGHVVANLFFEPSTRTRNSFEIAAKRLG AMVLNPNLKISAISKGETLFDTIKTLEAMGVYFFIVR HSENETPEQIAKQLSSGVVINAGDGNHQHPSQALI DLMTIKQHKPHWNKLCVTIIGDIRHSRVANSLMD GLVTMGVPEIRLVGPSSLLPDKVGNDSIKKFTELKP SLLNSDVIVTLRLQKERHDNSVDIDAFRGSFRLTPE KLYSAKPDAIVMHPGPVNREVEINSDVADNQQSVI LQQVRNGVA MRMAVLELFLLRDFRFF

Figure 130: Protein isoelectric point calculator. The number and prevalence of each amino acid in the protein coded from the BMW92_RS10855 gene of *Coxiella burnetii* (Kozlowski, Biology Direct, <http://isoelectric.org/>).

Sequence Similarity

BLAST

aspartate carbamoyltransferase catalytic subunit [Candidatus Coxiella mudrowiae]

Sequence ID: WP_048875734.1 **Length:** 316 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 5 to 305 GenPept Graphics Next Match Previou

Score	Expect	Method	Identities	Positives	Gaps
421 bits(1081)	3e-145	Compositional matrix adjust.	201/302(67%)	248/302(82%)	1/302(0%)

```
Query  3    ELPLHLLNMRSLTRDHIEKLIQRANYFLTQGMEKNSVFETLKGHVVANLFFEPSTRTRNS  62
            + P HLL M+SLTR+ I+ +++RAN FL + +++N VF+TLKG VVANLFFE STRTRNS
Sbjct  5    DFPYHLLGMQSLTRNEIDLILKRANDFL-RNIKENRVFDTLKGEVVANLFFETSTRTRNS  63

Query  63   FEIAAKRLGAMVLNPNLKISAISKGETLFDTIKTLEAMGVYFFIVRHSENETPEQIAKQL  122
            FEIAAKRL A+VL+P+LK+SA++KGE+L D  + L+AMG  FF++RH+EN  P  +A+ L
Sbjct  64   FEIAAKRLEAIVLSPDLKVSALNKGESLLDMARNLQAMGTRFFVIRHTENNRPRMLAEHL  123

Query  123  SSGVVINAGDGNHQHPSQALIDLMTIKQHKPHWNKLCVTIIGDIRHSRVANSLMDGLVTM  182
              G+VINAGDGNHQHP+Q LIDLMTI+QHKP W KLCVTIIGDI HSRVANSL+DGL+ M
Sbjct  124  EQGIVINAGDGNHQHPTQGLIDLMTIQQHKPDWTKLCVTIIGDIYHSRVANSLVDGLLIM  183

Query  183  GVPEIRLVGPSSLLPDKVGNDSIKKFTELKPSLLNSDVIVTLRLQKERHDNSVDIDAFRG  242
            GVPEIR+ GPS LLP+ V N  IKK  EL+ SL+NSDV+VTLRLQKERH N  +++ FR
Sbjct  184  GVPEIRITGPSQLLPETVKNPRIKKIPELEASLINSDVVVTLRLQKERHSNLTELNTFRQ  243

Query  243  SFRLTPEKLYSAKPDAIVMHPGPVNREVEINSDVADNQQSVILQQVRNGVAMRMAVLELF  302
             F L+ EK   AKPDAIVMHPGP+NRE+E+ S+VAD +QSVILQQV+NGVA+RMAVLEL
Sbjct  244  LFSLSAEKFALAKPDAIVMHPGPINREIEMTSEVADGKQSVILQQVQNGVAIRMAVLELL  303

Query  303  LL  304
             L
Sbjct  304  FL  305
```

Figure 131: BLAST first match for BMW92_RS10855 sequence from organism *Coxiella mudrowiae* with an e-value of 3e-145, 67% identity, 82% positives, 0% gaps, and an identity of aspartate cabamoyltransferase catalytic subunit (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

aspartate carbamoyltransferase catalytic subunit [Thiotrichales bacterium]

Sequence ID: MAJ10885.1 **Length:** 320 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 16 to 315 GenPept Graphics ▼ Next Match ▲ Previous

Score	Expect	Method	Identities	Positives	Gaps
290 bits(741)	2e-93	Compositional matrix adjust.	148/300(49%)	207/300(69%)	3/300(1%)

```
Query  7    HLLNMRSLTRDHIEKLIQRANYFLTQGMEKNSVFETLKGHVVANLFFEPSTRTRNSFEIA  66
            H L++  L++D +  ++  A  F + G   +     L+G  V NLFFE STRTR +FE+A
Sbjct  16   HFLSIDGLSKDLVTTILDHAETFTSVGERSSKKVPILRGKTVVNLFFESSTRTRTTFELA  75

Query  67   AKRLGAMVLNPNLKISAISKGETLFDTIKTLEAMGVYFFIVRHSENETPEQIAKQLSSGV  126
            AKRL A V+N NL+ SA  KGE+L DT+KTLEAM    F+VRH ++   E IA+Q++  +
Sbjct  76   AKRLSADVMNINLESSATKKGESLSDTLKTLEAMQADMFVVRHQDSGAAEFIARQVAQNI  135

Query  127  -VINAGDGNHQHPSQALIDLMTIKQHKPHWNKLCVTIIGDIRHSRVANSLMDGLVTMGVP  185
             VINAGDG+H HP+QA++D+ TI++HK  +++L V I+GDI HSRVA S +  L  +GVP
Sbjct  136  SVINAGDGSHSHPTQAMLDMFTIRKHKKTFDQLRVAIVGDIAHSRVARSEIHALQILGVP  195

Query  186  EIRLVGPSSLLPDKVGNDSIKKFTELKPSLLNSDVIVTLRLQKERHDNSV--DIDAFRGS  243
            EIRLVGP +L+P  V    ++ F  L+  +  +DVI+ LRLQ+ER  +++      F  +
Sbjct  196  EIRLVGPKTLIPTAVEKLGVRTFNSLESGIDKADVIIMLRLQQERMRSALLPSGREFFNT  255

Query  244  FRLTPEKLYSAKPDAIVMHPGPVNREVEINSDVADNQQSVILQQVRNGVAMRMAVLELFL  303
            F LT EKL SAK D IVMHPGP+NR +EI+S VAD+ +SVILQQV +G+A+RMAV+ L +
Sbjct  256  FGLTKEKLRSAKSDVIVMHPGPINRGIEIDSKVADSPESVILQQVTHGIAVRMAVMSLTM  315
```

Figure 132: BLAST second match for BMW92_RS10855 sequence from organism *Thiotrichales bacterium* with an e-value of 2e-93, 49% identity, 69% positives, 1% gaps, and an identity of aspartate cabamoyltransferase catalytic subunit (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

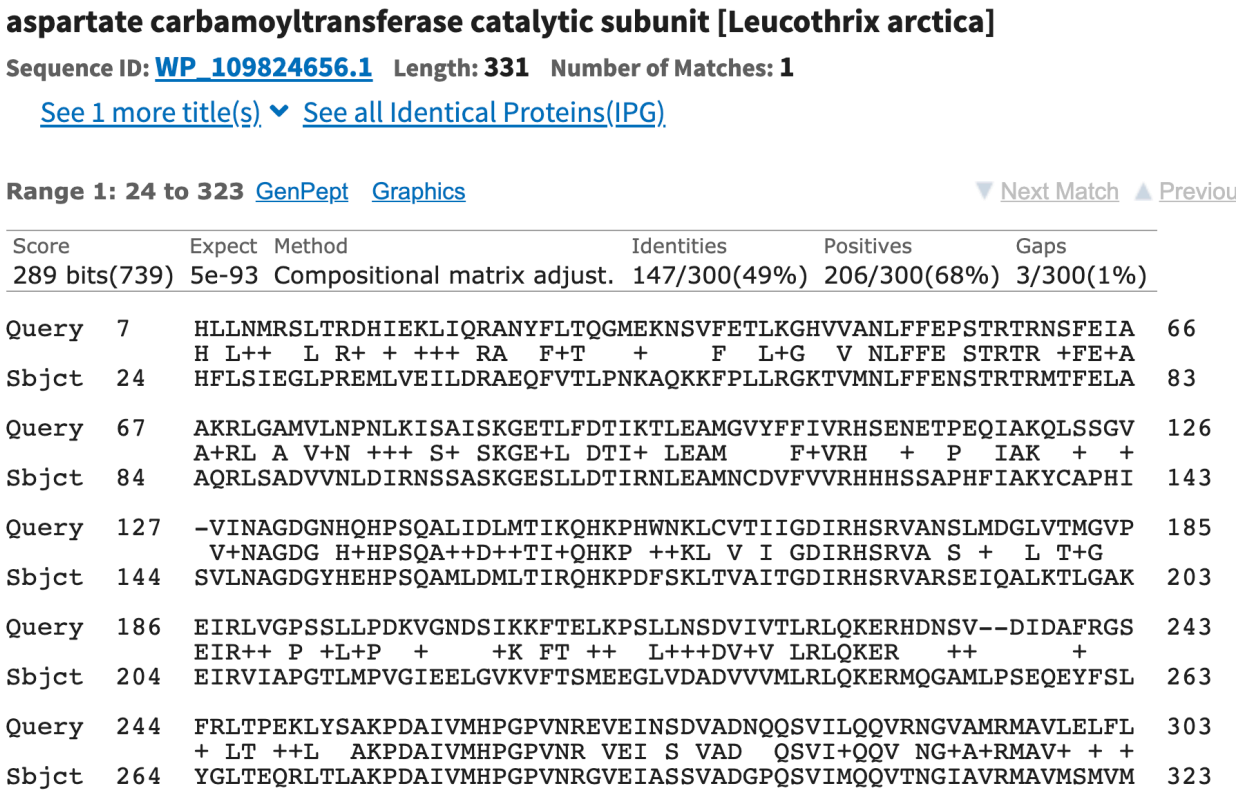


Figure 133: BLAST third match for BMW92_RS10855 sequence from organism *Leucothrix arctica* with an e-value of 5e-93, 49% identity, 68% positives, 1% gaps, and an identity of aspartate cabamoyltransferase catalytic subunit (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

aspartate carbamoyltransferase catalytic subunit [Alteromonadaceae bacterium]

Sequence ID: MAL98818.1 **Length:** 337 **Number of Matches:** 1

Range 1: 22 to 321 GenPept Graphics ▼ Next Match ▲ Previo

Score	Expect	Method	Identities	Positives	Gaps
288 bits(737)	1e-92	Compositional matrix adjust.	148/300(49%)	204/300(68%)	3/300(1%)

```
Query  7    HLLNMRSLTRDHIEKLIQRANYFLTQGMEKNSVFETLKGHVVANLFFEPSTRTRNSFEIA  66
            H L +  L RD + +++  A+ F+  G  +      L+G  V NLFFEPSTRTR++FE+A
Sbjct  22   HFLTIDGLGRDLLTEILDTADSFIEVGERRIKKVPLLRGRTVVNLFFEPSTRTRSTFELA  81

Query  67   AKRLGAMVLNPNLKISAISKGETLFDTIKTLEAMGVYFFIVRHSENETPEQIAKQLSSGV  126
            AKRL A VLN ++  SA SKGE+L DT+  LEAM    F+VRH+++     IA+ ++ GV
Sbjct  82   AKRLSADVLNLDISKSATSKGESLSDTLLNLEAMASDMFVVRHAQSGAAHFIARSVTPGV  141

Query  127  -VINAGDGNHQHPSQALIDLMTIKQHKPHWNKLCVTIIGDIRHSRVANSLMDGLVTMGVP  185
             +INAGDG H HP+QA++D++TI+QHK  +  L V I+GDI HSRVA S ++ L+T+G
Sbjct  142  AIINAGDGRHAHPTQAMLDMLTIRQHKERFEGLRVAIVGDILHSRVARSQVNALLTLGAE  201

Query  186  EIRLVGPSSLLPDKVGNDSIKKFTELKPSLLNSDVIVTLRLQKERHDNSV--DIDAFRGS  243
            E+RLVGP++L+P       +K  T ++  L ++DVI+ LRLQKER ++ +      F
Sbjct  202  EVRLVGPATLMPAAANQLGVKLCTTMEEGLADTDVIIMLRLQKERMESGLLPSEREFFKL  261

Query  244  FRLTPEKLYSAKPDAIVMHPGPVNREVEINSDVADNQQSVILQQVRNGVAMRMAVLELFL  303
            + LT EKL  AKPDAIVMHPGP+NR VEI S VAD  QSVIL QV NG+A+RMAV+ + +
Sbjct  262  YGLTREKLALAKPDAIVMHPGPINRGVEIESAVADGPQSVILSQVTNGIALRMAVMSMAM  321
```

Figure 134: BLAST fourth match for BMW92_RS10855 sequence from organism *Alteromonadaceae bacterium* with an e-value of 1e-92, 49% identity, 68% positives, 1% gaps, and an identity of aspartate cabamoyltransferase catalytic subunit (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

aspartate carbamoyltransferase catalytic subunit [Hydrocarboniclastica marina]

Sequence ID: WP_136546159.1 **Length:** 337 **Number of Matches:** 1

See 1 more title(s) ▾ See all Identical Proteins(IPG)

Range 1: 22 to 321 GenPept Graphics ▼ Next Match ▲ Previo

Score	Expect	Method	Identities	Positives	Gaps
287 bits(735)	3e-92	Compositional matrix adjust.	148/300(49%)	203/300(67%)	3/300(1%)

```
Query  7    HLLNMRSLTRDHIEKLIQRANYFLTQGMEKNSVFETLKGHVVANLFFEPSTRTRNSFEIA  66
            H L +  L RD + +++  A+ F+  G  +      L+G  V NLFFEPSTRTR++FE+A
Sbjct  22   HFLTIDGLGRDLLTEILDTADSFIEVGERRIKKVPLLRGRTVVNLFFEPSTRTRSTFELA  81

Query  67   AKRLGAMVLNPNLKISAISKGETLFDTIKTLEAMGVYFFIVRHSENETPEQIAKQLSSGV  126
            AKRL A VLN ++  SA SKGE+L DT+  LEAM    F+VRH+++     IA+ ++ GV
Sbjct  82   AKRLSADVLNLDISKSATSKGESLSDTLLNLEAMASDMFVVRHAQSGAAHFIARSVTPGV  141

Query  127  -VINAGDGNHQHPSQALIDLMTIKQHKPHWNKLCVTIIGDIRHSRVANSLMDGLVTMGVP  185
             +INAGDG H HP+QA++D++TI+QHK  +  L V I+GDI HSRVA S ++ L+T+G
Sbjct  142  AIINAGDGRHAHPTQAMLDMLTIRQHKERFEGLRVAIVGDILHSRVARSQVNALLTLGAE  201

Query  186  EIRLVGPSSLLPDKVGNDSIKKFTELKPSLLNSDVIVTLRLQKERHDNSV--DIDAFRGS  243
            E+RLVGP++L+P       +K  T ++  L  +DVI+ LRLQKER ++ +      F
Sbjct  202  EVRLVGPATLMPAAANQLGVKLCTTMEEGLAETDVIIMLRLQKERMESGLLPSEREFFKL  261

Query  244  FRLTPEKLYSAKPDAIVMHPGPVNREVEINSDVADNQQSVILQQVRNGVAMRMAVLELFL  303
            + LT EKL  AKPDAIVMHPGP+NR VEI S VAD  QSVIL QV NG+A+RMAV+ + +
Sbjct  262  YGLTREKLALAKPDAIVMHPGPINRGVEIESAVADGPQSVILSQVTNGIALRMAVMSMAM  321
```

Figure 135: BLAST fifth match for BMW92_RS10855 sequence from organism *Hydrocarboniclastica marina* with an e-value of 3e-92, 49% identity, 67% positives, 1% gaps, and an identity of aspartate cabamoyltransferase catalytic subunit (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

aspartate carbamoyltransferase catalytic subunit [Gammaproteobacteria bacterium]

Sequence ID: MAR77762.1 **Length:** 315 **Number of Matches:** 1

Range 1: 17 to 312 GenPept Graphics ▼ Next Match ▲ Previo

Score	Expect	Method	Identities	Positives	Gaps
286 bits(731)	4e-92	Compositional matrix adjust.	145/299(48%)	207/299(69%)	5/299(1%)

```
Query  7    HLLNMRSLTRDHIEKLIQRANYFLTQGMEKNSVFETLKGHVVANLFFEPSTRTRNSFEIA  66
            H LN+ +LT+ HI  ++  A+ F +   +KN  F+ L+G  VA+LFFEPSTRT+ +FE+A
Sbjct  17   HFLNIENLTKRHINDILNLADKFAS---DKNKKFKNLEGKTVASLFFEPSTRTKTTFELA  73

Query  67   AKRLGAMVLNPNLKISAISKGETLFDTIKTLEAMGVYFFIVRHSENETPEQIAKQLSSGV  126
            +KRL A  +N ++  S+  KGE++ D IKTLEAM    F+VRH+ + TP  IAK++   +
Sbjct  74   SKRLSADFINIDIANSSTLKGESIIDMIKTLEAMQCDMFVVRHANSGTPHFIAKEVDQKI  133

Query  127  -VINAGDGNHQHPSQALIDLMTIKQHKPHWNKLCVTIIGDIRHSRVANSLMDGLVTMGVP  185
             VINAGDG + HPSQA++D+ TIK++K  +N L V+I+GDI HSRVA SL+  L  + V
Sbjct  134  AVINAGDGTYAHPSQAMLDMYTIKKYKGGFNNLKVSIVGDILHSRVAKSLICSLKILCVD  193

Query  186  EIRLVGPSSLLPDKVGNDSIKKFTELKPSLLNSDVIVTLRLQKER-HDNSVDIDAFRGSF  244
            EI ++GP +L+PD      +  F +L+  + NSDVI+ LRLQKER HD  +  + +   +
Sbjct  194  EINIIGPENLMPDNKDVLGVNYFFDLEEGISNSDVIIMLRLQKERMHDALISTNDYYKKY  253

Query  245  RLTPEKLYSAKPDAIVMHPGPVNREVEINSDVADNQQSVILQQVRNGVAMRMAVLELFL  303
             LT +KL +AK D IVMHPGP+NR +EI+SDVAD   SVIL QV +G+++RMA++ L L
Sbjct  254  GLTSKKLLAAKKDVIVMHPGPINRGIEIDSDVADGSNSVILDQVSSGISIRMAIMSLLL  312
```

Figure 136: BLAST sixth match for BMW92_RS10855 sequence from organism *Gammaproteobacteria bacterium* with an e-value of 4e-92, 48% identity, 69% positives, 1% gaps, and an identity of aspartate cabamoyltransferase catalytic subunit (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

aspartate carbamoyltransferase catalytic subunit [Oceanococcus atlanticus]

Sequence ID: WP_083561731.1 **Length:** 319 **Number of Matches:** 1

See 2 more title(s) ▾ See all Identical Proteins(IPG)

Range 1: 13 to 312 GenPept Graphics ▼ Next Match ▲ Previo

Score	Expect	Method	Identities	Positives	Gaps
284 bits(726)	4e-91	Compositional matrix adjust.	148/300(49%)	198/300(66%)	3/300(1%)

```
Query  7    HLLNMRSLTRDHIEKLIQRANYFLTQGMEKNSVFETLKGHVVANLFFEPSTRTRNSFEIA  66
            HLL M+ L+ D+I  ++ RA  F++           LKG  + NLFFE STRTR++FE+A
Sbjct  13   HLLTMQGLSADNIHSILDRAESFVSSPGSGPRTSAELKGRTIVNLFFEASTRTRSAFELA  72

Query  67   AKRLGAMVLNPNLKISAISKGETLFDTIKTLEAMGVYFFIVRHSENETPEQIAKQLSSGV  126
             KRL A VLN ++  S+ SKGETL DT+KTLEAM V  FIVRH  +   + IA Q+  GV
Sbjct  73   GKRLSADVLNMDVATSSTSKGETLLDTLKTLEAMDVDMFIVRHHASGAAQFIANQVRPGV  132

Query  127  -VINAGDGNHQHPSQALIDLMTIKQHKPHWNKLCVTIIGDIRHSRVANSLMDGLVTMGVP  185
             V+NAGDG H HP+QAL+D+ TI++HKP +  L V I+GDI HSRVA S +  L  +GV
Sbjct  133  AVLNAGDGRHAHPTQALLDVFTIRRHKPDFASLSVAIVGDILHSRVARSEIRALRALGVR  192

Query  186  EIRLVGPSSLLPDKVGNDSIKKFTELKPSLLNSDVIVTLRLQKERHDNSV--DIDAFRGS  243
            ++R++GPS+LLP  +     +  T++   +  +DVI+ LRLQKER          D F
Sbjct  193  DLRVIGPSTLLPSGLAELGAQPETDMDRGIEGADVIIMLRLQKERMSGHFLPSADEFYQR  252

Query  244  FRLTPEKLYSAKPDAIVMHPGPVNREVEINSDVADNQQSVILQQVRNGVAMRMAVLELFL  303
            + LT  +L  A+ DA+VMHPGP+NR VEI S VAD  QSVILQQV +GVA+RMA++ + L
Sbjct  253  YGLTQSRLAGARADALVMHPGPINRGVEIESRVADGPQSVILQQVNHGVAVRMAIMSMIL  312
```

Figure 137: BLAST seventh match for BMW92_RS10855 sequence from organism *Oceanococcus atlanticus* with an e-value of 4e-91, 49% identity, 66% positives, 1% gaps, and an identity of aspartate cabamoyltransferase catalytic subunit (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

aspartate carbamoyltransferase catalytic subunit [Hahellaceae bacterium]

Sequence ID: MAM87066.1 Length: 323 Number of Matches: 1

Range 1: 11 to 310 GenPept Graphics ▼ Next Match ▲ Previous

Score	Expect	Method	Identities	Positives	Gaps
283 bits(725)	5e-91	Compositional matrix adjust.	147/300(49%)	203/300(67%)	3/300(1%)

```
Query  7    HLLNMRSLTRDHIEKLIQRANYFLTQGMEKNSVFETLKGHVVANLFFEPSTRTRNSFEIA  66
            HLL +  L+ + I +++  A  F+  G         L+G  VA LFFEPSTRTR +FE+A
Sbjct  11   HLLTLDGLSGELISQVLDTAESFIEVGSRSIKKVPLLRGKTVATLFFEPSTRTRTTFELA  70

Query  67   AKRLGAMVLNPNLKISAISKGETLFDTIKTLEAMGVYFFIVRHSENETPEQIAKQLSSGV  126
            AKRL A VLN N+  SA SKGE+L D ++ LEAM V  F+VRH+ +     IA++++  V
Sbjct  71   AKRLSADVLNININSSSATSKGESLSDMLRNLEAMAVDMFVVRHASSGAAHFIAREVTPEV  130

Query  127  -VINAGDGNHQHPSQALIDLMTIKQHKPHWNKLCVTIIGDIRHSRVANSLMDGLVTMGVP  185
             ++NAGDG H HP+QAL+D++TI+Q+KP +  L V IIGDI HSRVA S +  L  +GV
Sbjct  131  AIVNAGDGQHAHPTQALLDMLTIRQYKPDFPSLSVAIIGDILHSRVARSEIAALRALGVK  190

Query  186  EIRLVGPSSLLPDKVGNDSIKKFTELKPSLLNSDVIVTLRLQKERHDNSV--DIDAFRGS  243
            +IR+VGP +LLP  V +  +++    +   L  +DVI+TLRLQKER   S+      F
Sbjct  191  DIRVVGPDTLLPMAVESFGVRRCNRMSEGLDGADVIITLRLQKERMSGSLLPSEHEFYSL  250

Query  244  FRLTPEKLYSAKPDAIVMHPGPVNREVEINSDVADNQQSVILQQVRNGVAMRMAVLELFL  303
            + LT +KL +AK DAI+MHPGP+NR VEI S VAD+++SVIL+QV NG+A+RMAVL + +
Sbjct  251  YGLTTDKLATAKADAIIMHPGPINRGVEIESAVADSERSVILEQVTNGIAVRMAVLSMVM  310
```

Figure 138: BLAST eighth match for BMW92_RS10855 sequence from organism *Hahellaceae bacterium* with an e-value of 5e-91, 49% identity, 67% positives, 1% gaps, and an identity of aspartate cabamoyltransferase catalytic subunit (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

aspartate carbamoyltransferase catalytic subunit [Pseudolysobacter antarcticus]

Sequence ID: WP_129835034.1 **Length:** 315 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 14 to 313 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
283 bits(723)	8e-91	Compositional matrix adjust.	147/301(49%)	198/301(65%)	5/301(1%)

```
Query  7    HLLNMRSLTRDHIEKLIQRANYFLTQGMEKNSVFETLKGHVVANLFFEPSTRTRNSFEIA  66
            HLL +  L R  +E L+ RA               + L G  V NLFFEPSTRTR SF++A
Sbjct  14   HLLTLEGLPRATLEHLLDRAEQLRALSHNGTRRLDLLNGRTVINLFFEPSTRTRTSFDLA  73

Query  67   AKRLGAMVLNPNLKISAISKGETLFDTIKTLEAMGVYFFIVRHSENETPEQIAKQL-SSG  125
            AKRLGA V+N ++  S+  KGETL DT+ TLEAM    F+VRH E+ TPE IA+ L S+
Sbjct  74   AKRLGADVINFDIASSSTVKGETLLDTVHTLEAMHCDAFVVRHKESGTPEFIARHLRSNC  133

Query  126  VVINAGDGNHQHPSQALIDLMTIKQHKPHWNKLCVTIIGDIRHSRVANSLMDGLVTMGVP  185
             V+NAGDGN  HP+Q L+D +T+ +H+  +++LCV I GDIRHSRVA S +  L T+G+
Sbjct  134  AVLNAGDGNRAHPTQGLLDALTLLRHRADFSQLCVVICGDIRHSRVARSDVHALRTLGIG  193

Query  186  EIRLVGPSSLLPDKVGNDSIKKFTELKPSLLNSDVIVTLRLQKERHDNSV---DIDAFRG  242
            E+RL  P SLLPD          F++   +L  +D ++ LRLQKER + ++   + + FR
Sbjct  194  ELRLCAPESLLPDAAEMPGCLLFSDFDAALRGADAVIMLRLQKERMEGALIPSEAEYFR-  252

Query  243  SFRLTPEKLYSAKPDAIVMHPGPVNREVEINSDVADNQQSVILQQVRNGVAMRMAVLELF  302
             F L+ E+L  A PD +VMHPGP+NR+VEI +DVAD  +S+IL+QV NGV +RMAVL+
Sbjct  253  RFGLSSERLAQAAPDCLVMHPGPINRDVEIAADVADGPRSLILEQVGNGVFIRMAVLQEL  312

Query  303  L  303
            L
Sbjct  313  L  313
```

Figure 139: BLAST ninth match for BMW92_RS10855 sequence from organism *Pseudolysobacter antarcticus* with an e-value of 8e-91, 49% identity, 65% positives, 1% gaps, and an identity of aspartate cabamoyltransferase catalytic subunit (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

aspartate carbamoyltransferase catalytic subunit [Pseudomonas sabulinigri]

Sequence ID: WP_092287351.1 **Length:** 331 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 18 to 317 GenPept Graphics ▼ Next Match ▲ Previou

Score	Expect	Method	Identities	Positives	Gaps
283 bits(724)	1e-90	Compositional matrix adjust.	144/300(48%)	202/300(67%)	3/300(1%)

```
Query  7    HLLNMRSLTRDHIEKLIQRANYFLTQGMEKNSVFETLKGHVVANLFFEPSTRTRNSFEIA  66
            H L +  L+R+ + +L+  A+ FL  G          L+G  V N+FFE STRTR +FE+A
Sbjct  18   HFLTIEGLSRELLTELLDTADSFLEVGERAVKKVPLLRGKTVCNVFFENSTRTRTTFELA  77

Query  67   AKRLGAMVLNPNLKISAISKGETLFDTIKTLEAMGVYFFIVRHSENETPEQIAKQLSSGV  126
            AKRL A VLN N+  S+ SKGETL+DT++ LEAM    F+VRH ++     IA+++   V
Sbjct  78   AKRLSADVLNLNISTSSTSKGETLYDTLQNLEAMAADMFVVRHGDSGAAHFIAERVCPDV  137

Query  127  -VINAGDGNHQHPSQALIDLMTIKQHKPHWNKLCVTIIGDIRHSRVANSLMDGLVTMGVP  185
             VINAGDGNH HP+QA++D++TI++H+  + KL V I+GDI HSRVA S M  L T+G P
Sbjct  138  AVINAGDGNHAHPTQAMLDMLTIRRHRGDFEKLSVAIVGDILHSRVARSNMQALKTLGCP  197

Query  186  EIRLVGPSSLLPDKVGNDSIKKFTELKPSLLNSDVIVTLRLQKERHDNSV--DIDAFRGS  243
            +IR++ P +LLP+ +    ++ F +L+  L + DV++ LRLQKER  + +      F
Sbjct  198  DIRVIAPRTLLPEGIDQYGVRVFNDLRVGLRDVDVVIMLRLQKERMQSGLLPSEGEFYKL  257

Query  244  FRLTPEKLYSAKPDAIVMHPGPVNREVEINSDVADNQQSVILQQVRNGVAMRMAVLELFL  303
            + LT E L  AKPDA+VMHPGP+NR VEI S+VAD  QSVIL+QV  G+A+RMAV+ + +
Sbjct  258  YGLTRETLALAKPDALVMHPGPINRGVEIESEVADGPQSVILKQVTYGIAVRMAVMSMAM  317
```

Figure 140: BLAST tenth match for BMW92_RS10855 sequence from organism *Pseudomonas sabulinigri* with an e-value of 1e-90, 48% identity, 67% positives, 1% gaps, and an identity of aspartate cabamoyltransferase catalytic subunit (BLAST, <https:/blast.ncbi.nlm.nih.gov/Blast.cgi>).

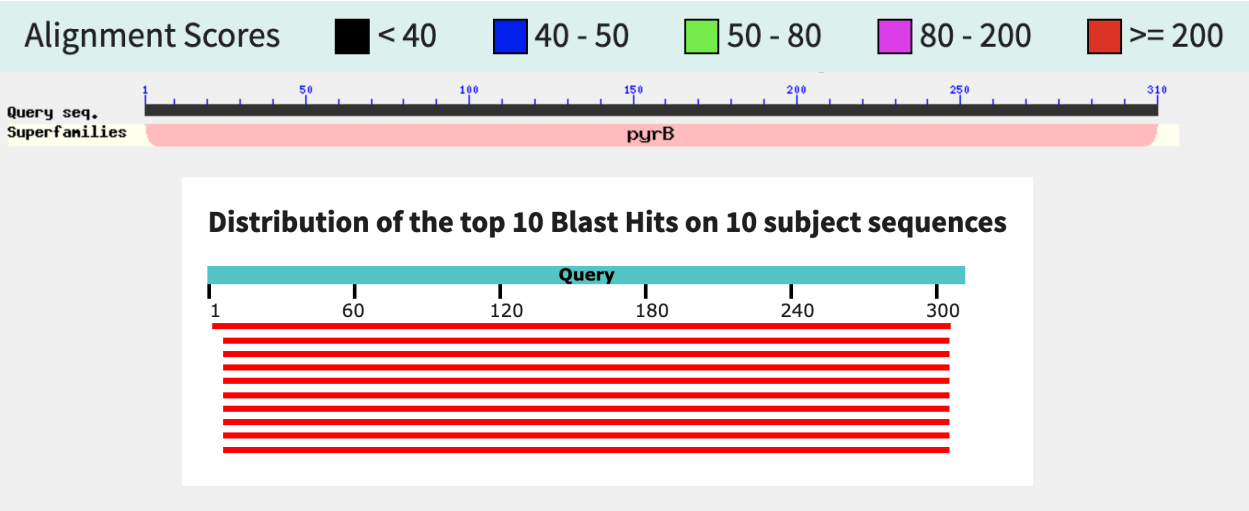
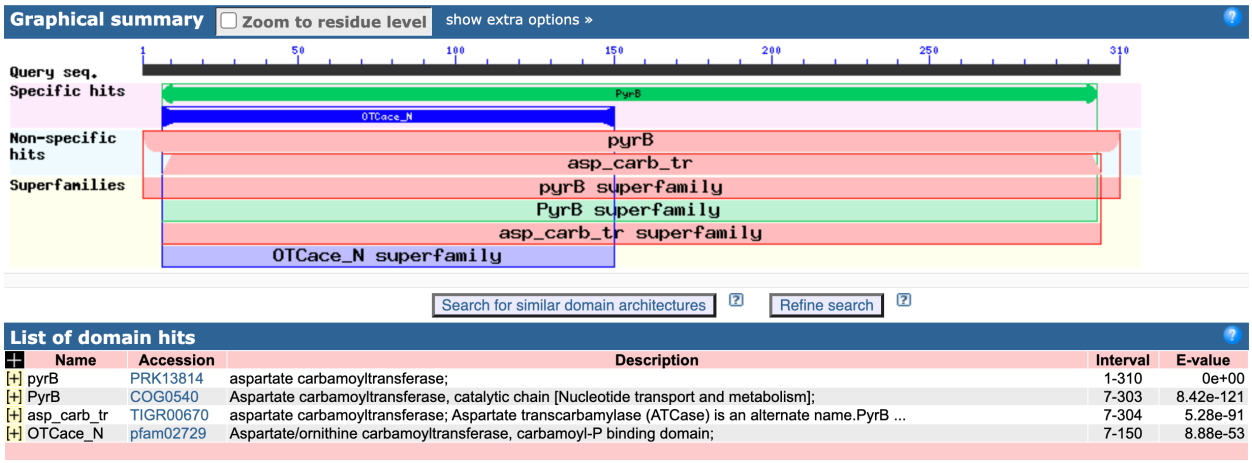


Figure 141: BLAST graphic summary of the top 10 organism sequences similarities selected aligned with *Coxiella burnetii* query sequence of gene BMW92_RS10855. Each of the alignment sequences selected are order from highest sequence similarity (top) to lowest sequence similarity (bottom). All organism sequences aligned with the query sequence have an alignment score of greater than 200 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

CDD



Name	Accession	Description	Interval	E-value
[+] pyrB	PRK13814	aspartate carbamoyltransferase;	1-310	0e+00
[+] PyrB	COG0540	Aspartate carbamoyltransferase, catalytic chain [Nucleotide transport and metabolism];	7-303	8.42e-121
[+] asp_carb_tr	TIGR00670	aspartate carbamoyltransferase; Aspartate transcarbamylase (ATCase) is an alternate name.PyrB ...	7-304	5.28e-91
[+] OTCace_N	pfam02729	Aspartate/ornithine carbamoyltransferase, carbamoyl-P binding domain;	7-150	8.88e-53

Figure 142: Conserved Domain Database output results for gene BMW92_RS10855. The top domain hit match was pyrB: aspartate carbamoyltransferase which aligned with the query sequence from amino acid residues 1-310 and had statistically significant e-value of 0e+00. The second domain hit match was PyrB: aspartate carbamoyltransferase which aligned with the query sequence from amino acid residues 7-303 and had a statistically significant e-value of 8e-121. The third domain hit match was asp_carb_tr: aspartate carbamoyltransferase which aligned with the query sequence from amino acid residues 7-304 and had a statistically significant e-value of 5.28e-91. The fourth domain hit match was OTCace_N: aspartate carbamoyltransferase which aligned with the query sequence from amino acid residues 7-150 and had a statistically significant e-value of 8.88e-53 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

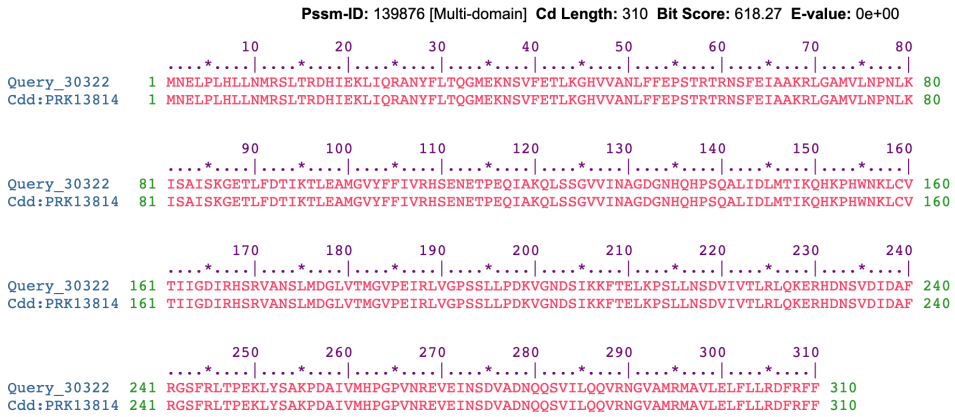


Figure 143: The first domain hit sequence pyrB: aspartate carbamoyltransferase aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 1-310 and had a statistically significant e-value of 0e+00 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

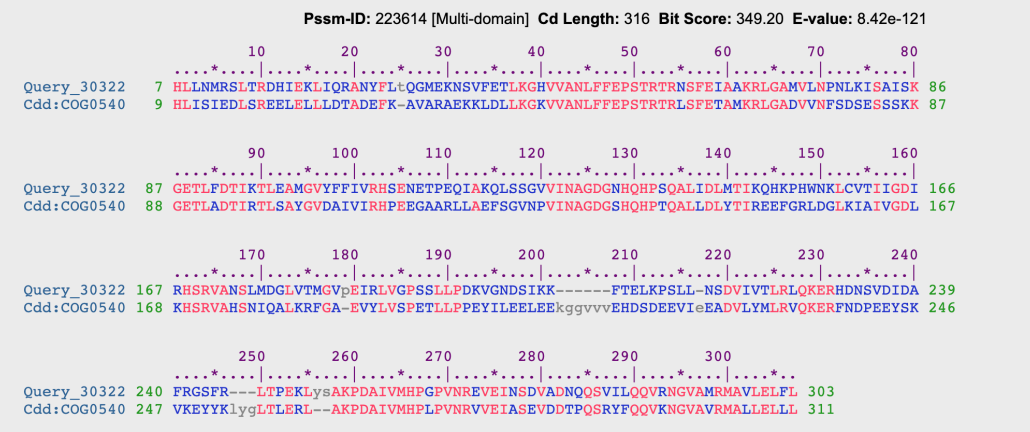


Figure 144: The second domain hit sequence PyrB: aspartate carbamoyltransferase aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 7-303 and had a statistically significant e-value of 8e-121 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

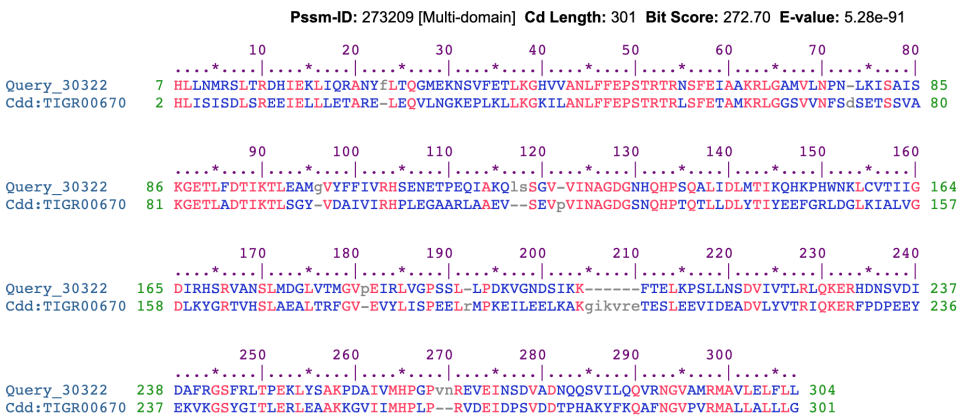


Figure 145: The third domain hit sequence asp_carb_tr: aspartate carbamoyltransferase (ATCase) aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 7-304 and had a statistically significant e-value of 5.28e-91 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

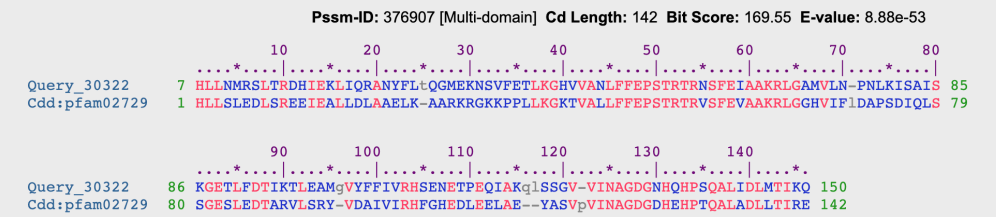


Figure 146: The fourth domain hit sequence OTCace_N: Aspartate/ornithine carbamoyltransferase aligned with the query sequence. The amino acid residues had an aligned interval from amino acids 7-150 and had a statistically significant e-value of 8.88e-53 (BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

MUSCLE

```
C.burnetii          ----------MNELPL-------HLLNMRSLTRDHIEKLIQRANYFLTQGMEKNSVFETL
C.mudrowiae         --------MVKKDFPY-------HLLGMQSLTRNEIDLILKRANDFLR-NIKENRVFDTL
P.antarcticus       ----------MPQPQLDEHGRLRHLLTLEGLPRATLEHLLDRAEQLRALSHNGTRRLDLL
G.bacterium         -------MVIDNNIQFDKNNKLKHFLNIENLTKRHINDILNLADKFAS---DKNKKFKNL
O.atlanticus        -----------MNPQIDEHGQFRHLLTMQGLSADNIHSILDRAESFVSSPGSGPRTSAEL
L.arctica           MKSKLVPGPLPSSIQLSEDGQLKHFLSIEGLPREMLVEILDRAEQFVTLPNKAQKKFPLL
T.bacterium         --------MGSQNIQLDQTGKLKHFLSIDGLSKDLVTTILDHAETFTSVGERSSKKVPIL
H.bacterium         -------------MQLNEAGELKHLLTLDGLSGELISQVLDTAESFIEVGSRSIKKVPLL
P.sabulinigri       ------MSQTPHHLQLNHQGQLRHFLTIEGLSRELLTELLDTADSFLEVGERAVKKVPLL
A.bacterium         --MNQGIKAISPGLQLTSGGQLKHFLTIDGLGRDLLTEILDTADSFIEVGERRIKKVPLL
H.marina            --MNQGIKAISPGLQLTSGGQLKHFLTIDGLGRDLLTEILDTADSFIEVGERRIKKVPLL
                                        *:* : .*    :  ::. *: :            *

C.burnetii          KGHVVANLFFEPSTRTRNSFEIAAKRLGAMVLNPNLKISAISKGETLFDTIKTLEAMGVY
C.mudrowiae         KGEVVANLFFETSTRTRNSFEIAAKRLEAIVLSPDLKVSALNKGESLLDMARNLQAMGTR
P.antarcticus       NGRTVINLFFEPSTRTRTSFDLAAKRLGADVINFDIASSSTVKGETLLDTVHTLEAMHCD
G.bacterium         EGKTVASLFFEPSTRTKTTFELASKRLSADFINIDIANSSTLKGESIIDMIKTLEAMQCD
O.atlanticus        KGRTIVNLFFEASTRTRSAFELAGKRLSADVLNMDVATSSTSKGETLLDTLKTLEAMDVD
L.arctica           RGKTVMNLFFENSTRTRMTFELAAQRLSADVVNLDIRNSSASKGESLLDTIRNLEAMNCD
T.bacterium         RGKTVVNLFFESSTRTRTTFELAAKRLSADVMNINLESSATKKGESLSDTLKTLEAMQAD
H.bacterium         RGKTVATLFFEPSTRTRTTFELAAKRLSADVLNINISSSATSKGESLSDMLRNLEAMAVD
P.sabulinigri       RGKTVCNVFFENSTRTRTTFELAAKRLSADVLNLNISTSSTSKGETLYDTLQNLEAMAAD
A.bacterium         RGRTVVNLFFEPSTRTRSTFELAAKRLSADVLNLDISKSATSKGESLSDTLLNLEAMASD
H.marina            RGRTVVNLFFEPSTRTRSTFELAAKRLSADVLNLDISKSATSKGESLSDTLLNLEAMASD
                     * .: .:*** ****. :*::*.:** * .:. ::  *:  ***:: *   .*:**

C.burnetii          FFIVRHSENETPEQIAKQLSSGV-VINAGDGNHQHPSQALIDLMTIKQHKPHWNKLCVTI
C.mudrowiae         FFVIRHTENNRPRMLAEHLEQGI-VINAGDGNHQHPTQGLIDLMTIQQHKPDWTKLCVTI
P.antarcticus       AFVVRHKESGTPEFIARHLRSNCAVLNAGDGNRAHPTQGLLDALTLLRHRADFSQLCVVI
G.bacterium         MFVVRHANSGTPHFIAKEVDQKIAVINAGDGTYAHPSQAMLDMYTIKKYKGGFNNLKVSI
O.atlanticus        MFIVRHHASGAAQFIANQVRPGVAVLNAGDGRHAHPTQALLDVFTIRRHKPDFASLSVAI
L.arctica           VFVVRHHHSSAPHFIAKYCAPHISVLNAGDGYHEHPSQAMLDMLTIRQHKPDFSKLTVAI
T.bacterium         MFVVRHQDSGAAEFIARQVAQNISVINAGDGSHSHPTQAMLDMFTIRKHKKTFDQLRVAI
H.bacterium         MFVVRHASSGAAHFIAREVTPEVAIVNAGDGQHAHPTQALLDMLTIRQYKPDFPSLSVAI
P.sabulinigri       MFVVRHGDSGAAHFIAERVCPDVAVINAGDGNHAHPTQAMLDMLTIRRHRGDFEKLSVAI
A.bacterium         MFVVRHAQSGAAHFIARSVTPGVAIINAGDGRHAHPTQAMLDMLTIRQHKERFEGLRVAI
H.marina            MFVVRHAQSGAAHFIARSVTPGVAIINAGDGRHAHPTQAMLDMLTIRQHKERFEGLRVAI
                     *::**  .  .  :*        ::*****   **:*.::*  *: .:.  :  * * *

C.burnetii          IGDIRHSRVANSLMDGLVTMGVPEIRLVGPSSLLPDKVGNDSIKKFTELKPSLLNSDVIV
C.mudrowiae         IGDIYHSRVANSLVDGLLIMGVPEIRITGPSQLLPETVKNPRIKKIPELEASLINSDVVV
P.antarcticus       CGDIRHSRVARSDVHALRTLGIGELRLCAPESLLPDAAEMPGCLLFSDFDAALRGADAVI
G.bacterium         VGDILHSRVAKSLICSLKILCVDEINIIGPENLMPDNKDVLGVNYFFDLEEGISNSDVII
O.atlanticus        VGDILHSRVARSEIRALRALGVRDLRVIGPSTLLPSGLAELGAQPETDMDRGIEGADVII
L.arctica           TGDIRHSRVARSEIQALKTLGAKEIRVIAPGTLMPVGIEELGVKVFTSMEEGLVDADVVV
T.bacterium         VGDIAHSRVARSEIHALQILGVPEIRLVGPKTLIPTAVEKLGVRTFNSLESGIDKADVII
H.bacterium         IGDILHSRVARSEIAALRALGVKDIRVVGPDTLLPMAVESFGVRRCNRMSEGLDGADVII
P.sabulinigri       VGDILHSRVARSNMQALKTLGCPDIRVIAPRTLLPEGIDQYGVRVFNDLRVGLRDVDVVI
A.bacterium         VGDILHSRVARSQVNALLTLGAEEVRLVGPATLMPAAANQLGVKLCTTMEEGLADTDVII
H.marina            VGDILHSRVARSQVNALLTLGAEEVRLVGPATLMPAAANQLGVKLCTTMEEGLAETDVII
                     *** *****.* : .*  :   ::.: .*  *:*             :  .:   *.::
```

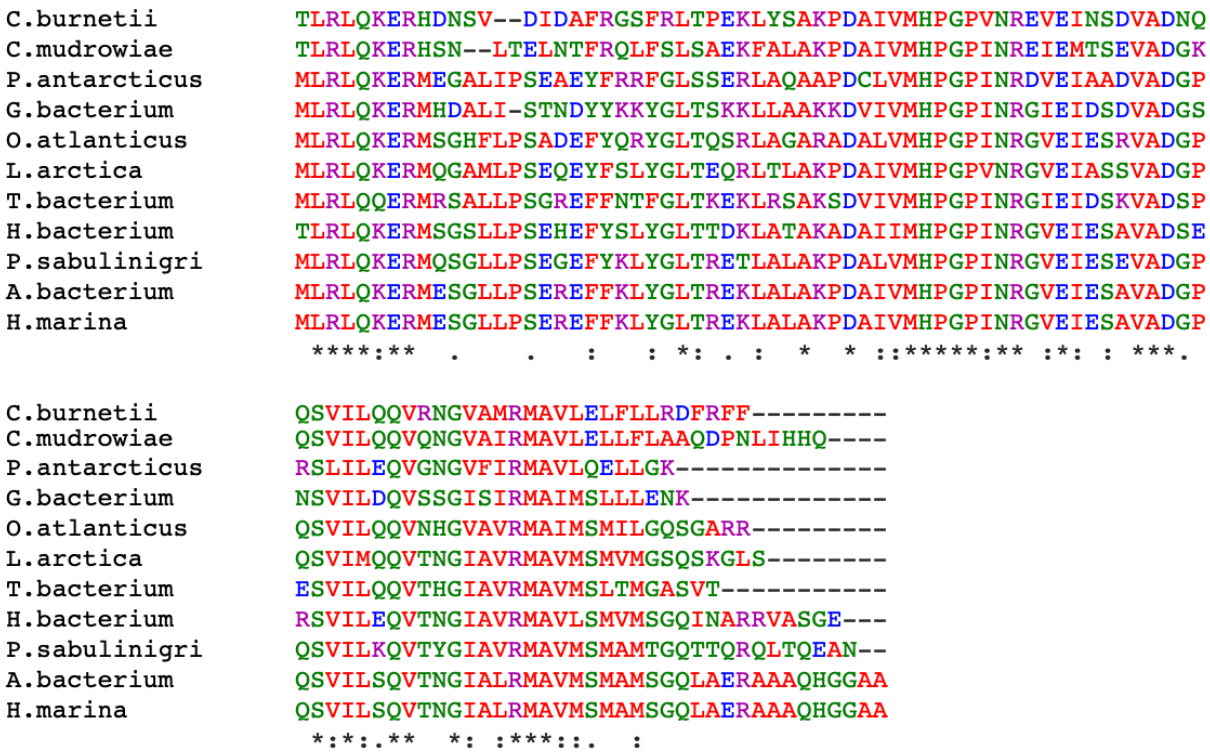


Figure 147: MUSCLE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10855 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered randomly and are listed from top to bottom as followed: *C. burnetii*, *C. mudrowiae*, *P. antarcticus*, *G. bacterium*, *O. atlanticus*, *L. arctica*, *T. bacterium*, *H. bacterium*, *P. sabulinigri*, *A. bacterium*, *H. marina.* Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

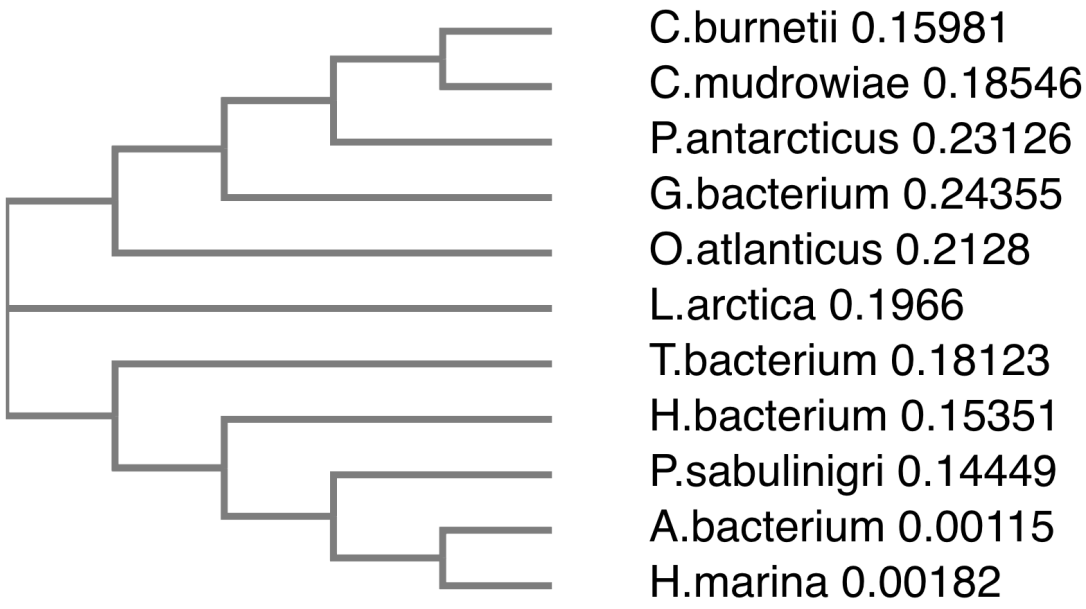


Figure 148: MUSCLE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10855 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *C. burnetii*, *C. mudrowiae*, *P. antarcticus*, *G. bacterium*, *O. atlanticus*, *L. arctica*, *T. bacterium*, *H. bacterium*, *P. sabulinigri*, *A. bacterium*, *H. marina.* The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

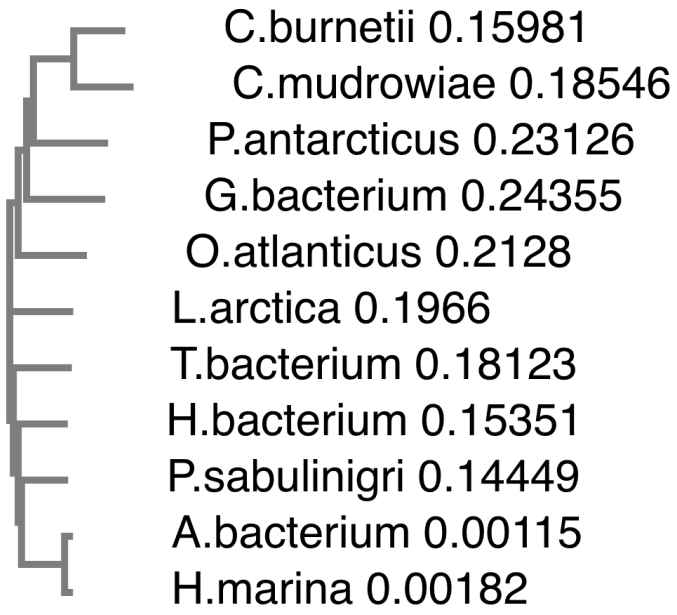
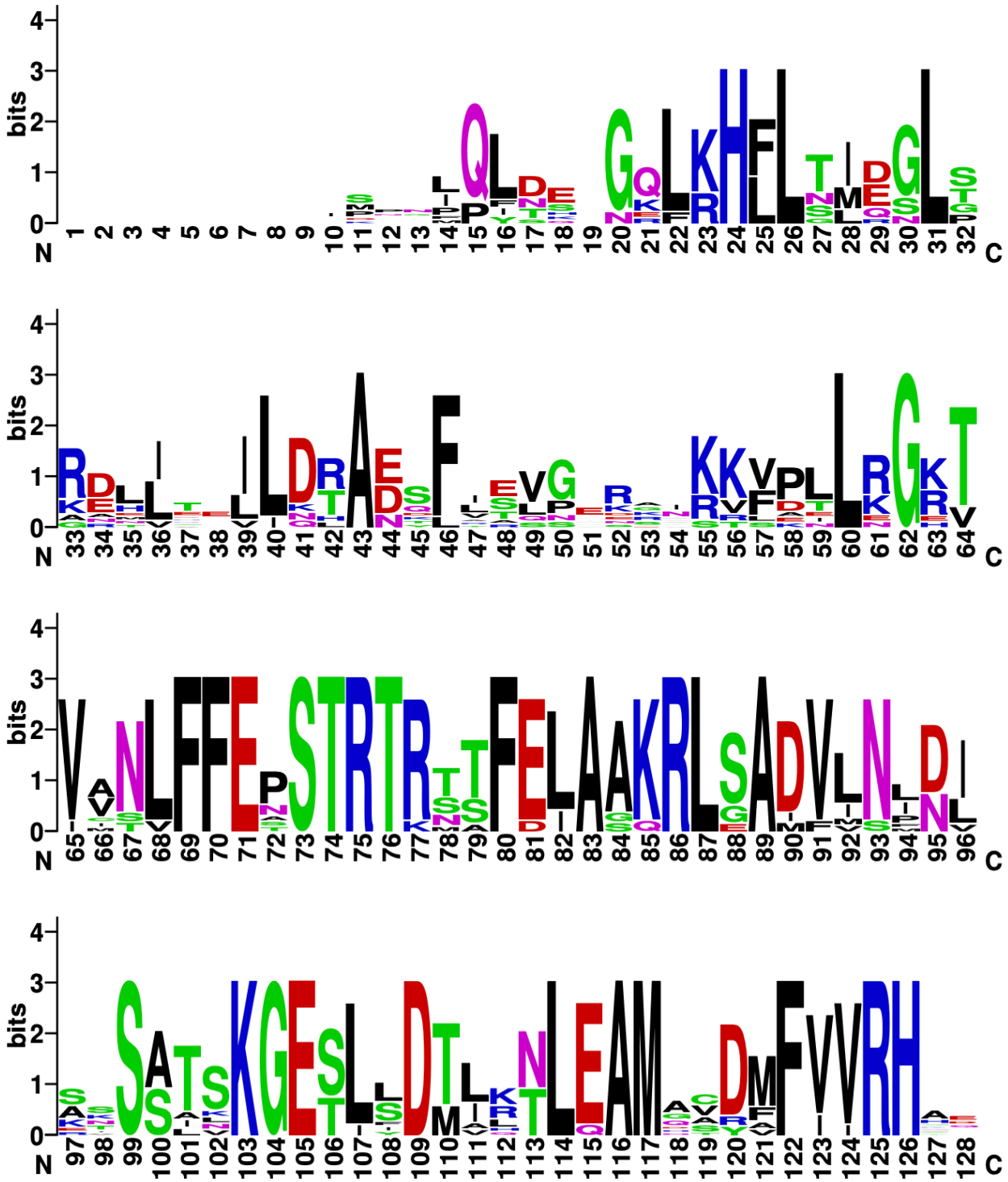
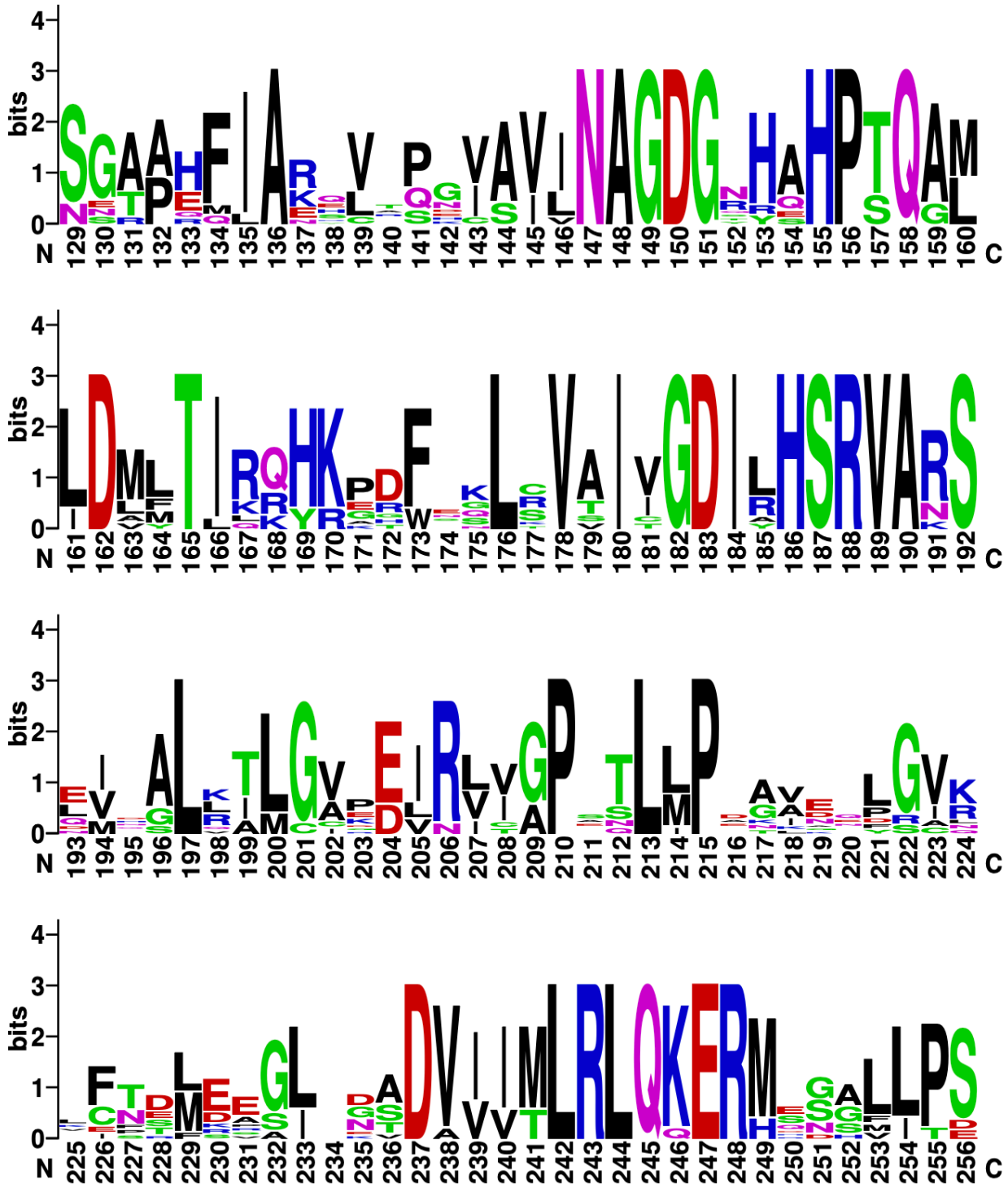


Figure 149: MUSCLE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10855 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *C. burnetii*, *C. mudrowiae*, *P. antarcticus*, *G. bacterium*, *O. atlanticus*, *L. arctica*, *T. bacterium*, *H. bacterium*, *P. sabulinigri*, *A. bacterium*, *H. marina.* The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (MUSCLE, <https://www.ebi.ac.uk/Tools/msa/muscle/>).

MUSCLE Sequence Logo




bits
4
3
2
1
0
N
C
129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192
193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224
225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256

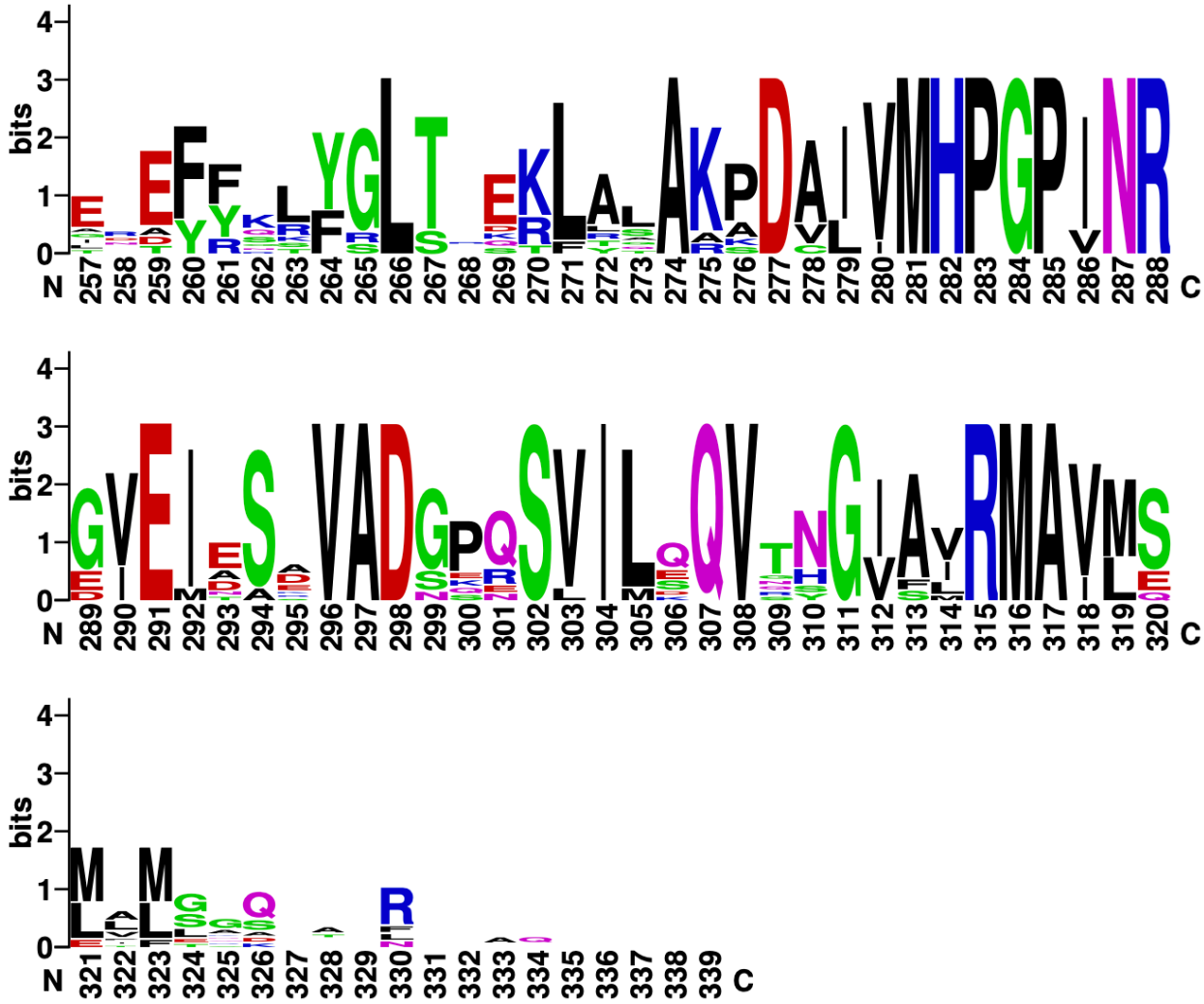


Figure 150: Sequence logo generated from the MUSCLE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10855 and the top 10 organism sequences similarities selected. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. The conservation of each amino acid among residue across similar sequences is represented by the height of each single letter abbreviation of the amino acid. (WebLogo, <https://weblogo.berkeley.edu/logo.cgi>).

T-COFFEE

```
A.ehrlichii       MSNNTLCFIGGGNMARSLIGGLLADGFDPQAVRVADPDAGKRDDLANRFG
A.mobile          MTMKTLCFIGGGNMARSLIGGLLTDGYDPQAIRVAEPDAGKREDLANRFG
C.bacterium       MKDVNIAFIGGGNMATSLIGGLLADHVSPARLCVADRDPAQREHLAAQFG
C.bacterium_1     MDTFTITFIGGGNMARSLIGGLIADGTPVDRIRVSDPSAEQRSQLQGLFG
C.burnetii        MNTSNITFIGGGNMARNIVVGLIANGYDPNRICVTNRSLDKLDFFKEKCG
C.mudrowiae       MRIANITFIGGGNMACNIVVGLLANGYDSNRICVTNPTSDKLTFFREKCK
N.halophilus      MNEKTLAFIGGGNMATSLIGGLIADGRNAQTIWVADPDRSKLDALHHRFS
N.mobilis         MAEESITFIGGGNMAYSLVGGLIADGYRAERVHVADPDPAKRMDLANRFR
O.beijerinckii    MQNATMAFLGAGNMCGSIIGGLIAEGYSPEQITATRRSEERLQAIKEEFG
T.denitrificans   MEQGIISFIGGGNMCSSLVGGLIADGYAPERIRVSDPGEETLASLRARFG
T.endolucinida    MSNNNITFIGGGNMATSLINGLIADGYEKQRITVSDPDAEKLAQLAARCG
                  *    : *:*.***. .:: **:::      : .:          :

A.ehrlichii       VRVYADNLEAAADADTVILAVKPQVVRTACEQLVAGSGDAGRLFISIAAG
A.mobile          VRVHEDNLEAAANAQAVILAVKPQVIRPVCEQLAGAEAGKGRVYISIAAG
C.bacterium       VRTSEDNAACAEDADVIVLAVKPQVLHEVCEALTDSVQRKQPLVVSVAAG
C.bacterium_1     IATFADNHDAIAGADVIVLAVKPQIMQAVATGLAPALSGVKPLLLSIAAG
C.burnetii        VHTTQDNRQGALNADVVVLAVKPHQIKMVCEELKDILSETKILVISLAVG
C.mudrowiae       VRTTQNNREGATNADAIILAVKPNQVKGVCEELKDIVNTLHPLIISVAVG
N.halophilus      VNTTPDNLQAAQEAEVVVLAVKPQQLRTVATGLKSVVTSSQPLWLTIAAG
N.mobilis         IHVHEDNRKAVLRAAAVVLAVKPQIIKSVLEPLGPILREQKSLVISIAAG
O.beijerinckii    VQTSTDNIAAVASHDVIILGVKPQMMKELCDQIKDQVQQSKPLVISVAAG
T.denitrificans   VHTTHDNREAAAGAGVVVLAVKPQVLPKVAAELAPVVQEHGTLVVSIAAG
T.endolucinida    VHTQSDNNSAISNAEVVVLAVKPQVLKSVAQDLAAAIQQVKPLVISIAAG
                  : .  :*         .::*.***: :      :          : :::*.*

A.ehrlichii       VREPDLTRWLGG-QAAVVRTMPNTPSLVGTGATALYANDRVKERQRELAE
A.mobile          VREPDLTRWLGG-SAAVVRTMPNTPSLVGTGATALYANPQVSEPQRELAE
C.bacterium       VRTDSLRRWLGGGDVAIVRAMPNTPALLQSGATGLYACTGVSEEQRDLAE
C.bacterium_1     IRSTDLHRWLGG-HVALVRTMPNTPALVRSGATGLFARKDVSREQRDLAE
C.burnetii        VTTPLIEKWLGK-ASRIVRAMPNTPSSVRAGATGLFANETVDKDQKNLAE
C.mudrowiae       VRVKLLQKWLQS-EPAIVRAMPNTPASVGAGATALFANEKATKEQRNLAE
N.halophilus      IRIPDLERWLGG-PAPIVRAMPNTPALVQAGATALFANAQTNPQQRQMAE
N.mobilis         VREPDISRWLGG-QIAVVRTMPNTPALVRAGATALYANEYVSQNQRDLAE
O.beijerinckii    LTTETLERWLGG-NVAVVRTMPNTPSLLRCGAAGLYANASVSDEQKQVAE
T.denitrificans   IRTTDLQRWLGA-GVALVRTMPNTPALVKSGATALFATAAVTAAQRDQAE
T.endolucinida    VKESSLRNWLGG-EVALVRSMPNTPAMIQSGATGLHAGPGVSEAQRNQAE
                  :    : .**       :**:*****: :  **:.*.*   .   *:: **

A.ehrlichii       SLMRAVGLVVWLDDEAQMDTVTAVSGSGPAYFFLLMEAIEDAARDLGLPG
A.mobile          SLMRAVGLVVWLDDETQMDTVTAVSGSGPAYFFLLMEAIEEAAREQGLPA
C.bacterium       AILRATGLTLWVDDEAQMDIVTALSGSGPAYFFRVMEGLEKAATELGLPA
C.bacterium_1     TILRAVGLTLWVDNEDLIDSVTAVSGSGPAYFFLIMEAMEEAAIQLGLDE
C.burnetii        SIMRAVGLVIWVSSEDQIEKIAALSGSGPAYIFLIMEALQEAAEQLGLTK
C.mudrowiae       SILRAVGLVVWLSLEDQIDEVAALSGSGPAYIFFVMEALQEAGEGLGLPK
N.halophilus      SVLRAVGLTLWLKDENLMEVVTALSGSGPAYFFLVMEAMEKAAIDLGLDD
N.mobilis         SLLRAVGIIQWLDDETLLDIVTALSGSGPAYFFLLMETLEAAAIELGLPE
O.beijerinckii    KMMQAVGIALWVDKEELMEAVTGVSGSGPAYFFLMMEAMQQAGVDVGLTP
T.denitrificans   SVLRAVGLTLWLENEEQMDAVTALSGSGPAYFFLVMEAMQGAAQAIGLPE
T.endolucinida    SILRAVGLTRWVEEESMMDAVTAVSGSGPAYFFLIMEAIESSARQMGLDE
                   :::*.*:  *:. *  :: ::.:********:* :** :: :.   **
```

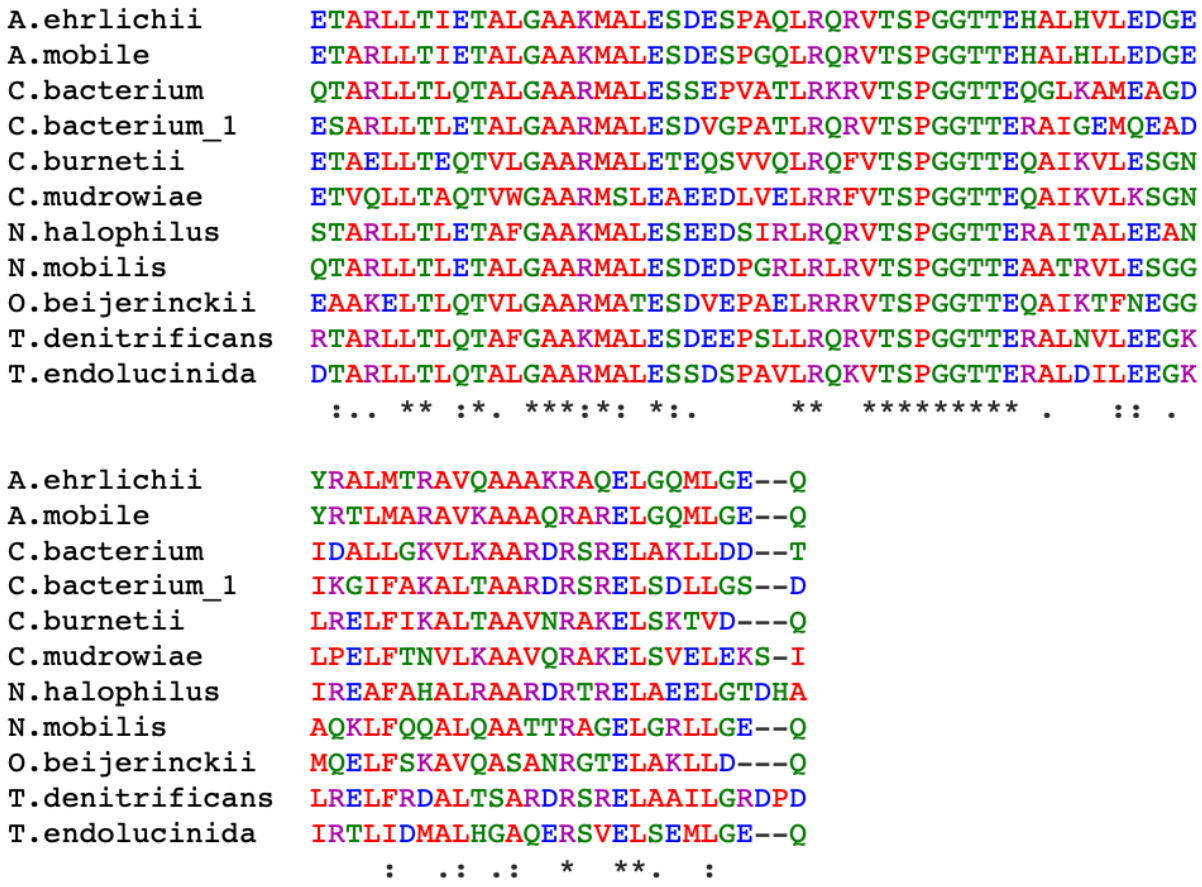


Figure 151: T-COFFEE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10855 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered randomly and are listed from top to bottom as followed: *A. ehrlichii*, *A. mobile*, *C. bacterium*, *C. bacterium_1*, *C. burnetii*, *C. mudrowiae*, *N. halophilus*, *N. mobilis*, *O. beijerinckii*, *T. denitrificans*, *T. endolucinida*. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

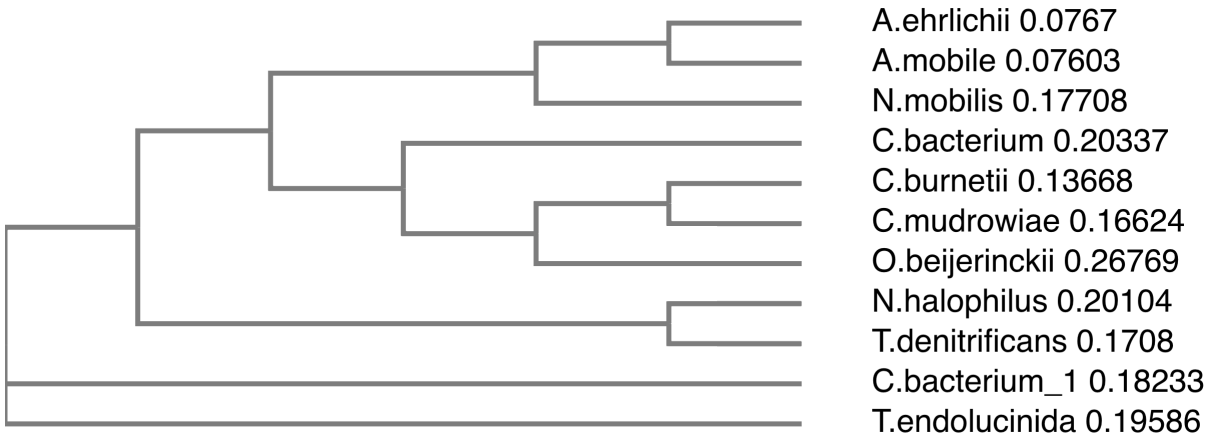


Figure 152: T-COFFEE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10855 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *A. ehrlichii*, *A. mobile*, *N. mobilis*, *C. bacterium*, *C. burnetii*, *C. mudrowiae*, *O. beijerinckii*, *N. halophilus*, *T. denitrificans*, *C. bacterium_1*, T. endolucinida. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

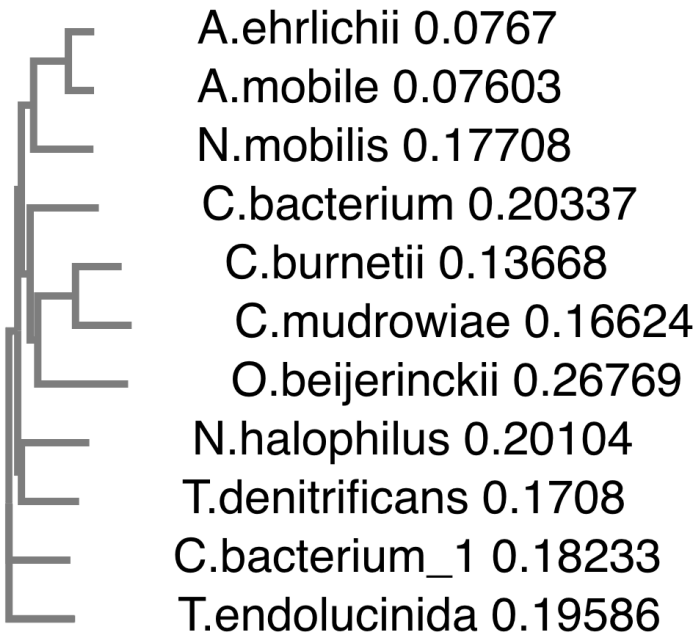
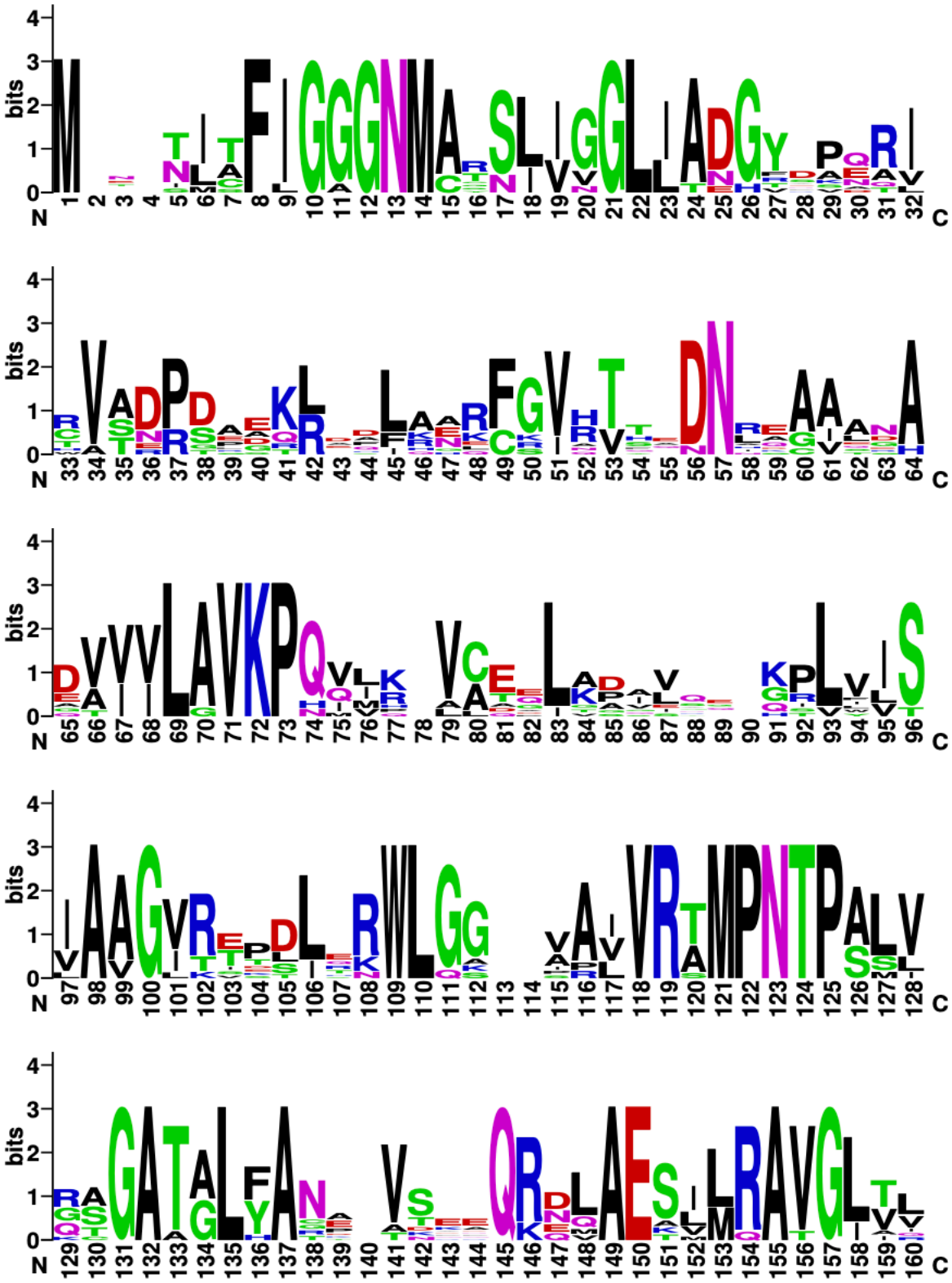


Figure 153: T-COFFEE multiple sequence alignment phylogenetic cladogram for *Coxiella burnetii* gene BMW92_RS10855 and the top 10 organism sequences similarities selected. Organism sequences were abbreviated by the genus and species in which the sequence similarity originated. Organism sequences were ordered from top to bottom as followed: *A. ehrlichii*, *A. mobile*, *N. mobilis*, *C. bacterium*, *C. burnetii*, *C. mudrowiae*, *O. beijerinckii*, *N. halophilus*, *T. denitrificans*, *C. bacterium_1*, T. endolucinida. The numbers followed behind each organism displays the correlation and closeness of each respective organism to a common ancestor shared between the organism and *Coxiella burnetii* (T-COFFEE, <https://www.ebi.ac.uk/Tools/msa/tcoffee/>).

T-COFFEE Sequence Logo



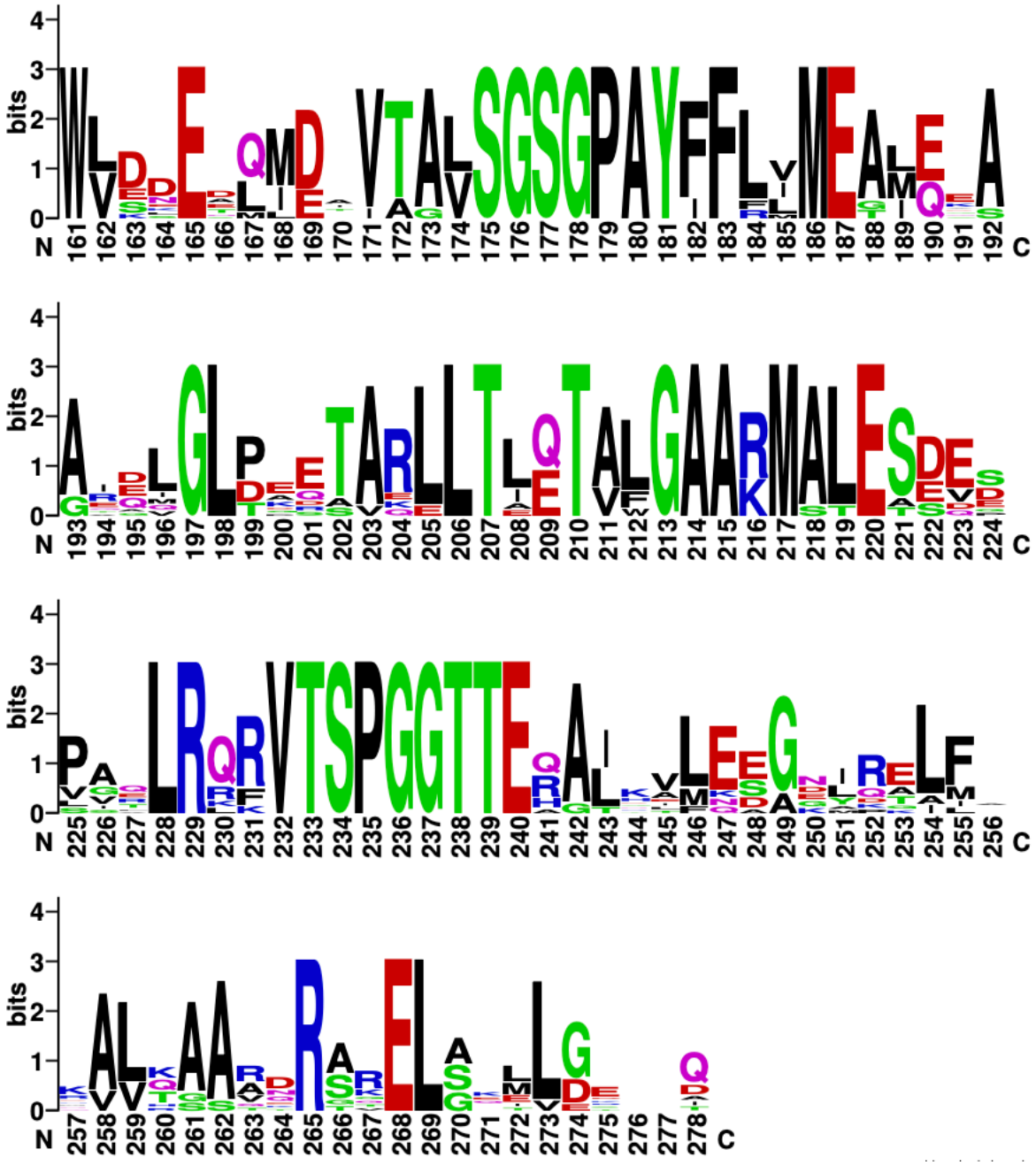
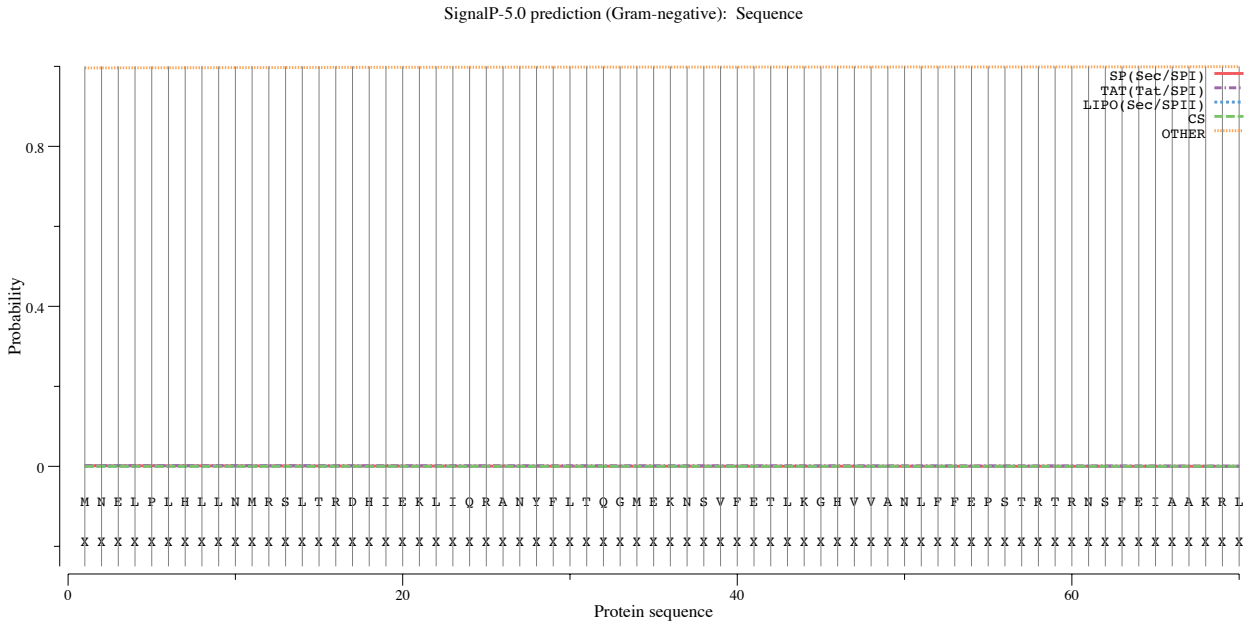


Figure 154: Sequence logo generated from the T-COFFEE multiple sequence alignments of *Coxiella burnetii* gene BMW92_RS10855 and the top 10 organism sequences similarities selected. Amino acids are represented by single letter abbreviations and distinct colors for each respective amino acid. The conservation of each amino acid among residue across similar

sequences is represented by the height of each single letter abbreviation of the amino acid. (WebLogo, <https://weblogo.berkeley.edu/logo.cgi>).

Protein Localization

SignalP



Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.0015	0.0016	0.0003	0.9966

Figure 155: SignalP 5.0 prediction (Gram-negative) for gene BMW92_RS10855 of *Coxiella burnetii*. The SP (Sec/SPI), TAT (Tat/SPI), LIPO (Sec/SPII), and CS probability scores combined were all less than a total 0.0034 (0.34%) which results in the likelihood of the protein being a signal peptide as highly unlikely and can confirm there is no signal peptide of these protein types. The program calculated the probability scores for OTHER as 0.9966 (99.66%). This probability score indicates the protein from the gene BMW92_RS10855 has another protein classification that is not related to similar function or type as a signal peptide (SignalP, <http://www.cbs.dtu.dk/services/SignalP/ >).

LipoP

```
# Sequence CYT score=-0.200913
# Cut-off=-3
Sequence        LipoP1.0:Best   CYT     1       1       -0.200913



# NO PLOT made - less than 4 putative cleavage sites predicted
```

Figure 156: LipoP 1.0 was unable to generate a plot graph due to there being less than four predicted putative cleavage sites. The best localization prediction resulted in the highest scoring class being the cytoplasmic protein class (LipoP, <http://www.cbs.dtu.dk/services/LipoP/>).

TMHMM

```
# WEBSEQUENCE Length: 310
# WEBSEQUENCE Number of predicted TMHs:  0
# WEBSEQUENCE Exp number of AAs in TMHs: 0.00546
# WEBSEQUENCE Exp number, first 60 AAs:  0
# WEBSEQUENCE Total prob of N-in:        0.00311
WEBSEQUENCE     TMHMM2.0        outside      1   310
```

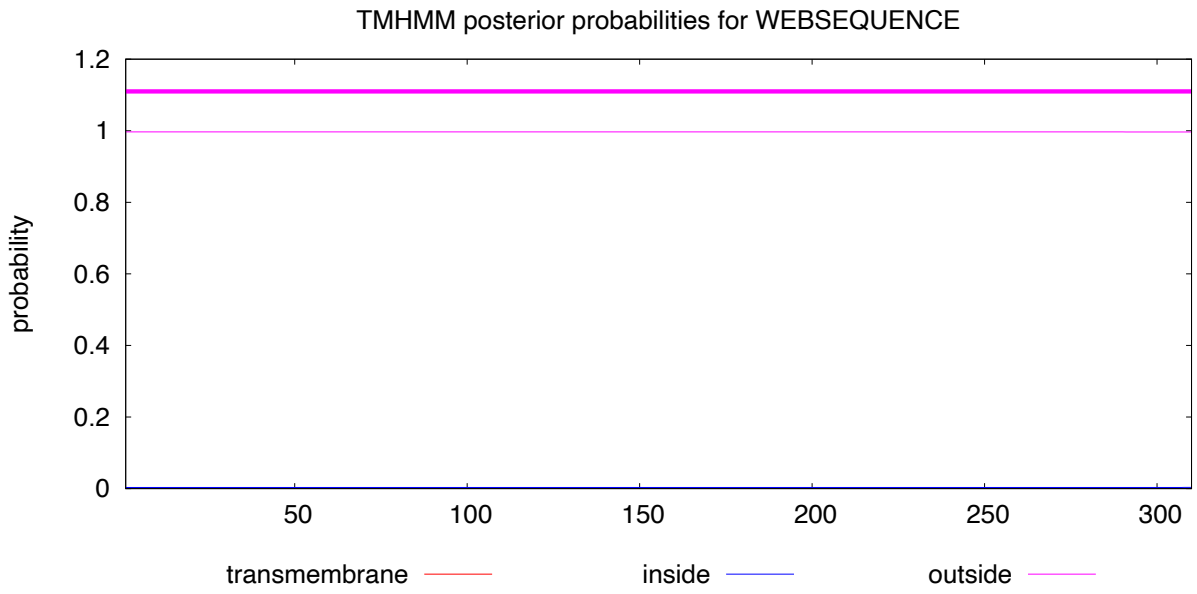


Figure 157: TMHMM posterior probability displayed a line graph that predicts the localization of the protein coded from BMW92_RS10855 as entirely outside the membrane. The red line, representative of the protein being located in the transmembrane, was 0.0 (0.0% probability) across the entirety of the line graph. This is indicative of the protein being located within the transmembrane as highly unlikely. The blue line, representative of the protein being located inside the membrane, was at 0.00 (0.0% probability). This is indicative of the protein being located inside of the membrane as highly unlikely. The magenta line, representative of the

protein being located outside the membrane, was at 1.0 (100% probability). This is indicative of the protein being located outside of the membrane as highly likely (TMHMM, <http://www.cbs.dtu.dk/services/TMHMM/ >).

BOMP

The total number of valid proteins submitted is: 1
The total number of integral ß-barrel outer membrane proteins predicted is: 0

Sequence name	Category	Best BLAST hit

Figure 158: The BOMP test result identified there are no integral beta-barrel outer membrane proteins for gene BMW92_RS10855 (BOMP, <http://services.cbu.uib.no/tools/bomp>).

PSORTb

```
SeqID: C.burnetii
  Analysis Report:
    CMSVM-            Unknown                    [No details]
    CytoSVM-          Cytoplasmic                [No details]
    ECSVM-            Unknown                    [No details]
    ModHMM-           Unknown                    [No internal helices found]
    Motif-            Unknown                    [No motifs found]
    OMPMotif-         Unknown                    [No motifs found]
    OMSVM-            Unknown                    [No details]
    PPSVM-            Unknown                    [No details]
    Profile-          Unknown                    [No matches to profiles found]
    SCL-BLAST-        Cytoplasmic                [matched 12230948: Ornithine carbamoyltransferase]
    SCL-BLASTe-       Unknown                    [No matches against database]
    Signal-           Unknown                    [No signal peptide detected]
  Localization Scores:
    Cytoplasmic            9.97
    CytoplasmicMembrane    0.01
    Periplasmic            0.01
    OuterMembrane          0.00
    Extracellular          0.00
  Final Prediction:
    Cytoplasmic            9.97
```

Figure 159: The PSORTb test resulted in an analysis report that identified no detectable internal helices, motifs, or signal peptides. The PSORTb localization scores resulted in a 9.97 value for the cytoplasmic location. The localization score for cytoplasmic membrane was 0.01. The localization score for periplasmic was 0.01. The localization score for the outer membrane location was 0.00. The localization score for the extracellular location was 0.00. The calculated localization scores for gene BMW92_RS10855 resulted in the final predictable location of the protein to be cytoplasmic (PSORTb, <https://www.psort.org/psortb/ >).

Phobius

```
ID   UNNAMED
FT   TOPO_DOM      1     310       NON CYTOPLASMIC.
//
```

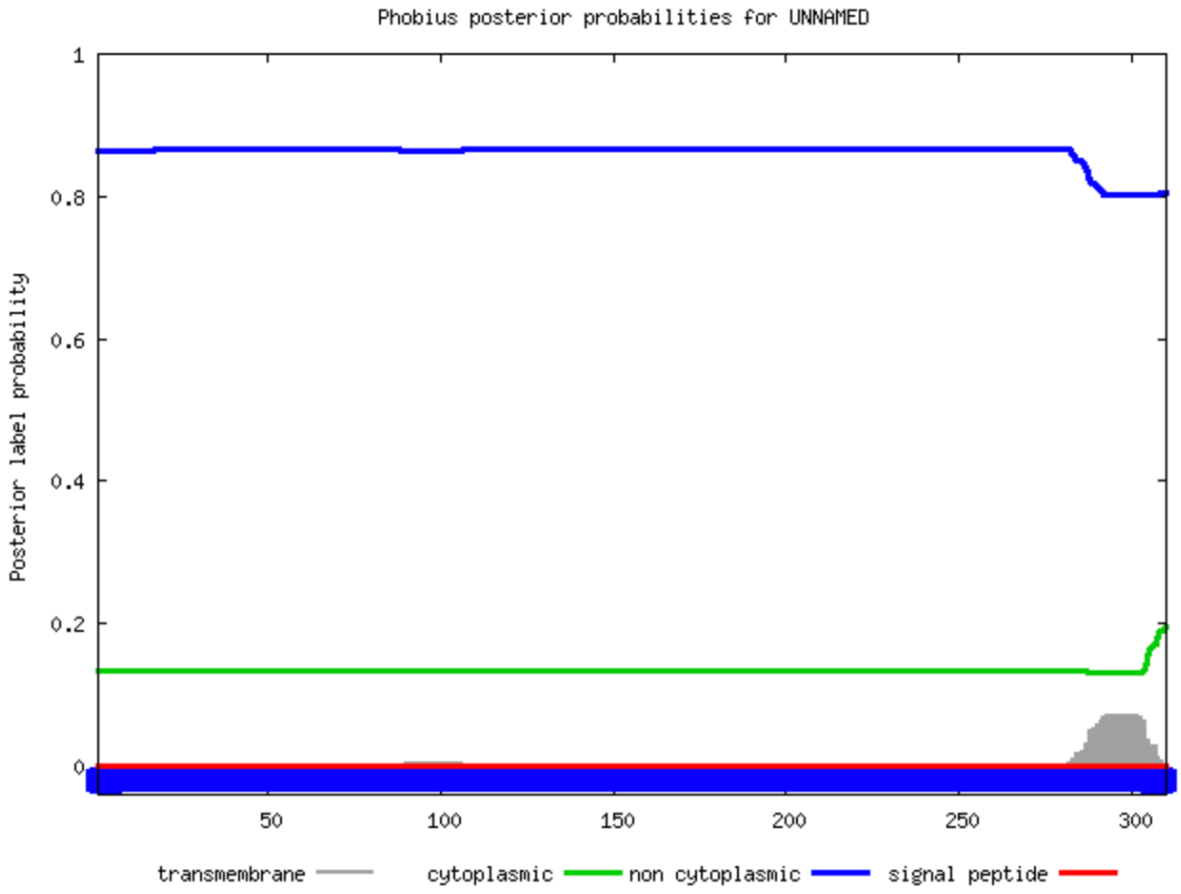


Figure 160: The Phobius posterior probability line graph generated for gene BMW92_RS10855 resulted in a calculated prediction that the whole sequence contains no membrane helices. The grey line, representative of the predicted transmembrane helices location, was around 0.1 (10%) posterior probability from amino acids 275-300. The green line, representative of the predicted cytoplasmic transmembrane helices location, was around 0.15 (15%) posterior probability from amino acids 0-302; the line changed to 0.2 (20%) from amino acids 302-310. The blue line, representative of the predicted non-cytoplasmic transmembrane helices location, was around

0.86 (86%) posterior probability from amino acids 0-280; the line changed to 0.8 (80%) from amino acids 281-310. The red line, representative of the presence or absence of a signal peptide, was 0.00 (0%) posterior probability (Phobius, <http://phobius.sbc.su.se>).

Discussion

Advancement in the field of bioinformatics and genomics has allowed for genomes of many organisms to be analyzed and understood. Continual mapping, sequencing, analyzing, and comparing of genomes continues this process of storing information and data of genes. The availability of public online bioinformatic programs used in this research allowed for the investigation of five hypothetical-protein coding genes from the bacterium *Coxiella burnetii*. Through the use of these bioinformatic programs, which are continuously updated, the five genes, BMW92_RS10760, BMW92_RS10830, BMW92_RS10835, BMW92_RS10840, BMW92_RS10855, were able to be analyzed and given a predicted function within the microorganism. Despite these genes undergoing analysis through various bioinformatic programs, further molecular and biochemical testing must be completed to fully assess and confirm the predicted function of each of the five genes selected from the bacteria *Coxiella burnetii*.

BMW92_RS10760

The first hypothetical protein coding gene of *Coxiella burnetii* that was examined was gene BMW92_RS10760. This gene was researched and predicted to encode uroporphyrinogen-III synthase. This enzyme catalyzes the asymmetrical cyclization of tetrapyrrole uroporphyrinogen-III, the fourth step in the biosynthesis of heme (Schubert et al. 2002). Tetrapyrroles have a general structure that consist of four pyrrolic rings bonded by methine bridges to one another in a cyclic form (Heinemann et al. 2008). The pyrrole rings are denoted A-D in a clockwise direction and have the ability to chelate divalent metal ions such as iron (Heinemann et al. 2008). Tetrapyrroles are derived from a common precursor, uroporphyrinogen

III, which begins the biosynthesis process as eight molecules of 2-aminoketone,5-aminolevulinic acid (Shoolingin-Jordan et al. 2003).

Every tetrapyrrole receives unique modifications around the ring periphery as the molecules are synthesized through a branched pathway. However, the unifying feature of all tetrapyrroles is the central metal ion that is capable of lying within the center of the cyclic ring structure (Schubert et al. 2002). This shared feature is only plausible due to the mechanism and action of the enzyme uroporphyrinogen-III synthase. The many biosynthetic steps of tetrapyrrole allow for modifications specific and tailored to organism species; but the final common step is critical whereby the D ring of hydroxymethylbilane is flipped during the ring closure. This allows the tetrapyrrole molecule to link to the A ring, thereby generating the asymmetrical structure of uroporphyrinogen-III (Schubert et al. 2002). This has led to an understanding that this final enzyme, uroporphyrinogen-III synthase, is essential and common to all living systems (Schubert et al. 2002).

The sequence alignments generated from BLASTp resulted in 10 outputs with high sequence similarity and level of conservation between the amino acid sequences of *Coxiella burnetii*, *Methylomarinum vadi*, *Methylovulum psychrotolerans*, *Methylomonas methanica*, *Gammaproteobacteria bacterium*, *Methylobacter tundripaludum*, *Methylobacter luteus*, *Thiohalophilus thiocyanatoxydans*, *Thiothrix nivea*, *Methylobacter oryzae*, *and Methylomonas lenta*. Each of the organisms, with respective genomes aligned with the BMW92_RS10760 gene from *C. burnetii*, were annotated to code for the same enzyme, uroporphyrinogen-III synthase. The regions of local similarity and agreement between each of the BLASTp output results support the prediction of the protein function of this specific gene.

The Conserved Domain Database (CDD) generated output results of the highest matched domain of protein to the selected gene of interest. The primary domain of this protein, projected from the top domain match when compared to the query sequence, was uroporphyrinogen-III synthase which is part of the HemD superfamily. This output result accentuates the sequence alignments, similarity, and proposed protein function examined earlier between the compared bacterial sequences.

The multiple sequence alignments and phylogenetic trees generated via MUSCLE and T-COFFEE gave visual representation of the high degree of conservation of this protein-coding sequence among similar organisms. These results further the proposition of the protein encoded and function of the protein as each of the organisms' sequences aligned and compared had e-values less than e^{-5}. This similarity is visible from the results generated via MUSCLE and T-COFFEE, displaying minimal discrepancies between each of the organisms. WEBLOGO provided a graphical representation of the level of conservation of each amino acid residue compared to the query sequence of the selected gene. The graphical representation displays high conservation of amino acids throughout the entirety of the sequence; however, the highest level of conservation of amino acid residues amongst all organisms was from amino acid residue positions 114-192. The minimal amounts of variation displayed from BLASTp, MUSLCE, T-COFFEE, and WEBLOGO output results suggests that the *Coxiella burnetii* gene BMW92_RS10760 encodes uroporphyrinogen-III synthase, which is similar in compared organisms.

These results support the predicted function of the protein and further emphasizes the necessity of this gene for the organism. This enzyme is critical for this obligate intracellular bacterium as its function to produce a precursor for the tetrapyrrole cofactor heme (Schubert et

al. 2002). The heme cofactor is the greatest reservoir of iron as it has the ability to coordinate iron atoms at the center of its porphyrin ring (Heinemann et al. 2008). When complexed together, heme and iron have increased abilities for electron transfer and redox activity (Heinemann et al. 2008). Cells rely heavily on heme molecules for the function of widely conserved heme-dependent-enzymes such as catalase, nitric oxide synthase, peroxidases, cytochromes, and hemoglobin (Choby and Skaar 2016). Furthermore, vital intracellular processes such as cellular respiration and the electron transport chain require heme to function as an electron shuttle for many of the enzymes used in these intracellular processes; without heme, the energy recovery would be minimal or non-existent (Choby and Skaar 2016). Heme has been linked to integral processes essential for living among organisms across many domains of life. Thus, organisms must either synthesize or acquire the heme molecule in order to survive, which leads to the speculation of genes that aid in this the synthesis of this molecule remain highly conserved and passed onto progeny. This further supports the prediction that this gene codes for uroporphyrinogen-III synthase, which aids in the production of a molecule, uroporphyrinogen-III, necessary for the synthesis of such an integral molecule for living.

Gram-negative bacteria, such as *C. burnetii*, contain an inner membrane (IM) and outer membrane (OM). The OM is distinct in composition as it is composed of an asymmetrical distribution of lipids, phospholipids, and lipopolysaccharides (Rollauer et al. 2015). This OM serves as a functional environment for outer membrane proteins (OMP) essential for cell functions. This double membrane composition serves as number of essential purposes; however, the cell still has to transported synthesized OMP, which are always synthesized in the cytoplasm, and transport them to locations throughout the cell, most often the OM (Rollauer et al. 2015). The uniqueness of OMP is attributed to the structure of the proteins molecules in which they lack

transmembrane alpha-helices (Koebnik et al 2000). OMP are integral membrane proteins which adopt a β-barrel structure with interchanging short and long loops on the periplasmic and extracellular sides (Rollauer et al. 2015). Each of the protein localization tests were conducted with the purpose of predicting the function of the protein and localizing the encoded protein, signal peptides, and transmembrane helices.

The first test, SignalP, identified no presence of signal peptides in the protein sequence. Signal peptides serve as a stop signal that aids in the creation of transmembrane proteins by anchoring proteins to the membrane (Coleman et al. 1985). Therefore, this result supports the idea that this protein does not require a signal peptide to attach to the membrane. The second test, LipoP, predicted the highest scoring class out of the four protein class types was cytoplasmic. There were no prediction outputs regarding the remaining three classes which further supports the data of the protein sequence lacking a signal peptide, lipoprotein signal peptide, or n-terminal transmembrane helix. Output results generated from TMHMM, Phobius, and PSORTb proposed the protein to be localized outside of the cytoplasm. Each test resulted in data with low localization scores and probabilities of the protein being localized in the cytoplasm, periplasm, and transmembrane. These findings were further supported by the BOMP test results whereby the predicted the number of integral β-barrel outer membrane proteins was zero. Therefore, this encoded protein has a low probability of being a transmembrane protein or localized in the cytoplasm of the cell.

BMW92_RS10830

The second hypothetical protein coding gene of *Coxiella burnetii* that was examined was gene BMW92_RS10830. This gene was researched and predicted to encode pyrroline-5-carboxylate reductase (P5C). This enzyme catalyzes the reduction of 1-pyrroline-5-carboxylate

(PCA) to L-proline, which is the final step of proline biosynthesis (Brandriss and Falvey 1992). L-proline is an important amino acid for many prokaryotes and eukaryotes microorganisms. Without ornithine cyclodeaminase in these microorganisms, the impeding result is amino acid starvation and blocking of protein synthesis if proline is not either synthesized or acquired (Forlani et al. 2011). Thus, the activity of P5C is absolutely necessary for many microorganisms.

Previous research has outlined the importance of this enzyme for pathogenic microorganisms. Inhibition of P5C, alongside select enzymes needed for amino acid biosynthesis, has been found to exert remarkable activity against bacteria (Harth and Horwitz 2003; Hutton et al. 2007; Forlani et al. 2011). The inhibition of enzymes like P5C, which catalyzes key reactions of amino acid biosynthesis and metabolism, have provided promising leads of control over pathogenic microorganisms (Pathania and Brown 2008). Furthermore, the amino acid proline has unique characteristics that contribute to distinct characteristics when incorporated into a protein sequence. Proline residues have been attributed with low configurational entropy due to the pyrrolidine ring hinderance; as a result, proline is important for unfolded and folded protein stability and structure (Ge and Pan 2009). Proline has been shown to play a role in stress tolerance and osmoregulation as a variety of microorganisms. Oxidative metabolism of proline allows bacteria to generate hydrogen peroxide, which implies that proline increases oxidative stress tolerance in bacteria (Zhang et al. 2014). Accumulation of compatible solutes, typically amino acids such as proline, preserve the positive turgor pressure required for cell division (Empadinhas and Costa 2008) Evidence has also suggested that an increase of intracellular proline allowed for prokaryotes to have increased protein and membrane stabilization even while under stress (Takagi 2008).

The sequence alignments generated from BLASTp resulted in 10 outputs with high sequence similarity and level of conservation between the amino acid sequences of *Coxiella burnetii*, *Coxiella mudrowiae*, *Thioalbus denitrificans*, *Nitrosococcus halophilus*, *Alkalilimnicola ehrlichii*, *Alkalispirillum mobile*, *Chromatiales bacterium, Nitrococcus mobilis*, *Nitrosococcus watsonii*, *Halobacteria archaeon*, *Aquicella lusitana*. Each of the organisms, with respective genomes aligned with the BMW92_RS10830 gene from *C. burnetii*, were annotated to code for the same enzyme, pyrroline-5-carboxylate reductase. The regions of local similarity and agreement between each of the BLASTp output results support the prediction of the protein function of this specific gene.

The Conserved Domain Database (CDD) generated output results of the highest matched domain of protein to the selected gene of interest. The primary domain of this protein, projected from the top domain match when compared to the query sequence, was pyrroline-5-carboxylate reductase which is part of the PRK11880 superfamily. This output result emphasizes the sequence alignments, similarity, and proposed protein function examined earlier between the compared bacterial sequences.

The multiple sequence alignments and phylogenetic trees generated via MUSCLE and T-COFFEE gave visual representation of the high degree of conservation of this protein-coding sequence among similar bacterial organisms. These results further the proposition of the protein encoded and function of the protein as each of the organisms' sequences aligned and compared had e-values less than e^{-5}. This similarity is visible from the results generated via MUSCLE and T-COFFEE, displaying minimal discrepancies between each of the organisms. WEBLOGO provided a graphical representation of the level of conservation of each amino acid residue compared to the query sequence of the selected gene. The graphical representation displays high

conservation of amino acids throughout the entirety of the sequence; however, the highest level of conservation of amino acid residues amongst all organisms was from positions 118-240. The minimal amounts of variation displayed from BLASTp, MUSLCE, T-COFFEE, and WEBLOGO output results suggests that the *Coxiella burnetii* gene BMW92_RS10830 encodes pyrroline-5-carboxylate reductase, which is similar in compared organisms.

These results support the predicted function of the protein and further emphasizes the necessity of this gene for the organism. Recent studies have outlined the necessity of proline for virulence during infection and a wide range of cellular processes. Some pathogens rely on proline as a critical substrate, whereas other pathogens exploit proline for stress protection or use proline as an energy source (Christgen and Becker 2019). Regardless of the biological function, proline has been linked to integral processes critical for functioning and living. Therefore, prokaryotes rely heavily on an important molecule, which suggest that genes that aid in this the synthesis of this molecule remain highly conserved and passed onto progeny. This further supports the prediction that this gene codes for pyrroline-5-carboxylate reductase, which catalyzes the reduction of 1-pyrroline-5-carboxylate (PCA) to L-proline, thereby aiding in the biosynthesis of an integral protein used for cell functioning.

Each of the protein localization tests were conducted with the purpose of predicting the function of the protein and localizing the encoded protein, signal peptides, and transmembrane helices. The first test, SignalP, identified no presence of signal peptides in the protein sequence. The predicted likelihood of the protein being classified as a protein type other than a signal peptide was the highest probability. Therefore, this result supports the idea that this protein does not require a signal peptide to attach to the membrane. The second test, LipoP, predicted the highest scoring class out of the four protein class types was cytoplasmic. There were no

prediction outputs regarding the remaining three classes which further supports the data of the protein sequence lacking a signal peptide, lipoprotein signal peptide, or n-terminal transmembrane helix. Output results generated from TMHMM, Phobius, and PSORTb proposed the protein to be localized in various regions. TMHMM predicted the highest probability of location to be extracellular. However, results also predicted a low probability of the protein localized inside the cell and transmembrane for the first 110 amino acid residues. This result was supported by Phobius which predicted the top domain as non-cytoplasmic while having an intermediate probability of the protein localized inside the cytoplasm and transmembrane for the first 110 amino acid residues. PSORTb generated a contradictory result which predicted the cytoplasm as the highest probability. This finding was further supported by the BOMP test results whereby the predicted the number of integral β-barrel outer membrane proteins was zero. Therefore, this encoded protein has a low probability of being a transmembrane protein or localized in the cytoplasm of the cell.

BMW92_RS10835

The third hypothetical protein coding gene of *Coxiella burnetii* that was examined was gene BMW92_RS10835. This gene was researched and predicted to encode pyridoxal phosphate-binding protein (PLPBP). This protein is highly uncharacterized but has been suggested to play a role in the homeostatic regulation of vitamin B_6 and amino acids. Deletion or absence of this protein causes pleotropic effects in many microorganisms, such as epilepsy in some eukaryotic mammals (Ito et al. 2019). Evidence has also displayed cell lines lacking PLPBP resulted in distinct proteomic changes, such as upregulation of several cytoskeleton and cell division associated proteins (Fux and Sieber 2019). Furthermore, recent studies have suggested the absence of this protein can result in the accumulation of pyridoxine 5′-phosphate

(PLP) and metabolites involved in the isoleucine and valine biosynthetic pathway (Ito et al. 2019).

The active form of vitamin B_6, pyridoxine 5′-phosphate (PLP), participates in many enzymatic processes (Fleischman et al. 2014). One of the main functions of vitamin B_6 is contributing as a highly versatile cofactor for diverse enzymatic reactions in basic metabolism (Richts et al. 2019). PLP has an important role during tryptophan synthase reaction. This molecule elicits an allosteric effect during the tryptophan synthase reaction as the enzymes involved require PLP (Jansonius 1998). Additionally, enzymes that catalyze decarboxylation, racemization, α-elimination, replacement, β- and γ-elimination or replacement reactions require PLP as a co-factor to regulate amino acid metabolism (Christen and Mehta 2001). Genes that code for PLPBP enable properly folded PLP to interact with PLP-dependent enzymes (Fleischman et al. 2014). This action of regulating PLP is critical to many organisms. This importance can be depicted by the number of PLP-dependent enzymes encoded by prokaryotes. Recent studies have revealed that PLP-dependent enzymes account for ∼1.5% of most prokaryotic genomes and are estimated to be involved in ∼4% of all catalytic reactions (Schneider et al. 2000). The predicted protein function of the gene has supportive evidence to be critical for maintaining homeostasis; thus, would be suggested to be highly conservative among organisms.

The sequence alignments generated from BLASTp resulted in 10 outputs with high sequence similarity and level of conservation between the amino acid sequences of *Coxiella burnetii*, *Rhipicephalus microplus*, *Coxiella mudrowiae*, *Amblyomma americanum*, *Amblyomma sculptum*, *Gammaproteobacteria bacterium*, *Beggiatoa*, *Thiotrichales bacterium*, *Nitrosococcus watsonii*, *Nitrosococcus halophilus*, *Nitrosococcus oceani*. Each of the organisms, with

respective genomes aligned with the BMW92_RS10835 gene from *C. burnetii*, were annotated to code for the same enzyme, pyridoxal phosphate-binding protein. The regions of local similarity and agreement between each of the BLASTp output results support the prediction of the protein function of this specific gene.

The Conserved Domain Database (CDD) generated output results of the highest matched domain of protein to the selected gene of interest. The primary domain of this protein, projected from the top domain match when compared to the query sequence, was pyridoxal phosphate-binding protein which is part of the PLPDE-III superfamily. This output result emphasizes the sequence alignments, similarity, and proposed protein function examined earlier between the compared bacterial sequences.

The multiple sequence alignments and phylogenetic trees generated via MUSCLE and T-COFFEE gave visual representation of the high degree of conservation of this protein-coding sequence among similar bacterial organisms. These results further the proposition of the protein encoded and function of the protein as each of the organisms' sequences aligned and compared had e-values less than e^{-5}. This similarity is visible from the results generated via MUSCLE and T-COFFEE, displaying minimal discrepancies between each of the organisms. WEBLOGO provided a graphical representation of the level of conservation of each amino acid residue compared to the query sequence of the selected gene. The graphical representation displays high conservation of amino acids throughout a majority of the entire sequence. The highest level of conservation of amino acid residues amongst all organisms was from positions 192-228. The lowest level of conservation of amino acid residues amongst all organisms was from positions 223-238; which displayed no amino acids conserved. This suggest that the amino acids from position 3-232 are conserved and involved with coding the protein. The minimal amounts of

variation displayed from BLASTp, MUSLCE, T-COFFEE, and WEBLOGO output results suggests that the *Coxiella burnetii* gene BMW92_RS10835 encodes pyridoxal phosphate-binding protein, which is similar in compared organisms.

Each of the protein localization tests were conducted with the purpose of predicting the function of the protein and localizing the encoded protein, signal peptides, and transmembrane helices. The first test, SignalP, identified no presence of signal peptides in the protein sequence. The predicted likelihood of the protein being classified as a protein type other than a signal peptide was the highest probability. Therefore, this result supports the idea that this protein does not require a signal peptide to attach to the membrane. The second test, LipoP, predicted the highest scoring class out of the four protein class types was cytoplasmic. There were no prediction outputs regarding the remaining three classes which further supports the data of the protein sequence lacking a signal peptide, lipoprotein signal peptide, or n-terminal transmembrane helix. Output results generated from TMHMM, Phobius, and PSORTb proposed the protein to be localized in various regions. TMHMM predicted the highest probability of location to be extracellular. This was not entirely supported by the Phobius result. Phobius predicted the location of the protein to primarily be non-cytoplasmic, which supports TMHMM results. However, there were equal probabilities for the protein to be localized in the cytoplasm as the probability for both locations had similar values with little deviation. PSORTb generated a result which predicted the cytoplasm as the highest probability. This prediction was given an extremely high localization score of 9.97, which supports the second highest probability Phobius result of the protein localized in the cytoplasm. This finding was further supported by the BOMP test results whereby the predicted the number of integral β-barrel outer membrane proteins was

zero. Therefore, this encoded protein has a low probability of being a transmembrane protein or localized outside of the cell.

BMW92_RS10840

The fourth hypothetical protein coding gene of *Coxiella burnetii* that was examined was gene BMW92_RS10840. This gene was researched and predicted to encode phosphoenolpyruvate carboxykinase (PEPCK). This enzyme catalyzes the phosphorylation and decarboxylation of oxaloacetate (OAA) to form phosphoenolpyruvate (PEP) using guanosine triphosphate (GTP) (Matte et al. 1997). PEPCK is a critical enzyme for gluconeogenesis that catalyzes the first committed step in the diversion of tricarboxylic acid cycle intermediates toward gluconeogenesis (Matte et al. 1997). Through a two-step process, oxaloacetate is first decarboxylated to yield pyruvate enolate anion intermediate. The second step involves the transfer of phosphoryl group from a GTP molecule to the intermediate molecule, which ultimately yields phosphoenolpyruvate (Matte et al. 1997). Regulation of the PEPCK enzyme is under control of divalent metal ions, which are necessary for the active functioning of the enzyme. This requirement for divalent metal ions can be met by magnesium, manganese, or calcium (Goldie and Sanwal 1980). Studies have suggested that organisms can interchangeably use ATP to donate the phosphoryl group needed to form PEP; however, strong evidence suggest that bacterial PEPCKs are monomeric which categorizes the enzymes as GTP-dependent (Goldie and Sanwal 1980). Furthermore, PEPCK has a unique mononucleotide-binding fold which allows for a sterically strained high-energy conformation that is capable of lowering energy of activation for phosphoryl transfer (Delbaere et al. 2004). In specific eukaryotes, PEPCK can be used in glycolytic pathways; whereas in mammals, the enzyme is critical for carbohydrate metabolism (Matte et al. 1997). In bacterial cells, the enzyme functions in the gluconeogenetic

direction which is common to almost all organisms. Thus, the selected gene could code for such an important enzyme that is absolutely necessary for many organisms.

The sequence alignments generated from BLASTp resulted in 10 outputs with high sequence similarity and level of conservation between the amino acid sequences of *Coxiella burnetii*, *Coxiella mudrowiae*, *Rhipicephalus microplus*, *Legionellales bacterium*, *Aquicella lusitana*, *Coxiellaceae bacterium*, *Pseudospirillum japonicum*, *Aquicella siphonis*, *Rickettsiella isopodorum*, *Rickettsiella viridis*, *Modicisalibacter wilcox*. Each of the organisms, with respective genomes aligned with the BMW92_RS10840 gene from *C. burnetii*, were annotated to code for the same enzyme, phosphoenolpyruvate carboxykinase. The regions of local similarity and agreement between each of the BLASTp output results support the prediction of the protein function of this specific gene.

The Conserved Domain Database (CDD) generated output results of the highest matched domain of protein to the selected gene of interest. The primary domain of this protein, projected from the top domain match when compared to the query sequence, was pyridoxal phosphate-binding protein which is part of the PEPCK-HprK superfamily. This output result emphasizes the sequence alignments, similarity, and proposed protein function examined earlier between the compared bacterial sequences.

The multiple sequence alignments and phylogenetic trees generated via MUSCLE and T-COFFEE gave visual representation of the high degree of conservation of this protein-coding sequence among similar bacterial organisms. These results further the proposition of the protein encoded and function of the protein as each of the organisms' sequences aligned and compared had e-values less than e^{-5}. This similarity is visible from the results generated via MUSCLE and T-COFFEE, displaying minimal discrepancies between each of the organisms. WEBLOGO

provided a graphical representation of the level of conservation of each amino acid residue compared to the query sequence of the selected gene. The graphical representation displays high conservation of amino acids throughout a majority of the entire sequence. The highest level of conservation of amino acid residues amongst all organisms was from positions 190-446. The lowest level of conservation of amino acid residues amongst all organisms was from positions 1-17; which displayed minimal or no amino acids conserved. This suggest that the amino acids from position 18-531 are conserved and involved with coding the protein. The minimal amounts of variation displayed from BLASTp, MUSLCE, T-COFFEE, and WEBLOGO output results suggests that the *Coxiella burnetii* gene BMW92_RS10840 encodes phosphoenolpyruvate carboxykinase, which is similar in compared organisms.

These results support the predicted function of the protein and further emphasizes the necessity of this gene for the organism. This enzyme is critical for the obligate intracellular bacterium, *C. burnetii*, to live (Chiba et al. 2015). PEPCK can be viewed as a cataplerotic enzyme as it has an important role of removing and recycling anions produced from the citric acid cycle to generate energy (Yang et al. 2009). Once OAA is converted to PEP, each subsequent product has a unique fate that allows for energy to be replenished or maintain necessary intracellular processes (Yang et al. 2009). The PEPCK enzyme serves a pivotal role of regulating carbon flow central metabolism, whether eukaryotic or prokaryotic (Chiba et al. 2015). Thus, the results lead to the idea that genes that aid such integral processes of the cell remain highly conserved and passed onto progeny. This further supports the prediction that this gene codes for phosphoenolpyruvate carboxykinase, an enzyme necessary for organism living.

Each of the protein localization tests were conducted with the purpose of predicting the function of the protein and localizing the encoded protein, signal peptides, and transmembrane

helices. The first test, SignalP, identified no presence of signal peptides in the protein sequence. The predicted likelihood of the protein being classified as a protein type other than a signal peptide was the highest probability. Therefore, this result supports the idea that this protein does not require a signal peptide to attach to the membrane. The second test, LipoP, predicted the highest scoring class out of the four protein class types was cytoplasmic. There were no prediction outputs regarding the remaining three classes which further supports the data of the protein sequence lacking a signal peptide, lipoprotein signal peptide, or n-terminal transmembrane helix. Output results generated from TMHMM, Phobius, and PSORTb proposed the protein to be localized in various regions. TMHMM predicted the highest probability of location to be extracellular. This was entirely supported by the Phobius result. Phobius predicted the location of the protein to primarily be non-cytoplasmic, which supports TMHMM results. PSORTb generated a result which predicted the cytoplasm as the highest probability. This prediction was given an extremely high localization score of 9.26, which is contradictory to both TMHMM and Phobius results. The BOMP test predicted the number of integral β-barrel outer membrane proteins as zero. This result is contradictory to each of the results that predicted the protein to be localized extracellularly with high significance. Therefore, this encoded protein has a low probability of being a signal protein or localized inside of the cell.

BMW92_RS10855

The fifth hypothetical protein coding gene of *Coxiella burnetii* that was examined was gene BMW92_RS10855. This gene was researched and predicted to encode aspartate carbamoyltransferase (ATCase). This enzyme catalyzes the first step in the biosynthesis of pyrimidine nucleotides (Patel et al. 2020). The enzymatic reaction occurs between carbamoyl phosphate and aspartate to form carbamoyl aspartate (Lehninger et al. 2013). ATCase has a

unique structure that allows it to act as an allosteric enzyme (Macol et al. 2001). Despite ATCase being ubiquitous and catalyzing the same reaction across many organisms, the enzyme remains polymorphic which allows for different oligomeric structures, compositions, and regulatory properties across diverse organisms (Patel et al. 2020). The composition of this enzyme consists of 12 polypeptide chains organized into six catalytic and six regulatory subunits (Lehninger et al. 2013). With such a unique structure, ATCase is capable of using its own regulatory subunits as either positive or negative regulators (Lehninger et al. 2013). This function of regulation is critical for all organisms. Any discrepancies in regulation of could result in upregulation or downregulation of nucleotides (Kantrowitz et al. 1988). Organisms must have their cells undergo RNA transcription and DNA replication to maintain homeostasis and survive. Both of these processes require ATCase which aids in the biosynthesis of pyrimidine nucleotides, the fundamental organic compound used for these absolutely necessary cellular processes. Pyrimidines are integral for forming hydrogen bonds with their complementary purines; thus, an absence of ATCase would result in a lack of nucleotide synthesis (Lehninger et al. 2013). Therefore, the selected gene could code for such an important enzyme that is absolutely necessary for all organisms.

The sequence alignments generated from BLASTp resulted in 10 outputs with high sequence similarity and level of conservation between the amino acid sequences of *Coxiella burnetii*, *Coxiella mudrowiae*, *Thiotrichales bacterium*, *Leucothrix arctica*, *Alteromonadaceae bacterium*, *Hydrocarboniclastica marina*, *Gammaproteobacteria bacterium*, *Oceanococcus atlanticus*, *Hahellaceae bacterium*, *Pseudolysobacter antarcticus*, *Pseudomonas sabulinigri*. Each of the organisms, with respective genomes aligned with the BMW92_RS10855 gene from *C. burnetii*, were annotated to code for the same enzyme, aspartate carbamoyltransferase. The

regions of local similarity and agreement between each of the BLASTp output results support the prediction of the protein function of this specific gene.

The Conserved Domain Database (CDD) generated output results of the highest matched domain of protein to the selected gene of interest. The primary domain of this protein, projected from the top domain match when compared to the query sequence, was aspartate carbamoyltransferase. This output result emphasizes the sequence alignments, similarity, and proposed protein function examined earlier between the compared bacterial sequences.

The multiple sequence alignments and phylogenetic trees generated via MUSCLE and T-COFFEE gave visual representation of the high degree of conservation of this protein-coding sequence among similar bacterial organisms. These results further the proposition of the protein encoded and function of the protein as each of the organisms' sequences aligned and compared had e-values less than e^{-5}. This similarity is visible from the results generated via MUSCLE and T-COFFEE, displaying minimal discrepancies between each of the organisms. WEBLOGO provided a graphical representation of the level of conservation of each amino acid residue compared to the query sequence of the selected gene. The graphical representation displays high conservation of amino acids throughout a majority of the entire sequence. The highest levels of conservation of amino acid residues amongst all organisms were from positions 60-126 and 147-215. The lowest levels of conservation of amino acid residues amongst all organisms were from positions 1-9 and 355-339; which displayed no amino acids conserved. This suggest that the amino acids from position 10-334 are conserved and involved with coding the protein. The minimal amounts of variation displayed from BLASTp, MUSLCE, T-COFFEE, and WEBLOGO output results suggests that the *Coxiella burnetii* gene BMW92_RS10855 encodes aspartate carbamoyltransferase, which is similar in compared organisms.

These results support the predicted function of the protein and further emphasizes the necessity of this gene for the organism. This enzyme is critical for the transcription and DNA replication in *C. burnetii* (Lehninger et al. 2013). The pyrimidines are vital constituents necessary for all living organisms. These molecules serve as the chemical molecules for transmission of genetic traits while also serving as sources of energy in the form of ATP, signaling molecules, and secondary messengers (Khedkar et al. 2016). Thus, the results lead to the idea that genes produce an enzyme necessary for such integral processes of the cell remain highly conserved and passed onto progeny. This further supports the prediction that this gene codes for aspartate carbamoyltransferase.

Each of the protein localization tests were conducted with the purpose of predicting the function of the protein and localizing the encoded protein, signal peptides, and transmembrane helices. The first test, SignalP, identified no presence of signal peptides in the protein sequence. The predicted likelihood of the protein being classified as a protein type other than a signal peptide was the highest probability. Therefore, this result supports the idea that this protein does not require a signal peptide to attach to the membrane. The second test, LipoP, predicted the highest scoring class out of the four protein class types was cytoplasmic. There were no prediction outputs regarding the remaining three classes which further supports the data of the protein sequence lacking a signal peptide, lipoprotein signal peptide, or n-terminal transmembrane helix. Output results generated from TMHMM, Phobius, and PSORTb proposed the protein to be localized in various regions. TMHMM predicted the highest probability of location to be extracellular. This was entirely supported by the Phobius result. Phobius predicted the location of the protein to primarily be non-cytoplasmic. Both TMHMM and Phobius results had significantly high probability and prediction scores of the protein localized extracellularly.

PSORTb generated a result which predicted the cytoplasm as the highest probability. This prediction was given an extremely high localization score of 9.97, which is contradictory to both TMHMM and Phobius results. The BOMP test predicted the number of integral β-barrel outer membrane proteins as zero. This result is contradictory to each of the results that predicted the protein to be localized extracellularly with high significance. Therefore, this encoded protein has a low probability of being a signal protein or localized inside of the cell.

Summary

Each of the five-hypothetical protein-coding genes selected from *Coxiella burnetii* were analyzed using various biochemical programs and tools that tested for sequence similarity and protein localization. The first gene, BMW92_RS10760, is projected to code for uroporphyrinogen-III synthase with no signal peptide present. The second gene, BMW92_RS10830, is projected to code for pyrroline-5-carboxylate reductase with no signal peptide present. The third gene, BMW92_RS10835, is projected to code for pyridoxal phosphate-binding protein with no signal peptide present. The fourth gene, BMW92_RS10840, is projected to code for phosphoenolpyruvate carboxykinase with no signal peptide present. The fifth gene, BMW92_RS10855, is projected to code for aspartate carbamoyltransferase with no signal peptide present. The limitations of this research include the accessibility to a limited number of internet programs and databases to test amino acid and protein sequences. Limitations regarding funding, time, and accessibility to open laboratory facilities led to hypothetically proposed conclusions generated from these sources. Additional laboratory testing involving biochemical tools and methods will be necessary to confirm each projected result. Potential directions of future research include transformation of each selected gene into well-characterized and understood bacteria followed by biochemical testing; or cell lines, with selected genes,

undergoing gene-knockout testing with biochemical analysis for attributed function. Proposed functions may be validated and tested using gene mapping techniques, comparative genomics, and biochemical analysis. Further testing may additionally include structure-based methods, sequence-based methods, and active sight computational programs (Mills et al. 2015). Furthermore, the use of oligonucleotide microarrays, mass spectrometry, and two-dimensional polyacrylamide gel electrophoresis may determine mRNA, protein, and gene expression that allows for comparative analysis and determination of protein function (Chen et al. 2002). Further laboratory testing must be conducted to confirm the predicted functions of the five-hypothetical protein-coding genes of *Coxiella burnetii* that were analyzed in this research.

Acknowledgements

First and foremost, all glory, praise, and thanks to God, Jehovah-Raah, for His showers of blessings and unending joy, peace, and guidance. I am incredibly grateful for God's loving heart to direct every step of mine while at Oral Roberts University and leading my life to encounter numerous mentors, professors, and friends.

I want to express my deep and sincere gratitude to my research advisor, Dr. Celestino Velásquez, Ph.D., Assistant Professor and Chair of Biology and Chemistry Department at Oral Roberts University, for providing me the opportunity to conduct research. His care, concern, and sincerity for his students are beyond profound and have continually motivated me to achieve my goals. He has taught me the fundamental skills and knowledge necessary for carrying out research and living life as a Christian scientist. It was a great privilege and honor to work and study under his guidance and expertise. I would also like to thank him for his friendship, compassion, and great sense of humor.

I am incredibly thankful to my parents for their love, prayers, and sacrifices that allowed me to have an excellent education and foundation for my future. The seeds of encouragement, support, and love you have sown into my life are things I am forever grateful for and will continually extend towards others as a reminder of what has been extended towards me. Also, I would like to express my thanks to my siblings, Jordan, Analynn, and Jesse, for their never-ending support and valuable prayers.

Lastly, I would like to acknowledge the countless friends I have had the privilege to meet and know while attending Oral Roberts University. The support and belief in my abilities, goals, future, and potential are greatly appreciated and have never gone unnoticed. Most importantly, I am extremely grateful for my best and closest friend, Karley Wills. Thank you for the countless

times you have encouraged and helped redirect my eyes onto the pioneer and perfecter of my faith, Jesus Christ. I am forever grateful for your continual patience, understanding, support, and love.

Literature Cited

Akporiaye ET, Rowatt JD, Aragon AA, Baca OG. Lysosomal response of a murine macrophage like cell line persistently infected with Coxiella burnetii. Infection and Immunity. 1983 [accessed 2020 Nov 11];40(3):1155–1162. doi:10.1128/iai.40.3.1155-1162.1983.

Amano K, Williams JC. Chemical and immunological characterization of lipopolysaccharides from phase I and phase II Coxiella burnetii. Journal of Bacteriology. 1984 [accessed 2020 Nov 11];160(3):994–1002. doi:10.1128/jb.160.3.994-1002.1984.

Angelakis E, Raoult D. Q fever. Veterinary microbiology. 2010 [accessed 2020 Nov 11];140(3–4):297–309. doi:10.1016/j.vetmic.2009.07.016.

Berven FS, Flikka K, Jensen HB, Eidhammer I. BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. Nucleic acids research. 2004 [accessed 2020 Nov 11];32:W394-9. http://dx.doi.org/10.1093/nar/gkh351. doi:10.1093/nar/gkh351.

Brandriss MC, Falvey DA. Proline biosynthesis in Saccharomyces cerevisiae: analysis of the PRO3 gene, which encodes delta 1-pyrroline-5-carboxylate reductase. Journal of Bacteriology. 1992 [accessed 2020 Nov 11];174(11):3782–3788. doi:10.1128/jb.174.11.3782-3788.1992.

Brock TD, Madigan MT, Bender KS, Buckley DH, Sattely MW, Stahl DA. Biology of microorganisms. 15th ed. San Francisco (CA): Pearson; 2017.

CBU - Computational Biology Unit. Uib.no. [accessed 2020 Aug 6]. http://services.cbu.uib.no/tools/bomp.

Chen G, Gharib TG, Huang C-C, Taylor JMG, Misek DE, Kardia SLR, Giordano TJ, Iannettoni MD, Orringer MB, Hanash SM, et al. Discordant protein and mRNA expression in lung adenocarcinomas. Molecular & cellular proteomics: MCP. 2002 [accessed 2020 Nov 11];1(4):304–313. http://dx.doi.org/10.1074/mcp.m200008-mcp200. doi:10.1074/mcp.m200008-mcp200.

Chiba Y, Kamikawa R, Nakada-Tsukui K, Saito-Nakano Y, Nozaki T. Discovery of PPi-type Phosphoenolpyruvate Carboxykinase Genes in Eukaryotes and Bacteria. Journal of Biological Chemistry. 2015 [accessed 2020 Nov 11];290(39):23960–23970. http://dx.doi.org/10.1074/jbc.m115.672907. doi:10.1074/jbc.m115.672907.

Christen P, Mehta PK. From cofactor to enzymes. The molecular evolution of pyridoxal-5'-phosphate-dependent enzymes. The Chemical Record. 2001 [accessed 2020 Nov 11];1(6):436–447. http://dx.doi.org/10.1002/tcr.10005. doi:10.1002/tcr.10005.

Choby JE, Skaar EP. Heme Synthesis and Acquisition in Bacterial Pathogens. Journal of Molecular Biology. 2016 [accessed 2020 Nov 10];428(17):3408–3428. doi:10.10 16/j.jmb.2016.03.018.

Christgen SL, Becker DF. Role of Proline in Pathogen and Host Interactions. Antioxidants & Redox Signaling. 2019 [accessed 2020 Nov 11];30(4):683–709. http://dx.doi.org/10.1089/ars.2017.7335. doi:10.1089/ars.2017.7335.

Coleman J, Inukai M, Inouye M. Dual functions of the signal peptide in protein transfer across the membrane. Cell. 1985 [accessed 2020 Nov 11];43(1):351–360. http://dx.doi.org/10.1016/0092-8674(85)90040-6. doi:10.1016/0092-8674(85)90040-6.

Coleman SA, Fischer ER, Howe D, Mead DJ, Heinzen RA. Temporal analysis of Coxiella burnetii morphological differentiation. Journal of bacteriology. 2004 [accessed 2020 Nov 11];186(21):7344–7352. http://dx.doi.org/10.1128/JB.186.21.7344-7352.2004. doi:10.1128/JB.186.21.7344-7352.2004.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. 2004 [accessed 2020 Sep 10]. http://dx.doi.org/10.1101/gr.849004. doi:10.1101/gr.849004.

Delbaere LTJ, Sudom AM, Prasad L, Leduc Y, Goldie H. Structure/function studies of phosphoryl transfer by phosphoenolpyruvate carboxykinase. Biochimica et biophysica acta. 2004 [accessed 2020 Nov 11];1697(1–2):271–278. http://dx.doi.org/10.1016/j.bbapap.2003.11.030. doi:10.1016/j.bbapap.2003.11.030.

Eldin C, Mélenotte C, Mediannikov O, Ghigo E, Million M, Edouard S, Mege J-L, Maurin M, Raoult D. From Q fever to Coxiella burnetii infection: A paradigm change. Clinical microbiology reviews. 2017 [accessed 2020 Nov 11];30(1):115–190. http://dx.doi.org/10.1128/cmr.00045-16. doi:10.1128/cmr.00045-16.

Emanuelsson O. Predicting protein subcellular localization from amino acid sequence information. Briefings in bioinformatics. 2002 [accessed 2020 Nov 11];3(4):361–376. http://dx.doi.org/ 10.1093/bib/3.4.361. doi:10.1093/bib/3.4.361.

EMBL-EBI. MUSCLE. Ebi.ac.uk. [accessed 2020 Jul 24]. https://www.ebi.ac.uk/Tools/msa/muscle/.

EMBL-EBI. T-Coffee. Ebi.ac.uk. [accessed 2020 Jul 24]. https://www.ebi.ac.uk/Tools/msa/tcoffee/.

Empadinhas N, da Costa MS. Osmoadaptation mechanisms in prokaryotes: distribution of compatible solutes. International microbiology: the official journal of the Spanish Society for Microbiology. 2008 [accessed 2020 Nov 11];11(3):151–161. https://www.ncbi.nlm.nih.gov/pubmed/18843593.

Fleischman NM, Das D, Kumar A, Xu Q, Chiu H-J, Jaroszewski L, Knuth MW, Klock HE, Miller MD, Elsliger M-A, et al. Molecular characterization of novel pyridoxal-5'-phosphate-dependent enzymes from the human microbiome: Microbiome PLP-Dependent Enzymes. Protein science: a publication of the Protein Society. 2014 [accessed 2020 Nov 10];23(8):1060–1076. http://dx.doi.org/10.1002/pro.2493. doi:10.1002/pro.2493.

Forlani G, Petrollino D, Fusetti M, Romanini L, Nocek B, Joachimiak A, Berlicki Ł, Kafarski P. δ1-Pyrroline-5-carboxylate reductase as a new target for therapeutics: inhibition of the enzyme from Streptococcus pyogenes and effects in vivo. Amino Acids. 2011 [accessed 2020 Nov 11];42(6):2283–2291. http://dx.doi.org/10.1007/s00726-011-0970-7. doi:10.1007/s00726-011-0970-7.

Fux A, Sieber SA. Biochemical and Proteomic Studies of Human Pyridoxal 5′-Phosphate-Binding Protein (PLPBP). ACS Chemical Biology. 2019 [accessed 2020 Nov

11];15(1):254–261. http://dx.doi.org/10.1021/acschembio.9b00857. doi:10.1021/acschembio.9b00857.

Gardy JL, Brinkman FSL. Erratum: Methods for predicting bacterial protein subcellular localization. Nature reviews. Microbiology. 2006 [accessed 2020 Aug 11];4(11):874–874. http://dx.doi.org/10.1038/nrmicro1539. doi:10.1038/nrmicro1539.

Ge M, Pan X-M. The contribution of proline residues to protein stability is associated with isomerization equilibrium in both unfolded and folded states. Extremophiles. 2009 [accessed 2020 Nov 11];13(3):481–489. http://dx.doi.org/10.1007/s00792-009-0233-7. doi:10.1007/s00792-009-0233-7.

GENI-ACT. n.d Bioinformatics Annotation of Coxiella burnetii Genes. Geni-act.org. [accessed 2020 Jul 13]. https://geni-act.org/communities/b1911ed2-f82f-43ca-add5-fc6f44cab5c9/group/c261eaa0-2dce-4264-bf64-efa3b7e33e39.

Goldie AH, Sanwal BD. Genetic and physiological characterization of Escherichia coli mutants deficient in phosphoenolpyruvate carboxykinase activity. Journal of Bacteriology. 1980 [accessed 2020 Nov 11];141(3):1115–1121. http://dx.doi.org/10.1128/JB.141.3.1115-1121.1980. doi:10.1128/JB.141.3.1115-1121.1980.

Gürtler L, Bauerfeind U, Blümel J, Burger R, Drosten C, Gröner A, Heiden M, Hildebrandt M, Jansen B, Offergeld R, et al. Coxiella burnetii - pathogenic agent of Q (query) fever. Transfusion medicine and hemotherapy. 2014 [accessed 2020 Nov 9];41(1):60–72. http://dx.doi.org/10.1159/000357107. doi:10.1159/000357107.

Harth G, Horwitz MA. Inhibition of Mycobacterium tuberculosis Glutamine Synthetase as a Novel Antibiotic Strategy against Tuberculosis: Demonstration of Efficacy In Vivo. Infection and Immunity. 2003 [accessed 2020 Nov 11];71(1):456–464. http://dx.doi.org/10.1128/iai.71.1.456-464.2003. doi:10.1128/iai.71.1.456-464.2003.

Heinemann IU, Jahn M, Jahn D. The biochemistry of heme biosynthesis. Archives of Biochemistry and Biophysics. 2008 [accessed 2020 Nov 8];474(2):238–251. doi: https://doi.org/10.10 16/j.abb.2008.02.015.

Hutton CA, Perugini MA, Gerrard JA. Inhibition of lysine biosynthesis: an evolving antibiotic strategy. Molecular Biosystems. 2007 [accessed 2020 Nov 11]. http://dx.doi.org/10.1039/b705624a. doi:10.1039/b705624a.

Ito T, Yamamoto K, Hori R, Yamauchi A, Downs DM, Hemmi H, Yoshimura T. Conserved Pyridoxal 5'-Phosphate-Binding Protein YggS Impacts Amino Acid Metabolism through Pyridoxine 5'-Phosphate in Escherichia coli. Applied and Environmental Microbiology. 2019 [accessed 2020 Nov 11];85(11). http://dx.doi.org/10.1128/AEM.00430-19. doi:10.1128/AEM.00430-19.

Jansonius JN. Structure, evolution and action of vitamin B6-dependent enzymes. Current Opinion in Structural Biology. 1998 [accessed 2020 Nov 11];8(6):759–769. http://dx.doi.org/10.1016/s0959-440x(98)80096-1. doi:10.1016/s0959-440x(98)80096-1.

Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein science: a publication of

the Protein Society. 2003 [accessed 2020 Nov 11];12(8):1652–1662. http://dx.doi.org/10.1110/ps.0303703. doi:10.1110/ps.0303703.

Kantrowitz E, Lipscomb W. Escherichia coli aspartate transcarbamylase: the relation between structure and function. Science. 1988 [accessed 2020 Nov 11];241(4866):669–674. http://dx.doi.org/10.1126/science.3041592. doi:10.1126/science.3041592.

Khedkar GD, Prakash B, Khedkar CD, Chopade BA. Nucleic Acids. In: Encyclopedia of Food and Health. Elsevier; 2016. p. 84–92. http://dx.doi.org/10.1016/b978-0-12-384947-2.00487-6. doi:10.1016/b978-0-12-384947-2.00487-6.

Klug WS, Cummings MR, Spencer CA, Palladino MA. Concepts of Genetics. 11th ed. Harlow (England): Pearson Education; 2015.

Koebnik R, Locher KP, Gelder PV. Structure and function of bacterial outer membrane proteins: barrels in a nutshell. Molecular Microbiology. 2000 [accessed 2020 Nov 10];37(2):239–253. http://dx.doi.org/10.1046/j.1365-2958.2000.01983.x. doi:10.1046/j.1365-2958.2000.01983.x.

Kozlowski LP. Calculation of protein isoelectric point. Isoelectric.org. [accessed 2020 Aug 13]. http://isoelectric.org/calculate.php.

Labedan B, Xu Y, Naumoff D, Glansdorff N. Using Quaternary Structures to Assess the Evolutionary History of Proteins: The Case of the Aspartate Carbamoyltransferase. Molecular Biology and Evolution. 2003 [accessed 2020 Nov 11];21(2):364–373. http://dx.doi.org/10.1093/molbev/msh024. doi:10.1093/molbev/msh024.

Lehninger AL, Nelson DL, Cox MM, Winslow S. Principles of biochemistry. 6th ed. New York, (NY): W.H. Freeman; 2013.

LipoP 1.0 Server. Dtu.dk. [accessed 2020 July 23]. http://www.cbs.dtu.dk/services/LipoP/

Macol CP, Tsuruta H, Stec B, Kantrowitz ER. Direct structural evidence for a concerted allosteric transition in Escherichia coli aspartate transcarbamoylase. 2001 [accessed 2020 Nov 12];8:423–426. http://dx.doi.org/10.1038/87582. doi:10.1038/87582.

Matte A, Tari LW, Goldie H, Delbaere LTJ. Structure and Mechanism of Phosphoenolpyruvate Carboxykinase. Journal of Biological Chemistry. 1997 [accessed 2020 Nov 11];272(13):8105–8108. http://dx.doi.org/10.1074/jbc.272.13.8105. doi:10.1074/jbc.272.13.8105.

Mills CL, Beuning PJ, Ondrechen MJ. Biochemical functional predictions for protein structures of unknown or uncertain function. Computational and Structural Biotechnology Journal. 2015 [accessed 2020 Nov 12];13:182–191. doi:10.1016/j.csbj.2015.02.003

McCaul TF, Williams JC. Developmental cycle of Coxiella burnetii: structure and morphogenesis of vegetative and sporogenic differentiations. Journal of Bacteriology. 1981 [accessed 2020 Nov 11];147(3):1063–1076. http://dx.doi.org/10.1128/JB.147.3.1063-1076.1981. doi:10.1128/JB.147.3.1063-1076.1981.

Minnick MF, Raghavan R. Genetics of Coxiella burnetii: on the path of specialization. Future Microbiology. 2011 [accessed 2020 Nov 12];6(11):1297–1314. http://dx.doi.org/10.2217/fmb.11.116. doi:10.2217/fmb.11.116.

Nakai K. Protein sorting signals and prediction of subcellular localization. In: Advances in Protein Chemistry. Elsevier; 2000. p. 277–344. http://dx.doi.org/10.1016/s0065-3233(00)54009-1. doi:10.1016/s0065-3233(00)54009-1.

Omsland A, Cockrell DC, Howe D, Fischer ER, Virtaneva K, Sturdevant DE, Porcella SF, Heinzen RA. Host cell-free growth of the Q fever bacterium Coxiella burnetii. Proceedings of the National Academy of Sciences. 2009 [accessed 2020 Nov 11];106(11):4430–4434. http://dx.doi.org/10.1073/pnas.0812074106. doi:10.1073/pnas.0812074106.

Patel C, Vaishnav A, Edwards BFP, Evans DR. Characterization and assembly of the Pseudomonas aeruginosa aspartate transcarbamoylase-pseudo dihydroorotase complex. PloS One. 2020 [accessed 2020 Nov 11];15(3). http://dx.doi.org/10.1371/journal.pone.0229494. doi:10.1371/journal.pone.0229494.

Pathania R, Brown E. Small and lethal: searching for new antibacterial compounds with novel modes of action. Biochemistry and Cell Biology. 2008 [accessed 2020 Nov 11]. http://dx.doi.org/10.1139/o08-011. doi:10.1139/o08-011.

Phobius. Sbc.su.se. [accessed 2020 Aug 7]. https://phobius.sbc.su.se/.

Protein BLAST: search protein databases using a protein query. Nih.gov. [accessed 2020 Jul 23]. https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome.

PSORTb Subcellular Localization Prediction Tool - version 3.0. Psort.org. [accessed 2020 Aug 6]. https://www.psort.org/psortb/.

Robertson P, Abdelhady H, Garduño RA. The many forms of a pleomorphic bacterial pathogen the developmental network of Legionella pneumophila. Frontiers in microbiology. 2014 [accessed 2020 Nov 11];5:670. http://dx.doi.org/10.3389/fmicb.2014.00670. doi:10.3389/fmicb.2014.00670.

Rollauer SE, Sooreshjani MA, Noinaj N, Buchanan SK. Outer membrane protein biogenesis in Gram-negative bacteria. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. 2015 [accessed 2020 Nov 9];370(1679). http://dx.doi.org/10.1098/rstb.2015.0023. doi:10.1098/rstb.2015.0023.

Schneider G, Fechner U. Advances in the prediction of protein targeting signals. Proteomics. 2004 [accessed 2020 Nov 9];4(6):1571–1580. http://dx.doi.org/10.1002/pmic.200300786. doi:10.1002/pmic.200300786.

Schneider G, Käck H, Lindqvist Y. The manifold of vitamin B6 dependent enzymes. Structure. 2000 [accessed 2020 Nov 9];8(1). http://dx.doi.org/10.1016/s0969-2126(00)00085-x. doi:10.1016/s0969-2126(00)00085-x.

Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic acids research. 1990 [accessed 2020 Aug 19];18(20):6097–6100. http://dx.doi.org/10.1093/nar/18.20.6097. doi:10.1093/nar/18.20.6097.

Schubert HL, Raux E, Matthews MA, Phillips JD, Wilson KS, Hill CP, Warren MJ. Structural diversity in metal ion chelation and the structure of uroporphyrinogen III synthase. Biochemical Society Transactions. 2002 [accessed 2020 Nov 9];30(4):595–600. http://dx.doi.org/10.1042/bst0300595. doi:10.1042/bst0300595.

Shoolingin-Jordan PM, Al-Daihan S, Alexeev D, Baxter RL, Bottomley SS, Kahari ID, Roy I, Sarwar M, Sawyer L, Wang SF. 5-Aminolevulinic acid synthase: mechanism, mutations and medicine. Biochimica et biophysica acta. Proteins and proteomics. 2003 [accessed 2020 Nov 9];1647(1–2):361–366. http://dx.doi.org/10.1016/s1570-9639(03)00095-5. doi:10.1016/s1570-9639(03)00095-5.

SignalP-5.0. Dtu.dk. [accessed 2020 July 23]. http://www.cbs.dtu.dk/services/SignalP/

Takagi H. Proline as a stress protectant in yeast: physiological functions, metabolic regulations, and biotechnological applications. Applied microbiology and biotechnology. 2008 [accessed 2020 Nov 9];81(2):211–223. http://dx.doi.org/10.1007/s00253-008-1698-5. doi:10.1007/s00253-008-1698-5.

TMHMM Server v. 2.0. Dtu.dk. [accessed 2020 Jul 23]. http://www.cbs.dtu.dk/services/TMHMM/

Tortora GJ, Funke BR, Case CL. Microbiology: an introduction. 12th ed. Harlow (England): Pearson Education; 2018.

Woldehiwet Z. Q fever (coxiellosis): epidemiology and pathogenesis. Research in veterinary science. 2004 [accessed 2020 Aug 22];77(2):93–100. http://dx.doi.org/10.1016/j.rvsc.2003.09.001. doi:10.1016/j.rvsc.2003.09.001.

Yang J, Kalhan SC, Hanson RW. What Is the Metabolic Role of Phosphoenolpyruvate Carboxykinase? Journal of Biological Chemistry. 2009 [accessed 2020 Nov 11];284(40):27025–27029. http://dx.doi.org/10.1074/jbc.R109.040543. doi:10.1074/jbc.R109.040543.

Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics (Oxford, England). 2010 [accessed 2020 July 24];26(13):1608–1615. http://dx.doi.org/10.1093/bioinformatics/btq249. doi:10.1093/bioinformatics/btq249.

Zhang L, Alfano JR, Becker DF. Proline metabolism increases katG expression and oxidative stress resistance in Escherichia coli. Journal of bacteriology. 2015 [accessed 2020 Nov 11];197(3):431–440. http://dx.doi.org/10.1128/JB.02282-14. doi:10.1128/JB.02282-14.