The Glycine and Proline Reductase Systems: An Evolutionary Perspective and Presence in Enterobacteriaceae

2013

Joshua Witt
University of Central Florida

Find similar works at: https://stars.library.ucf.edu/honorstheses1990-2015

University of Central Florida Libraries http://library.ucf.edu

Part of the Microbiology Commons, and the Molecular Biology Commons

Recommended Citation

https://stars.library.ucf.edu/honorstheses1990-2015/1546

This Open Access is brought to you for free and open access by STARS. It has been accepted for inclusion in HIM 1990-2015 by an authorized administrator of STARS. For more information, please contact lee.dotson@ucf.edu.
THE GLYCINE AND PROLINE REDUCTASE SYSTEMS:
AN EVOLUTIONARY PERSPECTIVE AND PRESENCE IN
ENTEROBACTERIACEAE

by

JOSHUA C. WITT

A thesis submitted in fulfillment of the requirements
for the Honors in the Major Program in Microbiology and Molecular Biology
in the College of Medicine
and in The Burnett Honors College
at the University of Central Florida
Orlando, Florida

Fall Term 2013

Thesis Chair: Dr. William Self
Abstract

The Glycine and Proline Reduction systems are two of the best characterized selenoenzymes in bacteria and have been found to occur in a wide variety of *clostridia* [1-5]. These enzymes are utilized to reduce glycine or D-proline to obtain energy via substrate level phosphorylation or membrane gradients, respectively [6, 7]. This includes the pathogens *C. difficile* and *C. botulinum* [5, 8]. Strains of *C. difficile* are activate toxigenic pathways whenever either of these pathways is active within the cell [5, 8]. Though evolutionary studies have been conducted on ammonia producing bacteria [9] none has been done to directly characterize these two system by themselves. This includes an understanding of whether or not this system is transferred between organisms, as many of the *clostridia* that are to be studied are known to have an “open genome.” [8, 10] With this information we were able to generate a phylogenetic model of the proline and glycine reduction systems. Through this analysis, we were able to account for many *clostridial* organisms that contain the system, but also many other organisms as well. These included *enterobacteriaceae* including a strain of the model organism, *Escherichia coli*. It was further concluded that Glycine Reductase was a much less centralized system and included a wide range of taxa while Proline Reductase was much more centralized to being within the phyla of firmicutes. It was also concluded that the strain of *E. coli* has a fully functional operon for Glycine Reductase.
ACKNOWLEDGMENTS

I would like to thank everyone in my honors committee for guiding me through this process as I have been going through a massive change in my life. I would like to first thank Dr. William Self for giving me my first chance to work in his lab and to understand the dynamics of microbiology and molecular biology lab. Without this experience, I doubt I would be where I am today with this thesis. I would also like to thank you for your continuing encouragement and guidance throughout this process as well as with respect to my own career goals in bioengineering, I don’t think I would have mustered the courage to change to that field without you. I would also like to thank Dr. Christopher Parkinson for making this project the best that it can be. I would also like to thank him for his work with the EXCEL program, if it weren’t for those efforts, I would not be where I am today.

I would also wish to acknowledge my parents, Jeffrey Witt, Pam Witt, and Lisa Colpe, and my girlfriend, Sarah Thorncroft, for their continuing love, encouragement, and efforts to help me be who I am today. Lastly, I would like to acknowledge my brothers from Phi Sigma Pi, especially Kelly Worthen and Shawn Gaulden, for their wisdom and guidance to being able to complete the Honors in the Major program.
# TABLE OF CONTENTS

INTRODUCTION.......................................................................................................................... 1

METHODS.................................................................................................................................. 4
  Establish the phylogeny of glycine and proline reductase...................................................... 4
  Genomic vs individual gene GC content.................................................................................. 5
  Further analyzing the Phylogenetic trees ............................................................................... 5

RESULTS ..................................................................................................................................... 7
  The occurrence of non-firmicutes expressing glycine reductase........................................... 7
  Further analysis of the Phylogenetic trees ............................................................................. 9
  A grd operon is discovered in an *E. coli* species................................................................. 10

DISCUSSION ............................................................................................................................... 11

APPENDIX: Figures ................................................................................................................... 14
  APPENDIX: Figures 1.1 ........................................................................................................ 15
    APPENDIX: Figures 1.1.1 .................................................................................................... Error! Bookmark not defined.
    APPENDIX: Figures 1.1.2 .................................................................................................... Error! Bookmark not defined.
  APPENDIX: Figures 1.2 ........................................................................................................ 16
    APPENDIX: Figures 1.2.1 .................................................................................................... Error! Bookmark not defined.
    APPENDIX: Figures 1.2.2 .................................................................................................... Error! Bookmark not defined.
  APPENDIX: Figures 1.3 ........................................................................................................ 17
  APPENDIX: Figures 1.4 ........................................................................................................ 18
  APPENDIX: Figures 2.1 ........................................................................................................ 19
  APPENDIX: Figures 2.2 ........................................................................................................ 20
  APPENDIX: Figures 2.3 ........................................................................................................ 21
  APPENDIX: Figures 2.4 ........................................................................................................ 22
  APPENDIX: Figures 3.1 ........................................................................................................ 23
  APPENDIX: Figures 3.2 ........................................................................................................ 24
  APPENDIX: Figures 3.3 ........................................................................................................ Error! Bookmark not defined.
  APPENDIX: Figures 3.4 ........................................................................................................ 26
  APPENDIX: Figures 4.1 ........................................................................................................ 27
INTRODUCTION

Stickland reactions, a type of amino acid fermentation, are a set of metabolic pathways that are found within a large variety of anaerobic eubacteria. The Stickland reaction is based on the use of an oxidative reaction (such as the oxidation of an amino acid) and the reduction of glycine and D-proline that is fed from the electrons and protons from the oxidation reaction [6, 7]. The most common electron acceptors are glycine and D-proline while alanine, valine, and isoleucine are common electron donors [6, 7]. This reaction is prevalent in many systems, including the rumen of many gastrointestinal tracts and the bacteria that use this process produce a large amount of ammonia[11]. The Stickland reaction and byproducts of amino acid fermentation are also thought to be very toxic and are one of the virulence factors that lead to the toxigenicity of Clostridium difficile[12].

The Stickland reaction has long been studied through many different species of microbes but has always been characterized by the use of a pair of amino acids to actively move protons and electrons [6, 7]. The reaction is as follows, using both proline and glycine as reactants in two separate and unique reactions:
The model organism for the Stickland reactions is the gram positive, spore forming, anaerobe, *C. sticklandii*. This is due to its prevalence in published biochemical analysis of the Stickland reaction [13, 14]. It is normally described as the prototypical species for study of Stickland reactions [14]. This species uses the following amino acids for fermentation purposes: threonine, arginine, serine, cysteine, proline, and glycine [14]. It also tends to excrete the following amino acids: glutamic acid, aspartic acid, and alanine [14].

These proteins contain three seleno-proteins within the general subset: GrdA, GrdB and PrdB [5, 15, 1]. The process by which selenium is transferred into the cell is unknown, but the process by which the bacteria incorporates the selenium into the proteins is known. This process uses the stop codon that is inserted into the protein through a specific tRNA (sec) which is known by its
genetic name *selC* [16, 17]. The process is mediated by the elongation factor for the selenium tRNA, or SelB [16, 18]. There are many bacterial and eukaryotic organisms that use selenium in certain types of proteins, but these are limited in their scope [19].

The way in which we describe whether or not a system is of parallel evolutionary origin or through horizontal gene transfer is through a GC comparison analysis [20]. The commonly associated bacterial systems that use glycine and proline reduction, the *clostridia*, are well known to have open genomes and will frequently transfer genetic material [8, 10]. These include the pathogenic species such as strains of *C. botulinum* and *C. difficile* [6, 7]. GC content analysis can reveal information to allow for a proper evolutionary analysis when dealing with such a large degree of bacteria that are known to exchange genetic material [20]. Fortunately for this study, the GC content of the clostridia and the firmicutes is known to very low in comparison to many other species of bacteria [21].

Genome context can give further insight into how a specific enzyme is used within a specific organism. Genes that have a metabolic relationship with each other are normally in close proximity to each other within the genome of the host [22]. This close proximity allows for rapid transcription of a system and also the components that may be necessary for the proliferation of that system [22]. For this analysis, it will be important to understand what sort of associations can be established with respect to both the glycine and proline reduction pathways, as well as whether or not these are related within themselves.
METHODS

Establish the phylogeny of glycine and proline reductase. We inferred the phylogeny of proteins involved in the Stickland reactions using phylagrams. Phylogenic trees were constructed from proteins that are specific to the reactions: GrdA, GrdB, PrdA, PrdB. The family of proteins for each target (GrdA, GrdB, PrdA, PrdB) was obtained using the search tool of the NCBI protein database. The results that were produced were then extracted into FASTA format onto a Word document. The amino acid sequences were then screened for non-repeating sequences and fragments. The .txt file that is created for each protein is to then be placed into the alignment program ClustalX. The amino acid sequences of all the species are aligned using this program [23]. The protein alignments are carried out using the Neighbor Joining algorithm, which produces an unrooted phylogenetic tree [23, 24]. The branch length of each entry within the tree is also calculated using this algorithm as well [23, 24].

The ClustalX program also contains a command to produce phylogenetic trees with the alignment data that has been input. Once the tree file (.phy) was produced, it was viewed via the Figtree program [25, 26]. Within the program, the firmicutes and non-firmicutes were distinguished by representing the firmicutes in black and non-firmicutes in red [25, 26]. Once these corrections were made, the output file was a PDF that could be further utilized.

The next procedure was to track the phylogeny of each species that appears in the search for each of the proteins based on taxa. When conducting the search on the NCBI Protein Database a phyla report is provided on the side of the window that allows for the user to see a taxonomic breakdown of the species present in the search. This information is then manually entered into an
EXCEL file for each protein. Once all of the data has been entered for each protein, a phylogenetic analysis can then be carried out. This was first done by measuring the percent of each phyla present that accounts for the total amount of species present. This was carried out to another level where the amount of entries in each class is then measured against the total as well as their amount accounting towards their individual phyla.

**Genomic vs individual gene GC content.** We analyzed GC content to calculate the potential for horizontal gene transfer. This can easily be observed with this case because of there being two distinct groups of high and low genome GC content. This was done by using the nucleotide database search tool to find the genomic profile for each species that has been accounted for. The total GC content, as well as the genome size (in MBp), is recorded for each species and entered into an EXCEL file. The gene GC content was found by using the gene database of the NCBI database. This was done by using the nucleotide and gene database search tool to find the nucleotide sequences for each protein of interest for all the species accounted for. The GC content and gene size (in base pairs) was able to be calculated through a GC content calculator [27]. These values were then entered into an EXCEL file corresponding to their species genome size and GC content. With both the genomic and each gene’s GC content for all the protein and their species, these corresponding values were then plotted with the individual gene being the dependent variable while the genome GC content was the independent variable.

**Further analyzing the Phylogenic trees.** This step is a further analysis of a smaller subset of each gene according to different factors which include: evolutionary age according the phylogram, virulence, historic models for the system, and anomalies. This subgroup was further
analyzed by the relationship of their proteins of interest to each other as well as the genetic context of the genes of interest in their individual genomes.

Using the amino acid sequences that were gathered from the first method, the sequences for each subset were collected and then loaded on the CLUSTALX alignment tool [23]. This data was then formatted onto a Word document using a postscript function within the CLUSTALX program to produce protein alignments that are viewable [23].

For each species, the genetic context was collected via the biocyc database [28] through the genomic view of the gene of interest. The area that was looked upon was a 15 KB radius around the gene of interest. The genes that were located within this radius were recorded as well as their order, length, and direction of transcription. This information was able to be formulated on a Word document.
RESULTS

The occurance of non-firmicutes expressing glycine reductase. When the phylogeny was established via the CLUSTALX [23] alignment tool with a phylagram, there was a noticeable trend that was apparent with respect to both protein families. Since the evolutionary age of the proteins that could be deciphered by the distance away from the origin or stem point of the tree, evolutionary age could then be inferred. The branch length of each phylogram was calculated by the J-N algorithm within the CLUSTALX program. The age of the selection would be more closely related to the origin if the branch length is at its smallest.

The GrdA protein contains two distinct groups that vary greatly with respect to their change from the common ancestral sequence (Figure 1.1). There is primarily a larger group of proteins that are closely related to the common ancestor that contains mostly bacteria from the firmicutes and the clostridia. This tree contains a large amount of non-firmicutes within it that have a wide range of phylogeny across the domain of eubacteria (Figure 1.1). On the other hand, in the clade that is less related to the common ancestor, there is a mix of both firmicutes and non-firmicutes. There is also the presence of non-firmicutes within the clade of entries that are more closely related to the common ancestor. This may also be due to the presence of smaller proteins that are present within the glycine produced set.

GrdB seems to exhibit the same sorts of characteristics as the GrdA tree but with some key differences (figure 1.2). One being that the branch that is much less related to the common ancestor is not nearly as inhabited by as many entries as compared to the GrdA tree. This group still contains many non-firmicutes in it as well (figure 1.1, 1.2). The other main difference
between the two trees is that there is a high volume of non-firmicutes in a region that is much more closely related to the ancestral sequence. This could suggest two potential options, one being that this is a case of parallel evolution, or that this protein set is in fact very ancient and these are the last remaining species with a glycine reduction pathway.

The Proline Reductase proteins exhibit many of the same characteristics that the Glycine Reductase proteins. The PrdA protein contains no non-firmicutes within its subset but seems to exhibit the same pattern of phylogeny as GrdA with the only discrepancy being that there is a branch that is highly related to the common ancestor (Figure 1.1, 1.3, Table 1.1). While the PrdB protein, which contains a small subset of non firmicutes, expresses similar branching patterns to that of PrdA (Figure 1.3, 1.4, Table 1.1). This includes a highly related branch to the common ancestor, another much less related and a much larger subset that falls in between both groups. The key differences being that the most common branch in PrdA consists entirely of *C. difficile* strains while the same branch on the PrdB tree consists entirely of *C. botulinum* strains (Figure 1.3, 1.4).

Since there was a large contingency of non-firmicutes, it was decided to test whether or not there was a horizontal gene transfer between these two distinct groups. This was done using a GC comparison described in the methods section above. This type of analysis was done because of the firmicutes having a distinctly low GC content (~20%) while most other bacteria have a GC content around 50%. When plotted, the two graphs of interest, *grdA* and *grdB* genes, yielded a positive direct relationship between the GC content of the gene and the GC content of the genome (Figure 2.1, 2.2). This implies two separate two scenarios: gene transfer in the deep past
that allowed for point mutations to resonate with the genome or parallel evolution of the proteins of interest. As for the prdA and prdB genes, there was a direct relationship between these as well but there was a large degree of grouping because the GC content was similar between the group since this was almost an entirely firmicutes based group (Figure 2.3,2.4).

**Further analysis of the Phylogenetic trees.** Further analysis at the trees and their constituents as well, there were many surprises that appeared. The first being that the model organism, *E. coli*, has a constituent strain that appeared in both the GrdA and GrdB protein subsets. Not only was it present, but relative species were present in both the GrdA and GrdB proteins as well (Figure 1.1, 1.2). The phyla, proteobacteria, for *E. coli* was also strongly represented with regards to the appearance of the GrdA and GrdB proteins (Table 1.1, 1.2). Also found in the GrdA and GrdB proteins was a non-firmicute species containing a protein similar to the common ancestor, *Brachyspira pilosicoli* (Figure 1.1, 1.2). This bacteria is found in the spirochetes phyla, which is also shown to have a large representation within both glycine reductase proteins (Table 1.1, 1.2). Also, the presence of strains of *Salmonella enterica* strains was also surprising as well seeing as it is a model organism as well.

The species that were selected for further analysis for both glycine reductase were: *B. pilosicoli*, *E. coli* TA206, *C. difficile* R20291, *C. sporogenes*, *E. raffinous*, *S. enterica Newport Strain*. The *Salmonella, Escherichia* samples were selected because of their being model organisms that are not thought of containing this pathway as well as being less related to the common ancestral protein sequence. The *Enterococcus* and *Brachyspira* were selected for being "more distant" proteins while also not being considered model organisms. The clostridium species were selected
for their age, being model organisms for the system, as well as being pathogenic. The proline set was constructed along similar guidelines: *C. difficile* R20291, *C. sporogenes*, *C. sticklandii*, *C. botulinum* B1, *L. antri*, *P. acidipropionici*, *K. racemifer*, *M. micronuciformis*.

**A grd operon is discovered in an *E. coli* species.** With the groups constructed, further research was needed to determine any possible functionality to these proteins and what their context was within the organism. The first step being done by sequence alignment of their amino acid sequences. Generally there was a high degree of similarity between the amino acid sequences of the protein alignments though the length of the proteins could vary in some instances. For the Protein alignment of the GrdA protein shows some differentiation of the sequences with an insertion but high degree of similarity between the set besides for the *S. enterica* strain (Figure 3.3). The GrdB protein alignment shows a high degree of similarity between all the constituents of the glycine reductase subset (Figure 3.4). Protein alignment of the PrdA protein shows high degree of similarity of the protein between the proline reductase subset in the first portion of the protein but less further on (Figure 3.1). The protein alignment of the PrdB protein alignment shows a relatively high degree of similarity between the members of the subset (Figure 3.2).

The other protein group of proline reduction also had the same parameters placed onto it. For *C. difficile* it was all of the genes upstream of prdA and prdB within the 12Kb radius and ending with sspB gene. While for *C. sticklandii* all of the proteins that were included within the 30Kb diameter were included. For *C. botulinum* B1: prdR, prdE, prdD, and an electron transporter. In *L. antri* all of the genes downstream from lemA to the end of the 15Kb radius were sound to be relevant. While for *M. micronuiformis* all of the genes found within the 30Kb diameter were found to be relevant. For the entries that were not from clostridia (*M. micronuciformis, L. antri*) there was once again the presence of selenium integration factors that were proximal to our genes of interest.
DISCUSSION

With both firmicutes and non-firmicutes having been present in all groups within the glycine reductase system, as well as corollary evidence from a positive GC content trend two conclusions can be formed. One is that glycine reduction is a parallel evolutionary system with roots in many origins across eubacteria. It can be further assessed that this is likely a niche type system for bacterial systems that are present within the gut of animals or in protein rich sediment. This leads to another possible scenario, one in which the genetic material was exchanged long ago within the gut of animals between two distantly related species of bacteria [22]. The reason for such a long time period would be that according to the GC content analysis done in this study, it would take quite a long time for the GC of a firmicute (which are normally very low) to acclimate to that of normal levels of bacteria. Though most of the species that seemed to contain the glycine reductase system were of firmicute origin, it can be said that this is not a system that is characterized entirely as being such.

While it can be said that glycine reductase is not primarily a firmicute system, when studied further, the species E. coli TA206 was found to have a relevant protein that has a crucial role in its function. It was found to be near the lac operon as well as the selenium incorporation system within its genome. Though the presence of glycine reduction within this strain of E. coli has not been thoroughly described with this study, further physiological testing can be done to determine whether or not it is present and active. This is important because this can establish a new model organism that is not within the clostridium or firmicutes to further study the physiology of the glycine reduction system as well as Stickland reactions as a whole.
As for the proline reduction system, it can be said that there was a high occurrence in the firmicutes and the common ancestor could have been derived from within the firmicutes. Though there were some entries within PrdB that were not within the firmicutes, and these entries should be examined further as there was evidence to describe them as functional with their genomic context. Though it remains to be seen as to how functional this pathway is in those entries because of the lack of PrdA. Further physiological studies can be done to further analyze the rate of reaction within the organisms outlined above that have the B subunit but not the A subunit of proline reductase. With respect to the other organisms present, there were many that would be described as being oral pathogens such as the many strains of *Oribacterium*, and this could be described by the high levels of proline within the environments of the mouth and also the presence of delta-aminovaleric acid in plaque [30]. These findings can help target possible oral pathogens and allow for better understanding as well as better methods of treatment for dentists. *Oribacterium*, and this could be described by the high levels of proline within the environments of the mouth and also the presence of delta-aminovaleric acid in plaque [30]. These findings can help target possible oral pathogens and allow for better understanding as well as better methods of treatment for dentists.

Another possible scenario is that this pathway has been selected out of most prokaryotes over an extremely long period of time. This theory has already been postulated that the Stickland fermentation pathways were the key metabolic step to the origin of life on earth [29]. This theory utilizes the property of ribosomes to act as catalytic enzymes and complementarity of tRNA anticodons of amino acids that are Stickland fermentation pairs [29]. The reason this pathway may
have been selected against may have been due in part to the inefficiency of this pathway compared to that of other forms of oxidative phosphorylation and membrane gradient control.

These findings can help give a better context to the whole microbiome as a whole because of the products and results of these reactions have on their host organisms. As stated earlier, the end products of Stickland reaction can trigger the virulent effects of *C. difficile* [5, 8] and the same physiological effects could apply to other pathogens as well. This study and other studies like this can help find and further understand the species that may have this system. As well as for virulence, the by-products of glycine reduction, especially acetate, has a large scale impact on humans with regards to obesity [28] and the same physiological effects could apply to other pathogens as well. This study and other studies like this can help find and further understand the species that may have this system. As well as for virulence, the by-products of glycine reduction, especially acetate, has a large scale impact on humans with regards to obesity [31].

The results of this analysis are not solid through the end of times. This is because of the ever changing nature of the genome databases and the large amount of data that is being processed. Further hits will appear for both glycine and proline reduction will appear in the future and as we study more bacterial environments.
APPENDIX: Figures
The GrdA protein sequences were collected from the NCBI protein database. These sequences were then aligned by the CLUSTALX protein alignment program, and produced the phylogenetic tree which was viewed with the Figtree program. Branch distance calculated using the J-N Algorithm in ClustalX.
APPENDIX: Figures 1.2

Figure 1.2: Phylogenic tree (phylogram) of the GrdB protein.

The GrdB protein sequences were collected from the NCBI protein database. These sequences were then aligned by the CLUSTALX protein alignment program, and produced the phylogenic tree which was viewed with the Figtree program. Branch distance calculated using the J-N Algorithm in ClustalX.
The PrdA protein sequences were collected from the NCBI protein database. These sequences were then aligned by the CLUSTALX protein alignment program, and produced the phylogenetic tree which was viewed with the Figtree program. Branch distance calculated using the J-N Algorithm in ClustalX.
APPENDIX: Figures 1.4

Figure 1.4: Phylogenic tree (phylogram) of the PrdB protein.

The PrdB protein sequences were collected from the NCBI protein database. These sequences were then aligned by the CLUSTALX protein alignment program, and produced the phylogenic tree which was viewed with the Figtree program. Branch distance calculated using the J-N Algorithm in ClustalX.
Figure 2.1: This table shows the wide variety of organisms that contain the grdA gene as well as the lack of horizontal gene transfer due to the positive correlation in gene and genome GC content.

The GC content of both the gene and the genome of the host were collected by the NCBI gene and genome database.
APPENDIX: Figures 2.2

**Figure 2.2:** This table shows the relative diversity of the grdB gene and also the lacking of horizontal gene transfer with the positive correlation between the GC content of the gene and the genome.

The GC content of both the gene and the genome of the host were collected by the NCBI gene and genome database.
APPENDIX: Figures 2.3

Figure 2.3: This table shows the relative small distribution of the prdA gene across species and only within the firmicutes and clostridia due to the small distribution of the GC content of the gene compared to the genome.

<table>
<thead>
<tr>
<th>prdA GC content</th>
<th>0.00%</th>
<th>10.00%</th>
<th>20.00%</th>
<th>30.00%</th>
<th>40.00%</th>
<th>50.00%</th>
<th>60.00%</th>
<th>70.00%</th>
<th>80.00%</th>
<th>90.00%</th>
<th>100.00%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome GC content</td>
<td>20.00%</td>
<td>30.00%</td>
<td>40.00%</td>
<td>50.00%</td>
<td>60.00%</td>
<td>70.00%</td>
<td>80.00%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The GC content of both the gene and the genome of the host were collected by the NCBI gene and genome database.
Figure 2.4: This table shows the GC content of the prdB primarily concentrating between within the low levels of the firmicutes but also being more dispersed compared to that of the prdA GC content comparison.

The GC content of both the gene and the genome of the host were collected by the NCBI gene and genome database.
APPENDIX: Figures 3.1

Figure 3.1: Protein alignment of the PrdA protein shows high degree of similarity of the protein between the proline reductase subset in the first portion of the protein but less further on.

The protein alignment of the PrdA protein of the glycine reductase subset of organisms was collected using the protein sequences from Figure 1.1 and the alignment was done on the CLUSTALX protein alignment program.
The protein alignment of the PrdB protein of the glycine reductase subset was collected using the protein sequences used in Figure 1.2 and the alignment done on the CLUSTALX protein alignment program.
Figure 3.3: Protein alignment of the GrdA protein shows some differentiation of the sequences with a insertion but high degree of similarity between the set besides for the *S. enterica* strain.

The protein alignment of the GrdA protein of the proline reductase subset of organisms used the protein sequences from Figure 1.3 and were aligned using the CLUSTALX protein alignment program.
The GrdB protein alignment of the proline reductase subset of organisms used the protein sequences from Figure 1.4 and were aligned using the CLUSTALX protein alignment program.
APPENDIX: Figures 4.1

Figure 4.1: Genomic context of the glycine reductase shows high affinity with thioredoxin in all species as well as selenium incorporation within the non-firmicutes.

Clostridium difficile R20291

The genome context of *C. difficile R20291* was collected through the Biocyc database and the set organism being this organism. The diameter of the area of the genome is 35kb (kilobases) long with the center being the glycine reductase operon.
APPENDIX: Figures 4.1.2

The genome context of *B. pilosicoli 95/1000* was collected through the Biocyc database and the set organism being this organism. The diameter of the area of the genome is 35kb (kilobases) long with the center being the glycine reductase operon.
The genome context of *E. coli TA206* was collected through the Biocyc database and the set organism being this organism. The diameter of the area of the genome is 35kb (kilobases) long with the center being the glycine reductase operon.
APPENDIX: Figures 4.2

Figure 4.2: Genomic context for the proline reductase system shows a different operons for the prdA and prdB genes in some species as well as selenium incorporation elements in non-firmicutes.

**Clostridium difficile R20291**

The genome context of *C. difficile R20291* was collected through the Biocyc database and the set organism being this organism. The diameter of the area of the genome is 35kb (kilobases) long with the center being the proline reductase operon.
APPENDIX: Figures 4.2.2

*Clostridium sticklandii DSM 519*

1: serine-tRNA

The genome context of *C. sticklandii DSM 519* was collected through the Biocyc database and the set organism being this organism. The diameter of the area of the genome is 35kb (kilobases) long with the center being the proline reductase operon.
The genome context of *Clostridium botulinum* b1 strain Okra was collected through the Biocyc database and the set organism being this organism. The diameter of the area of the genome is 35kb (kilobases) long with the center being the proline reductase operon.
**APPENDIX: Figures 4.2.4**

*Lactobacillus antri DSM 16041*

The genome context of *L. antri DSM 16041* was collected through the Biocyc database and the set organism being this organism. The diameter of the area of the genome is 35kb (kilobases) long with the center being the proline reductase operon.
The genome context of *M. micronuciformis F0359* was collected through the Biocyc database and the set organism being this organism. The diameter of the area of the genome is 35kb (kilobases) long with the center being the proline reductase operon.
APPENDEX: Tables 1.1

Table 1.1: The phylogenetic distribution shows a large degree of variation with the glycine reductase systems in both the GrdA and GrdB proteins as a large amount of non-firmicutes are present. The proline reductase proteins meanwhile have much less variation and are almost entirely from the firmicutes. Though the PrdB protein seems to have constituents while PrdA does not.

<table>
<thead>
<tr>
<th>Phylum</th>
<th>GrdA</th>
<th>GrdB</th>
<th>PrdA</th>
<th>PrdB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firmicutes</td>
<td>61.16% (137)</td>
<td>64.16% (111)</td>
<td>100% (60)</td>
<td>85% (51)</td>
</tr>
<tr>
<td>Proteobacteria</td>
<td>10.95% (22)</td>
<td>15.03% (26)</td>
<td>-</td>
<td>1.67% (1)</td>
</tr>
<tr>
<td>Synergistetes</td>
<td>9.95% (20)</td>
<td>6.36% (11)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Tenericutes</td>
<td>1.0% (2)</td>
<td>0.58% (1)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Spirochetes</td>
<td>8.46% (17)</td>
<td>12.14% (21)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>0.5% (1)</td>
<td>0.58% (1)</td>
<td>-</td>
<td>8.33% (5)</td>
</tr>
<tr>
<td>Fusobacteria</td>
<td>1.0% (2)</td>
<td>0.58% (1)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Chloroflexi</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.33% (2)</td>
</tr>
<tr>
<td>Unclassified</td>
<td>-</td>
<td>1.16% (2)</td>
<td>-</td>
<td>1.67% (1)</td>
</tr>
<tr>
<td>Total</td>
<td>201</td>
<td>173</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

The phylogenetic breakdown was generated from the phylogenic trees of Figures 1.1, 1.2, 1.3, and 1.4. This is a count of the total number of constituent organisms of each phylum were found to have any of the four genes of interest.
APPENDIX: Tables 1.2

Table 1.2: The distribution of the class of each bacterial entry for each protein with respect to the whole set as well as a comparison to that class’s phyla shows that mostly clostridia inhabit all four proteins. The non-firmicutes the glycine reductase proteins are mostly centered in the proteobacteria and also have divergence in there as well.

<table>
<thead>
<tr>
<th>Class</th>
<th>GrdA  (against phyla)</th>
<th>GrdB  (against phyla)</th>
<th>PrdA  (against phyla)</th>
<th>PrdB  (against phyla)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clostridia</td>
<td>58.21% (117)</td>
<td>85.40% (117/137)</td>
<td>85.59% (95/111)</td>
<td>81.67% (49)</td>
</tr>
<tr>
<td></td>
<td>85.40% (117/137)</td>
<td>85.40% (117/137)</td>
<td>85.59% (95/111)</td>
<td>81.67% (49/60)</td>
</tr>
<tr>
<td>Bacilli</td>
<td>9.95% (20)</td>
<td>14.6% (20/137)</td>
<td>8.09% (14)</td>
<td>12.61% (14/111)</td>
</tr>
<tr>
<td>Negativicutes</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.67% (4)</td>
</tr>
<tr>
<td>Erysipelotrichia</td>
<td>-</td>
<td>1.16% (2)</td>
<td>1.80% (2/111)</td>
<td>-</td>
</tr>
<tr>
<td>Alpha Proteobacteria</td>
<td>-</td>
<td>1.16% (2)</td>
<td>7.69% (2/26)</td>
<td>-</td>
</tr>
<tr>
<td>Gamma Proteobacteria</td>
<td>6.47% (13)</td>
<td>59.09% (13/22)</td>
<td>5.78% (10)</td>
<td>38.46% (10/26)</td>
</tr>
<tr>
<td>Delta Proteobacteria</td>
<td>4.48% (9)</td>
<td>40.91% (9/22)</td>
<td>8.09% (14)</td>
<td>53.85% (14/26)</td>
</tr>
<tr>
<td>Synergistia</td>
<td>9.95% (20)</td>
<td>6.36% (11)</td>
<td>100%</td>
<td>-</td>
</tr>
<tr>
<td>Spirchetia</td>
<td>8.46% (17)</td>
<td>12.14% (21)</td>
<td>100%</td>
<td>-</td>
</tr>
<tr>
<td>Fusobacteria</td>
<td>1.0% (2)</td>
<td>0.58% (1)</td>
<td>100%</td>
<td>-</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>0.5% (1)</td>
<td>0.58% (1)</td>
<td>100%</td>
<td>-</td>
</tr>
<tr>
<td>Mollicutes</td>
<td>1.0% (2)</td>
<td>100%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ktedonbacteria</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.67% (1)</td>
</tr>
<tr>
<td>Caldilineae</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.67% (1)</td>
</tr>
<tr>
<td>Unclassified</td>
<td>-</td>
<td>1.16% (2)</td>
<td>100%</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>201</td>
<td>173</td>
<td>60</td>
<td>60</td>
</tr>
</tbody>
</table>

This phylogenetic distribution was collected by further breaking down the data from Table 1.1 into the amount of organisms present are of a specific class.
WORKS CITED


26. Rambaut, A.(2006). FigTree (v1.4) [Software].


29. de Vladder H. P., *Amino acid fermentation at the origin of the genetic code*. Biology Direct 2012, **7**:6
