Published in final edited form as:

J Mol Graph Model. 2017 January ; 71: 135–153. doi:10.1016/j.jmglm.2016.10.011.

Structure-based Analysis of Bacilli and Plasmid Dihydrofolate Reductase Evolution

Mona Alotaibi^{a,b}, Ben Delos Reyes^a, Tin Le^a, Phuong Luong^a, Faramarz Valafar^c, Robert P. Metzger^a, Gary B. Fogel^d, and David Hecht^{a,e,*}

^aDepartment of Chemistry and Biochemistry, San Diego State University, San Diego, CA 92182-1030, USA

^bKing Saud University, P.O. box 245714, Riyadh 11312

^cBioinformatics and Medical Informatics Research Center, San Diego State University, San Diego, CA 92182-1030

^dNatural Selection, Inc., 6480 Weathers Place, Suite 350, San Diego, CA 92121 USA, tel: (858) 455-6449, fax: (858) 455-1560

^eDepartment of Chemistry, Southwestern College, 900 Otay Lakes Rd., Chula Vista, CA 91910 USA, tel: (619) 421-6700 x5461

Abstract

Dihydrofolate reductase (DHFR), a key enzyme in tetrahydrofolate-mediated biosynthetic pathways, has a structural motif known to be highly conserved over a wide range of organisms. Given its critical role in purine and amino acid synthesis, DHFR is a well-established therapeutic target for treating a wide range of prokaryotic and eukaryotic infections as well as certain types of cancer. Here we present a structural-based computer analysis of bacterial (Bacilli) and plasmid DHFR evolution. We generated a structure-based sequence alignment using 7 wild-type DHFR x-ray crystal structures obtained from the RCSB Protein Data Bank and 350 chromosomal and plasmid homology models we generated from sequences obtained from the NCBI Protein Database. We used these alignments to compare active site and non-active site conservation in terms of amino acid residues, secondary structure and amino acid residue class. With respect to amino acid sequences and residue classes, active-site positions in both plasmid and chromosomal DHFR are significantly more conserved than non-active site positions. Secondary structure conservation was similar for active site and non-active site positions. Plasmid-encoded DHFR proteins have greater degree of sequence and residue class conservation, particularly in sequence positions associated with a network of concerted protein motions, than chromosomally encoded DHFR proteins. These structure-based were used to build DHFR specific phylogenetic trees from which evidence for horizontal gene transfer was identified.

*Corresponding Author.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Graphical abstract

Keywords

Dihydrofolate reductase; DHFR; structure-based alignment; phylogeny; evolution; plasmid evolution; horizontal gene transfer; network of coupled protein motion

1. Introduction

Both prokaryotes and eukaryotes require reduced folate cofactors for the biosynthesis of a broad range of cellular components [1]. More specifically, tetrahydrofolate cofactors are essential for the synthesis of purines, thymidine and glycine [2]. While most bacteria and plants produce folate cofactors through *de novo* biosynthesis, a variety of mammals, and some prokaryotes lack this ability and must convert dietary folates to reduced folates through the use of their enzymatic salvage pathways for survival. Most bacterial, plant, and animal species possess a pathway (Figure 1), by which tetrahydrofolate which has been oxidized to dihydrofolate during its use in biosynthetic processes, is recycled back to tetrahydrofolate through use of NADPH and the activity of the enzyme dihydrofolate reductase (DHFR).

Folate effects the efficiency of DNA replication, repair, and methylation. Rapidly proliferating cells including leukocytes, and enterocytes require high amounts of folate. Epidemiological studies have indicated that folate deficiency is often associated with increased risk of breast cancer and birth defects [3]. Indeed, low folate homeostasis may induce hypomethylation of DNA which may promote development of cancer [4].

There is a stark difference in tetrahydrofolate salvage pathways between bacteria and humans which makes the bacterial pathways ideal targets for antibacterial agents [5]. The development of bacterial DHFR inhibitors that will *not* affect human DHFR continues to be an area of great interest for treatment of bacterial infections [2, 6, 7]. Such folate analogs, (e.g., methotrexate and trimethoprim), are known to inhibit DHFR, resulting in the death of cancerous human cells and bacterial cells, respectively. These agents bind tightly to the catalytic site of the enzyme in a region known to form a hydrophobic dihydrofolate-binding site [8].

However, the ability of bacteria to adapt rapidly to new conditions has led to the development of resistance to antifolate agents. The most common resistance results from mutations leading to the synthesis of an altered form of DHFR with reduced affinity for an inhibitor or inhibitors [9]. Acquired resistance is also associated with mobile elements temporarily located on the chromosome by transposon movement, referred to as plasmid-encoded DHFR [10–14].

Plasmids containing genes encoding drug resistant DHFR variants were first discovered in *Escherichia coli* and *Klebsiella aerogenes* by Fleming [15]. To date, more than 30 different plasmid-encoded DHFR genes have been identified. Plasmid DHFR genes encode two distinct DHFR types, I and II, and have therefore been divided into two families, *dfrA* and *dfrB* [16]. The majority of the *dfrA* family is carried as a gene cassette encoding DHFR variants ranging between 157 to 187 amino acid residues. There is a high degree of conservation within the N-terminus, with 12 residues strictly conserved across all enzymes [17]. The second family (*dfrB1*, *dfrB2*, *dfrB3*, *dfrB4*, *dfrB5* and *dfrB6*) also forms part of a gene cassette, and encodes the enzymes DHFR IIa, IIb and IIc, respectively. These enzymes have significantly shorter amino acid sequences of 78 residues which is roughly 42–50% of the amino acid residues of the enzymes encoded by the *dfrA* family. Members of the *dfrB* family are completely nonresponsive to trimethoprim, methotrexate, and other folate analogs and appear to have a distinctive evolutionary origin [18, 19].

Phylogenetic analyses of DHFR [20–23] as well as studies and simulations of evolutionary processes [14, 24–30] have been performed to study the rapid development of resistance to antifolate agents in response to the excessive use of antibiotics [31]. Our efforts focus on predictive models of the evolution of drug resistance in bacterial pathogens having the goal of developing novel therapeutics targeting drug resistant strains *in anticipation* of their development in the wild [26, 27, 32]. This proactive approach to drug discovery requires a thorough understanding of the evolutionary processes specific to the gene of interest. To this end we have performed a structural based analysis of sequence and structure variation in *Bacilli* DHFR. *Bacilli* are a prokaryotic class belonging to the taxonomic phylum *Firmicutes* [33–35]. Members of this class are generally Gram-positive bacteria. The *Bacilli* class consists of two orders *Bacillales* and *Lactobacillales* (Figure 2).

Here we present an analysis of DHFR evolution based on a structure-based alignment of 273 *Bacilli* species and 77 *dfrA* family DHFR protein sequences of which five come from *Bacilli* species (*Planococcus citreus*, *Listeria monocytogenes*, *Staphylococcus aureus* (2), *Enterococcus faecium*). The remainder of the plasmid sequences include non-*Bacilli* sequences of which *E. coli* (30) and *Klebsiella pneumoniae* (8) were the most common. Using this structure-based alignment we evaluated the degree of site-specific variability/conservation of amino acid sequence, secondary structure, and residue class. Using the structure-based alignment, we generated phylogenetic trees and explored the evolution of *Bacilli* DHFR in terms of sequence and structure conservation between chromosomal and plasmid DHFR, *Bacillales* and *Lactobacillales*, in order to attempt to correlate the observed differences to their environment.

2. Materials and Methods

2.1 Bacilli and Plasmid DHFR sequences

Following a comprehensive search of the NCBI Protein database, 15,542 DHFR sequences were downloaded in FASTA format, and imported into BioEdit [36] where they were sorted based on their species and identification as either a mutant sequence or wild-type (wt) sequence. These were truncated to include only 273 unique *Bacilli* wt-chromosomal and 77 plasmid-encoded *dfrA* gene family DHFR sequences. A “canonical” phylogenetic tree of the *Bacilli* genera was generated from publically available tools at www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwcmt.cgi (Figure 2).

2.2 Homology Modeling

29 prokaryotic and plasmid DHFR x-ray crystal structures were downloaded from the Protein Data Bank (PDB). Using DeepView (Swiss-PDBviewer [37]) each structure was first truncated to include only a monomer of DHFR and then imported into MOE (www.chemcomp.com). The LigX function in MOE was used to prepare each structure for homology modeling through optimization of the protonation as well as removal of any water and other non-protein molecules. Pairwise sequence alignments were then performed matching each of the 350 DHFR sequences to the 29 DHFR x-ray structures to select the crystal structure for each sequence that had the largest value of pairwise percentage identity (indicating the greatest sequence homology) for use as a structural template for *each* sequence. Homology models were then generated for each sequence using the Amber 12:EHT potential [38] and default MOE settings. RMSD values for each homology model from the corresponding template x-ray crystal structure ranged between 0.27 and 1.94 Å, with the average RMSD ($\pm$ confidence interval, 95%) being 0.75 ± 0.09 Å. Finally Ramachandran plots were generated to identify models having unusual or unreasonable geometries (none were identified).

2.3 Structure-based Sequence Alignments

Seven *Bacilli* DHFR x-ray crystal structures (Table 1) were imported into MOE and structurally aligned with the 350 homology models using MOE’s “Superpose” function. For purposes of this paper, this structure-based alignment will be called the Total DHFR alignment (Figure 3a). This Total DHFR alignment was truncated into alignments based on class and genus. (Figures 4 and 5). A script written in support vector language (SVL) was run on the aligned sequences in MOE to generate masks labeling every residue in each sequence for the type of secondary structure present: alpha helix (H), beta-sheet (S), turn (T) and unstructured (–). Another SVL script was run to generate masks labeling every residue in each sequence for the amino acid residue class: acid (A), base (B), non-aromatic hydrophobic (N), non-charged polar (Q), aromatic (R), proline (P), glycine (G).

2.4 Active Site & Non-Active Site Analysis

The Total DHFR alignment was divided into two new structure-based alignments, one that included residues 6 Å from the NADPH and substrate binding site residues (active site alignment) and one that included residues 6 Å from the NADPH and substrate binding site

residues (non-active site alignment). Three representative active-site alignments are presented in Figures 6–8. Complete total sequence and active site alignments are included in the supplemental material.

The active site sequence alignment was exported as a FASTA file and uploaded into WebLogo [39,40] in order to generate a graphical analysis of sequence conservation on a position-by-position basis (Figure 9). These three alignments Total, Active-Site and Non-Active Site were further truncated into class and genus for further analyses.

2.5 Determination of Sequence and Structural Conservation

All structural alignments were analyzed in attempt to quantify the degree of sequence and structural conservation. BioEdit was used to calculate the frequency of finding a particular amino acid at each position in the alignments. Positions with gaps were assigned null values. The SVL codes used for generating the secondary structure and residue class masks also generated site-specific frequencies of finding a residue in an α -helix, a β -sheet, or a turn conformation, as well as the frequency of finding a particular residue in each of the seven residue class types (acid, base, non-aromatic hydrophobic, non-charged polar, aromatic, proline, and glycine).

The *average* of the average frequency *per position* was then calculated for amino acid occurrence, secondary structure conformation, and residue class. These values provide a metric for the relative degree of conservation or variability at each site in the DHFR alignment. For purposes of this paper these values will be called Amino Acid Conservation, Secondary Structure Conservation, and Residue Class Conservation, respectively.

The Amino Acid Conservation metric ranges from a value of 4.5%, in which each of the DHFR sequences has a different amino acid at each position, to a value of 100.0%, which represents complete conservation. The Secondary Structure Conservation values range from 25.0% to 100.0% and the Residue Class Conservation values range from 14.3% to 100.0% respectively. T-tests ($\alpha=0.05$) were performed to compare the relative degree of active-site, nonactive site and overall sequence site-specific conservation. These analyses were extended to evaluate the degree of conservation/variability between the chromosomal and plasmid DHFR sequences, as well as the degree of conservation/variability between different families and genera belonging to the prokaryotic class *Bacilli* (Figure 3; Tables 2 and 3).

2.6 DHFR Phylogenetics

Phylogenetic trees were generated for the Total, *Lactobacillales*, and *Bacillales* alignments with MEGA6 [41] using the Neighbor-Joining method with the following parameters: Jones-Taylor-Thornton (JTT) substitution model; uniform rate of site-specific mutation and complete deletion of missing data and gaps. Each tree was evaluated using the bootstrap method and interior-branch test, with 100 replications. As ten of the DHFR sequences were truncated, they were not able to be included in these analyses (Figures 10–12).

3. Results and Discussion

3.1 Comparison of Sequence and Structure Conservation: Chromosomal Bacilli vs Plasmid-Encoded DHFR

For the total DHFR structure-based alignment, composed of both chromosomal and plasmid encoded DHFR sequences, the Amino Acid Conservation was $9.7\% \pm 0.6\%$ (Figure 3a). The calculated value of 9.7% is indicative of a relative lack of site-specific sequence conservation. The Secondary Structure Conservation was determined to be, $36.9\% \pm 1.0\%$ and the Residue Class Conservation, $21.2\% \pm 0.9\%$ (Figure 3a). These results suggest that although a significant degree of site-specific sequence variability exists across the 350 DHFR sequences, there is less tolerance for mutations that result in changes in secondary structure or residue class relative to amino acid replacement.

A similar analysis was performed for the active site residues and the non-active site residues (Figure 3a). Not surprisingly, the Amino Acid Conservation for the active site residues, $15.5\% \pm 1.7\%$, was significantly higher than the Amino Acid Conservation for the entire structure-based alignment, $9.7\% \pm 0.6\%$. The value for Residue Class Conservation for the active site residues, $28.8\% \pm 2.3\%$, was also significantly higher. However, there was no significant difference between the values of Secondary Structure Conservation for the active site residues compared to the residues of the entire alignment.

For the non-active site residues, the values for Amino Acid Conservation, $7.4\% \pm 0.3\%$, and Residue Class Conservation, $18.3 \pm 0.8\%$, were both significantly lower than those of the active site residues. There was no significant difference in the values of Secondary Structure Conservation between non-active site and active site residues. The same procedure used to analyze the separate structure-based alignments of the chromosomal and plasmid DHFR sequences (Figure 3.b and 3.c) showed that the degree of Amino Acid Conservation and Residue Class Conservation was higher in the active site positions than in the non-active site positions.

When evaluating differences in site-specific sequence and structure variability between plasmid-encoded DHFR proteins and chromosomal DHFR proteins, slight, statistically significant increases in the Amino Acid Conservation values of plasmid-encoded DHFR sequences ($17.5\% \pm 1.3\%$) compared to chromosomal DHFR sequences ($10.8\% \pm 0.6\%$) and Residue Class Conservation for plasmid-encoded DHFR ($30.5\% \pm 1.9\%$) compared to chromosomal DHFR ($22.5\% \pm 1.0\%$) were observed. Interestingly, the values of Secondary Structure Conservation values for chromosomal encoded DHFR sequences were significantly greater than the plasmid encoded sequences ($43.5\% \pm 1.2\%$ vs. $38.5\% \pm 1.6\%$ respectively).

These observations are surprising, and contrary to what we had expected, (e.g., greater variability in the plasmid sequences). However, it is important to note that approximately half (38) of the 77 plasmid sequences originate from two organisms, *E. coli* and *Klebsiella pneumoniae*. As most of the plasmid species come from human and animal pathogens that have developed resistance to antifolate therapeutics, it is possible to hypothesize that adaptation to this selection pressure contributes to the higher degree of sequence

conservation. Another possible hypothesis may be that amino acid substitutions are less tolerated in plasmid DHFR proteins because of their effect on concerted protein motions associated with catalysis. There are observed differences in the overall protein folds and secondary and tertiary structures between the plasmid and chromosomal proteins and it may be that chromosomal DHFR proteins have greater structural flexibility to absorb these amino acid substitutions (Figure 3). We discuss conservation of residues involved with protein motion in more detail in Section 3.5.

WebLogo was used to perform a graphical analysis of position specific sequence conservation for the active site residues of the Total DHFR alignment (Figure 9). Key positions containing glycine, tryptophan, proline and aspartic or glutamic acid were highly conserved, even in the plasmid sequences. Although there was variation in the other positions with regards to amino acid type, the residue class was highly conserved throughout.

3.2 Comparison of Sequence and Structure Conservation: Bacillales vs. Lactobacillales

In attempt to evaluate the differences in site-specific sequence and structure variability between the two orders *Bacillales* and *Lactobacillales*, the values of Amino Acid Conservation, Secondary Structure Conservation and Residue Class Conservation were calculated and compared (Table 2). There was a slight increase in the degree of Amino Acid Conservation when separating out the plasmid encoded DHFR sequences in both orders *Bacillales* and *Lactobacillales*, versus the Total Alignment. Likewise, there was also a slight increase in *Bacillales* and *Lactobacillales* Residue Class Conservation as compared to the overall, *Bacilli* Residue Class Conservation. These results are consistent with divergent paths of DHFR evolution in the two orders.

Interestingly, the overall Secondary Structure Conservation values varied significantly for the *Lactobacillales* and *Bacillales*, with even greater differences in these values for active site residues. There were no significant differences for the Amino Acid Conservation and Residue Class Conservation values for the *Bacillales* and *Lactobacillales* alignments. These results are suggestive of the presence of a selection pressure(s) against amino acid replacements that result in changes in *Bacillales* DHFR secondary structure.

3.3 DHFR Phylogeneny

A phylogenetic tree of the chromosomal and plasmid-encoded DHFR sequences was generated using the Total DHFR alignment (Figure 10). Although the separation of the two *Bacilli* orders, *Bacillales* and the *Lactobacillales*, was not as apparent as observed in the canonical taxonomic tree (Figure 2), it remains relatively traceable. The majority of plasmid-encoded DHFR sequences (53) clustered close together as did the majority of the *Lactobacillales* DHFR sequences (with the exception of one branch of 22 species). The majority of *Bacillales* sequences also clustered together, with approximately 23 sequences being grouped with the plasmids. 11 plasmid DHFR sequences of types 12, 21, 22, and 23 create a coherent well separated clade. (Figure 11) Type 1, 7, and 17 plasmid DHFR sequences, which for the most part originate from *E. coli*, cluster on separate clades. It is

worth noting that the majority of the plasmid-encoded DHFR sequences, including the 32 plasmids from *E. coli*, originated from prokaryotic classes other than class *Bacilli*.

Phylogenetic trees were then generated for each order (Figure 12). In this analysis, the *Lactobacillales* tree appears to be closer to the canonical taxonomic tree (Figure 2) with families clustered in more clearly defined clades (Figure 12a) whereas the families in the *Bacillales* tree do not occur in well-defined groups (Figure 12b). For example, the two genera *Leuconostoc* and *Weissella* of the *Leuconostocaceae* family (order *Lactobacillales*) each form a monophyletic group. As would be expected, the conservation level of both of these genres is relatively high compared to that of the other families (Table 3). Likewise, 56 species (out of 60 total) from the *Streptococcus* family, cluster into a single clade. Not surprisingly, the *Lactobacillus* and *Enterococcus* families, each having relatively low conservation values, are scattered in several clades within the phylogenetic tree (Figure 12a). Similarly, the members of the families *Bacillaceae* and *Paenibacillaceae*, both of which have relatively low conservation were dispersed throughout the *Bacillales* tree. Finally, the *Geobacillus* sequences, having the largest values of DHFR conservation, show an interesting placement where all of 7 family members appear in a well separated, monophyletic clade.

3.4 Evidence for Horizontal Gene Transfer in Bacilli DHFR Evolution

Horizontal Gene Transfer (HGT) is thought to play a significant role in the evolution of single-celled organisms and to be one of the major mechanisms for the spread of antibiotic resistance in bacteria [42–44]. Typically HGT has been identified through the use of sequence composition based (parametric) as well as phylogenetic based methods [44–46]. Careful analysis of the phylogenetic trees derived from the total DHFR structure-based alignment (Figures 11 and 12) shows some evidence of HGT in *Bacilli* DHFR evolution.

For example DHFR type 1 *Planococcus citerius* clusters with chromosomal *Planococcus citerius* DHFR yet both are clustered relatively far from the other three *Planococcus* species: *P. antarcticus*, *P. donghaensis* and *P. halocryophilis*. (Figure 10) Similarly, type 1L *Listeria monocytogenes* clusters close to chromosomal *Listeria monocytogenes* yet relatively far from the other six *Listeria* species: *L. grayi*, *L. innocua*, *L. ivanovii*, *L. marthii*, *L. seeligeri* and *L. welshimeri*. *L. monocytogenes* also clusters with *Staphylococcus haemolyticus*, a major component of the flora of the epidermis and associated with various skin infections.

An interesting example of potential cross-species HGT is the clustering of type K *Staphylococcus aureus* (plasmid) with two *Enterococcus faecium* sequences. Both species are associated with food poisoning and are found in the intestine where it possible for HGT to occur. A second similar example is the clustering of *Enterococcus faecium* B1249 with the plasmid type 1 *S. aureus* and sixteen *Staphylococcus* sequences, including *S. aureus*.

Other potential examples of HGT identified in our analysis include the clustering of the type 13 *Enterococcus faecium* plasmids with four species of *Brevibacillus* (often found in air, soil and decaying material) and type 9 and type 24 *E. coli* plasmids clustering with *Bacillus thuringiensis*. *B. thuringiensis* is found in soil and is often used in farming as a natural

pesticide. The remaining (non-*Bacilli*) plasmid sequences all cluster together with *Enterococcus faecalis* and *Macroccoccus caseolyticus* (Figure 11).

3.5 Patterns of Amino Acid Conservation in a Network of Coupled Protein Motion

Recent studies have shown that a network of coupled protein motions is a critical component of dihydrofolate catalysis [47–50]. This network consists of a set of a highly conserved non-active site residues that facilitate hydride transfer. Not surprisingly, many of these residues are found in proximity to the NADPH binding site. (Figure 13).

Using the x-ray structure of *E. coli* (1RX2.pdb) as a template for identifying the residue positions in the structure-based DHFR alignment we performed an analysis of the percent amino acid conservation at key positions in this network. (Table 4). As expected there was a high degree of conservation across all species, including the plasmids, for 7 of the 21 positions analyzed consistent with the literature [47–50]. These positions include: Ile14, Gly15, Pro21, Phe31, Gly43, Gly121 and Asp122. In general, the plasmid sequences seemed to have a much greater degree of conservation than the chromosomal sequences (see bold and italicized values in Table 5). As discussed in section 3.1, it is possible that the observed differences in the secondary and tertiary structures of the plasmid versus chromosomal DHFR proteins makes them less tolerant to amino acid substitutions.

Another interesting observation is that most of the highly conserved positions in plasmid sequences differ from both *E. coli* as well as from chromosomal *Bacilli* sequences. These include: M16N; E17G; N18P; A19D; M20I; V40L; I41L; M42V; I60A; I61V; L62V; S63T; T123V. (First letter is the *E. coli* residue, last letter is the plasmid residue). As the vast majority of plasmid sequences are from *E. coli* and closely related species, this is another surprising observation and again may originate in the differences in tertiary structures of plasmid vs chromosomal DHFR proteins.

With several exceptions there was very little difference in the patterns of amino acid conservation between *Bacillales* and *Lactobacilles* sequences. These exceptions include positions 18, 19, 20, 41, 60 and 61. These observations will be the subject of future studies.

3.6 Impact of Environment on DHFR Conservation and Variability

Most *Bacillales* are rod-shaped organisms which form dormant and resistant endospores in challenging environments, including extreme temperatures, high or low pH, lack of nutrients, or aridity [51, 52]. When the environmental conditions become suitable, growth is reactivated through the process of germination [51]. The majority of *Lactobacillales* species are non-spore forming facultative anaerobic bacteria [53]. The lactic acid produced as the end product results in increased acidity in their environment, giving them an advantage over their competitors.

DHFR proteins from extremophiles have been used to study prokaryotic adaptation to ecosystems with extreme temperatures [54–56], pH [57], salt concentrations [58] and variations in pressures [59, 60]. In general proteins from extremophiles tend to have more stable protein folds than proteins from organisms living in more moderate environments (mesophiles) [61–63]. Not surprisingly, proteins from thermophiles have different patterns

of amino acid expression than those coming from mesophiles [60,64–67]. A positive correlation has recently been identified between the relative frequencies of glutamic acid (E), isoleucine (I), lysine (K), tyrosine (Y) and valine (V) (taken together as “EIKTV”) in proteins found in thermophiles as compared to those found in mesophiles [62]. Conversely, thermophiles tend to have lower levels of alanine (A), cysteine (C), histidine (H), glutamine (Q) and threonine (T) (taken together as “ACHQT”) than mesophiles [61].

Using the structure-based alignment, we performed a similar analysis for *Bacilli* DHFR and found that there is a statistically significant difference between *Bacillales* vs. *Lactobacillales* species for both EIKYV and ACHQT (t-test, $p < 0.05$). (Table 5) DHFR from *Bacillales* species, which are often found in challenging environments, have significantly higher frequencies of EIKYV and significantly lower levels of ACHQT relative to DHFR from *Lactobacillales* species. An analysis comparing each of the genera resulted in no significant differences within each order.

Although it has been difficult to correlate sequence differences in DHFR with environmental conditions, here we make an attempt to do so, albeit recognizing the limitations of the analysis due to sparseness of information about the environmental variables for each species [68]. While *Bacilli* species inhabit a wide range of environments, our structure-based analysis, with several important exceptions, has revealed only slight differences in DHFR sequence/structure variability. These exceptions include 5 *Bacilli* genera: *Geobacillus*, found in soil; *Weissella*, found in meat, vegetables, cheese and soil; *Leuconostoc*, found in meat, milk, and green vegetables; *Listeria*, inhabiting wide range of environments; and *Staphylococcus*, found in skin, skin glands, mucous membranes as well as the intestines of humans, animals or birds. All of these genera have significantly greater conservation values than the other families. It is important to note that the observed increase in the conservation of sequence and structure in these five groups could be attributed to the small sampling size as only seven DHFR sequences were analysed in case of *Geobacillus*, nine in case of *Leuconostoc*, seven in case of *Listeria* and six in case of *Weissella*.

Geobacillus species, usually found in soil, had the highest conservation levels of all families: $98.1\% \pm 0.8\%$ for active site Amino Acid Conservation and 100% conservation of active site Secondary Structure. (Table 3). Although *Bacillus* and *Paenibacillus* species are also found in soil, these families had some of the lowest conservation values. One important difference between *Bacillus* and *Paenibacillus* species compared to *Geobacillus* species is that *Geobacillus* species require temperatures between 35–75°C with the optimal temperature range being 55–65°C for growth. Species that grow at high temperatures often have adaptations in DHFR sequences that allow molecular stability and structural flexibility [51,64–71]. These high values of Amino Acid and Residue Class conservation, $82.6\% \pm 2.0\%$, and $90.8\% \pm 1.6\%$ respectively, suggest the presence of a strong selection pressure on *Geobacillus* DHFR that may be related to temperature. This hypothesis can be tested through thermal stability and function studies. Interestingly, the DHFR proteins from *Bacillus* species were 35% EIKYV as compared to 33% for *Geobacillus* and 30% for *Paenibacillus* species. The higher level of EIKYV expression for *Bacillus* DHFR may reflect the greater diversity of environments of its species.

Although species in the genus *Weissella* are found in a wide range of variable habitats ranging from meat (*W. weissellathailandensis* and *W. weissellaaramesenteroides*), vegetables (*W. weissellacibaria* and *W. weissellakoreensis*) and the human intestinal microbiome (*W. confuse*), their DHFR sequences and structures are highly conserved [72]. Like *Weissella*, members of the genus *Leuconostoc*, also inhabit high nutrient environments such as milk, meat, green vegetables, and fermented drinks and have relatively high levels of conservation [73]. In fact, the conservation values of amino acid residue, class and secondary structure for species from these two genera appear relatively close, with no significant differences detected. This may be related to similarities in their phenotypic characteristics requiring the use of molecular methods to distinguish between species from these two closely related genera [72]. For both genera there does seem to be a strong selection pressure present. Although the species inhabit a wide range of environments, all seem to be nutrient rich.

Listeria and *Staphylococcus* are second to *Geobacillus* in terms of conservation values. Staphylococci species are known for containing several significant pathogens [74, 75]. They are usually found in skin, skin glands and mucous membranes of animals, including humans [76, 77]. *Listeria* species inhabit a wider range of environments (e.g., soil, animal feed, vegetation, sewage, water, poultry, feces) and can infect humans and animals through food contamination [78]. Like species from *Weissella* and *Leuconostoc*, these species living in high nutrient environments have a strong selection pressure on amino acid substitutions resulting in changes in DHFR sequence and structure [79–81].

The conservation values for genus *Enterococcus* DHFR, known for tolerating extreme environmental conditions [82], are significantly similar to the conservation values of genus *Streptococcus* whose species inhabit relatively moderate environments [83, 84]. Many *Streptococcus* are considered part of the human microbiome as they usually exist in the intestine, skin, mouth, and respiratory tract. However, under suitable conditions, *Streptococci* can cause disease [83].

Lactobacillus species grow best in acidic media with a temperature range of 30–40°C [85]. In addition to carbohydrates, *Lactobacillus* require amino acids, nucleotides and vitamins as energy sources [86], whereas most *Enterococcus* species are part of the intestinal flora of mammals [87]. One of the distinguishing characteristics of the majority of enterococcus is their ability to survive in a wide range of extreme ecological conditions including high alkalinity (pH ~9.6), temperature (10–45°C), as well as high salt concentrations [82] which helps them to compete more effectively for resources. Additional studies are required to study the correlations (if any) between the observed variation in the degree of *Bacilli* DHFR sequence and structure conservation and environmental conditions.

4. Conclusions

Bacteria continue to evolve resistance against many antibiotics including those targeting DHFR [88]. This resistance is acquired either through mutations in chromosomal DNA or through transfection via plasmids. In order to better understand the future evolutionary progression of antifolate resistance, it is first necessary to understand historical natural evolution without the additional pressures of drug affinities. To this end we generated a

structure-based alignment of *Bacilli* and plasmid encoded DHFR sequence in order to evaluate the degree of site-specific conservation/variability in terms of both sequence and structure.

As expected, the active-sites in both plasmid and chromosomal DHFR were significantly more conserved with respect to amino acid and residue class but not with regards to secondary structure. These observations suggest that there is less tolerance for changes in secondary structure or residue class than the amino acid replacement itself. We also found high levels of conservation in key residues involved in the network of coupled protein motions involved with the catalysis of dihydrofolate into tetrahydrofolate.

Interestingly we found that plasmid-encoded DHFR proteins are less tolerant of amino acid substitutions in the active site than chromosomal-encoded DHFR proteins and have less tolerance of amino acid substitutions residues involved in the network of coupled protein motions. We hypothesize that the greater degree of sequence conservation in plasmid DHFR may arise from differences in the tertiary structures of plasmid versus chromosomal proteins. This will be explored in future studies.

Phylogenetic trees generated from *Lactobacillales* DHFR sequences were closer to the canonical taxonomic tree than trees generated from *Bacillales* DHFR sequences. The differences seemed to correlate with the relative degree of DHFR sequence and structure conservation observed among the families and genera. These analyses provided evidence for HGT including intra-species exchange of DHFR. Although these analyses have provided some new insights into the evolution of DHFR in bacteria and plasmids, many new questions arose, particularly with respect to the environment specific selection pressures that are present.

In future studies we plan to model potential amino acid replacements that could lead to resistance against anti-folates with the goal of using the predicted new active sites to propose second generation antifolate analogs [26, 27, 32]. In these studies, we will use the structure-based alignment presented here in order to generate a DHFR specific amino acid substitution matrix, similar to the PAM matrix. This DHFR specific amino acid substitution matrix will then be used to generate a population of variant DHFR sequences. The highly conserved active site positions containing glycine, tryptophan, proline and aspartic or glutamic acid identified from the structure-based alignment presented here will be of great importance to these models. A selection pressure based on predicted binding affinities (from *in silico* modeling) of anti-folate therapeutics will then be applied to select potential amino acid substitutions that could lead to resistance. We are currently performing experimental validation of several variant *Plasmodium falciparum* DHFR sequences and we plan to extend these studies to other pathogens including the various plasmids, *Staphylococcus aureus*, *Bacillus anthracis* and *Streptococcus pneumoniae*.

Acknowledgments

This research was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number SC3GM100791. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to thank Rocky Goldsmith from the Chemical Computing Group, who provided a SVL script for automating the statistical

calculations of the structure-based alignment. Finally we would also like to thank the reviewers for their insightful comments and suggestions.

References

1. Blakley, RL. Dihydrofolate reductases. In: Blakley, RL.; Benkovic, SJ., editors. *Folates and pterins: Chemistry and biochemistry of folates*. Vol. 1. New York, NY: John Wiley & Sons, Ltd; 1984. p. 191-244.
2. Hawser S, Lociuro S, Islam K. Dihydrofolate reductase inhibitors as antibacterial agents. *Biochemical Pharmacology*. 2006; 71(7):941–948. DOI: 10.1016/j.bcp.2005.10.052 [PubMed: 16359642]
3. Terry P, Jain M, Miller AB, Howe GR, Rohan TE. Dietary intake of folic acid and colorectal cancer risk in a cohort of women. *International Journal of Cancer*. 2002; 97(6):864–867. DOI: 10.1002/ijc.10138 [PubMed: 11857369]
4. Rossi M, Amaretti A, Raimondi S. Folate production by probiotic bacteria. *Nutrients*. 2011; 3(1): 118–134. DOI: 10.3390/nu3010118 [PubMed: 22254078]
5. Sefton AM. Mechanisms of antimicrobial resistance. *Drugs*. 2002; 62(4):557–566. [PubMed: 11893225]
6. Sharma M, Chauhan PMS. Dihydrofolate reductase as a therapeutic target for infectious diseases: Opportunities and challenges. *Future Medicinal Chemistry*. 2012; 4(10):1335–1365. [PubMed: 22800373]
7. Zhou W, Scocchera EW, Wright DL, Anderson AC. Antifolates as effective antimicrobial agents: New generations of trimethoprim analogs. *Med Chem Comm*. 2013; 4(6):908–915.
8. Simonsen CC, Chen EY, Levinson AD. Identification of the type I trimethoprim-resistant dihydrofolate reductase specified by the *Escherichia coli* Rplasmid R483: Comparison with procaryotic and eucaryotic dihydrofolate reductases. *Journal of bacteriology*. 1983; 155(3):1001–1008. [PubMed: 6411680]
9. Assaraf YG. Molecular basis of antifolate resistance. *Cancer Metastasis Reviews*. 2007; 26(1):153–181. DOI: 10.1007/s10555-007-9049-z [PubMed: 17333344]
10. Sundström L, Jansson C, Bremer K, Heikkilä E, Olsson-Liljequist B, Sköld O. A new dhfrVIII trimethoprim-resistance gene, flanked by IS26, whose product is remote from other dihydrofolate reductases in parsimony analysis. *Gene*. 1995; 154(1):7–14. [PubMed: 7867952]
11. Ojo KK, Kehrenberg C, Schwarz S, Odelola HA. Identification of a complete dfrA14 gene cassette integrated at a secondary site in a resistance plasmid of uropathogenic *Escherichia coli* from Nigeria. *Antimicrobial Agents and Chemotherapy*. 2002; 46(6):2054–2055. [PubMed: 12019141]
12. Yachnin BJ, Colin DY, Volpato JP, Ebert M, Pelletier JN, Berghuis AM. Novel crystallization conditions for tandem variant R67 DHFR yield a wild-type crystal structure. *Acta Crystallogr Sect F Struct Biol Cryst Commun*. 2011; 67(11):1316–22.
13. Buwembo W, Aery S, Rwenyonyi CM, Swedberg G, Kironde F. Point Mutations in the folP Gene Partly Explain Sulfonamide Resistance of Streptococcus mutants. *Int J Microbiol*. 2013; Epub 2013 Feb 25. doi: 10.1155/2013/367021
14. Ebert MC, Morley KL, Volpato JP, Schmitzer AR, Pelletier JN. Asymmetric mutations in the tetrameric R67 dihydrofolate reductase reveal high tolerance to active-site substitutions. *Protein Sci*. 2015; 24(4):495–507. [PubMed: 25401264]
15. Fleming MP, Datta N, Gruneberg RN. Trimethoprim resistance determined by R factors. *British Medical Journal*. 1972; 1(5802):726–728. [PubMed: 4552466]
16. Šeputien V, Povilonis J, Ružauskas M, Pavilonis A, Sužiedien E. Prevalence of trimethoprim resistance genes in *Escherichia coli* isolates of human and animal origin in Lithuania. *Journal of Medical Microbiology*. 2010; 59(3):315–322. [PubMed: 20007760]
17. Grape M, Sundstrom L, Kronvall G. Two new dfr genes in trimethoprim-resistant integron-negative *Escherichia coli* isolates. *Antimicrobial Agents and Chemotherapy*. 2007; 51(5):1863–864. [PubMed: 17307981]
18. Smith SL, Stone D, Novak P, Baccanari DP, Burchall JJ. R plasmid dihydrofolate reductase with subunit structure. *Journal of Biological Chemistry*. 1979; 254(14):6222–6225. [PubMed: 376528]

19. Fling ME, Elwell LP. Protein expression in *Escherichia coli* minicells containing recombinant plasmids specifying trimethoprim-resistant dihydrofolate reductases. *Journal of Bacteriology*. 1980; 141(2):779–785. [PubMed: 6988391]
20. Iwagami M, Higo H, Miura S, Yanagi T, Tada I, Kano S, Agatsuma T. Molecular phylogeny of *Trypanosomacruzi* from central America (Guatemala) and a comparison with south American strains. *Parasitology Research*. 2007; 102(1):129–134. [PubMed: 17828552]
21. O'Neil RH, Lilien RH, Donald BR, Stroud RM, Anderson AC. Phylogenetic classification of protozoa based on the structure of the linker domain in the bifunctional enzyme, dihydrofolate reductase-thymidylate synthase. *Journal of Biological Chemistry*. 2003; 278(52):52980–52987. [PubMed: 14555647]
22. Hecht D, Tran J, Fogel GB. Structural-based analysis of dihydrofolate reductase evolution. *Mol Phylogenet Evol*. 2011; 61:212–30. [PubMed: 21704717]
23. Zhao S, Shalchian-Tabrizi K, Klaveness D. Sulcozoa revealed as a paraphyletic group in mitochondrial phylogenomics. *Mol Phylogenet Evol*. 2013; 69(3):462–8. [PubMed: 23973893]
24. Vartanian JP, Henry M, Wain-Hobson S. Simulating pseudogene evolution in vitro: Determining the true number of mutations in a lineage. *Proceedings of the National Academy of Sciences*. 2001; 98(23):13172–13176.
25. Iwakura M, Maki K, Takahashi H, Takenawa T, Yokota A, Katayanagi K, Gekko K. Evolutional design of a hyperactive cysteine- and methionine-free mutant of *Escherichia coli* dihydrofolate reductase. *Journal of Biological Chemistry*. 2006; 281(19):13234–13246. [PubMed: 16510443]
26. Hecht, D.; Fogel, GB. Towards predictive structure-based models of evolved drug resistance; 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology; San Diego. 2012. p. 120-126.
27. Hecht D, Fogel GB. Modeling the evolution of drug resistance in malaria. *J Comput Aided Mol Des*. 2012; 26:1343–53. [PubMed: 23179493]
28. Bhabha G, Ekiert DC, Jennewein M, Zmasek CM, Tuttle Lisa M, Kroon Gerard, Dyson H Jane, Godzik Adam, Wilson Ian A, Wright Peter E. Divergent evolution of protein conformational dynamics in dihydrofolate reductase. *Nature Structural & Molecular Biology*. 2013; 20:1243–1249.
29. Masterson JE, Schwartz SD. Evolution alters the enzymatic reaction coordinate of dihydrofolate reductase. *J Phys Chem B*. 2015; 119(3):989–96. [PubMed: 25369552]
30. Ogbunugafor CB, Wylie CS, Diakite I, Weinreich DM, Hartl DL. Adaptive Landscape by Environment Interactions Dictate Evolutionary Dynamics in Models of Drug Resistance. *PLoS Comput Biol*. 2016; 12(1):e1004710. 2016 Jan 25. eCollection 2016 Jan. doi: 10.1371/journal.pcbi.1004710 [PubMed: 26808374]
31. Arav-Boger R, Shapiro TA. Molecular mechanisms of resistance in antimalarial chemotherapy: the unmet challenge. *Annu Rev Pharmacol Toxicol*. 2005; 45:565–85. [PubMed: 15822189]
32. Fogel, GB.; Islas, C.; Hecht, D. Modeling the Evolution of Drug Resistance in *Plasmodium falciparum* DHFR. 2013 IEEE Congress on Evolutionary Computation; Cancun, Mexico. 2013. p. 206-212.
33. Ludwig, W.; Schleifer, KH.; Whitman, WB. Family IV. *Enterococcaceae*. In: De Vos, P.; Garrity, G.; Jones, D.; Krieg, NR.; Ludwig, W.; Rainey, FA.; Whitman, W., editors. *Bergey's manual of systematic bacteriology: The firmicutes*. Vol. 3. New York, NY: Springer; 2009a. p. 594
34. Ludwig, W.; Schleifer, KH.; Whitman, WB. Family VI. *Planococcaceae* Krasil'nikov 1949, 328AL. In: De Vos, P.; Garrity, G.; Jones, D.; Krieg, NR.; Ludwig, W.; Rainey, FA.; Whitman, W., editors. *Bergey's manual of systematic bacteriology: The firmicutes*. Vol. 3. New York, NY: Springer; 2009b. p. 348
35. Ludwig, W.; Schleifer, KH.; Whitman, WB. Revised road map to the phylum firmicutes. In: De Vos, P.; Garrity, G.; Jones, D.; Krieg, NR.; Ludwig, W.; Rainey, FA.; Whitman, W., editors. *Bergey's manual of systematic bacteriology: The firmicutes*. Vol. 3. New York, NY: Springer; 2009c. p. 1-13.
36. Hall TA. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*. 1999; 41:95–98.

37. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*. 1997; 18(15):2714–2723. [PubMed: 9504803]
38. Gerber PR, Müller K. MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. *J Comput Aided Mol Des*. 1995; 3:251–68.
39. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*. 1990; 18:6097–6100. [PubMed: 2172928]
40. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Research*. 2004; 14:1188–1190. [PubMed: 15173120]
41. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*. 2013; 30:2725–2729. [PubMed: 24132122]
42. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature*. 2000; 405:299–304. DOI: 10.1038/35012500 [PubMed: 10830951]
43. Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*. 2008; 9:605–618. DOI: 10.1038/nrg2386
44. Ravenhall M, Škunca N, Lassalle F, Dessimoz C. Inferring Horizontal Gene Transfer. *PLOS Computational Biology*. 2015; :1–16. DOI: 10.1371/journal.pcbi.1004095
45. Goldman N, Anderson JP, Rodrigo AG. Likelihood-Based Tests of Topologies in Phylogenetics. *Syst Biol*. 2000; 49:652–670. [PubMed: 12116432]
46. Abby SS, Tannier E, Gouy M, Daubin V. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *Bioinformatics*. 2010; 11:324–327. [PubMed: 20550700]
47. Agarwal PK, Billeter SR, Rajagopalan PTR, Benkovic SJ, Hammes-Schiffer S. Network of coupled promoting motions in enzyme catalysis. *Proc Natl Acad Sci USA*. 2002; 99:2794–2799. [PubMed: 11867722]
48. Ravi Rajagopalan PTR, Lutz S, Benkovic SJ. Coupling Interactions of Distal Residues Enhance Dihydrofolate Reductase Catalysis: Mutational Effects on Hydride Transfer Rates. *Biochemistry*. 2002; 41:12618–12628. [PubMed: 12379104]
49. Hammes-Schiffer S, Benkovic SJ. Relating protein motion to catalysis. *Annu Rev Biochem*. 2006; 75:519–541. [PubMed: 16756501]
50. Singh P, Kevin Francis K, Kohen A. Network of Remote and Local Protein Dynamics in Dihydrofolate Reductase Catalysis. *ACS Catalysis*. 2015; 5:3067–3073. [PubMed: 27182453]
51. Paredes-Sabja D, Setlow P, Sarker MR. Germination of spores of *Bacillales* and *Clostridiales* species: Mechanisms and proteins involved. *Trends in Microbiology*. 2011; 19(2):85–94. [PubMed: 21112786]
52. Logan, NA.; De Vos, P. Family I. *Bacillaceae*. In: De Vos, P.; Garrity, G.; Jones, D.; Krieg, NR.; Ludwig, W.; Rainey, FA.; Whitman, W., editors. *Bergey's manual of systematic bacteriology: The firmicutes*. Vol. 3. New York, NY: Springer; 2009. p. 20-128.
53. Salminen, S. *Lactic acid bacteria*. Boca Raton, FL: CRC Press, Taylor & Francis; 2012.
54. Kim HS, Damo SM, Lee SY, Wemmer D, Klinman JP. Structure and hydride transfer mechanism of a moderate thermophilic dihydrofolate reductase from *Bacillus stearothermophilus* and comparison to its mesophilic and hyperthermophilic homologues. *Biochemistry*. 2005; 44(34): 11428–11439. [PubMed: 16114879]
55. Loveridge EJ, Rodriguez RJ, Swanwick RS, Allemann RK. Effect of dimerization on the stability and catalytic activity of dihydrofolate reductase from the hyperthermophile *Thermotogamaritima*. *Biochemistry*. 2009; 48(25):5922–5933. [PubMed: 19453185]
56. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge: Making sense of genetic and ecological diversity. *Science*. 2009; 323(5915):741–746. [PubMed: 19197054]
57. Redecke L, Brehm MA, Bredehorst R. Cloning and characterization of dihydrofolate reductase from a facultative alkaliphilic and halotolerant *Bacillus* strain. *Extremophiles*. 2007; 11(1):75–83. [PubMed: 17021659]
58. Wright DB, Banks DD, Lohman JR, Hilsenbeck JL, Gloss LM. The effect of salts on the activity and stability of *Escherichia coli* and *Haloflexvolcanii* dihydrofolate reductases. *Journal of Molecular Biology*. 2002; 323(2):327–344. [PubMed: 12381324]

59. Xu Y, Feller G, Gerday C, Glansdorff N. Moritella cold-active dihydrofolate reductase: Are there natural limits to optimization of catalytic efficiency at low temperature? *Journal of Bacteriology*. 2003; 185(18):5519–5526. [PubMed: 12949104]
60. Murakami C, Ohmae E, Tate SI, Gekko K, Nakasone K, Kato C. Cloning and characterization of dihydrofolate reductases from deep-sea bacteria. *Journal of Biochemistry*. 2010; 147(4):591–599. [PubMed: 20040594]
61. Cherry JL. Highly Expressed and Slowly Evolving Proteins Share Compositional Properties with Thermophilic Proteins. *Mol Biol Evol*. 2007; 27:735–741. DOI: 10.1093/molbev/msp270
62. Zeldovich KB, Chen P, Shakhnovich Boris E, Shakhnovich Eugene I. A First-Principles Model of Early Evolution: Emergence of Gene Families, Species, and Preferred Protein Folds. *PLoS Computational Biology*. 2007; 3:1124–1238.
63. Serohijos AWR, Rimas, Shakhnovich EI. Protein Biophysics Explains Why Highly Abundant Proteins Evolve Slowly. *Cell Rep*. 2012; 2:249–256. DOI: 10.1016/j.celrep.2012.06.022 [PubMed: 22938865]
64. Singer GA, Hickey DA. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene*. 2003; 317:39–47. [PubMed: 14604790]
65. Berezovsky IN, Zeldovich KB, Shakhnovich EI. Positive and Negative Design in Stability and Thermal Adaptation of Natural Proteins. *PLoS Computational Biology*. 2007; 3:498–507. www.ploscompbiol.org March 2007, Volume 3, Issue 3,e52.
66. Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA Sequence Determinants of Thermophilic Adaptation. *PLoS Computational Biology*. 2007; 3:62–72. www.ploscompbiol.org January 2007, Volume 3, Issue 1,e5.
67. Bershtein S, Serohijos AWR, Bhattacharyya Sanchari, Manhart M, Choi J-M1, Mu W, Zhou J, Shakhnovich EI. Protein Homeostasis Imposes a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria. *PLOS Genetics*. 2015; October 20, 2015. doi: 10.1371/journal.pgen.1005612
68. Murakami C, Ohmae E, Tate SI, Gekko K, Nakasone K, Kato C. Comparative study on dihydrofolate reductases from *Shewanella* species living in deep-sea and ambient atmospheric-pressure environments. *Extremophiles*. 2011; 15(2):165–175. 40. [PubMed: 21181485]
69. Alvarez M, Wouters J, Maes D, Mainfroid V, Rentier-Delrue F, Wyns L, Martial JA. Lys13 plays a crucial role in the functional adaptation of the thermophilic triose-phosphate isomerase from *Bacillus stearothermophilus* to high temperatures. *Journal of Biological Chemistry*. 1999; 274(27):19181–19187. [PubMed: 10383424]
70. Kawamura S, Abe Y, Ueda T, Masumoto K, Imoto T, Yamasaki N, Kimura M. Investigation of the structural basis for thermostability of DNA-binding protein HU from *Bacillus stearothermophilus*. *Journal of Biological Chemistry*. 1998; 273(32):19982–19987. [PubMed: 9685334]
71. Perl D, Mueller U, Heinemann U, Schmid FX. Two exposed amino acid residues confer thermostability on a cold shock protein. *Nature Structural & Molecular Biology*. 2000; 7(5):380–383.
72. Fusco V, Quero GM, Cho G-S, Kabisch J, Meske D, Neve H, Bockelmann W, Franz CMAP. The genus *Weissella*: taxonomy, ecology and biotechnological potential. *Front Microbiol*. 2015; 6:155. Published online 2015 Mar 17. doi: 10.3389/fmicb.2015.00155 [PubMed: 25852652]
73. Nieminen, TT.; Säde, E.; Endo, A.; Johansson, P.; Björkroth, J. The family *Leuconostocaceae*. In: Rosenberg, E.; DeLong, EF.; Lory, S.; Stackebrandt, E.; Thompson, F., editors. *The Prokaryotes: Firmicutes and Tenericutes*. Heidelberg, Germany: Springer; 2014. p. 215-240.
74. Gotz, F.; Bannerman, T.; Schleifer, KH. The genera *Staphylococcus* and *Macrococcus*. In: Dworkin, M.; Falkow, S.; Rosenberg, E.; Schleifer, K-H.; Stackebrandt, E., editors. *The prokaryotes: Firmicutes, cyanobacteria*. Vol. 4. New York, NY: Springer; 2006. p. 5-75.
75. Schleifer, KH.; Bell, JA. Family VIII. *Staphylococcaceae* fam. nov. In: De Vos, P.; Garrity, G.; Jones, D.; Krieg, NR.; Ludwig, W.; Rainey, FA.; Whitman, W., editors. *41 Bergey's manual of systematic bacteriology: The Firmicutes*. Vol. 3. New York, NY: Springer; 2009. p. 392
76. Noble, WC.; Somerville, DA. *Microbiology of human skin*. London, England: W. B. Saunders; 1974.

77. Kloos WE, Musselwhite MS. Distribution and persistence of *Staphylococcus* and *Micrococcus* species and other aerobic bacteria on human skin. *Applied Microbiology*. 1975; 30(3):381–395. [PubMed: 810086]
78. Collins MD, Wallbanks S, Lane DJ, Shah J, Nietupski R, Smida J, Stackebrandt E. Phylogenetic analysis of the genus *Listeria* based on reverse transcriptase sequencing of 16S rRNA. *International Journal of Systematic Bacteriology*. 1991; 41(2):240–246. [PubMed: 1713054]
79. Feresu SB, Jones D. Taxonomic studies on *Brochothrix*, *Erysipelothrix*, *Listeria* and atypical *Lactobacilli*. *Journal of General Microbiology*. 1988; 134(5):1165–1183. [PubMed: 3143805]
80. Wilkinson BJ, Jones D. A numerical taxonomic survey of *Listeria* and related bacteria. *Journal of General Microbiology*. 1977; 98(2):399–421. [PubMed: 856940]
81. Farber JM, Peterkin PI. *Listeria monocytogenes*, a food-borne pathogen. *Microbiological Reviews*. 1991; 55(3):476–511. [PubMed: 1943998]
82. Lory, S. The family *Enterococcaceae*. In: Rosenberg, E.; DeLong, EF.; Lory, S.; Stackebrandt, E.; Thompson, F., editors. *The prokaryotes: Firmicutes and Tenericutes*. New York, NY: Springer; 2014. p. 75–77.
83. Jenkinson HF, Lamont RJ. *Streptococcal* adhesion and colonization. *Critical Reviews in Oral Biology & Medicine*. 1997; 8(2):175–200. [PubMed: 9167092]
84. Toit, MD.; Huch, M.; Cho, GS.; Franz, CM. The family *Streptococcaceae*. In: Holzapfel, WH.; Wood, BJB., editors. *Lactic acid bacteria: Biodiversity and taxonomy*. New York, NY: John Wiley & Sons, Ltd; 2014. p. 445–446.
85. Schleifer, KH. Family I. *Lactobacillaceae* Winslow, Broadhurst, Buchanan, Krumwiede, Rogers and Smith 1917, familia. In: De Vos, P.; Garrity, G.; Jones, D.; Krieg, NR.; Ludwig, W.; Rainey, FA.; Whitman, W., editors. *Bergey's manual of systematic bacteriology: Volume 3: The Firmicutes*. New York, NY: Springer; 2009. p. 465–532.
86. Elli M, Zink R, Rytz A, Reniero R, Morelli L. Iron requirement of *Lactobacillus* spp. in completely chemically defined growth media. *Journal of Applied Microbiology*. 2001; 88(4):695–703.
87. Huycke MM, Sahm DF, Gilmore MS. Multiple-drug resistant *Enterococci*: The nature of the problem and an agenda for the future. *Emerging Infectious Diseases*. 1998; 4(2):239–249. [PubMed: 9621194]
88. Amyes SG, Gemmell CG. Antibiotic resistance in bacteria. *Journal of Medical Microbiology*. 1992; 36(5):366. Arav-Boger R, Shapiro T. Molecular mechanisms of resistance in antimalarial chemotherapy: The unmet challenge. *Annual Review of Pharmacology and Toxicology*. 2005; 45(1):565–585.

1. A structure-based alignment of *Plasmodium* DHFR from 350 Bacilli and plasmid DHFR sequences was performed.
2. Using this structure-based alignment of *Plasmodium* DHFR, the degree of conservation or variability was determined at each site in the alignment – not only in terms of amino acid residue, but also secondary structure and residue class.
3. Phylogenic trees were generated and differences were found when compared to the canonical phylogenetic trees. Evidence for horizontal gene transfer was identified.
4. As expected, there was significantly increased conservation at active site residues across the species. Surprisingly though, these active-site residues appear to be more highly conserved in plasmid DHFR sequences than Bacilli chromosomal DHFR sequences.
5. We explore the conservation of residues involved with concerted protein motions and find greater degree of conservation in plasmid versus chromosomal DHFR proteins.

Figure 1.
Reactions in the folate pathway [1,2].

Figure 2.

Canonical taxonomic tree of 38 families representing 273 *Bacilli* genera, generated using NCBI website of taxonomic relationship: www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/ www.cmt.cgi.

Figure 3.

Site-specific frequencies of amino acid, secondary structure type and residue class calculated for the entire structural based alignment positions, the active site positions, and the non-active site positions of 3 DHFR alignments: **a.** Representing the total DHFR alignment with overall rmsd = 1.51Å, **b.** Representing 273 chromosomal DHFR sequences structurally aligned with overall rmsd of 1.50Å, and **c.** Representing 77 plasmid-encoded DHFR sequences structurally aligned with overall rmsd of 1.78Å.

Figure 4. Structure-based alignments of (a) *Bacillales* and (b) *Lactobacillales* DHFR. **a.** The structure-based alignment of 124 chromosomal DHFR sequences representing members of the order *Bacillales* (rmsd = 1.50 Å). **b.** The structure-based alignment of 149 chromosomal DHFR sequences representing members of the order *Lactobacillales* (rmsd = 1.75 Å).

I. Bacillales Families:

II. Lactobacillales Families:

Figure 5.

Structure-based alignments representing 8 families of *Bacilli*. **a.** 59 *Bacillaceae* DHFR (rmsd = 1.46 Å). **b.** 19 *Staphylococcaceae* DHFR (rmsd = 0.85 Å). **c.** 10 *Planococcaceae* DHFR (rmsd = 0.72 Å). **d.** 21 *Paenibacillaceae* DHFR (rmsd = 0.1.64 Å). **e.** 59 *Streptococcaceae* DHFR (rmsd = 0.72 Å). **f.** 33 *Lactobacillaceae* DHFR (rmsd = 1.76 Å). **g.** 18 *Leuconostocaceae* DHFR (rmsd = 1.65 Å). **h.** 20 *Enterococcaceae* DHFR (rmsd = 2.11 Å).

	5	20	25	30	35	40	45	50	54	90	95	110
												
<i>Staphylococcus arlettae</i>	IAA-NDIPWRIP	----	KDWEYVKNTT	-IILGRKNLESI	-G-----	RAL-FIFGGEQIY	-YIT					
<i>Staphylococcus aureus</i>	LVA-NQLPWHLP	----	NDLKHKVKKLS	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus capitis</i>	LVA-NQLPWHLP	----	NDLKHKVKKLS	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus caprae</i>	LVA-NQLPWHLP	----	NDLKHKVKKLS	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus carnosus</i>	LVA-NQLPWHLP	----	NDLKHKVKKLS	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus epidermidis</i>	LVA-NQLPWHLP	----	NDLKHKVKKLS	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus equorum subsp.</i>	LVA-NQLPWHLP	----	NDLKHKVKKLS	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus haemolyticus</i>	IVA-NDIPWRIS	----	SDWEYVKNTT	-IILGRKNLQSI	-G-----	RAL-FIFGGEQIY	-YVT					
<i>Staphylococcus hominis</i>	LVA-NALPWHIP	----	NDLKHKVKKLS	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus lugdunensis</i>	LVA-NALPWHLP	----	NDLKHKVKKLS	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus massiliensis</i>	IVA-NQLPWHIP	----	NDLKHKVKKLT	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus pettenkoferi</i>	LVA-NQLPWHLP	----	NDLKHKVKKLS	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus pseudintermedius</i>	LVA-NQLPWHLP	----	NDLKHKVKKLT	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus saprophyticus</i>	LVA-NQLPWHLP	----	NDLKHKVKKLS	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus simiae</i>	LVA-NQLPWHLP	----	NDLKHKVKKLS	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus simulans</i>	LVA-NQLPWHLP	----	NDLKHKVKKLS	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>Staphylococcus warneri</i>	LVA-NQLPWHLP	----	NDLKHKVKKLT	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-FIT					
<i>Staphylococcus xylosus</i>	LVA-NQLPWHLP	----	NDLKHKVKKLT	-LVMGRKTFESI	-G-----	KPL-FIFGGQTLF	-YIT					
<i>S. arlettae</i>	SSS-----	HHHHHHHH	----	SSS-----	HHHHHH	-----	SSS-----	HHH-SSS				
<i>S. aureus</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. capitis</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. caprae</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. carnosus</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. epidermidis</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. equorum subsp. equorum Mu2</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. haemolyticus</i>	SSS-----	HHHHHHHHHH	SSS-HHH--H	-----	SSS-----	HHH-SSS						
<i>S. hominis</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. lugdunensis</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. massiliensis</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. pettenkoferi</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. pseudintermedius</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. saprophyticus</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. simiae</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. simulans</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. warneri</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>S. xylosus</i>	SSS-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS-----	HHHH-SSS						
<i>Staphylococcus arlettae</i>	NNN-QANPRBNP	----	BARARNBQQQ	-NNGBBQNAQN	-G-----	BNN-RNRGGAQNR	-RNQ					
<i>Staphylococcus aureus</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus capitis</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus caprae</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus carnosus</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus epidermidis</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus equorum subsp.</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus haemolyticus</i>	NNN-QANPRBNQ	----	QARARNBQQQ	-NNGBBQNAQN	-G-----	BNN-RNRGGAQNR	-RNQ					
<i>Staphylococcus hominis</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus lugdunensis</i>	NNN-QANPRBNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus massiliensis</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus pettenkoferi</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus pseudintermedius</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus saprophyticus</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus simiae</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus simulans</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus warneri</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					
<i>Staphylococcus xylosus</i>	NNN-QQNPBPNP	----	QANBBNBBNQ	-NNQGBBQRAQN	-G-----	BPN-RNRGGQQNR	-RNQ					

Figure 6.

Subset of the complete structural alignment consisting of 18 *Staphylococcus* DHFR sequences truncated to residues 6 Å from bound NADPH and the active site. The first 18 sequences represent the actual residues in the sequence. The second set of 18 sequences are labeled for the type of secondary structure present: **H** for α -helix, **S** for β -sheet and **T** for turn. The last set of 18 sequences are labeled for the type of amino acid residue class: **A** for acid, **B** for base, **N** for non-aromatic hydrophobic, **Q** for non-charged polar, **R** for aromatic,

P for proline, *G* for glycine. Residues $>6 \text{ \AA}$ from bound NADPH and the active site are not shown and are designated by “–”.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

	5	20	25	30	35	40	45	50	54	90	95	110
												
<i>Bacillus alcalophilus</i>	IFA-NDMPWHL	----	NDLKYFKKVT	-VVMGRKTYESL	-G-----	RAL-FVIGGAGLY	-YVT					
<i>Bacillus amyloliquefaciens</i>	IFA-NDLPWHL	----	DDLAYFKKVT	-IVMGRKTFESI	-G-----	RPL-FVIGGAQLY	-YMT					
<i>Bacillus anthracis</i>	MVA-NNLPWRL	----	SELQYVKKTT	-LIMGRKNYEA	-G-----	RPL-FIFGGAQY	-YIT					
<i>Bacillus atrophaeus</i>	IFA-NDLPWRL	----	NDLAYFKKVT	-IVMGRKTFESI	-G-----	RPL-FVIGGAQLY	-YMT					
<i>Bacillus azotoformans</i>	VVA-NQLPWH	----	ADLAFFKKVT	-IVMGRKTYESI	-G-----	RPL-CVIGGAEIF	-YIT					
<i>Bacillus bataviensis</i>	IVA-NQLPWH	----	EDLKFFKRV	-IAMGRKTHDS	-G-----	RVL-FVIGGAEIF	-YIT					
<i>Bacillus cellulosilyticus</i>	IAA-GEMPWHL	----	NDLKFFKKT	-VLMGRKTYESI	-G-----	RPL-FVIGGATLY	-YIT					
<i>Bacillus cereus</i>	MVA-NKLPWH	----	SELQYVKKTT	-LIMGRKNYEA	-G-----	RPL-FIFGGAQY	-YIT					
<i>Bacillus clausii</i>	VYA-NGLPWH	----	ADLRHFKRT	-IVMGRKTFDS	-G-----	GPL-YVIGGAEIF	-YVT					
<i>Bacillus coagulans</i>	LWA-NALPWH	----	EDLKYFKNT	-VVMGRRTFESI	-G-----	KPL-FVTGGAEIF	-YRT					
<i>Bacillus coahuilensis</i>	IWA-NSMPWH	----	KDLKHFKRT	-IIMGRKTFES	-G-----	KPL-MVIGGSQY	-EVT					
<i>Bacillus cytotoxicus</i>	MVA-NKLPWH	----	SELQYVKKTT	-LIMGRKNYEA	-G-----	RPL-FVFGGAQY	-YIT					
<i>Bacillus halodurans</i>	IVA-NDLPWR	----	ADLAHFKRV	-VIMGRKTYESI	-G-----	KPL-FVIGGSSEL	-YVT					
<i>Bacillus isronensis</i>	IVA-NDLPWY	----	GDLKYFKMT	-MIMGRKTYDS	-G-----	RPL-MIIGGAQY	-YIT					
<i>Bacillus licheniformis</i>	IAA-NDLPWH	----	QDLKYFKVT	-VIMGRKTFESI	-G-----	RPL-FVIGGSSTLY	-YVT					
<i>Bacillus macauensis</i>	VVA-NELPWH	----	ADLQHFKVT	-IVMGRKTYESI	-G-----	RPL-CVIGGAHL	-YIT					
<i>Bacillus megaterium</i>	IVA-NDLPWR	----	EDLKYFKRT	-IIMGRKTFESI	-G-----	RPL-FVIGGATLY	-YIT					
<i>Bacillus methanolicus</i>	IWA-NKLPWH	----	EDLKFFKRV	-IVMGRKTHDS	-G-----	RNL-FVIGGAEIF	-YCT					
<i>Bacillus mojavensis</i>	IFA-NDLPWH	----	NDLAYFKKVT	-IIMGRKTYESI	-G-----	RPL-FVIGGAQY	-YMT					
<i>Bacillus nealsonii</i> AAU1	IVA-NDIPWR	----	RDWEYVKKT	-IILGRKNFQSI	-G-----	KAL-FIFGGEELY	-YIT					
<i>Bacillus pseudofirmus</i>	IVA-NDLPWH	----	ADLAHFKRV	-IVMGRKTYESI	-G-----	RPL-FIIGGAEL	-YIT					
<i>Bacillus pumilus</i>	IVA-NQMPWH	----	ADLAYFKKVT	-IVMGRKTFESI	-G-----	RAL-FIIGGSKLY	-YIT					
<i>Bacillus selenitireducens</i>	IVA-GQMPWH	----	NDLKHFKA	-VIMGRKTFESI	-G-----	RPL-FIIGGGELY	-YVT					
<i>Bacillus smithii</i>	IVA-NRLPWR	----	EDLKFFKRT	-IVMGRKTFESI	-G-----	KPL-FIIGGAEL	-YIT					
<i>Bacillus sonorensis</i>	IFA-NDLPWH	----	NDLKYFKNT	-VIMGRKTYESI	-G-----	RPL-FVIGGSSTLY	-YMT					
<i>Bacillus</i> sp. 10403023	LVA-NDLPWS	----	ADLAYFKRV	-IIMGRKTYDS	-G-----	RAL-FVIGGAEIF	-YIT					
<i>Bacillus</i> sp. 1NLA3E	IWA-NQLPWH	----	EDLKFFKKT	-IAMGRKTHVS	-G-----	RPL-FIIGGAEL	-YIT					
<i>Bacillus</i> sp. 2_A_57_CT2	MWA-NKLPWH	----	EDLKFFKRT	-IAMGRKTHVS	-G-----	RPL-FVIGGAEIF	-YLT					
<i>Bacillus</i> sp. 5B6	IFA-NDLPWH	----	DDLAYFKKVT	-IVMGRKTFESI	-G-----	RPL-FVIGGAQY	-YMT					
<i>Bacillus</i> sp. B14905	IFA-NDLPWH	----	DDLAYFKKVT	-IVMGRKTFESI	-G-----	RPL-FVIGGAQY	-YMT					
<i>Bacillus</i> sp. HYC-10	IVA-NQMPWH	----	ADLAYFKKVT	-IVMGRKTFESI	-G-----	RAL-LIIGGSKLY	-YIT					
<i>Bacillus</i> sp. JS	IFA-NDLPWH	----	NDLAYFKKVT	-IIMGRKTFESI	-G-----	RPL-FVIGGAQY	-YMT					
<i>Bacillus</i> sp. M 2-6	----	HRLDWL	-GTE-GEEDTRYEEFY	-LIMGRNTYEQV	-L-KLSPDEFPY	-WILGGGELL	-YLT					
<i>Bacillus</i> sp. m3-13	IVA-NDLPWR	----	ADLAYFKKVT	-IVMGRKTYESI	-G-----	KPL-FIIGGAHIF	-YLT					
<i>Bacillus</i> sp. NRRL B-14911	LWA-NDLPWR	----	EDLKYFKRV	-IAMGRKTHVS	-G-----	RPL-FIIGGAEIF	-YLT					
<i>Bacillus</i> sp. SG-1	VWA-NKLPWR	----	EDLKFFKET	-IVMGRKTFESI	-G-----	KPL-MVTGGSEVF	-YVT					
<i>Bacillus stratosphericus</i>	YGA-HRLDWL	-GTE-GEEDTRYEEFY	-LIMGRNTYEQV	-L-KLSPDEFPY	-WIVGGGELL	-YLT						
<i>Bacillus subtilis</i>	IFA-NDLPWH	----	NDLAYFKKVT	-IIMGRKTFESI	-G-----	RPL-FVIGGAQY	-YMT					
<i>Bacillus thuringiensis</i>	IVA-NKLLCR	----	DDMKHFIKKT	-IIMGYKTFASL	-GH-----	KPL-VVIGGQTY	-YLT					
<i>Bacillus vallismortis</i>	IFA-NDLPWR	----	NDLAYFKKVT	-IIMGRKTFESI	-G-----	RPL-FVIGGAQY	-YMT					
<i>B. alcalophilus</i>	SSS-----	H-----	HHHHHHHH	-----SSSSHHHHHH	-----	SSS--HHHH	-SSS					
<i>B. amyloliquefaciens</i>	SSS-----	H-----	HHHHHHHH	-----SSSSHHHHHH	-----	SSS--HHHH	-SSS					
<i>B. Anthracis</i>	SSS-----	H-----	HHHHHHHH	-----SSSSHHHHHH	-----	SSS--HHH	-SSS					
<i>B. atrophaeus</i>	SSS-----	H-----	HHHHHHHH	-----SSSSHHH	-----	SSS--HHHH	-SSS					
<i>B. azotoformans</i>	SSS-----	H-----	HHHHHHHHHH	SSS-HHHHHHH	-----	SSS--HHHH	-SSS					
<i>B. bataviensis</i>	SSS-----	H-----	HHHHHHHHHH	SSS-HHHHHH	-----	SSS--HHHH	-SSS					
<i>B. cellulosilyticus</i>	SSS-----	H-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS--HHHH	-SSS					
<i>B. cereus</i>	SSS-----	H-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS--HHHH	-SSS					
<i>B. clausii</i>	SSS-----	H-----	HHHHHHHHHH	SSS-HHHHHHH	-----	SSS--HHHH	-SSS					
<i>B. coagulans</i>	SSS-----	H-----	HHHHHHHHHH	SSS-----	-----	SSS--HHHH	-SSS					
<i>B. coahuilensis</i>	SSS-----	H-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS--HHHH	-SSS					
<i>B. cytotoxicus</i>	SSS-----	H-----	HHHHHHHH	-----SSSSHHHHHH	-----	SSS--HHHH	-SSS					
<i>B. halodurans</i>	SSS-----	H-----	HHHHHHHHHH	SSS-HHH	-----H-----	SSS--HHHH	-SSS					
<i>B. isronensis</i>	SSS-----	H-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS--HHHH	-SSS					
<i>B. licheniformis</i>	SSS-----	H-----	HHHHHHHHHH	SSS-HHHHHHH	-----	SSS--HHHH	-SSS					
<i>B. macauensis</i>	SSS-----	H-----	HHHHHHHHHH	SSSSHHHHHH	-----	SSS--HHHH	-SSS					
<i>B. megaterium</i>	SSS-----	H-----	HHHHHHHHHH	SSS-HHHHHH	-----	SSS--HHHH	-SSS					
<i>B. methanolicus</i>	SSS-----	H-----	HHHHHHHHHH	SSS-HHHHHHH	-----	SSS--HHHH	-SSS					
<i>B. mojavensis</i>	SSS-----	H-----	HHHHHHH	-----SSSSHHH	-----H-----	SSS--HHHH	-SSS					

<i>B. nealsonii</i> AAU1	SSS-----H-----HHHHHHH---SSS-HHHHHH-----SSS---HHH-SSS
<i>B. pseudofirmus</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHHH-SSS
<i>B. pumilus</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHHH-SSS
<i>B. selenitireducens</i>	SSS-----H-----HHHHHHHHH-SSSSSHHHHH-----S---HHHHHH-SSS
<i>B. smithii</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHHH-SSS
<i>B. sonorensis</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHHH-SSS
<i>B. sp. 10403023</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHHH-SSS
<i>B. sp. 1NLA3E</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHHH-SSS
<i>B. sp. 2_A_57_CT2</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHHH-SSS
<i>B. sp. 5B6</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHHH-SSS
<i>B. sp. B14905</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHHH-SSS
<i>B. sp. HYC-10</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHH-SSS
<i>B. sp. JS</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHHH-SSS
<i>B. sp. M 2-6</i>	-----HH-----HHHHH-SSSSSHHH-----SSS---HHHH-SSS
<i>B. sp. m3-13</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHHH-SSS
<i>B. sp. NRRL B-14911</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHHH-SSS
<i>B. sp. SG-1</i>	SSS-----H-----HHHHHHHHH-SSSSSHHHHH-----SSS---HHH-SSS
<i>B. stratosphericus</i>	SSS---HH-----HHHHH-SSSSSHHH-----HH-----SSS---HHHH-SSS
<i>B. subtilis</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHH-SSS
<i>B. thuringiensis</i>	SSS---SSS---H-----HHHHHHHHH-SSS---HHHH-----SSS---HHHH-SSS
<i>B. vallismortis</i>	SSS-----H-----HHHHHHHHH-SSS-HHHHHH-----SSS---HHH-SSS
<i>Bacillus alcalophilus</i>	NRN-QAQRBNP-----QANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RNQ
<i>Bacillus amyloliquefaciens</i>	NRN-QANPRBNP-----AANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RQQ
<i>Bacillus anthracis</i>	QNN-QQNPRBNP-----QANQRNBBQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RNQ
<i>Bacillus atrophaeus</i>	NRN-QANPRBNP-----QANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RQQ
<i>Bacillus azotoformans</i>	NNN-QQNPRBNP-----NANRRBBNQ-NNQGBBQRAQN-G-----BPN-QNNGGNANR-RNQ
<i>Bacillus bataviensis</i>	NNN-QQNPRBNP-----AANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNANR-RNQ
<i>Bacillus cellulosilyticus</i>	NNN-QAQRBNP-----QANRRBBQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RNQ
<i>Bacillus cereus</i>	QNN-QBNPRBNP-----QANQRNBBQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RNQ
<i>Bacillus clausii</i>	NRN-QGNPRBNP-----NANBRBBQ-NNQGBBQRAQN-G-----GPN-RNNGGNANR-NNQ
<i>Bacillus coagulans</i>	NRN-QNNPRBNP-----AANRRBBQ-NNQGBBQRAQN-G-----BPN-RNNGGNANR-RBQ
<i>Bacillus coahuilensis</i>	NRN-QQQPRBNP-----BANBRBBQ-NNQGBBQRAQN-G-----BPN-QNNGGQNR-ANQ
<i>Bacillus cytotoxicus</i>	QNN-QBNPRBNP-----QANQRNBBQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RNQ
<i>Bacillus halodurans</i>	NNN-QANPRBNP-----NANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGQNR-RNQ
<i>Bacillus isronensis</i>	NNN-QANPRBNP-----QANRRBAQ-NNQGBBQRAQN-G-----BPN-QNNGGNQNR-RNQ
<i>Bacillus licheniformis</i>	NNN-QANPRBNP-----QANRRBANQ-NNQGBBQRAQN-G-----BPN-RNNGGQNR-RNQ
<i>Bacillus macauensis</i>	NNN-QANPRBNP-----NANQRBBNQ-NNQGBBQRAQN-G-----BPN-QNNGGNBNR-RNQ
<i>Bacillus megaterium</i>	NNN-QANPRBNP-----AANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RNQ
<i>Bacillus methanolicus</i>	NRN-QBNPRBNP-----AANRRBBQ-NNQGBBQRAQN-G-----BPN-RNNGGNANR-RQQ
<i>Bacillus mojavensis</i>	NRN-QANPRBNP-----QANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RQQ
<i>Bacillus nealsonii</i> AAU1	NNN-QANPRBNP-----BARARNBBNQ-NNGBBQRAQN-G-----BPN-RNNGGAANR-RNQ
<i>Bacillus pseudofirmus</i>	NNN-QANPRBNP-----NANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNANR-RNQ
<i>Bacillus pumilus</i>	NNN-QQQPRBNP-----NANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGQBNR-RNQ
<i>Bacillus selenitireducens</i>	NNN-QQQPRBNP-----QANBRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGQANR-RNQ
<i>Bacillus smithii</i>	NNN-QBNPRBNP-----AANBRBBQ-NNQGBBQRAQN-G-----BPN-RNNGGNANR-RNQ
<i>Bacillus sonorensis</i>	NRN-QANPRBNP-----QANRRBBQ-NNQGBBQRAQN-G-----BPN-RNNGGQNR-RQQ
<i>Bacillus sp. 10403023</i>	NNN-QANPRBNP-----NANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNANR-RNQ
<i>Bacillus sp. 1NLA3E</i>	NRN-QQNPRBNP-----AANRRBBQ-NNQGBBQRAQN-G-----BPN-RNNGGNANR-RNQ
<i>Bacillus sp. 2_A_57_CT2</i>	QNN-QBNPRBNP-----AANRRBBQ-NNQGBBQRAQN-G-----BPN-RNNGGNANR-RNQ
<i>Bacillus sp. 5B6</i>	NRN-QANPRBNP-----AANRRBBQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RQQ
<i>Bacillus sp. B14905</i>	NRN-QANPRBNP-----AANRRBBQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RQQ
<i>Bacillus sp. HYC-10</i>	NNN-QQQPRBNP-----NANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGQBNR-RNQ
<i>Bacillus sp. JS</i>	NRN-QANPRBNP-----QANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RQQ
<i>Bacillus sp. M 2-6</i>	---BBNARNN-GQA-GAAAQBRAARR-NNQGBBQRAQN-N-BNQPAARPR-RNNGGGANN-RNQ
<i>Bacillus sp. m3-13</i>	NNN-QANPRBNP-----NANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNBNR-RNQ
<i>Bacillus sp. NRRL B-14911</i>	NRN-QANPRBNP-----AANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNANR-RNQ
<i>Bacillus sp. SG-1</i>	NRN-QBNPRBNP-----AANRRBAQ-NNQGBBQRAQN-G-----BPN-QNNGGQANR-RNQ
<i>Bacillus stratosphericus</i>	RGN-BBNARNN-GQA-GAAAQBRAARR-NNQGBBQRAQN-N-BNQPAARPR-RNNGGGANN-RNQ
<i>Bacillus subtilis</i>	NRN-QANPRBNP-----QANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RQQ
<i>Bacillus thuringiensis</i>	NNN-QBNQBNNB-----AAQBBRBBQ-NNQGBBQRAQN-G-----BPN-RNNGGQQR-RNQ
<i>Bacillus vallismortis</i>	NRN-QANPRBNP-----QANRRBBNQ-NNQGBBQRAQN-G-----BPN-RNNGGNQNR-RQQ

Figure 7.

Subset of the complete structural alignment consisting of 40 *Bacillus* DHFR sequences truncated to residues 6 Å from bound NADPH and the active site. The first 40 sequences represent the actual residues in the sequence. The second set of 40 sequences are labeled for the type of secondary structure present: **H** for α -helix, **S** for β -sheet, and **T** for turns. The last set of 40 sequences are labeled for the type of amino acid residue class: **A** for acid, **B** for base, **N** for non-aromatic hydrophobic, **Q** for non-charged polar, **R** for aromatic, **P** for

proline, **G** for glycine. Residues >6 Å from bound NADPH and the active site are not shown and are designated by “–”.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

	5	20	25	30	35	40	45	50	54	90	95	110
												
<i>Paenibacillus alvei</i>	IWA-NGMPWHL	-----	NDMKFFRENT	-VVMGRKTYESL	-N-----	GAL-MIIGGAQIY	-LVT					
<i>Paenibacillus curdolanolyticus</i>	IWA-NRIPWRL	-----	ADMAFFKAYT	-VVMGRKTFESF	-P-----	KAL-VIIGGSEIY	-LVT					
<i>Paenibacillus dendritiformis</i>	IWA-NGMPWRL	-----	SDMAFFKAMT	-VVMGRKTYESL	-G-----	KAL-MVMGGAQIY	-LVT					
<i>Paenibacillus elgii</i>	ILA-NTLPWRL	-----	ADLVYFKQTT	-VLMGRKTFDSI	-G-----	KPL-FVIGGAEVY	-YIT					
<i>Paenibacillus lactis</i>	IWA-GTMPWHL	-----	RDLAFFKEKT	-IVMGRKTWDSF	-GS-----	KPL-MVIGGSQIY	-YCT					
<i>Paenibacillus mucilaginosus</i>	MFA-NRMPWYL	-----	ADFAYFKQTT	-VLMGRKTFDSL	-GG-----	KSL-FVIGGAEIF	-YIT					
<i>Paenibacillus peoriae</i>	IWA-NKLPWRL	-----	RDMAFFKEQT	-VLMGRKTWESF	-GG-----	RSL-MVIGGAEIY	-IVT					
<i>Paenibacillus polymyxa</i>	IWA-NKLPWRL	-----	RDMAFFKEQT	-VLMGRKTWESFGG	-----	KSL-MVIGGAEIY	-IVT					
<i>Paenibacillus popilliae</i>	IWA-NGMPWRL	-----	SDMAFFKAMT	-VVMGRNTYESL	-G-----	KAL-MVIGGTQIY	-LVT					
<i>Paenibacillus sp. Aloe-11</i>	IWA-NQLPWRL	-----	RDMAFFKEQT	-VLMGRKTWESFGG	-----	KSL-MVIGGAEIY	-IVT					
<i>Paenibacillus sp. HGF5</i>	IWA-GTMPWRL	-----	RDMAFFKEKT	-ILMGRKTWDSFGG	-----	KPL-MVIGGSQIY	-YCT					
<i>Paenibacillus sp. HGF7</i>	IFA-NKLPWHL	-----	ADLKYFKKMT	-ILMGRKTYDSI	-G-----	RPL-FVIGGTEVF	-YVT					
<i>Paenibacillus sp. JC66</i>	IAA-NDMPWSL	-----	KDLAYFMRTT	-VVSGRKNFEAM	-K-----	GPI-FIIGGAEIY	-LLT					
<i>Paenibacillus sp. JDR-2</i>	LWA-PDNPPKI	-----	EALHDFATHA	-FIVGRRTFEGMRNG	-----	GPG-LISGGAEIH	-IFN					
<i>Paenibacillus sp. oral taxon 7</i>	IWA-NQLPWRL	-----	ADLKFFKANT	-IVMGRKTWESMG	-----	KPL-MIIGGAGVF	-IVT					
<i>Paenibacillus terrae</i>	LWA--PDNPPYK	-----	ETLDDFATHA	-FIVGRRTFEGMRNGGG	-----	GPF-LISGGADIH	-IFN					
<i>Paenibacillus vortex</i>	IWA-GTMPWRL	-----	RDMAFFKEKT	-ILMGRKTWESF	-GS-----	KPL-MVIGGAQIY	-YCT					
<i>P. alvei</i>	SSS-----	H----	HHHHHHHH	--SSSSHHHHHH	-----	SSS--HHHH	-SSS					
<i>P. curdolanolyticus</i>	SSS-----	H----	HHHHHHHH	--SSSSHHHHHH	-----	SSS--HHHH	-SSS					
<i>P. dendritiformis</i>	SSS-----	H----	HHHHHHHH	--SSS-HHHHHH	-----	SSS--HHHH	-SSS					
<i>P. elgii</i>	SSS-----	H----	HHHHHHHH	--SSS-HHHHHH	-----	SSS--HHH	-SSS					
<i>P. lactis</i>	SSS-----	H----	HHHHHHHH	--SSS-HHHHHH	-----	SSS--HHHH	-SSS					
<i>P. mucilaginosus</i>	SSS-----	H----	HHHHHHHH	--SSS-HHHH	-----	SSS--HHH	-SSS					
<i>P. peoriae</i>	SSS-----	H----	HHHHHHHH	--SSS-HHHH	-----	SSS--HHHH	-SSS					
<i>P. polymyxa</i>	SSS-----	H----	HHHHHHHH	--SSS-HHHH	-----	SSS--HHH	-SSS					
<i>P. popilliae</i>	SSS-----	H----	HHHHHHHH	--SSSSHHHHHH	-----	SSS--HHHH	-SSS					
<i>P. sp. Aloe-11</i>	SSS-----	H----	HHHHHHHH	--SSS-HHHHHH	-----	SSS--HHH	-SSS					
<i>P. sp. HGF5</i>	SSS-----	H----	HHHHHHHH	--SSS-HHHHH	-----	SSS--HHHH	-SSS					
<i>P. sp. HGF7</i>	SSS-----	H----	HHHHHHHH	--SSS-HHHHH	-----	SSS--HHH	-SSS					
<i>P. sp. JC66</i>	SSS-----	H----	HHHHHHHH	--SSS-HHH	-H-----	SSS--HHH	-SSS					
<i>P. sp. JDR-2</i>	SSS-----	H----	HHHHHHHH	--SSS-HHHH	-----	SS--HHH	-SSS					
<i>P. sp. oral taxon 786</i>	SSS-----	H----	HHHHHHHH	--SSS-HHHHHH	-----	SSS--HHHH	-SSS					
<i>P. terrae</i>	SSS-----	H----	HHHHHHHH	--SSSSHHH	-----	SS--HHHH	-SSS					
<i>P. vortex</i>	SSS-----	H----	HHHHHHHH	--SSS-HHHH	-----	SSS--HHHH	-SSS					
<i>Paenibacillus alvei</i>	NRN-QGQPRBN	-----	QAQBRBAQQ	-NNQGBBQRAQN	-Q-----	GNN-QNNGGNQNR	-NNQ					
<i>Paenibacillus curdolanolyticus</i>	NRN-QBNPRBN	-----	NAQNRRBNRQ	-NNQGBBQRAQR	-P-----	BNN-QNNGGQANR	-NNQ					
<i>Paenibacillus dendritiformis</i>	NRN-QGQPRBN	-----	QAQNRRBNQ	-NNQGBBQRAQN	-G-----	BNN-QNNGGNQNR	-NNQ					
<i>Paenibacillus elgii</i>	NRN-QQNPRBN	-----	NANNRRBQQ	-NNQGBBQRAQN	-G-----	BPN-QNNGGNANR	-RNQ					
<i>Paenibacillus lactis</i>	NRN-GQQPRBN	-----	BANNRRBABQ	-NNQGBBQRAQR	-GQ-----	BPN-QNNGGQQR	-RQQ					
<i>Paenibacillus mucilaginosus</i>	QRN-QBQPRBN	-----	NARNRRBQQ	-NNQGBBQRAQN	-GG-----	BQN-QNNGGNANR	-RNQ					
<i>Paenibacillus peoriae</i>	NRN-QBNPRBN	-----	BAQNRRBAQQ	-NNQGBBQRAQR	-GG-----	BQN-QNNGGNANR	-NNQ					
<i>Paenibacillus polymyxa</i>	NRN-QBNPRBN	-----	BAQNRRBAQQ	-NNQGBBQRAQR	-GG-----	BQN-QNNGGNANR	-NNQ					
<i>Paenibacillus popilliae</i>	NRN-QGQPRBN	-----	QAQNRRBNQ	-NNQGBBQRAQN	-G-----	BNN-QNNGGQQR	-NNQ					
<i>Paenibacillus sp. Aloe-11</i>	NRN-QQNPRBN	-----	BAQNRRBAQQ	-NNQGBBQRAQR	-GG-----	BQN-QNNGGNANR	-NNQ					
<i>Paenibacillus sp. HGF5</i>	NRN-GQQPRBN	-----	BAQNRRBABQ	-NNQGBBQRAQR	-GG-----	BPN-QNNGGQQR	-RQQ					
<i>Paenibacillus sp. HGF7</i>	NRN-QBNPRBN	-----	NANBRRBBQ	-NNQGBBQRAQN	-G-----	BPN-QNNGGQANR	-RNQ					
<i>Paenibacillus sp. JC66</i>	NRN-QAQPRQN	-----	BANNRRBQQ	-NNQGBBQRAQN	-B-----	GPN-QNNGGNANR	-NNQ					
<i>Paenibacillus sp. JDR-2</i>	NRN-PAQPPBN	-----	ANNBARNQB	-RNNGBBQRAQBQ	-GG-----	GPG-QNNGGNANB	-NRQ					
<i>Paenibacillus sp. oral taxon 7</i>	NRN-QQNPRBN	-----	NANBRRBNQ	-NNQGBBQRAQR	-GQ-----	BPN-QNNGGNANR	-NNQ					
<i>Paenibacillus terrae</i>	NRN--PAQPPRB	-----	AQNAARNQB	-RNNGBBQRAQBQ	-GG-----	GPR-QNNGGNANB	-NRQ					
<i>Paenibacillus vortex</i>	NRN-GQQPRBN	-----	BAQNRRBABQ	-NNQGBBQRAQR	-GQ-----	BPN-QNNGGNQNR	-RQQ					

Figure 8.

Subset of the complete structural alignment consisting of 17 *Paenibacillus* DHFR sequences truncated to residues 6 Å from bound NADPH and the active site. The first 18 sequences represent the actual residues in the sequence. The second set of 17 sequences are labeled for the type of secondary structure present: **H** for α -helix, **S** for β -sheet and **T** for turns. The last set of 17 sequences are labeled for the type of amino acid residue class: **A** for acid, **B** for base, **N** for non-aromatic hydrophobic, **Q** for non-charged polar, **R** for aromatic, **P** for

proline, **G** for glycine. Residues >6 Å from bound NADPH and the active site are not shown and are designated by “–”.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 9.

WebLogo representation of site-by-site sequence conservation for DHFR alignments truncated to residues 6Å from bound NADPH and the active site.

Figure 10.

Phylogenetic tree generated from structural alignment of complete DHFR sequences from 146 *Lactobacillales* species (blue), 123 *Bacillales* species (black), and 71 plasmid sequences (red) using MEGA6. 5 plasmids sequences originating from *Bacilli* species are tagged and labelled.

Subset of the phylogenetic presenting the cluster of non-*Bacilli* plasmid species and two chromosomal sequences (e.g., *M. caseolyticus* and *E. faecalis*). Several plasmids of unknown origin are included and are listed without a species.

Figure 12.

Phylogenetic tree generated from structural alignment of complete DHFR sequences from **a.** 146 *Lactobacillales* species and **b.** 123 *Bacillales* species using MEGA6.

Figure 13.
Key residues involved in the network of coupled DHFR protein motions [47–50].

Table 1

DHFR x-ray structures used in the total structure-based alignment.

PDB ID	Organism
1ZDR	<i>Geobacillus stearothermophilus</i>
2QK8	<i>Bacillus anthracis str. Sterne</i>
2W9S	<i>Staphylococcus aureus</i>
3IX9	<i>Streptococcus pneumoniae</i>
3E0B	<i>Bacillus anthracis</i>
3JTW	<i>Pediococcus pentosaceus ATCC 25745</i>
3SQY	<i>Staphylococcus aureus</i>

Table 2

Site-specific frequencies of amino acid, secondary structure type and residue class calculated for the entire structural based alignment positions, the active site alignment positions and the non-active alignment positions of 124 DHFRs belong to *Bacillales* families and 149 DHFRs belong to *Lactobacillales* families (Figure 6).

Prokaryotic order	Complete Alignment (% conservation $\pm$ % standard deviation)	Active Site Alignment (% conservation $\pm$ % standard deviation)	Non-Active Site Alignment (% conservation $\pm$ % standard deviation)
i. <i>Bacillales</i>			
Amino Acid Conservation	15.2 $\pm$ 0.9	24.0 $\pm$ 2.5	11.7 $\pm$ 0.5
Secondary Structure Conservation	60.8 $\pm$ 3.8	72.9 $\pm$ 3.8	55.6 $\pm$ 2.4
Residue Class Conservation	28.1 $\pm$ 1.4	38.8 $\pm$ 3.1	23.7 $\pm$ 1.2
ii. <i>Lactobacillales</i>			
Amino Acid Conservation	13.6 $\pm$ 1.1	22.9 $\pm$ 3.2	10.2 $\pm$ 0.5
Secondary Structure Conservation	44.7 $\pm$ 1.3	49.5 $\pm$ 2.5	42.8 $\pm$ 1.5
Residue Class Conservation	27.3 $\pm$ 1.4	38.9 $\pm$ 3.4	21.9 $\pm$ 1.0

Table 3

Site-specific frequencies of amino acid, secondary structure type and residue class calculated for the entire structural based alignment positions, the active site positions and the nonactive positions of *Bacillales* and *Lactobacillales* genera.

Prokaryotic Genera	Complete Alignment (% conservation $\pm$ % standard deviation)	Active Site Alignment (% conservation $\pm$ % standard deviation)	Non-Active Site Alignment (% conservation $\pm$ % standard deviation)
<i>Bacillales</i>:			
(i) <i>Bacillaceae</i>			
<i>Bacillus</i> (n= 40)			
Amino Acid Conservation	22.7 $\pm$ 1.9	32.4 $\pm$ 2.8	19.7 $\pm$ 1.0
Secondary Structure Conservation	73.6 $\pm$ 2.2	78.9 $\pm$ 3.8	71.7 $\pm$ 2.6
Residue Class Conservation	37.00 $\pm$ 1.9	51.8 $\pm$ 4.1	32.2 $\pm$ 1.8
<i>Geobacillus</i> (n= 8)			
Amino Acid Conservation	82.6 $\pm$ 2.0	89.6 $\pm$ 3.0	79.5 $\pm$ 2.5
Secondary Structure Conservation	98.1 $\pm$ 0.8	100 $\pm$ 0.0	97.4 $\pm$ 1.1
Residue Class Conservation	90.8 $\pm$ 1.6	94.8 $\pm$ 2.2	88.8 $\pm$ 2.1
<i>Staphylococcaceae</i>			
<i>Staphylococcus</i> (n= 18)			
Amino Acid Conservation	53.0 $\pm$ 2.3	64.2 $\pm$ 4.4	48.3 $\pm$ 2.6
Secondary Structure Conservation	84.8 $\pm$ 2.0	94.4 $\pm$ 2.4	81.2 $\pm$ 2.5
Residue Class Conservation	67.3 $\pm$ 2.4	81.4 $\pm$ 3.9	61.7 $\pm$ 2.8
<i>Listeriaceae</i>			
<i>Listeria</i> (n= 7)			
Amino Acid Conservation	55.7 $\pm$ 2.2	66.5 $\pm$ 3.9	51.3 $\pm$ 2.5
Secondary Structure Conservation	76.7 $\pm$ 1.9	78.1 $\pm$ 2.7	76.1 $\pm$ 2.4
Residue Class Conservation	66.7 $\pm$ 2.2	76.1 $\pm$ 3.3	62.9 $\pm$ 2.7
<i>Paenibacillaceae</i>			
<i>Paenibacillus</i> (n= 17)			
Amino Acid Conservation	28.3 $\pm$ 1.5	40.5 $\pm$ 3.7	23.8 $\pm$ 1.4
Secondary Structure Conservation	64.1 $\pm$ 2.5	80.0 $\pm$ 4.0	58.3 $\pm$ 2.9
Residue Class Conservation	41.4 $\pm$ 2.0	58.2 $\pm$ 4.1	35.1 $\pm$ 2.0
(ii) <i>Lactobacillales</i>:			
<i>Streptococcaceae</i>			
<i>Streptococcus</i> (n= 56)			
Amino Acid Conservation	29.7 $\pm$ 1.7	43.5 $\pm$ 3.8	24.2 $\pm$ 1.7
Secondary Structure Conservation	68.7 $\pm$ 2.1	80.1 $\pm$ 3.6	63.2 $\pm$ 2.4
Residue Class Conservation	44.4 $\pm$ 2.2	62.8 $\pm$ 4.5	37.1 $\pm$ 2.2
<i>Lactobacillaceae</i>			
<i>Lactobacillus</i> (n= 31)			
Amino Acid Conservation	22.53 $\pm$ 1.5	35.0 $\pm$ 4.0	17.6 $\pm$ 1.0
Secondary Structure Conservation	52.89 $\pm$ 1.8	56.5 $\pm$ 3.2	51.0 $\pm$ 2.2
Residue Class Conservation	35.94 $\pm$ 1.8	50.8 $\pm$ 4.4	30.0 $\pm$ 1.5

Prokaryotic Genera	Complete Alignment (% conservation $\pm$ % standard deviation)	Active Site Alignment (% conservation $\pm$ % standard deviation)	Non-Active Site Alignment (% conservation $\pm$ % standard deviation)
<i>Leuconostocaceae</i>			
<i>Leuconostoc</i> (n=9)			
Amino Acid Conservation	58.6 $\pm$ 2.5	74.7 $\pm$ 4.5	52.5 $\pm$ 2.7
Secondary Structure Conservation	80.1 $\pm$ 2.0	83.9 $\pm$ 3.5	79.2 $\pm$ 2.4
Residue Class Conservation	71.3 $\pm$ 2.3	86.1 $\pm$ 3.5	65.7 $\pm$ 2.8
<i>Weissella</i> (n= 6)			
Amino Acid Conservation	58.8 $\pm$ 2.5	78.1 $\pm$ 4.3	50.8 $\pm$ 2.7
Secondary Structure Conservation	77.6 $\pm$ 2.0	89.1 $\pm$ 3.2	73.7 $\pm$ 2.4
Residue Class Conservation	68.3 $\pm$ 2.3	84.5 $\pm$ 3.6	62.0 $\pm$ 2.7
<i>Enterococcaceae</i>			
<i>Enterococcus</i> (n= 22)			
Amino Acid Conservation	27.2 $\pm$ 1.5	36.8 $\pm$ 3.5	43.1 $\pm$ 2.0
Secondary Structure Conservation	63.9 $\pm$ 2.1	69.7 $\pm$ 3.8	58.0 $\pm$ 4.4
Residue Class Conservation	24.9 $\pm$ 1.4	61.5 $\pm$ 2.5	36.5 $\pm$ 1.9

Table 4

Percent amino acid identity at each sequence position identified from the x-ray crystal structure of *E. coli* DHFR (1RX2.pdb) as being part of a network of coupled motions [47–50]. Highly conserved positions are in red. Bold and italics are used to highlight percent amino acid identity values that are of interest.

Position	Secondary Structure	Amino Acid	Chromosomal + Plasmid (%)	Percent Chromosomal (%)	Percent Bacillales (%)	Percent Lactobacillales (%)	Percent Plasmid (%)
14	β sheet	Ile (<i>E. coli</i>)	93.2	94.1	94.4	94.0	90.0
15	β sheet	Gly (<i>E. coli</i>)	93.5	93.0	94.4	92.0	95.0
16		Met (<i>E. coli</i>)	0.0	0.0	0.0	0.0	0.0
		Lys	38.8	48.0	51.6	45.0	7.5
		Asn	15.0	3.7	3.2	4.0	53.8
17		Glu (<i>E. coli</i>)	11.3	13.6	13.0	14.1	3.8
		Asp	33.4	42.1	54.8	31.5	3.8
		Gly	19.8	1.8	0.8	2.7	81.3
18		Asn (<i>E. coli</i>)	40.2	49.8	85.5	20.1	7.5
		Pro	17.3	1.1	1.6	0.7	72.5
		Gly	26.3	31.4	8.1	50.3	10.0
19		Ala (<i>E. coli</i>)	1.4	1.5	2.4	0.7	1.3
		Asp	30.3	17.2	32.3	4.7	73.8
20		Met (<i>E. coli</i>)	11.1	14.3	23.4	6.7	0.0
		Leu	56.4	70.7	62.9	77.2	7.5
		Ile	27.5	9.9	8.9	10.7	87.5
21		Pro (<i>E. coli</i>)	94.1	93.4	92.7	94.0	96.3
31	α helix	Phe (<i>E. coli</i>)	86.1	83.9	71.8	94.0	93.8
40	β sheet	Val (<i>E. coli</i>)	16.4	15.0	17.7	12.8	21.3
		Ile	41.9	52.8	54.8	51.0	5.0
		Leu	30.0	18.3	21.0	16.1	70.0
41	β sheet	Ile (<i>E. coli</i>)	24.1	27.8	37.9	19.5	11.3
		Val	34.8	38.5	50.8	28.2	22.5
		Leu	38.8	31.1	7.3	51.0	65.0
42	β sheet	Met (<i>E. coli</i>)	76.8	90.5	92.7	88.6	30.0
		Val	16.2	1.8	1.6	2.0	65.0

Position	Secondary Structure	Amino Acid	Chromosomal + Plasmid (%)	Percent Chromosomal (%)	Percent Bacillales (%)	Percent Lactobacillales (%)	Percent Plasmid (%)
43	β sheet	Gly (<i>E. coli</i>)	89.6	98.5	99.2	98.0	98.8
60	β sheet	Ile (<i>E. coli</i>)	40.2	47.3	55.7	40.3	16.3
		Leu	19.0	23.4	1.6	41.6	3.8
		Val	19.8	21.6	35.5	10.1	13.8
		Ala	14.5	0.0	0.0	0.0	63.8
61	β sheet	Ile (<i>E. coli</i>)	25.2	31.1	19.4	40.9	5.0
		Val	73.7	68.5	79.8	59.1	91.3
62	β sheet	Leu (<i>E. coli</i>)	56.7	68.1	54.8	79.2	18.0
		Val	26.6	13.6	26.6	2.7	71.3
63	β sheet	Ser (<i>E. coli</i>)	17.3	11.7	7.3	15.4	36.3
		Thr	79.3	86.5	81.9	81.9	55.0
100	α helix	Tyr (<i>E. coli</i>)	52.1	39.9	45.2	35.6	93.8
		Phe (<i>E. coli</i>)	39.4	50.6	51.6	49.7	1.3
121		Gly (<i>E. coli</i>)	89.1	87.5	84.7	89.8	94.7
122	β sheet	Asp (<i>E. coli</i>)	95.1	95.6	96.0	95.2	93.4
123	β sheet	Thr (<i>E. coli</i>)	62.3	76.8	78.2	75.5	10.5
		Val	17.9	10.0	2.4	16.3	46.1

Table 5

Percent frequencies of the amino acids alanine (A), cysteine (C), histidine (H), glutamine (Q) and threonine (T) versus the percent frequencies of the amino acids glutamic acid (E), isoleucine (I), lysine (K), tyrosine (Y) and valine (V). EIKYV tend to be overrepresented and ACHQT underrepresented in the proteins of thermophiles as compared to those from mesophiles [61].

	<i>Bacilli</i> (%)	<i>Plasmid</i> (%)
ACHQT	19	17
EIKYV	32	33

	<i>Bacillales</i> (%)	<i>Bacillus</i> (%)	<i>Geobacillus</i> (%)	<i>Staphylococcus</i> (%)	<i>Listeria</i> (%)	<i>Paenibacillus</i> (%)	<i>Brevibacillus</i> (%)
ACHQT	17	16	17	17	19	17	16
EIKYV	33	35	33	32	35	30	35

	<i>Lactobacillales</i> (%)	<i>Streptococcus</i> (%)	<i>Leuconostoc</i> (%)	<i>Weissella</i> (%)	<i>Enterococcus</i> (%)	<i>Lactobacillus</i> (%)
ACHQT	20	21	22	22	20	21
EIKYV	31	33	29	30	31	29