A Computational Method that Uses Protein Structures to Detect Horizontal Gene Transfer

Peter Z. Revesz, Swetha Billa, Mark A. Griep and Venkat R. B. Santosh

***Abstract*—** No current method to detect *Horizontal Gene Transfer* (HGT) events between organisms is accurate in all cases. This paper presents a novel computational method that uses protein structures to detect HGT. The use of protein structures instead of protein sequences increases the accuracy of our method. The new method uses Z-score similarities between the protein structures and DaliLite to search efficiently for similarities in both protein sequences and structures. In addition, the java viewer tool Jmol is used for visual structural comparisons and sequence alignment. The experimental results include six previously unreported cases of HGTs between various *Firmicutes* and *Proteobacteria* bacteria. Various methods of handling false positives are also described.

***Keywords*—** Bacteria, COG, horizontal gene transfer, PDB, protein database.

I. Introduction

IN nature, most gene transfer occurs between two similar or closely related species via typical routes of reproduction, such as cross pollination of plants and interbreeding of animals. Such transfer is also called *vertical gene transfer*, since traits are vertically passed from parent to offspring. However, sometimes genes also move between different species, such as bacteria and plants, through a process unrelated to reproduction that is known as *horizontal gene transfer* (HGT). HGT can also occur between two closely related species.

HGT was first described in 1959 in a Japanese publication about the transfer of antibiotic resistance from one bacterium to another (Akiba et al. [1]). The phenomenon of HGT is quite significant in prokaryotes and some unicellular eukaryotes. Most work on biological evolution has focused on the role of genetic mutations [12] and consequent changes in the active sites of proteins [11], whereas the importance of HGTs has not been extensively studied [7] even though it plays a major role in evolution and medicine. For instance, it is the major contributor of bacterial evolution, enabling species to acquire genes to adapt to new environments. Therefore, further study of HGT and its implications is necessary to understand the effects of HGT in biology and to study techniques to enable or disable the process based on its effects.

This paper is organized as follows. Section II describes the basic concepts of horizontal gene transfer. Section III reviews previously proposed horizontal gene transfer methods. Section IV describes our new method for identifying horizontally transferred genes. Section V describes the experimental results and analysis. Finally, Section VI presents some conclusions and future work.

Peter Z. Revesz is a professor in the Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588 USA (phone: 402-472-3488; fax: 402-472-7767; email: revesz@cse.unl.edu). Mark A. Griep is a professor in the Department of Chemistry, University of Nebraska-Lincoln, NE 68588, USA. Swetha Billa and Venkat Santosh are former graduate students at the University of Nebraska-Lincoln.

II. Basic Concepts of Horizontal Gene Transfer

A. How to determine HGT?

For a successful natural horizontal gene transfer, it would require stable integration of the gene into the genome, no disturbance of regulatory or genetic structures, expression and successive production of a functional protein (Susanna et al. [23]). There are two approaches to determine Horizontal Gene Transfer in a genome: (i) Phylogenetic Comparison and (ii) Parametric Comparison. In Phylogenetic Comparison, different organisms are compared to find the similarity or dissimilarity. While in Parametric Comparison, genes that appear to be anomalous in their current genome contexts are considered transferred from a foreign source (Lawrence and Ochman [10]).

B. Why is it important to study HGT?

HGT plays a major role in bacterial evolution. *Antibiotic resistance (AR)* or antimicrobial resistance is a type of drug resistance where a microorganism is able to survive exposure to an antibiotic. The development of antibiotic resistance characteristics is often observed to develop much more rapidly than simple vertical inheritance of traits. Hence it is believed that development of antibiotic resistance among different bacteria is the result of HGT, as one bacterial cell acquires resistance and transfers those genes to other bacterial species (Frank-Kamenetskii [5]).

Antibiotic resistance (AR) poses a significant problem for the public health in the world. As more and more bacterium develop resistance to drugs, the need for alternative treatments

increases. Controlling *antibiotic resistance (AR)* in bacteria requires investigation of the antibiotic resistance mechanism (Song et al. [22]). Hence studies on HGT will help provide a greater insight on how this can be curbed.

C. Mechanisms of HGT

Horizontal gene transfer could occur by several mechanisms between organisms. These basic mechanisms are the following:

- Transformation - The uptake of naked DNA is a common mode of horizontal gene transfer that can mediate the exchange of any part of a chromosome; this process is most common in bacteria that are naturally transformable; typically only short DNA fragments are exchanged.
- Conjugation - The transfer of DNA mediated by conjugal plasmids or conjugal transposons; requires cell to cell contact but can occur between distantly related bacteria or even bacteria and eukaryotic cells; can transfer long fragments of DNA.
- Transduction - The transfer of DNA by phage requires that the donor and recipient share cell surface receptors for phage binding and thus is usually limited to closely related bacteria; the length of DNA transferred is limited by the size of the phage head. Gene transfer agents, virus-like elements encoded by the host, that are found in the *alphaproteobacteria* order *Rhodobacterales*.

Each of these methods of genetic exchange can introduce sequences of DNA that share little homology with the remaining DNA of the recipient cell. If there are homologous sequences shared between the donor DNA and the recipient chromosome, the donor sequences can be stably incorporated into the recipient chromosome by genetic recombination. If the homologous sequences flank sequences that are absent in the recipient, the recipient may acquire an insertion from another strain of unrelated bacteria. Such insertions can be small or quite large. Large insertions that have been acquired from another bacterium (often inferred from differences in GC content or codon usage) and are absent from related strains of bacteria are called "islands."

III. Review of Previous HGT Detection Methods

A. Compositional Methods

A horizontally transferred gene can contain recognizable signatures of its previous location. Compositional methods use atypical nucleotide (Lawrence and Ochman [8]), atypical codon usage patterns (Lawrence and Ochman [9]) or their combination (Tsirigos and Rigoutsos [27]) to detect horizontally transferred genes. Since the horizontally transferred genes adopt the signatures of their new host genome over time, these methods are most useful for identifying genes that have been transferred fairly recently. These methods are easily applied to completely sequenced genomes. However, high rates of false positives and negatives have been observed in these methods.

B. Phylogeny-Based Method

Phylogeny-based detection of HGT is one of the most commonly used approaches. It is based on discrepancies in the gene trees generated by phylogenetic algorithms, such as those shown in Revesz [14, 15]. When the gene and species trees are compared, HGT events are invoked to explain the discrepancies because the evolutionary history of the gene does not agree with the species phylogeny.

One drawback of using this method is the absence of firm criteria for uniquely identifying the HGT scenario. In addition, the bacterial phylogenetic trees are not definitive. The quality of the phylogenetic reconstruction is usually done statistically, which has an impact on the HGT detection and sometimes underestimates or overestimates the number HGT events.

C. Distance-Based Detection of HGT

The Distance-Based method incorporates distances typically used in the Phylogeny-based detection of HGT rather than the trees themselves. This method has many of the strengths of Phylogenetic approaches but avoids some of their pitfalls.

This method uses only the pair-wise distance instead of building the whole trees as in the Phylogeny-based approach, which makes the distance-based approach run much more quickly, allowing scanning of whole genomes. As there is no 'consensus' tree in this method, it does not suffer in the cases where no tree matches all of the given data. Instead it just compares the pair-wise distance between species and thus called the Distance-Based method for detecting HGT.

D. Composition-Based Detection of HGT

Although the Phylogeny-Based detection methods are more powerful than the Composition-based methods, especially when the donor is closely related to the recipient genome, they are very time consuming. The four methodologies commonly employed by Composition-based methods to detect HGT are based on:

- The codon adaptation index, codon usage, and GC percentage (CAI/GC)
- The distributional profile
- The Bayesian model
- The first-order Markov model

All these methods attempt to identify genes with anomalous compositions. The genomic DNA of different organisms has a particular mean G+C content. Genes in a given genome use the same coding strategy for choices among synonymous codons. That is, the bias in codon usage is species-specific.

Statistical methods have been developed to use these anomalies in the GC content to detect HGT.

One notable problem with the compositional approaches is that the codon usage and GC content give different results, each detecting a different set of possible horizontal gene transfers that do not match with each other.

A study on these methods shows that both the Bayesian models and the Markov models can detect HGT when closely related species are studied, though the Markov model is more effective. The CAI/GC method appears to be a less effective approach in the detection of HGT but is very effective in detecting HGT when the foreign genes are from a phylogenetically distant species. The distribution profile method exhibited an average detection level of approximately 50% for foreign genes but failed to go beyond 80% threshold of detection.

If a compositional method with an accurate detection level of horizontally transferred genes can be developed, it could avoid the application of exhaustive processes and slow phylogenetic reconstructions used in the phylogeny-based approach. The compositional method can be used together with fast bacterial genome sequencing [13, 18].

IV. A New Method to Detect HGT

We recently devised a method based on approximate search on protein structures to detect HGT among bacteria (see the preliminary publications Billa et al. [3] and Santosh et al. [20]). Below we extend the applicability of our earlier method to be generally applicable to detect horizontal gene transfer and show many more examples of HGTs.

Our method makes use our observation (Shortridge et al. [21]) that protein function and structure is conserved at a much higher level that protein and DNA sequences. Hence, our hypothesis is that horizontally transferred genes can be reliably identified through an examination of protein structure comparisons between organisms. Although the transferred gene sequence can change, it changes in a way that leaves the resulting protein structure and function intact. Hence, our experiment is to identify structural anomalies between proteins with conserved function that are isolated from divergent organisms.

To identify these protein structure anomalies, we make use of the Cluster of Orthologous Group (COG) classification [24, 25]. They are identified as the single-best-hit within each organism after a protein sequence comparison of all genes against all proteins coded by other organisms in the dataset. According to the COG classification, proteins with similar functionality share a COG number. According to evolutionary theory, proteins with the same COG number should have similar structures.

In our experiments, we consider two phyla of bacteria: (i) *Firmicutes,* and (ii) *Proteobacteria*. Most of the *Firmicutes* bacteria are Gram-positive. They are found in various environments and the group includes some notable pathogens. *Proteobacteria* is the largest and most diverse in the domain bacteria. This is an environmentally, geologically and evolutionarily important group. Most of the bacteria in *Proteobacteria* group are Gram-negative. *Firmicutes* and *Proteobacteria* diverged millions of years ago, and underwent random mutations during which they retained most of their native characteristics (Shortridge et al. [21]). Evidence of protein characteristics of bacteria belonging to one phyla being similar to the protein characteristics of bacteria in another phyla would indicate horizontal gene transfer.

A. Description of the Method

We compared the *Firmicutes* bacterium *Bacillus subtilis* with *Proteobacteria* bacteria. We chose *Bacillus subtilis* because it has a large number of identified structures in the biological databases that were available for our research.

Stage 1: As the first stage of the method, we needed information about all the proteins that were studied in each of these bacteria. To get this data we made use of the PROFESS database (Triplet et al. [26]). Querying the PROFESS database, we obtained a list of proteins from each of the bacteria in our set and the COGs to which they belong. There were 494 proteins for *Bacillus subtilis* and 3264 proteins for *Escherichia coli* documented in the PDB database. When we perform structural comparison for these two bacteria we are interested only in the common COGs between them. There are 88 common COGs among them.

Stage2: As the second stage of the method, we performed a structural comparison of the proteins. To perform pairwise structural comparison of proteins within each organism within the same COG, we would have $n\frac{(n-1)}{2}$ pairs of PDB IDs, where n is the number of proteins in a given COG for a given organism. The DaliLite program was used to compare the protein structures. It takes the input of two PDB ids and applies structural comparison algorithms and reports a Z-score, which is the index for measuring structural similarity in proteins. For comparison of proteins within a COG number in the two different organisms under consideration, we would have the cross product of the number of PDB IDs in that particular COG in each of the organisms. This has to be repeated for all the common COGs in the two organisms.

For all pairs of PDB IDs obtained above, an alignment algorithm is applied to get a Z-score measure for each pair. The DaliLite tool is used to obtain this. When a pair-wise comparison is done using DaliLite it gives results based on multiple variations in the alignments of the two proteins. We choose the result set with the highest Z-score. In other words

Table 1. Example of Documented Data.

COG Number	*B. subtilis*	*E. coli*	Comparison Z-score	*B. subtilis* Z-score Normalized	*E. coli* Z-score Normalized	Comparison Z-score Normalized
454	12.09	35.7	9.71	0.34	1	0.27

we use the score from the best alignment. The average Z-score is calculated within each COG. These average Z-scores are then normalized. By analyzing these normalized values, we can identify anomalous structures.

Since the average Z-scores are calculated within the same COGs, we expect the average Z-score for the same COG in two different organisms to be equal or have very little difference. If any large difference in the values of the average Z-score with in a same COG appears in the two organisms under consideration then it is unusual and further inspection of the proteins in that particular COG is required. For our research the threshold value for identifying this anomalous behavior is chosen to be 75%. If the average Z-score value of the first organism is less than or equal to 75% of the average Z-score value of the second organism then that particular COG is identified as an anomaly. After identifying all such COGs further analysis of structures needs to be done to identify a possible candidate of HGT. Table 1 shows sample data resulting from the comparison of *Bacillus subtilis* and *Escherichia coli*. In this example, COG 454 is considered anomalous because the average Z-score of *Bacillus subtilis* is only 39% of the average Z-score of *Escherichia coli*, which falls below our considered threshold value.

V. Experimental Results

Analysis of proteins from *Bacillus subtilis*, which is gram positive, with other Gram-negative organisms needs to be done. The protein structures of *Bacillus subtilis* were compared with all the *Proteobacteria* (Gram negative) bacteria having more than 40 crystallized proteins in the PDB. There were 19 Gram-negative organisms with number of crystallized proteins in them greater than 40. Of these nineteen Gram-negative organisms only five organisms had matching COG numbers with the ones in *Bacillus subtilis*. The Gram-negative organisms were:

1. *Escherichia coli*
2. *Pseudomonas aeruginosa*
3. *Pseudomonas putida*
4. *Haemophilus influenzae*
5. *Helicobacter pylori*

The protein structures of *Bacillus subtilis* are compared with the above gram-negative organisms. This comparison is performed only for the common COGs between the two different classes of bacteria, i.e., one Gram positive and five Gram negative organisms. Table 2 gives a summary of the proteins structure comparisons performed in our preliminary analysis.

A. Summary of Suspected Horizontal Gene Transfers

Table 3 summarizes the findings of a further detailed analysis of all the proteins in these candidate HGTs. In particular, the proteins 1VI0 in COG-1309 and 2GGE in COG-4948 were identified as likely HGTs to *Bacillus subtilis* and proteins 2DY0 in COG-503, 1M33 in COG-596, 1O98 & 1O8C in COG-604 and 3MEF in COG-1278 as possible HGTs from *Bacillus subtilis*. The ΔZ-score shown in Table 3 is the difference of the average comparison Z-scores of the HGT suspected protein with all the proteins in the opposite Gram organism and the average Z-scores of all the other proteins in the same COG as the suspected protein with all the proteins in the opposite Gram organism.

B. Detailed Analysis of COG-503

COG-503 from *E. coli* includes five structures of Xanthine Transferase (1A95, 1A96, 1A97, 1A98, 1NUL) and one structure of Adenine Transferase (2DY0). Among these the Adenine Transferase had the most divergent structure according to the Z-score comparison; an average of 10 compared to an average of 25 for all the others. COG-503 from *Bacillus subtilis* includes four structures, one repressor (1O57) and three Xanthine Transferase (1P96, 1Y0B, 2FXV). As Table 4 shows, all of the four proteins were closely related according to their Z-scores. The *E. coli* protein 2DY0 was more similar to the four Bacillus subtilis proteins than it was to the *E. coli* proteins. Therefore, it is an excellent candidate to be a new horizontally transferred gene product.

To further confirm this is a genuine case of HGT, we compared visually the 3-D structure of the protein 2DY0 and a sequence alignment with the proteins in *Bacillus subtilis* and other proteins in *E. coli* in the COG-503 as shown in Figs. 1 and 2.

C. Detailed Analysis of COG-596

The suspected PDB-IDs were more similar to the proteins in *Bacillus subtilis* than it had been to the *E. coli* proteins. COG-596 from *E. coli* includes two structures one of a BioH protein (1M33) and one of a C-C bond hydrolase (1U2E). COG-596 from *Bacillus subtilis* includes two structures of the same Sigma factor SigB regulation protein. *E. coli* protein 1M33 was more similar to the protein in *Bacillus subtilis* than it had been to the *E. coli* protein.

Table 2 Summary of the HGT candidates among the compared protein structures.

COG	Number of Structures in Bacterial		Findings
	E. coli	*Bacillus subtilis*	
500	2	2	Statistically promising example of HGT, provided there were more structures.
503	6	4	Most likely a good example of HGT.
526	38	13	Substrate diversity.
596	2	2	Most likely a good example of HGT.
604	3	2	Most likely a good example of HGT.
789	6	2	A closer examination revealed it was the result of protein fragments in *E. coli*.
840	2	2	Two Gram-positive protein structures are not similar to any of the Gram-negatives.
1278	2	4	Most likely a good example of HGT.
1609	42	2	Substrate diversity.
	E. coli	*Staphylococcus aureus*	
441	9	2	Protein fragments in *E. coli* and the two Gram-positive proteins are not different.
526	38	4	Substrate diversity.
614	8	2	The two Gram-positive protein structures are not different and have similar Z-scores to all the protein structures in Gram-negative.
5640	15	3	The three Gram-positive protein structures are not different and have similar Z-scores to all the protein structures in Gram-negative.
	E. coli	*Bacillus*	
80	30	2	NULL values of Z-scores, Substrate diversity, Protein fragments*.
266	6	12	Substrate diversity, confirmation changes.
508	6	8	NULL values of Z-scores, Protein domains & fragments*.
522	33	2	Substrate diversity, NULL values of Z-scores, Protein domains/ fragments*.
	E. coli	*Streptococcus*	
745	16	4	NULL values of Z-scores, Protein domains & fragments*, same protein crystallized more than once.
	E. coli	*Lactococcus lactis*	
266	6	7	Conformation changes.
2376	5	2	Different subunit of a multi-subunit enzyme, structures are unrelated but not a HGT.
	E. coli	*Bacillus anthracis*	
5126	3	10	Same proteins with and without ligand. Substrate diversity, HGT not from any Gram-positive bacteria.
	E. coli	*Bacillus megaterium*	
1028	7	4	Substrate diversity.
1609	42	9	Substrate diversity.
1925	8	4	Protein domains & fragments*.
	Bacillus	*E.coli*	
236	2	6	False hit because of protein complex.
454	5	2	The Gram-positive protein structures are same with different ligands and the two Gram-negative proteins are same proteins crystalized twice.
745	2	16	Substrate diversity.
1057	2	2	The two Gram-positive protein structures are same and the two Gram-negative protein structures are same.
1309	3	8	Substrate diversity.
1925	7	8	False positive due to multiple protein conformations.
	Bacillus	*Pseudomonas*	
1057	2	3	The two Gram-positive protein structures are of the same protein and the three Gram-negative proteins are same with different ligands.
	Bacillus	*Pseudomonas putida*	
1309	3	5	Most likely a good example of HGT.
4948	3	5	Most likely a good example of HGT.
	Bacillus	*Haemophilus influenzae*	
2050	2	2	The two Gram-positive protein structures are of the same protein. One protein structure of the Gram-negative organism is a protein fragment.
	Bacillus	*Helicobacter pylori.*	
745	2	4	Two Gram-positive proteins are completely dissimilar. One is a protein fragment.

Table 3 Summary of proteins suspected as horizontal gene transfers.

COG	PDB-ID	ΔZ-score	Receiving Bacteria	Donor Bacteria
503	2DY0	11.85	*Escherichia coli*	*Bacillus subtilis*
596	1M33	4.95	*Escherichia coli*	*Bacillus subtilis*
604	(1O98, 1O8C)	15.45	*Escherichia coli*	*Bacillus subtilis*
1278	3MEF	5.28	*Escherichia coli*	*Bacillus subtilis*
1309	1VI0	3.49	*Bacillus subtilis*	*Pseudomonas putida*
4948	2GGE	8.49	*Bacillus subtilis*	*Unknown*

Table 4 COG- 503 in Comparison between *Escherichia coli* and *Bacillus subtilis.*

		E. coli						*Bacillus subtilis*			
		1A95	1A96	1A97	1A98	1NUL	2DY0	1O57	1P4A	1Y0B	2FXV
E. coli	1A95		29.7	28.3	22.7	26	11.3	9.9	10.8	10.3	10.1
	1A96	29.7		28.3	22.7	26	11.3	9.9	10.9	10.3	10.1
	1A97	28.3	28.3		23.2	26.2	11	9.7	10.6	10	9.8
	1A98	22.7	22.7	23.2		23.5	9.6	8.5	9.3	9.6	8.9
	1NUL	26	26	26.2	23.5		10.2	9.1	9.9	9.4	9.3
	2DY0	11.3	11.3	11	9.6	10.2		20.5	20.3	23.7	22.2
Bacillus subtilis	1O57	9.9	9.9	9.7	8.5	9.1	20.5		39.9	23	23.6
	1P4A	10.8	10.9	10.6	9.3	9.9	20.3	39.9		22.9	23.6
	1Y0B	10.3	10.3	10	9.6	9.4	23.7	23	22.9		32.8
	2FXV	10.1	10.1	9.8	8.9	9.3	22.2	23.6	23.6	32.8	

Table 5 COG-596 comparison between *Escherichia coli* and *Bacillus subtilis*

		E. coli		*Bacillus subtilis*	
		1M33	1U2E	1WO	1WPR
E. coli	1M33		24.8	30.5	0.7
	1U2E	24.8		25.7	25.6
Bacillus subtilis	1WOM	30.5	25.7		47.6
	1WPR	30.7	25.6	47.6	

Table 5 shows that it is an excellent candidate to be a horizontally transferred gene product.

To further confirm this is a genuine case of HGT, we compared visually, the 3-D structure of the protein 1M33 and a sequence alignment with the proteins in *Bacillus subtilis* and other proteins in *E. coli* in the COG-596 (see Figs. 3 and 4).

D. Detailed Analysis of COG-604

COG-604 from *Bacillus subtilis* includes two structures of the same YhfP hypothetical protein without and with NAD bound (1TT7, 1Y9E).

Table 6 shows that *E. coli* proteins 1O89 and 1O8C are more similar to all the structures of the protein in *Bacillus subtilis* than they had been to *E. coli* protein 1QOR. Therefore, they are excellent candidates to be horizontally transferred gene products. In this case we cannot really distinguish between the proteins 1O89, 1O8C and pinpoint one of them as the candidate for HGT as they are of the same protein.

To further confirm this is a genuine case of HGT, we compared the 3-D structure of the protein 1O89 (chose one of the two similar proteins) and a sequence alignment with the proteins in *Bacillus subtilis* and other protein in *E. coli* in the COG-604 as shown in Figs. 5 and 6

Table 6 COG-604 comparison between *E. coli* and *Bacillus subtilis*

		E. coli			*Bacillus*	
		1O89	1O8C	1QOR	1TT7	1Y9E
E. coli	1O89		49.4	29.1	44.8	44.9
	1O8C	49.4		31.8	47.7	47.6
	1QOR	29.1	31.8		30.9	30.7
Bacillus subtilis	1TT7	44.8	47.7	30.9		55
	1Y9E	44.9	47.6	30.7	55	

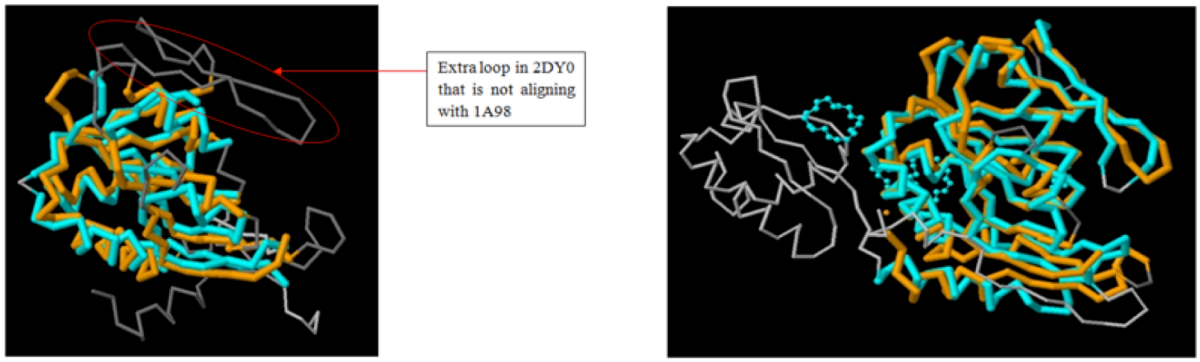

Fig. 1. Pre-calculated jFATCAT-rigid structure alignment results (left) 2DY0 (*E. coli*) vs.1A98 (*E. coli*) and (right) 2DY0 (*E. coli*) vs.1O57 (*Bacillus subtilis*).

```
27:A         40:A      50:A       60:A      70:A      80:A       90:A
LFRDVTSLLEDPKAYALSIDLLVERYKNA-GITKVVGTEARGFLFGAPVALGLGV-GFVPVRKPGKLPRE
EKYIV-----TWDMLQIHARKLASRLMPSEQWKGIIAVSRGGLVPGALLARELGIRHVDTVAI-------
3:A                  20:A      30:A      40:A      50:A      60:A

     100:A     110:A     120:A     130:A     140:A     150:A     160:A     170:A     180:A
TISETYDLEYGTDQLEIHVDAIKPGDKVLVVDDLLATGGTIEATVKLIRRLGGEVADAAFIINLFDLGGEQRLEKQGITSYSLVPFPGH
----------------------GDGEGFIVIDDLVD---TAVAIREMYP-----KAHFVTIFAKP--AGRPLV------DDYVVDIPQD
                       90:A                       110:A      120:A      130:A

2:A              20:A      30:A      40:A      50:A      60:A      70:A
TATAQQLEYLKNSIKSIQDYPKPGILFRDVTSLLEDPKAYALSIDLLVERYKNAGITKVVGTEARGFLFG
AEAEEFVQTLGQSLANPERILP--GGYVYLTDILGKPSVLSKVGKLFASVFAEREIDVVMTVATKGIPLA
78:A        90:A       100:A     110:A     120:A     130:A     140:A

        80:A      90:A      100:A       110:A     120:A     130:A
APVALGLGVGFVPVRKPGKLPRETISETYDL--EYGIDQLEIHVDAIKPGDKVLVVDDLLATGGTIEATV
YAAASYLNVPVVIVRKD----GSTVSINYVSGSSNRIQTMSLAKRSMKTGSNVLIIDDFMKAGGTINGMI
    150:A     160:A              180:A     190:A     200:A     210:A

140:A     150:A     160:A     170:A     180:A
KLIRRLGGEVADAAFIINLFDLGGEQRLEKQGITSYSLVPFPGH
NLLDEFNANVAGIGVLVEAEG--VDERLV---DEYMSLLTLSTI
   220:A     230:A       240:A       250:A

    | ... Structurally equivalent and identical residues
    : ... Structurally equivalent and similar residues
    . ... Structurally equivalent, but not similar residues.
```

Fig. 2. (top) Sequence alignment results 2DY0 (*E. coli*) vs. 1A98 (*E. coli*). Long part of the sequence is not aligning here, corresponding to the extra grey loop in the 3-D structural comparison in Fig. 1 (left), and (bottom) Sequence alignment results 2DY0 (*E. coli*) vs. 1O57 (*Bacillus subtilis*).

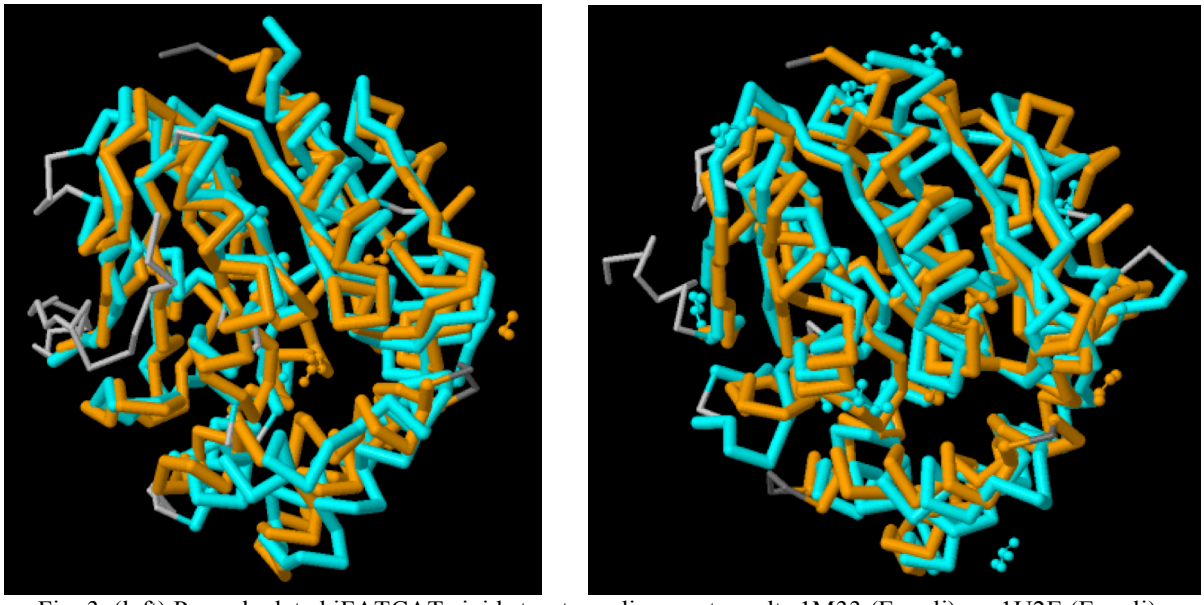

Fig. 3. (left) Pre-calculated jFATCAT-rigid structure alignment results 1M33 (E. coli) vs. 1U2E (E. coli). (right) Pre-calculated jFATCAT-rigid structure alignment results 1M33 (E. coli) vs. 1WOM (Bacillus subtilis). Note that there are more grey regions in this alignment of 1M33 & 1U2E compared to the alignment of 1M33 & 1WOM

Fig. 4. (left) Sequence alignment results 1M33 (*E. coli*) vs. 1U2E (*E. coli*) and (right) Sequence alignment results 1M33 (*E. coli*) vs. 1WOM (*Bacillus subtilis*). The sequence alignment clearly shows more gaps in the alignment of 1M33 vs. 1U2E than in the alignment of 1M33 vs. 1WOM.

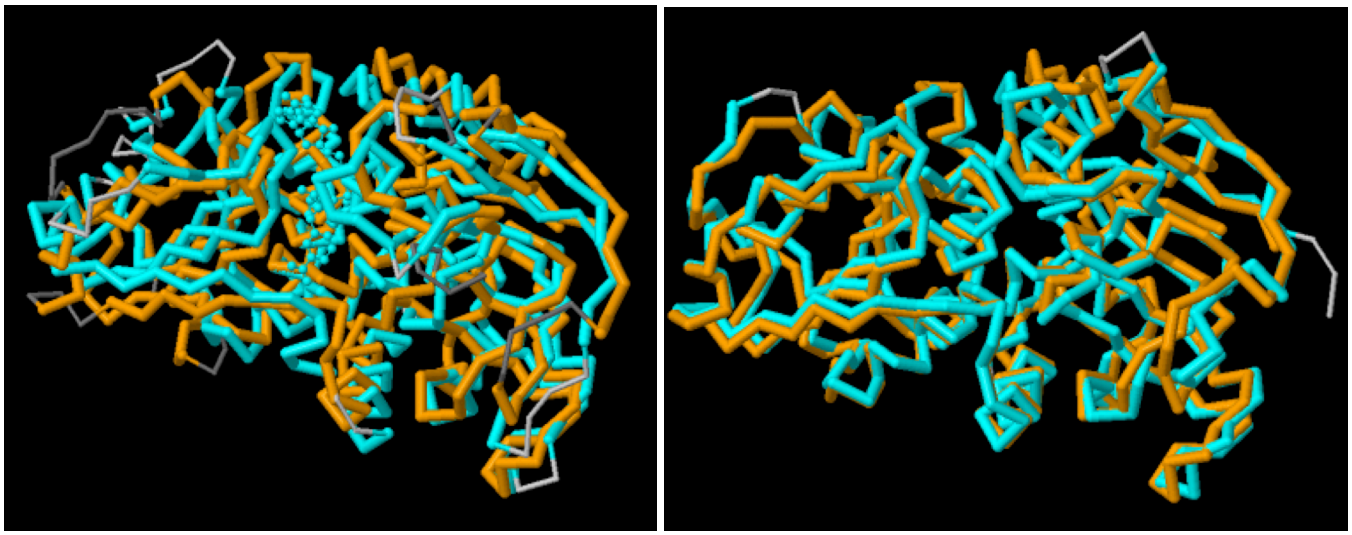

Fig. 5. (left) Pre-calculated jFATCAT-rigid structure alignment results 1O89 (*E. coli*) vs. 1QOR (*E. coli*). Note the numerous small grey loops spread throughout the alignment. (right) Pre-calculated jFATCAT-rigid structure alignment results 1O89 (*E. coli*) vs. 1TT7 (*Bacillus subtilis*). In this case there are very few loops in the alignment.

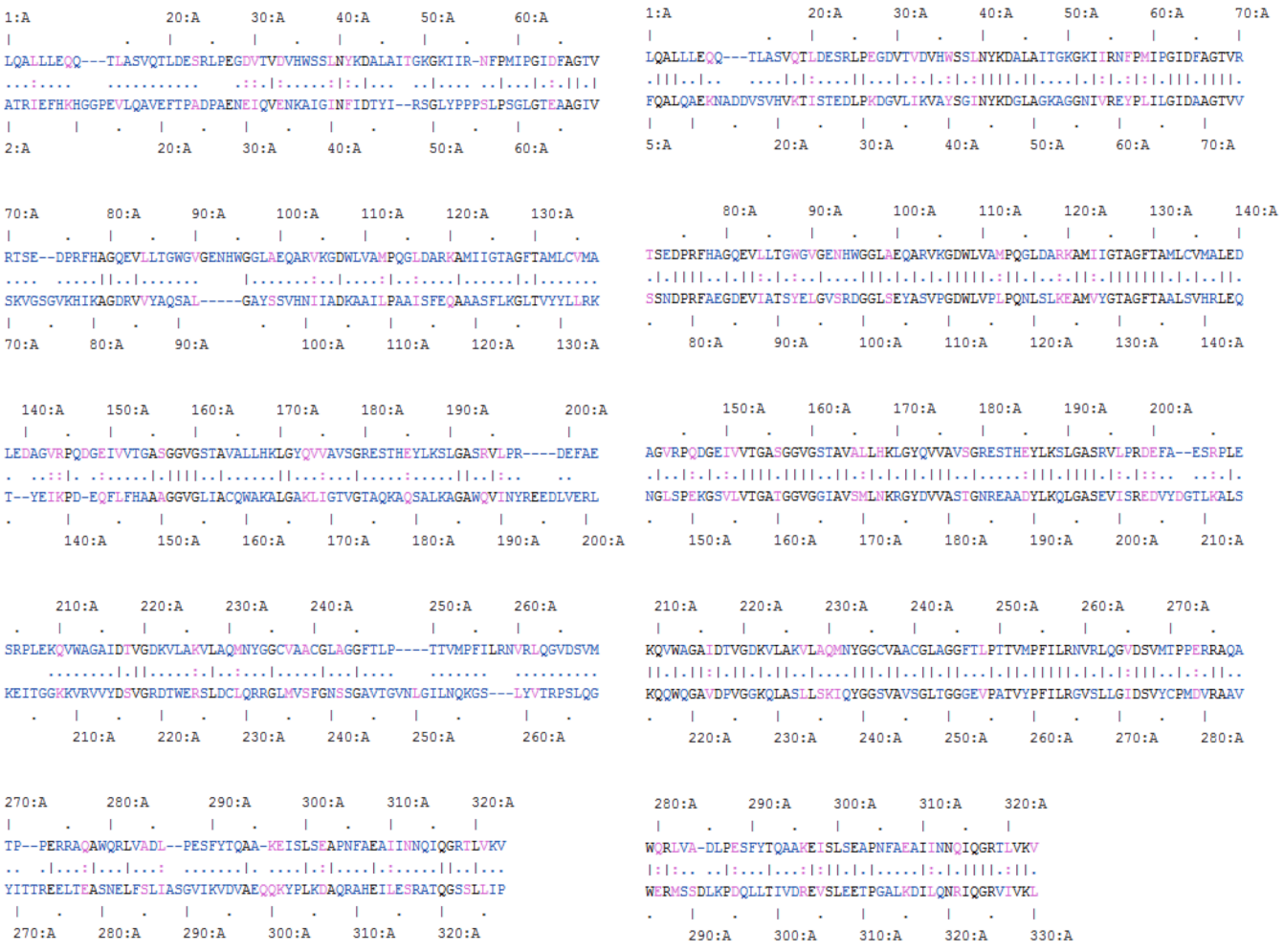

Fig. 6. (left) Sequence alignment results 1O89 (*E. coli*) vs. 1QOR (*E. coli*) clearly showing many small mismatches, which correspond to the numerous small grey regions in the 3-D structure alignment. (right) Sequence alignment results 1O89 (*E. coli*) vs. 1TT7 (*Bacillus subtilis*) showing very few mismatches.

E. Detailed Analysis of COG-1278

COG-1278 from *E. coli* includes two structures of two different cold shock proteins (2BH8, 3MEF). COG-1278 from *Bacillus subtilis* includes four protein structures of the same protein, two of which are with different ligands (2ES2, 2F52) and two of which are site mutants of the same protein (2I5L, 2I5M). *E. coli* protein 3MEF is more similar to all the different structures of the proteins in *Bacillus subtilis* than it had been to the *E. coli* protein 2BH8. Therefore, it is an excellent candidate to be a new horizontally transferred gene product. To further confirm this is a genuine case of HGT, we compared visually the 3-D structure of the protein 3MEF and a sequence alignment with the proteins in *Bacillus subtilis* and other proteins in *E. coli* in the COG-1278.

Table 7 COG-1278 comparison between *E. coli* and *Bacillus subtilis*.

		E. coli		*Bacillus subtilis*			
		2BH8	3MEF	2ES2	2F52	2I5L	2I5M
E. coli	2BH8		5.1	4.7	4	4.8	4.8
	3MEF	5.1		10.4	9	10.2	9.8
Bacillus subtilis	2ES2	4.7	10.4		12.1	15	14.6
	2F52	4	9	12.1		11.7	10.8
	2I5L	4.8	10.2	15	11.7		15
	2I5M	4.8	9.8	14.6	10.8	15	

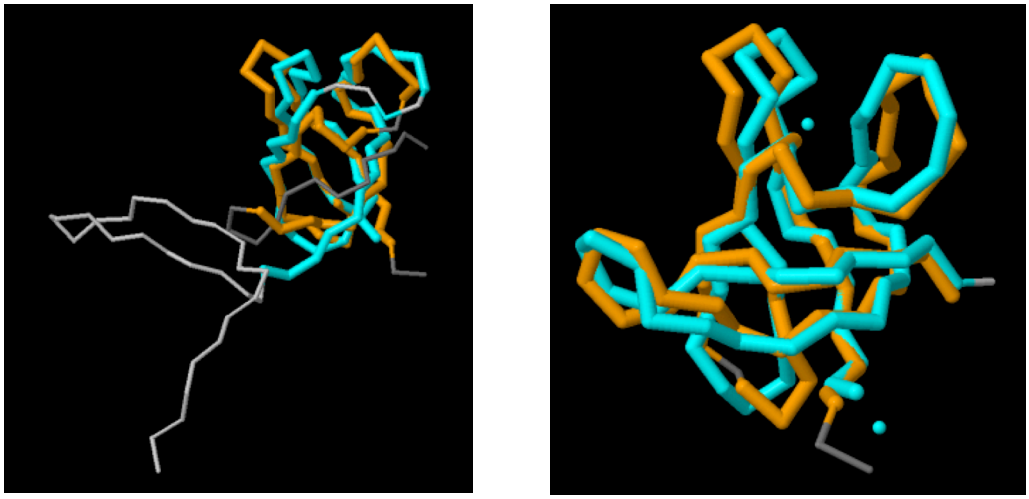

Fig. 7. (left) Pre-calculated jFATCAT-rigid structure alignment results 3MEF (*E. coli*) vs. 2BH8 (*E. coli*). A detailed examination of the 2BH8 protein reveals possible errors in the backbone assignment, but these do not alter the location of the small grey loops. (right) Pre-calculated jFATCAT-rigid structure alignment results 3MEF (*E. coli*) vs. 2ES2 (*Bacillus subtilis*). Note that the structure of 3MEF better matches with 2ES2 than 2BH8.

Fig. 8. (top) Sequence alignment results 3MEF (*E. coli*) vs. 2BH8 (*E. coli*) and (bottom) Sequence alignment results 3MEF (*E. coli*) vs. 2ES2 (*Bacillus subtilis*). Clearly, the sequence of 3MEF better matches with 2ES2 than with 2BH8.

F. Detailed Analysis of COG-1309

COG-1309 from *Bacillus subtilis* includes 2 structures of putative transcriptional regulators (1RKT, 1SGM) and one structure of transcriptional regulator (1VI0). Among these the 1VI0 had the most divergent structure according to the Z-score comparison shown in Table 4. COG-1309 from *Pseudomonas putida* includes five structures, all which are transcriptional regulators. All of the five proteins were closely related according to their Z-scores.

Table 8 shows that *Bacillus subtilis* protein 1VI0 was more similar to the five *Pseudomonas putida* proteins than it was to the other *Bacillus subtilis* proteins. Therefore, it is an excellent candidate to be a horizontally transferred gene product.

To further confirm that this is a genuine case of HGT, we compare the 3-D structure of the protein 1VI0. Sequence alignments with all the proteins in *Pseudomonas putida* with all other proteins in *Bacillus subtilis* in the COG-1309 are done using the Jmol tool. The results are shown in Figs. 9 and 10. The results indicate a strong alignment between the IV10 (*Bacillus subtilis*) and the 2UXH (*Pseudomonas putida*) proteins, further supports the case of horizontal gene transfer between the two bacteria.

Table 8 COG-1309 in Comparison between *Bacillus subtilis* and *Pseudomonas putida*.

		Bacillus subtilis			*Pseudomonas putida*				
		1RK	1SG	1VI0	2UXH	2UXI	2UX	2UXP	2UX
Bacillus subtilis	1RKT		12.3	15.5	15.8	15.8	15.9	15.8	15.9
	1SGM	12.3		15	11.9	11.9	11.9	11.9	11.9
	1VI0	15.5	15		17.3	17.2	17.3	17.5	17.5
Pseudomonas putida	2UXH	15.8	11.9	17.3		32	32.4	32.5	32.3
	2UXI	15.8	11.9	17.2	32		32.3	32.4	32.3
	2UXO	15.9	11.9	17.3	32.4	32.3		32.8	32.5
	2UXP	15.8	11.9	17.5	32.5	32.4	32.8		32.5
	2UXU	15.9	11.9	17.5	32.3	32.3	32.5	32.5	

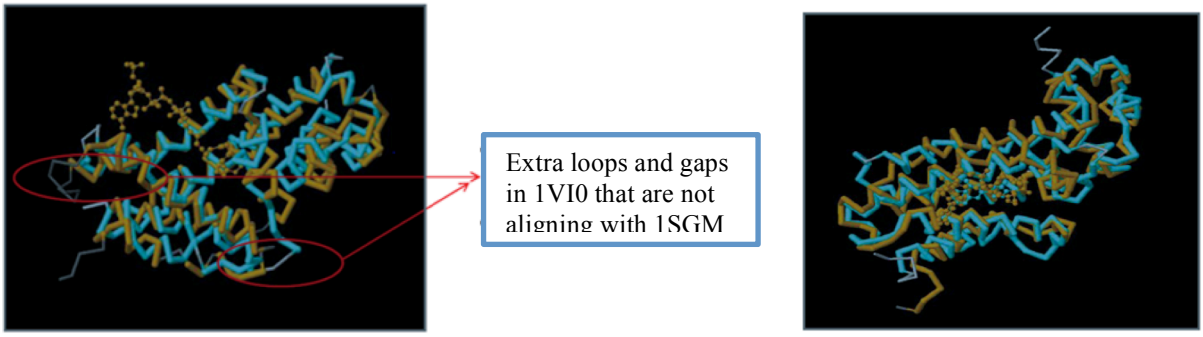

Fig. 9. (left) Pre-calculated jFATCAT-rigid structure alignment results 1VI0 (*Bacillus subtilis*) vs. 1SGM (*Bacillus subtilis*). (right) Pre-calculated jFATCAT-rigid structure alignment results 1VI0 (*Bacillus subtilis*) vs. 2UXH (*Pseudomonas putida)*.

Fig. 10. (top) Sequencing alignment results 1VI0 (*Bacillus subtilis*) vs. 1SGM (*Bacillus subtilis*) and (bottom) sequence alignment results 1VI0 (*Bacillus subtilis*) vs. 2UXH (*Pseudomonas putida*).

Table 9 COG-4948 in comparison between *Bacillus subtilis* and *Pseudomonas putida*.

		Bacillus subtilis			*Pseudomonas putida*				
		1JP	1TK	2GGE	1BKH	1F9C	1MU	2MU	3MU
Bacillus subtilis	1JPM		59.2	36.5	46.7	46.5	46.7	46.6	46.5
	1TKK	59.2		38.6	46.3	46.5	46.5	46.5	46.3
	2GGE	36.5	38.6		38.1	37.8	38	38.1	38.1
Pseudomonas putida	1BKH	46.7	46.3	38.1		63	64.6	64.4	64.3
	1F9C	46.5	46.5	37.8	63		63.4	63.4	62.5
	1MUC	46.7	46.5	38	64.6	63.4		65.3	65.2
	2MUC	46.6	46.5	38.1	64.4	63.4	65.3		65.5
	3MUC	46.5	46.3	38.1	64.3	62.5	65.2	65.5	

G. Detailed Analysis of COG-4948

Table 9 shows a detailed comparison of the COG-4948 proteins in *Bacillus subtilis* and *Pseudomonas putida*. Here we can observe that protein 2GGE does not fit well into the set of COG-4948 proteins of *Bacillus subtilis* because its similarity score with the other proteins is much lower than the similarity score between 1JPM and 1TKK, the other two COG-4948 proteins of *Bacillus subtilis*. However, 2GGE does not show an unusually high similarity with the set of COG-4948 proteins of *Pseudomonas putida*. Hence 2GGE was likely transferred horizontally to *Bacillus subtilis* but not from *Pseudomonas putida*.

H. False Positives

A false positive occurs when a positive result is erroneously observed. During our detailed analysis, we noted several situations that routinely generated false positives:

1. **Protein Fragments:** Many of the PDB-ids in the Protein Data Bank (PDB) database [2, 4] correspond to protein domains and protein fragments. The structural comparison of these domains and protein fragments with the whole protein sometimes leads to falsely suspecting a protein for HGT. Good examples of this case are COG-2050 and COG-745.
2. **Substrate Diversity:** The COG's enzyme specificity is fixed within the COG but the substrate specificity is diverse. Good examples for this case are COG-745 and COG-1309.
3. **Conformation changes:** There are two or more conformations of the same protein. Example: COG-1925 and COG-745.
4. **HGT from other sources:** There are some cases in which a protein is identified as possible HGT but not exactly from the organism with which we are comparing. Example: Protein 2GGE in COG-4948.
5. **Different Subunits:** Different subunits of a multi subunit enzyme have very dissimilar structures and with the structure-based method these could look like a possible candidate of HGT, but they are not.

VI. Conclusions and Future Work

We extended our earlier method [3, 20] and presented an improved protein structure-based method to detect horizontal gene transfer. We tried to identify possible HGT in *Firmicutes* from *Proteobacteria*. Various cases of false positives have been identified and documented. This method cannot be evaluated for efficiency over other methods for two reasons. First, because it uses a completely different approach to identify HGT, uses protein structures rather than the complete genomes used in other techniques. Secondly, each of the techniques used to identify HGT do not yield the same result set. Automation of the procedure to identify HGT was possible only to a certain extent after which the data had to be analyzed manually, which took substantial amount of time. Automation of the entire procedure would be complex to implement as careful analysis and structural visualization of each candidate for HGT was required to zero in on a participant of HGT.

In the future, we plan to improve our algorithm by eliminating false positives. The accuracy of our method also depends on the accuracy of sources from which data is collected for various organisms. Unfortunately, we cannot guarantee this. The main source of data for this research was the PDB database [2, 4]. Many underlying problems exist with this database, some of which are as follows:

1. Like any other biological database, the PDB is incomplete because it does not contain complete protein structure information for all the organisms. Although, it is a constantly growing collection of protein structure data, there is a limitation in choosing organisms.
2. Since it relies on entries from various biologists and biochemists, same proteins may be crystallized multiple times, resulting in duplicated entries (multiple PDB IDs for the same protein).
3. Some proteins have been crystallized with and without ligands and substrates, each appear with a unique PDB-id.
4. Protein domains and protein fragments appear with unique PDB-id.

5. Some proteins have been mutated at only one or a few residues, but each structure has a unique PDB-id.

As the quality of the biological databases used increases, so can the efficiency of our method be improved.

This research was based on COG classification, which is a generalized classification. But researchers are moving away from this classification to more specific types of classification of proteins such as GO and eggNOG. Some of the databases have already gotten rid of this classification. Our method can also be applied and tested with these classifications to prove its efficiency. Following similar procedures to identify HGT with these new classifications might provide interesting results.

The DaliLite tool used in this research for structural comparison of proteins can be replaced with CPASS program which compares ligand defined active sites to determine sequence and structural similarity.

This research can be scaled to other organisms belonging to other classifications of phyla. As more genomic data of organisms becomes available in the biological databases, this research can be used to identify more cases of HGT. Scalability of this research might help to answer other intriguing questions, such as the following:

1. Which proteins have more probability of being horizontally gene transferred?
2. What is the functionality of such proteins?
3. Which organism has the highest percentage of HGT proteins?
4. What are the conditions that would enable a horizontal gene transfer?
5. What is rate of occurrence of the HGT?

Identifying the reasons and causes behind the occurrence of HGT can be an interesting way to extend this research. Each method to detect HGT follows a different approach. Comparison and statistical analysis to see the accuracy of each of the methods could also provide interesting results.

HGTs can play a role in epidemics. Hence a better understanding of HGTs could aid in quickly identifying newly developing epidemics and in better tracking their movements [17, 19].

Our method is also applicable beyond gene evolution. For example, studying the evolution of script families, such as the Cretan Script Family [16], can also benefit from the principles of our method because there may be horizontal influences, that is, the transfer of individual script symbols, from one script to another.

REFERENCES

[1] T. Akiba, K. Koyama, Y. Ishiki, S. Kimura, and T. Fukushima, "On the mechanism of the development of multiple-drug resistant clones of Shigella." *Japanese Journal of Microbiology,* vol. 4, pp. 219–27, 1960.

[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235-42, 2000.

[3] S. Billa, M. A. Griep, and P. Z. Revesz, "Approximate search on protein structures for identification of horizontal gene transfer in bacteria," In: *Proc. 9th International Symposium on Abstraction, Reformulation and Approximation*, AAAI Press, pp. 18-25, 2011.

[4] P. E. Bourne, K. J. Addess, W. F. Bluhm, L. Chen, N. Deshpande, Z. Feng, W. Fleri, R. Green, J. C. Merino-Ott, W. Townsend-Merino, H. Weissig, J. Westbrook, and H. M. Berman, "The distribution and query systems of the RCSB Protein Data Bank," *Nucleic Acids Research,* vol. 32, pp. 223-225, 2004.

[5] M. D. Frank-Kamenetskii, and L. Liapin, L. *Unraveling DNA: The Most Important Molecule of Life*. Reading, Mass.: Perseus Books, 1993.

[6] P. C. Kanellakis, G. M. Kuper, and P. Z. Revesz, "Constraint query languages," *Journal of Computer and System Sciences*, vol. 51, no. 1, pp. 26-52, 1995.

[7] E. V. Koonin, T. G. Senkevich, and V. V. Dolja, "The ancient virus world and evolution of cells," *Biology Direct,* vol. 1, no. 29, 2006.

[8] J. G. Lawrence, and H. Ochman, "Amelioration of bacterial genomes: rates of change and exchange," *Journal of Molecular Evolution*, pp. 383-97, 1997.

[9] J. G. Lawrence, and H. Ochman, "Molecular archaeology of the Escherichia coli genome," *Proceedings of the National Academy of Sciences USA*, pp. 9413-9417, 1998.

[10] J. G. Lawrence, and H. Ochman, "Reconciling the many faces of lateral gene transfer". *Trends in Microbiology*, vol. 10, pp. 1-4, 2002.

[11] R. Powers, J. C. Copeland, K. Germer, K. A. Mercier, V. Ramanathan, and P. Z. Revesz, "Comparison of Protein Active Site Structures for Functional Annotation of Proteins and Drug Design," *PROTEINS: Structure, Function, and Bioinformatics*, vol. 65, no. 1, pp. 124-135, 2006.

[12] J. Ramanan, P. Z. Revesz, "Testing the independence hypothesis of accepted mutations for pairs of adjacent amino acids in protein sequences," *International Journal of Biology and Biomedical Engineering*, vol. 11, pp. 170-179, 2017.

[13] P. Z. Revesz, "Refining Restriction Enzyme Genome Maps," *Constraints*, vol. 2, no. 3-4, pp. 361-375, 1997.

[14] P. Z. Revesz, *Introduction to Databases: From Biological to Spatio-Temporal*, Springer, New York, NY, 2010.

[15] P. Z. Revesz, "An algorithm for constructing hypothetical evolutionary trees using common mutations similarity matrices," in *Proceedings of the 4th ACM International Conference on Bioinformatics and Computational Biology,* ACM Press, pp. 731-734, 2013.

[16] P. Z. Revesz, "Bioinformatics evolutionary tree algorithms reveal the history of the Cretan Script Family," *International Journal of Applied Mathematics and Informatics*, vol. 10, pp. 67-76, 2016.

[17] P. Z. Revesz and T. Triplet, "Temporal data classification using linear classifiers," Information Systems, vol. 36, no. 1, pp. 30-41, 2011.

[18] P. Z. Revesz, D. Singh, "Fast virus and bacteria genome sequencing by compatible restriction enzyme fingerprinting," *International Journal of Biology and Biomedical Engineering*, vol. 12, pp. 18-27, 2018.

[19] P. Z. Revesz and S. Wu, "Spatiotemporal reasoning about epidemiological data," *Art. Int. in Medicine*, vol. 38, no. 2, pp. 157-170, 2006.

[20] V. Santosh, M. A. Griep, P. Z. Revesz, "Protein Structure-Based Method for Identifying Horizontal Gene Transfer," In: *Proceedings of the 4th International C* Conference on Computer Science and Software Engineering,* ACM Press, pp. 9-16, 2011.

[21] M. Shortridge, T. Triplet, P. Z. Revesz, M. A. Griep, and R. Powers, "Bacterial protein structures reveal phylum dependent divergence," Computational Biology and *Chemistry*, vol. 35, no. 1, pp. 24-33, 2011.

[22] L. Song, Y. Ning, Q. Zhang, C. Yang, G. Gao, and J. Han, "Studies on antimicrobial resistance transfer in vitro and existent selectivity of avian antimicrobial-resistant enterobacteriaccae in vivo," *Agricultural Sciences in China*, vol. 7, pp. 636-640, 2008.

[23] K. A. Susanna, C. D. den Hengst, L. W. Hamoen, and O. P. Kuipers, "Expression of transcription activator ComK of Bacillus subtilis in the heterologous host Lactococcus lactis leads to a genome-wide repression

pattern: A case study of horizontal gene transfer," *Applied Environmental Microbiology,* vol. 72, no. 1, pp. 404–411, 2006.
[24] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, "The COG database: A tool for genome-scale analysis of protein functions and evolution," *Nucleic Acids Research*, vol. 28, pp. 33-36, 2000.
[25] R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin, "The COG database: New developments in phylogenetic classification of proteins from complete genomes," *Nucleic Acids Research*, vol. 29, pp. 22-28, 2001.
[26] T. Triplet, M. Shortridge, M. A. Griep, J. Stark, R. Powers, and P. Z. Revesz, "PROFESS: A protein function, evolution, structure and sequence database," *Database - The Journal of Biological Databases and Curation, DOI*=10.1093/baq011, 2010. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2911846/#lpo=2.6158
[27] A. Tsirigos, and I. Rigoutsos, "A new computational method for the detection of horizontal gene transfer events," *Nucleic Acids Research*, vol. 33, pp. 922–933, 2005.

Peter Z. Revesz (Ph.D.'91) holds a Ph.D. degree in Computer Science from Brown University and was a postdoctoral fellow at the University of Toronto.

He is an expert in constraint databases [6], data mining, big data analytics and bioinformatics. He is the author of *Introduction to Databases: From Biological to Spatio-Temporal* (Springer, 2010) and *Introduction to Constraint Databases* (Springer, 2002). He is currently a professor in the Department of Computer Science and Engineering at the University of Nebraska-Lincoln, Lincoln, NE 68588, USA.

Dr. Revesz also held visiting appointments at the Aquincum Institute of Technology, the IBM T. J. Watson Research Center, INRIA, the Max Planck Institute for Computer Science, the University of Athens, the University of Hasselt, the U.S. Air Force Office of Scientific Research and the U.S. Department of State. He is a recipient of an AAAS Science & Technology Policy Fellowship, a J. William Fulbright Scholarship, an Alexander von Humboldt Research Fellowship, a Jefferson Science Fellowship, a National Science Foundation CAREER award, and a "Faculty International Scholar of the Year" award by *Phi Beta Delta*, the Honor Society for International Scholars.

Swetha Billa (M.S.'11) earned a M.S. degree in Computer Science at the University of Nebraska-Lincoln. She is currently working as a Technical Implementation Consultant at CPA Global in Omaha, Nebraska.

Mark A. Griep (Ph.D.'86) holds a Ph.D. degree in Biochemistry from The University of Minnesota and was a postdoctoral fellow at the University of Colorado Health Sciences Center. He is an enzymologist working on bacterial DNA replication proteins with a focus on primase, helicase, and antibiotic drug discovery. In his work on chemical education and outreach, he is an author with Marjorie L Mikasen of the book *ReAction! Chemistry in the Movies* (Oxford University Press, 2009) and is the recipient of the 2017 American Chemical Society Helen M. Free Award for Public Outreach. He is currently an associate professor in the Department of Chemistry at the University of Nebraska-Lincoln, Lincoln, NE 68588, USA.

Venkat R. B. Santosh (M.S.'11) earned a M.S. degree in Computer Science at the University of Nebraska-Lincoln. He is currently working as a senior MySQL Database Administrator at Salesforce in the San Francisco Bay area.