Protein Structure-Based Method for Identifying Horizontal Gene Transfer

Venkat R. B. Santosh
Department of Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, NE, 68588
+1-402-617-3578
vsantosh@cse.unl.edu

Mark A. Griep
Department of Chemistry
University of Nebraska-Lincoln
Lincoln, NE, 68588
+1-402-472-3429
mgriep1@unl.edu

Peter Z. Revesz
Department of Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, NE, 68588
+1-571-201-5639
revesz@cse.unl.edu

ABSTRACT
Genetics has traditionally focused on vertical gene transfer, which is the passing of the genetic material of an organism to its offspring. However, recent studies in genetics increased the awareness that horizontal gene transfer, which is the passing of the genetic material of an organism to another organism that is not its offspring, is also a significant phenomenon. Horizontal gene transfer is thought to play a major role in the natural evolution of bacteria, such as, when several different types of bacteria all suddenly develop the same drug resistance genes. Artificial horizontal gene transfer occurs in genetic engineering.

This paper provides methods to detect horizontal gene transfer among bacteria using BLAST and DaliLite measures of protein sequence and structural similarities. This research is novel and unique because no previous horizontal gene transfer study worked directly on protein sequences and structures. The main method is a computer algorithm to detect horizontal gene transfer among different COG classifications of proteins. The paper also considers visual structural comparisons and sequence alignments using the 'Jmol' tool. Finally, the paper considers the possibility that the methods yield false positives.

Keywords
Keywords are your own designated keywords.

1. INTRODUCTION
Horizontal gene transfer (HGT), which is also called lateral gene transfer, is any process in which an organism incorporates genetic material from another organism without being the offspring of that organism. In contrast, vertical gene transfer occurs when an organism receives genetic material from its ancestor, e.g. its parent or a species from which it evolved. Growing study in genetics has acknowledged that horizontal gene transfer is also a highly significant phenomenon, and among single-celled organisms perhaps the dominant form of genetic transfer. There is some evidence that even higher plants and animals have been affected and this has raised concerns for safety. Due to the increasing amount of evidence suggesting the importance of these phenomena for evolution, molecular biologists have described horizontal gene transfer as a "new paradigm for biology". It should also be noted that the process may be a hidden hazard of genetic engineering, as it may allow dangerous transgenic DNA which is optimized for transfer to spread from species to species.

C3S2E'11, May 16–18, 2011, Montreal, Quebec, Canada.

1.1 Mechanisms of HGT
Horizontal gene transfer could occur by several mechanisms between organisms. There are three basic mechanisms as described below.

- 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."

1.2 Methods Currently Used to Detect HGT

During the past decade, different approaches have been proposed for the detection of HGT, which can be classified in two major categories: (a) the phylogeny-based methods and (b) the composition-based methods. Some of them are described here which helps us understand the uniqueness of the new approached which uses protein structures to detect HGT.

1.2.1 Phylogeny-Based Detection of HGT

Phylogeny-based detection of HGT is one of the most commonly used approaches for detecting HGT. It is based on the fact that HGT causes discrepancies in the gene tree as well as create conflict with the species phylogeny. So the methods that use this approach would compare the gene and species tress which would come up with a set of HGT events to explain the discrepancies among these trees.

When HGT occurs, the evolutionary history of the gene would not agree with the species phylogeny. The gene trees get reconstructed and their disagreements are used to estimate how many events of HGT could have occurred and the donors and recipients of the gene transfer. Some of the issues when using this method for HGT detection are, determining if the discrepancy is actually a HGT and uniquely identifying the HGT scenario.

The Phylogenetic trees are only partially known and they are reconstructed using Phylogeny reconstruction techniques. The quality of this reconstruction which is usually done statistically has an impact on the HGT detection and sometimes could underestimate or overestimate the number HGT events. Eliminating these statistical errors is possible but this will lead to non-binary Phylogenetic trees. But this method works with Binary Phylogenetic trees only. So this method will need to be modified to accommodate non-Binary Phylogenetic trees as well.

1.2.2 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.

1.2.3 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.

2. METHODOLOGY

Instead of using the traditional methods for identifying HGT, we devised a novel protein structure-based method (HGT-SBM). When a protein is acquired by HGT, the structure of the protein remains fairly similar to that of the donor organism as it tries to retain close similarities to the function of the donor protein.

COG classification of protein function was considered to look for protein structure anomalies. All proteins under the same COG classification are supposed to have similar function, which evolutionary theory indicates they should have similar structures.

For this research, we try to identify HGT between the bacterial phyla *Firmicutes* and *Proteobacteria*. Most medically important bacteria fall into these two phyla, which diverged hundreds of millions of years ago. During their subsequent evolutions, the proteins in all *Firmicutes* bacteria acquired random mutation but still remained more similar to the other *Firmicutes* bacterial proteins than to the *Proteobacteria* bacterial proteins and vice-versa. Hence any anomalous proteins (i.e. proteins having characteristics of the other phyla's protein) in either of the phylum would be a very good candidate for a horizontal gene transfer that occurred fairly recently.

2.1 The Method

We chose *E. coli* from *Proteobacteria* and *Bacillus subtilis* from *Firmicutes* as candidates from the two phyla as they have the most number of studied structures from their respective phyla.

PROFESS database was used to get the list of proteins that have been studied in *E. coli* and *Bacillus subtilis* and also the COGs to which they belong to. The COG classification enables us to identify the proteins which are functionally similar.

The DaliLite program was used for the structure comparison of the proteins. We first determine the extent of structural similarity of all the proteins in a particular COG within each organism chosen from the two phyla (in this case *E. coli* and *Bacillus subtilis*). Then a pair-wise structural comparison of the proteins between the two organism in each COG is done (in this case the *E. coli* proteins are compared with the *Bacillus subtilis* proteins).

Figure 1. Flow chart of the method.

We have approximately 3264 unique PDB IDs in *E. coli* and 494 unique PDB IDs in *Bacillus subtilis*. There are about 88 COGs common in both these organisms. This would result in n * (n-1)/2 pairs of PDB IDs for each COG, where n is the number of proteins in a COG. For the pair-wise comparison between the two organisms within the same COG the number of pairs of PDB IDs would be the cross product of number of proteins in that COG in each organism. For all these cases the averages of the Z-scores for all the pair wise comparison within a COG (for all the common COGs) are documented in a table.

DaliLite gives different Z-scores values for a pair of proteins corresponding to different alignments. We use the best alignment i.e. the highest Z-score value that DaliLite outputs for a given pair. A normalization process is done on the documented Z-score. This is done by choosing the maximum of the 3 average Z-scores values obtained for a COG (one average Z-score from the comparison of proteins within the *Proteobacteria*, one average Z-score from the comparison of proteins within the *Firmicutes* and one average Z-score from the comparison of proteins between *Proteobacteria* & *Firmicutes*.) All the three average Z-scores for a COG are divided by this max Z-score value.

Now we try to compare and look for Z-score anomalies as this will identify protein structure anomalies. Usually the average values of Z-scores for proteins within a COG for both the organisms in comparison are supposed to be pretty similar. So we try to identify those COGs which have high average Z-scores in one organism and a low average Z-score in the other organism. A threshold of 75% was chosen for the average Z-score values to identify as an anomaly. Table 1 shows the average Z-scores.

Table 1. Example of documented data**.**

Common COG	*E. coli*	*Bacillus subtilis*	Comparison Z-Score	*E. coli* normalized Z-Score	*Bacillus subtilis* normalized Z-score	Comparison Z-Score normalized
500	11	39.5	15.4	0.28	1	0.39

In the above example, the COG 500 is identified as having an anomaly because the average Z-score in COG 500 in *E. coli* is only 11 which is 27.8% of the average Z-score in *Bacillus subtilis* which is 39.5.

Most of the times the reason this happens is, there are one or more proteins in the COG that have dissimilar protein structures compared to the other proteins in the same COG. These proteins are candidates for HGT. Not all anomalous protein structures can be identified as HGT. A careful and a systematic hand curation of the Z-scores must be done to identify or eliminate different PDB structures for the same protein, some of them bound to ligands and some with different conformation. It is also necessary to examine enzyme names to ensure the PDBs are for different proteins with the same COG. Finally, it was necessary to compare super imposed structures to verify that HGT had occurred. We compared *E. coli* with all the *Firmicutes* bacteria to detect possible HGTs in *E. coli* from *Firmicutes* bacteria. An automation program greatly reduced the data set to be analyzed by hand.

3. ANALYSIS AND RESULTS

We needed to do our analysis on proteins from two bacterial phyla. *E. coli* was chosen as the representative organism from *Proteobacteria* because it is among the most extensively studied bacteria and has the most number of crystallized proteins.

The protein structures of *E. coli* were compared with all the *Firmicutes* (Gram positive) bacteria having greater than forty of crystallized proteins in the PDB. There were fifteen Gram positive organisms with crystallized proteins greater than 40. But *E. coli* could be compared to only seven of them that had COG numbers matching with the ones *E. coli* has.

The Gram positive organisms compared with *E. coli* are:

1. *Bacillus subtilis*
2. *Staphylococcus aureus*
3. *Bacillus stearothermophilus*
4. *Streptococcus pneumonia*
5. *Lactococcus lactis*
6. *Bacillus anthracis*
7. *Bacillus megaterium*

The comparison of protein structure within the common COGs of *E. coli* and the other Gram positive organism is tabulated. The COGs of our interest are those that have Z-score values less than or equal to 75% of the average Z-score value in the other organism within the same COG. Detailed study of these suspicious COGs gave the results shown in Table 2.

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

COG	No. of Structures in Bacterial Pairs		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	Most likely a good example of HGT. But a closer examination revealed it was the result of protein fragments in *E. coli*.
840	2	2	The two Gram-positive protein structures are not different and not similar to any of the Gram-negative protein structures.
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 stearothermophilus***	
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 pneumoniae***	
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, so the structures are unrelated but is 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*.

* In cases where protein fragments are involved, other methods can be used instead of the Z-score comparison. For example, we could use Revesz's sequence tilting method, which approximately reconstructs the entire sequence of a protein using fragments of another protein. The measure of the goodness of the tiling between two strings a and b, called the tiling similarity, is defined as:

$$TS(a,b) = \frac{\text{sum of the similarities in the alignments}}{\text{number of tiles in the tiling}}$$

If there are several possible tilings, we need to choose the tiling that yields the highest tiling similarity score.

3.1 Summary of Suspected HGTs

A further detailed analysis of all the proteins in these candidate HGTs resulted in identification of the proteins 2DY0 in COG-503, 1M33 in COG-596, 1O98 & 1O8C in COG-604 and 3MEF in COG-1278 as possible HGT to *E. coli* from *Bacillus subtilis*.

Table 3: Summary of Proteins suspected as HGTs.

PDB-ID	COG	ΔZ-score*	Receiving Bacteria	Donor Bacteria
2DY0	503	11.85	*Escherichia coli*	*Bacillus subtilis*
1M33	596	4.95	*Escherichia coli*	*Bacillus subtilis*
(1O98, 1O8C)	604	15.45	*Escherichia coli*	*Bacillus subtilis*
3MEF	1278	5.28	*Escherichia coli*	*Bacillus subtilis*

* The ΔZ-score 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.

3.2 Detailed Analysis of suspected COGs

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 3 Xanthine Transferase (1P96, 1Y0B, 2FXV). All of the four proteins were closely related according to their Z-scores.

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 horizontally transferred gene product. This example has not been reported in the literature.

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

E. coli proteins versus each other						
	1A95	1A96	1A97	1A98	1NUL	2DY0
1A95		29.7	28.3	22.7	26	11.3
1A96			28.3	22.7	26	11.3
1A97				23.2	26.2	11
1A98					23.5	9.6
1NUL						10.2
2DY0						

Bacillus subtilis proteins versus each other				
	1O57	1P4A	1Y0B	2FXV
1O57		39.9	23	23.6
1P4A			22.9	23.6
1Y0B				32.8
2FXV				

E. coli versus *Bacillus subtilis* proteins				
	1O57	1P4A	1Y0B	2FXV
1A95	9.9	10.8	10.3	10.1
1A96	9.9	10.9	10.3	10.1
1A97	9.7	10.6	10	9.8
1A98	8.5	9.3	9.6	8.9
1NUL	9.1	9.9	9.4	9.3
2DY0	20.5	20.3	23.7	22.2

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.

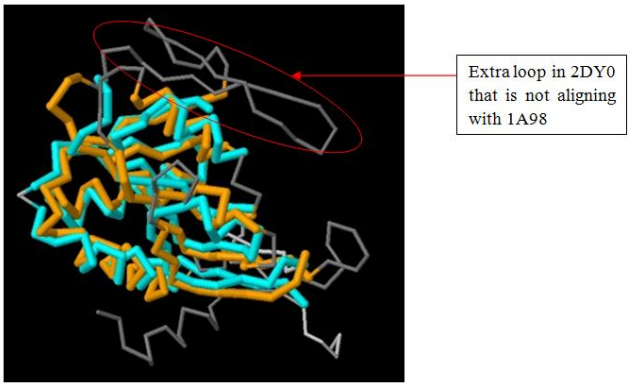

Figure 2: Pre-calculated jFATCAT-rigid structure alignment results 2DY0 (*E. coli*) vs. 1A98 (*E. coli*).

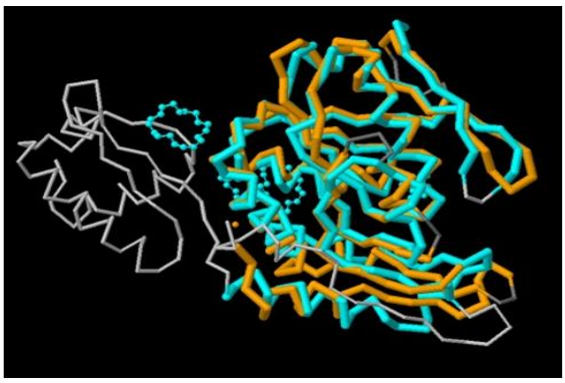

Figure 3: Pre-calculated jFATCAT-rigid structure alignment results 2DY0 (*E. coli*) vs. 1O57 (*Bacillus subtilis*).

```
2:A                 20:A       30:A       40:A       50:A       60:A       70:A
|     |     .     |     .     |     .     |     .     |     .     |     .     |
TATAQQLEYLKNSIKSIQDYPKPGILFRDVTSLLEDPKAYALSIDLLVERYKNAGITKVVGTEARGFLFG
....:..:.|..|:...:....  ..:....|.:|..|.........|....:....|..|......:....
AEAEEFVQTLGQQSLANPERILP--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--EYGTDQLEIHVDAIKPGDKVLVVDDLLATGGTIEATV
...|..|.|..|.|||.   ..|:|..|..  ......:.:.....|.|..||::||.:..|||||...:
YAAASYLNVPVVIVRKD----GSTVSINYVSGSSNRIQTMSLAKRSMKIGSNVLIIDDFMKAGGTINGMI
     |     .     |     .     |     .     |     .     |     .     |     .
    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.
```

Figure 4: Sequence alignment results 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
```

Figure 5: 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 Figure 2.

A similar detailed analysis was done on the COGs 596, 604 and 1278 and the suspected PDB-IDs were more similar to the proteins in *Bacillus subtilis* than it had been to the *E. coli* proteins.

3.3 False Positives

Initially the analysis on these COGs with suspicious HGTs seemed to have found a very a large number of HGTs. However, an intensive analysis proved that many of these were false positives. There were the following reasons for false positives:

1. Protein Fragments: Many of the PDB-ids in the Protein Data Bank 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-1925 and COG-2376.
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-526 and COG-1609.
3. Conformation changes: There are two or more conformations of the same protein. Example: COG-266
4. NULL values: Comparison of structures with no significant similarity should be considered a 'NULL'. This disturbs the statistical analysis greatly.
5. 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 1BJF in COG-5126.
6. 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.

4. CONCLUSIONS AND FUTURE WORK

4.1 Conclusions

Identifying HGTs is a difficult process. No process or method proposed so far is capable of identify perfectly all cases of HGTs. Each process has its own advantages and disadvantages. This research devised a novel protein structure-based method for identifying HGTs and has proved that working directly with the proteins and their structures is a good option and an innovative approach for identifying HGTs. The various possibilities of false positives also have been studied and documented.

4.2 Future Work

The process of identifying HGTs using whole organism protein structures is the first of its kind and has a vast scope for improvements and advancements. In particular, ways to eliminate each one of the cases for false positives discussed in Chapter 4 would be the highest priority for improving our method.

PDB is the best database available for the various crystallized proteins, their structures etc. However, some of the problems encountered when using PDB are:

1. There is some redundancy in the PDB i.e. some proteins that have been crystallized more than once and each appear with a unique PDB-id.
2. Some proteins have been crystallized with and without ligands and substrates, each appear with a unique PDB-id.
3. Protein Domains and Protein fragments appear with unique PDB-id.
4. Some proteins have been mutated at only one or a few residues, but each structure has a unique PDB-id.

These issues cause considerable deviation in the analysis as well as the results. Some of the false positive cases can be eliminated when the PDB gets cleaned.

There are millions of proteins in various organisms. Not all the proteins have been crystallized and their structures are not available. This is one of the main limitations of using the protein structure based approach for identifying HGT. As more protein structures are crystallized and the PDB expands, the efficiency of this protein structure-based method for detecting HGT will only get better.

COG classification is more of a generalized classification of proteins and there are various other protein classifications that can be used instead of the COG. Some of the fairly recent classification like GO classification, eggNOG classification etc. would be a good choice to experiment this process on. The results of the same process with a different classification could give better and more interesting results.

This research has a great potential for scalability. As more analysis is done with the other organisms and as we find more cases of HGT it would be very interesting to look into the statistics. This could include, which organism has higher percentage of HGT proteins? Which type of protein has higher cases of a HGT, etc? For all these cases we could look into the reason and this might drive us into very interesting causes and reasons. The statistics of this method could be compared to the statistics of the other methods for detecting HGT, but these statistics might not match each other because each method works in a different way.

5. REFERENCES

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

[2] Barlow M. 2009. "What antimicrobial resistance has taught us about horizontal gene transfer". *Methods in Molecular Biology* (Clifton, N.J.) 532: 397–411.

[3] Bergey, D. H., Holt, J.G., Krieg, N.R., Sneath, P.H.A., 1994. *Bergey's Manual of Determinative Bacteriology, (9th ed.)*, Lippincott Williams & Wilkins.

[4] Consortium, T. G. O. 2006. The Gene Ontology (GO) project. *Nucleic Acids Research*, 34, D326.

[5] Cortez, D., Delaye, L., Lazcano, A., Becerra, A., 2009. Composition-based methods to identify horizontal gene transfer. *Methods Molecular Biology*, 532:215-25. PMID: 19271187.

[6] Creighton, T. H., 1993. *Proteins: structures and molecular properties*. San Francisco: W. H. Freeman.

[7] DaliLite. 2000. DaliLite workbench for protein structure comparison. *Bioinformatics*, 16 (6):566-7.

[8] Garcia-Vallve, S., Romeu, A., Palau, J., 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Research*, 10: 1719–1725.

[9] Griffiths A.J.F., Miller J.H., Suzuki D.T., Lewontin R.C., Gelbart W.M., 2000, *An Introduction to Genetic Analysis.* W.H. Freeman and Company.

[10] Jensen, L.J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., Bork, P., 2008. eggNOG: Automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, 36:D250–D254.

[11] Jmol. 2010. Jmol: An open-source Java viewer for chemical structures in 3D. http://www.jmol.org/

[12] La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., Raoult, D., 2008. "The virophage as a unique parasite of the giant mimivirus".

[13] Syvanen, P., 1984. Cross-species Gene Transfer; Implications for a New Theory of Evolution. Harvard Medical School, Boston, MA.

[14] Gogarten, P., 2000. "Horizontal Gene Transfer - A New Paradigm for Biology" Esalan Center for theory and research.

[15] Revesz, P., 2010. *Introduction to databases: From biological to spatio-temporal*, Springer, New York.

[16] Salton, M.J.R., Kim, K.S., 1996. Structure in*: Baron's Medical Microbiology* (Baron S et al., eds.) (4th ed.). Univ of Texas Medical Branch.

[17] Tatusov, R. L. et al., 2003. The COG database: An updated version including eukaryotes. *BMC Bioinformatics*, 4:41.

[18] Than C, Ruths D, Innan H, Nakhleh L. 2006. Identifiability issues in phylogeny-based detection of horizontal gene transfer. *Proceedings of Comparative Genomics*, 4205:215-229.

[19] Triplet, T., Shortridge, M., Griep, M., Stark, J. L., Powers, R., Revesz, P., 2010. PROFESS: A protein function, evolution, structure and sequence database. *Database – The Journal of Biological Databases and Curation*, doi number 10.1093.

[20] Wei, S., Cowen, L., Brodley C., Brady, A., Sculley, D., Slonim, D.K.. 2008. A distance-based method for detecting horizontal gene transfer in whole genomes.