University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Computer Science and Engineering: Theses, Dissertations, and Student Research

Computer Science and Engineering, Department of

5-1-2011

Protein Structure – Based Method for Identification of Horizontal Gene Transfer in Bacteria

Swetha Billa
University of Nebraska-Lincoln, swethabilla@gmail.com

Follow this and additional works at: http://digitalcommons.unl.edu/computerscidiss

Part of the Bioinformatics Commons, Computer Engineering Commons, and the Computer Sciences Commons

Billa, Swetha, "Protein Structure – Based Method for Identification of Horizontal Gene Transfer in Bacteria" (2011). *Computer Science and Engineering: Theses, Dissertations, and Student Research.* Paper 23.
http://digitalcommons.unl.edu/computerscidiss/23

This Article is brought to you for free and open access by the Computer Science and Engineering, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Computer Science and Engineering: Theses, Dissertations, and Student Research by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

PROTEIN STRUCTURE – BASED METHOD FOR IDENTIFICATION OF HORIZONTAL GENE TRANSFER IN BACTERIA

By

Swetha Billa

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Peter Revesz and Professor Mark A. Griep

Lincoln, Nebraska

May, 2011

PROTEIN STRUCTURE – BASED METHOD FOR IDENTIFICATION OF HORIZONTAL GENE TRANSFER IN BACTERIA

Swetha Billa, M.S.

University of Nebraska, 2011

Advisor: Peter Z. Revesz and Mark A. Griep

Horizontal Gene Transfer is defined as the movement of genetic material from one strain of species to another. Bacteria, being an asexual organism were always believed to transfer genes vertically. But recent studies provide evidence that shows bacteria can also transfer genes horizontally.

HGT plays a major role in evolution and medicine. It is the major contributor in bacterial evolution, enabling species to acquire genes to adapt to the new environments. Bacteria are also believed to develop drug resistance to antibiotics through the phenomenon of HGT. 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.

Methods to detect HGT events have been studied extensively but no method can accurately detect all the transfers between the organisms. This thesis discusses the various methods to detect HGT that were studied earlier and provides a new unique protein structure-based method to detect HGT in bacteria. This method makes use of Z-score similarities between the protein structures. This method uses functions of BLAST

and DaliLite to work with protein sequence and structural similarities. Also ‘Jmol’, a java viewer tool is used for visual structural comparisons and sequence alignment. This thesis is an interdisciplinary effort, using both biological tools and computer algorithm to detect Horizontal Gene Transfer in bacteria.

Acknowledgements

I am heartily thankful to my supervisor Dr. Peter Revesz for accepting me as his student and for providing me the opportunity to work with Bioinformatics. I would also like to thank him for his encouragement, guidance and unfailing support throughout the completion of this thesis.

I owe my deepest gratitude to Prof. Mark Griep for taking time out of his busy schedule and providing his valuable expertise and guiding me through the completion of this thesis. I would also like to thank Prof. Jitender Deogun for agreeing to be on my committee and for providing valuable feedback and comments on my thesis.

I am grateful for my dearest friend Venkat Ram Santosh for his motivation and encouragement in a number of ways. I would also like to thank Dipty Singh for always being there for me.

I would like to thank my parents for always being supportive of me and believing that I could get through this, my brother and sister for giving me inspiration.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of this project.

Contents

List of Figures

List of Tables

List of Abbreviations

Introduction 1
1.1 Horizontal gene transfer 1
1.2 How to determine HGT ? 2
1.3 Why is it important to study HGT ? 2
Background 4
2.1 Basic Biology Overview 4
2.1.1 Amino Acids 4
2.1.2 Proteins 5
2.1.3 Nucleotides 6
2.1.4 Nucleic Acids 6
2.1.5 Codons 8
2.1.6 Bacteria 9
2.1.7 Virus 11
2.2 Basics of HGT 11
Mechanisms of HGT 11
2.3 Biological Databases Used 12
2.3.1 Protein Data Bank: 12
2.3.2 COG database: 13
2.3.3 GO database: 14
2.3.4 PROFESS Database: 14
2.5 Overview of Existing HGT Detection Methods 15
2.5.1 Compositional Methods 15

2.5.2 Phylogeny-Based Methods 16
2.5.3 Distance-Based Detection of HGT 16
Methodology 17
3.1 Method 18
3.2 Automation 21
3.2.1 Data modeling: 22
Analysis and Results 25
4.1 Summary of Suspected HGT 32
4.2 Detailed Analysis of COG-1309 33
4.3 Detailed Analysis of COG-4948 37
4.4 False Positives 38
Conclusion and Future Work 39
5.1 Conclusion 39
5.2 Future Work 40

List of Figures

Figure 2.1: DNA and RNA nucleotide structure……. 7
Figure 2.2: Cell walls of gram-positive and gram-negative bacteria…….. 10
Figure 3.1: E-R Diagram of the database…… 23
Figure 3.2: Module of structurally comparing opposite gram bacterial proteins…… 24
Figure 4.1: Pre-calculated jFATCAT-rigid structure alignment results 1VI0 (*Bacillus subtilis*) vs. 1SGM (*Bacillus Subtilis*)……... 34
Figure 4.2: Pre-calculated jFATCAT-rigid structure alignment results 1VI0 (*Bacillus subtilis*) vs. 2UXH (*Pseudomonas putida*)…… 34
Figure 4.3: Sequence alignment results 1VI0 (*Bacillus subtilis*) vs. 1SGM (*Bacillus subtilis*)…….. 35
Figure 4.4: Sequence alignment results 1VI0 (*Bacillus subtilis*) vs. 2UXH (*Pseudomonas putida*)…….. 36

List of Tables

Table 2.1: List of standard proteins……... 5

Table 2.2: Codes for primary nucleobases……. 6

Table 3.1: Example of anomalous COG identified in the preliminary analysis…… 20

Table 3.2: Sample result set from PROFESS…… 22

Table 3.3: Sample result set from DaliLite…… 22

Table 4.1: Z-score structural comparison between *Bacillus subtilis* and *Escherichia coli*……..... 27

Table 4.2: Z-score structural comparison between *Bacillus subtilis* and *Pseudomonas aeruginosa*……. 28

Table 4.3: Z-score structural comparison between *Bacillus subtilis* and *Pseudomonas putida*……. 28

Table 4.4: Z-score structural comparison between *Bacillus subtilis* and *Haemophilus influenzae*…….. 29

Table 4.5: Z-score structural comparison between *Bacillus subtilis* and *Helicobacter pylori*……... 29

Table 4.6: Summary of candidates for HGT among the compared protein structures……... 30

Table 4.7: Summary of Proteins suspected as HGT…….. 32

Table 4.8: COG-1309 in Comparison between *Bacillus subtilis* and *Pseudomonas putida*…… 33

Table 4.9: COG-4948 in Comparison between *Bacillus subtilis* and *Pseudomonas putida*…… 37

List of Abbreviations

HGT- Horizontal Gene Transfer

LGT- Lateral Gene Transfer

AR-Antibiotic resistance

DNA- Deoxyribonucleic acid

RNA- Ribonucleic acid

CPASS- Comparison of Protein Active Site Structures

PDB- Protein Data Bank

COG- Clusters of Orthologous Groups

GO- Gene Ontology

NCBI- National Center for Biotechnology Information

PROFESS- PROtein Functions, Evolution, Structures and Sequences Database

BLAST- Basic Local Alignment Search Tool

Chapter 1

Introduction

1.1 Horizontal gene transfer

Horizontal gene transfer (HGT) or *lateral gene transfer* is the passing of genetic material from one organism to another, other than by descent in which genetic information travels through the generations as the cell divides. In nature, gene transfer occurs between two same species 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 passed on from parent to the offspring vertically.

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 has first been described in a Japanese publication in 1959, which describes about the transfer of antibiotic resistance from one bacterium to another [1]. The phenomenon of HGT is quite significant in prokaryotes and some unicellular eukaryotes. Importance of HGT in the evolution of multicellular organisms has not been extensively studied.

1.2 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 [2]. 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 context are thought to have been transferred or introduced from a foreign source [3].

1.3 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 [4] [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 of antibiotic resistance (AR) in bacteria requires

investigation of the antibiotic resistance mechanism [6]. Hence studies on HGT will help provide a greater incite on how this can be curbed.

Chapter 2

Background

2.1 Basic Biology Overview

2.1.1 Amino Acids

Amino acids are molecules containing an amine group, a carboxylic acid group and a side chain that varies between different amino acids. The key elements of an amino acid are carbon, hydrogen, oxygen, and nitrogen. Amino acids play a major role in metabolism. One or more amino acids together form a Protein. The International Union of Pure and Applied Chemistry (IUPAC) has a system for giving codes to identify long sequences of amino acids. This would allow for these sequences to be compared to try to find homologies. These codes consist of either a one letter code or a three letter code.

For example: Alanine: Single letter code is ‘A’, Three letter code is ‘Ala’. These codes make it easier and shorter to write down the amino acid sequences that make up proteins. The 20 standard proteins and their codes are tabulated as follows:

Table 2.1: List of standard proteins.

Amino Acid	3-Letter Code	1-Letter Code
Alanine	*Ala*	*A*
Arginine	*Arg*	*R*
Asparagine	*Asn*	*N*
Aspartic acid	*Asp*	*D*
Cysteine	*Cys*	*C*
Glutamine	*Gln*	*Q*
Glutamic acid	*Glu*	*E*
Glycine	*Gly*	*G*
Histidine	*His*	*H*
Isoleucine	*Ile*	*I*
Leucine	*Leu*	*L*
Lysine	*Lys*	*K*
Methionine	*Met*	*M*
Phenylalanine	*Phe*	*F*
Proline	*Pro*	*P*
Serine	*Ser*	*S*
Threonine	*Thr*	*T*
Tryptophan	*Trp*	*W*
Tyrosine	*Tyr*	*Y*
Valine	*Val*	*V*

2.1.2 Proteins

These are linear chains of amino acids typically folded into a globular of fibrous form in a biologically functional way. Amino acids are linked together in various combinations to form a wide range of proteins. Since there are 20 standard amino acids, there are lot of different protein chains that can be built. Many of the proteins that make up our body may contain hundreds of amino acids. The sequence of amino acids in a protein is defined by the sequence of a gene.

The folding of proteins to form a defined structure is variable. Some proteins function without any folding, while some fold in rigid structures with minimum or no changes at

all. These proteins therefore have a single structure. There are other proteins which undergo rearrangements in their structures, so they exist in different conformations.

2.1.3 Nucleotides

These are molecules that, when joined together, make up the structural units of RNA and DNA. A nucleotide is composed of a *nucleobase* (nitrogenous base), a five-carbon sugar (either ribose or 2'-deoxyribose), and one to three phosphate groups. The International Union of Pure and Applied Chemistry (IUPAC) also has a system for giving codes to identify nucleotide bases. The codes for the primary nucleobases are given below.

Table 2.2: Codes for primary nucleobases.

IUPAC nucleotide code	Base
A	*Adenine*
C	*Cytosine*
G	*Guanine*
T	*Thymine*
U	*Uracil*

The nucleotide bases are made up of purines (adenine and guanine) and pyrimidines (cytosine and thymine). These nucleotide base codes make the genome of an organism much smaller and easier to read.

2.1.4 Nucleic Acids

These are linear chains of nucleotides. Nucleic acids are divided into two major forms Deoxyribonucleic acid (DNA) and Ribonucleic acid (RNA). Both of these nucleic acids are present in all kinds of living organisms.

2.1.4.1 Deoxyribonucleic acid (DNA)

It is a hereditary material. DNA contains the pyrimidine bases thymine and cytosine and the purine bases adenine and guanine. Also if we know what the DNA sequence is, we can work out which amino acids the protein must contain and in what order. HGT occurs at the DNA level [7]. DNA has a double helical structure. The structure of DNA was first proposed by James Watson and Francis in 1953.

2.1.4.2 Ribonucleic acid (RNA)

RNA is similar to DNA except that the thymine is replaced by uracil. In some viruses where DNA is not available, RNA acts as the hereditary material.

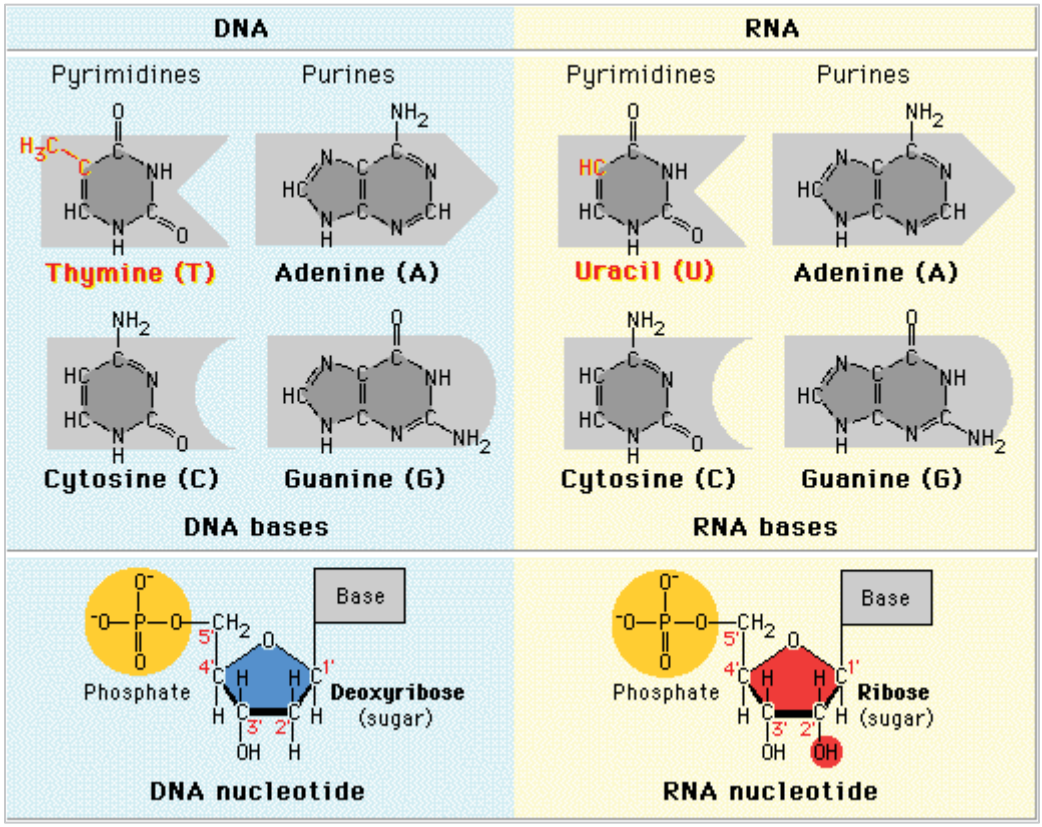

Figure 2.1: DNA and RNA nucleotide structure.

The central dogma of Biology as proposed by Francis Crick in 1958 states that information flows from DNA to RNA to Protein [7].

2.1.5 Codons

A codon codes for a single amino acid, each codon consists of 3 nucleotides. Information for the genetic code is stored in the sequence of three nucleotide bases of DNA called base triplets, which act as a template for which messenger RNA (mRNA) is transcribed. A sequence of three successive nucleotide bases in the transcript mRNA is called a codon.

Codons are complimentary to base triplets in the DNA. For example, if the base triplet in the DNA sequence is GCT, the corresponding codon on the mRNA strand will be CGA. Because there are four possible nucleotide bases to be incorporated into a three base sequence codon, there are 64 possible codons ($4^3 = 64$). Sixty-one of the 64 codons signify the 20 known amino acids in proteins. These codons are ambiguous codons, meaning that more than one codon can specify the same amino acid. For example, in addition to GCA, five additional codons specify the amino acid arginine. Because the RNA/DNA sequence cannot be predicted from the protein, and more than one possible sequence may be derived from the same sequence of amino acids in a protein, the genetic code is said to be degenerate. The remaining three codons are known as stop codons and signal one of three termination sequences that do not specify an amino acid, but rather stop the synthesis of the polypeptide chain.

2.1.6 Bacteria

Bacteria are single celled microscopic organisms. They do not have a membrane enclosed nucleus nor other membrane-enclosed organelle like mitochondria and chloroplasts. The study of bacteria is called bacteriology, which is a branch of microbiology.

2.1.6.1 Classification of Bacteria

Until recently classification of bacteria has been done on the basis of traits such as:

- shape
    - **bacilli**: rod-shaped
    - **cocci**: spherical
    - **spirilla**: curved walls
- ability to form spores
- method of energy production (glycolysis for anaerobes, cellular respiration for aerobes)
- nutritional requirements
- reaction to the Gram stain.

2.1.6.1: The Gram Staining Procedure

The Gram stain is a *differential stain* which allows most bacteria to be divided into two groups, Gram-positive bacteria and Gram-negative bacteria. The *Gram stain* is named after the 19th century Danish bacteriologist Christian Gram who developed it in 1884. The bacterial cells are first stained with a purple dye called crystal violet. Then the preparation is treated with alcohol or acetone. This washes the stain out of *Gram-*

negative cells. To see them now requires the use of a counterstain of a different color (e.g., the pink of safranin). Bacteria that are not decolorized by the alcohol/acetone wash are *Gram-positive*.

The gram stain procedure distinguishes between two fundamentally different kinds of bacterial cell walls which are made up of peptidoglycan and reflects a natural division among the bacteria. The technique is based on the fact that the *Gram positive* cell wall has a stronger attraction for crystal violet when Gram's iodine is applied than does the *Gram negative* cell wall [8]. Gram's iodine is known as a *mordant*. It is able to form a complex with the crystal violet that is attached more tightly to the *Gram-positive* cell wall than to the *Gram-negative* cell wall. This complex can easily be washed away from the Gram-negative cell wall with ethyl alcohol. Gram-positive bacteria, however, are able to retain the crystal violet and therefore will remain purple after decolorizing with alcohol. Since Gram-negative bacteria will be colorless after decolorizing with alcohol, counterstaining with safranin will make them appear pink.

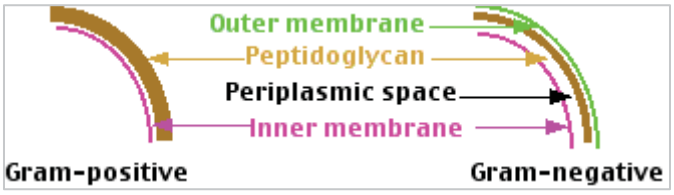

Figure 2.2: Cell walls of gram-positive and gram-negative bacteria

The chemical basis of the gram staining procedure was not understood by Gram and is still not fully understood today. It is known, however, that the two groups of bacteria

have very different cell walls and that the type of cell wall dictates the way a bacterium responds to the Gram stain. The Gram stain is probably the most commonly used staining procedure in microbiology. The two phyla of bacteria that we concentrate on in this research are *Firmicutes* which is gram positive and *Proteobacteria* which is gram negative.

2.1.7 Virus

A virus is a small infectious agent that can replicate only inside the living cells of organisms. Most viruses are too small to be seen directly with a light microscope. Viruses infect all types of organisms, from animals and plants to bacteria and archaea [9]. Group of viruses that infect bacteria are called bacteriophages also called phages, or bacterial viruses. Thousands of varieties of phage exist, each of which may infect only one type or a few types of bacteria

2.2 Basics of HGT

Mechanisms of HGT

Exchange of genetic material can occur in 3 different ways in bacteria: Transformation, Conjugation and Transduction.

Transformation: A process of alteration of the gene by introducing foreign genetic material. This is more common in bacteria than in eukaryotes. This is the most common method of HGT used in laboratories to insert genes into bacteria for experimental purposes. Only short DNA can be exchanged through this process.

Conjugation: A process in which a bacterial cell transfers genetic material to another cell through cell-cell contact. This can occur between distantly related bacteria or between a bacteria and eukaryotic cell. This process can transfer long fragments of DNA. The genes required for conjugation are usually found on a plasmid DNA.

Transduction: A process in which a DNA is moved from one bacterium to another by a bacterial virus. This bacterial virus is called a *bacteriophage* or simply *phage*. A phage inserts its DNS into a recipient and modifies its DNA. This method requires the donor and recipient to share the cell surface receptors. Hence it is usually seen in closely related bacteria. The length of the DNA transferred depends on the size of the phage head.

2.3 Biological Databases Used

2.3.1 Protein Data Bank: (Website: *http://www.pdb.org/)*

PDB [10] is a worldwide repository containing information about experimentally determined 3D structures of large biological molecules including proteins and nucleic acids. The data of these molecules is derived experimentally primarily from X-ray crystallography, nuclear magnetic resonance (NMR) and cryo-electron microscopy. These molecules are part of all living organisms like bacteria, yeast, insects, plants, animals and humans. Study of the structure and shape of the molecule provides us an insight into the functioning of the molecule. Hence PDB provides its users tools with which a structure's role in human health and disease can be deduced and thus help in drug development.

PDB provides accurate and timely structural information to a worldwide community of users regardless of local hardware and software and geographic location [11]. PDB archive is available to users free of cost. The archive consists of structures that range from that of tiny proteins and bits of DNA to complex molecular machines like the ribosome. PDB also has a website where users can perform queries on the data based on sequence, structure and function, analyze and visualize the results. As of this writing there are 68139 structures in the PDB archive.

2.3.2 COG database: (Website: http://www.ncbi.nlm.nih.gov/COG)

Clusters of Orthologous Groups of proteins database [12] is maintained and updated by the National Center for Bio-technology Information (NCBI). It phylogenetically classifies the proteins encoded by the complete genomes. Each COG includes proteins that are thought to be connected through vertical evolutionary descent. The COGs are generated by comparing the protein sequences of complete proteins. Each COG is a group of three or more proteins. The COG database is updated periodically as new genomes become available. The updated version of COG database consists of eukaryotes too. This database serves as a useful tool for studies on genome evolution.

The COG database collection currently consists of 138,458 proteins from 66 genomes. The database also consists of a program called COGNITOR which assigns new proteins from newly sequenced genomes to the COGs already in the database [13].

2.3.3 GO database: (Website: http://amigo.geneontology.org/)

Gene Ontology database [14] is a relational database, consisting of GO ontologies and the annotations of genes and gene products to the terms in the GO. It provides a controlled vocabulary of terms for describing genes product characteristics and gene product data. It addresses the need to have consistent descriptions of gene products in various databases. The GO database is populated with data from the most recent version of the ontology and annotation files contributed by the members of the GO consortium. It is currently being maintained as a MySQL database. The database can be accessed online using the AmiGO browser and search engine. Along with enabling the users to download terms and annotations, Amigo provides tools for analyzing and data processing.

2.3.4 PROFESS Database: (Website: http://cse.unl.edu/~profess)

PROtein Function, Evolution, Structure and Sequence database [15] is a framework that integrates various biological databases. It was developed at University of Nebraska-Lincoln, to assist in the functional and evolutionary analysis of the proteins. A predecessor system of PROFESS is the CPASS system, which enabled the comparison of protein active sites based on the structural similarity of the active sites of proteins [16].

Some of the databases integrated into PROFESS are : CATH (Class Architecture Topology and Homologous superfamily) database, COG (Clusters of Orthologous Groups) of proteins database, Gene Ontology, Protein Data Bank(PDB), Structural Classification of Proteins (SCOP), UniProt Knowledge Base, Protein Families (PFAM) database and Pancreatic Cell 'omics' Data (PCOD). In addition to that PROFESS also

includes ’all-against-all’ pairwise structural comparisons for all protein structures within their respective orthologous cluster.

With about 1100 molecular biological databases freely available online for users, PROFESS provides a unique interface for biologists and other users who are required to use more than one biological database to perform their studies, without worrying about designing their own database that fits their requirement. Data from the various core databases is updated every four months. This ongoing project promises to incorporate other biological databases based on the user feedback and their requirement.

2.5 Overview of Existing HGT Detection Methods

2.5.1 Compositional Methods

A gene which is horizontally transferred can contain recognizable signatures of its previous location since it comes from a different genomic background. Compositional methods use atypical nucleotide [17], atypical codon usage patterns [18] or their combination [19] to detect which genes in a genome have been horizontally gene transferred. Since over time the horizontally transferred genes adopt the signatures of the new genome, these methods can be used only on genes which have been transferred fairly recently. These methods are easily applicable to completely sequenced genomes. However, high rates of false positives and negatives have been observed in these methods.

2.5.2 Phylogeny-Based Methods

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.

2.5.3 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 drawbacks.

Chapter 3

Methodology

A protein structure - based method is devised in this thesis to identify HGT among organisms. This method makes use of the fact that similar structure of a protein would mean similar functionality. And when a protein is horizontally gene transferred from another organism, the structure of the protein would remain fairly similar to the protein from the donor organism, since it is trying to retain its functionality. The structure of the protein transferred may be different from proteins with similar functionality in the recipient organism. Hence to detect HGT, the goal would be identify anomalies in the structures of the proteins in an organism, with similar functionalities.

To identify these protein structure anomalies, we make use of the Cluster of Orthologous Group (COG) classification. According to this classification all proteins with similar functionality are categorized under the same COG number. And according to evolutionary theory they should have similar structures.

For this research we consider two phyla of bacteria i) *Firmicutes* and ii) *Proteobacteria*. Most of *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 [20]. 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.

3.1 Method

For this research we compare bacteria in each of the two phyla. For *Firmicutes* we chose *Bacillus subtilis* and from *Proteobacteria* we chose *Escherichia coli*. These two bacteria have the most number of identified structures in their respective phyla, as documented by the biological databases that we have used in this 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. Querying the PROFESS database we get the list of proteins studied in each of the bacteria and the COGs to which they belong to. The COG number uniquely identifies groups of proteins that have functional similarity.

Stage 2

As the second stage of the method we perform a structural comparison of the proteins. This again is a two-step process, as in we first structurally compare proteins in each of the COGs with in each organism and then we structurally compare proteins in each of the COGs among the two organisms. DaliLite program was used for the structural comparisons. The DaliLite program takes the input of two PDB ids and applies structural comparison algorithms and provides a result in the form of a Z- score which is the index for measuring structural similarity in proteins.

There are 494 proteins for *Bacillus subtilis* and 3264 proteins for *Escherichia Coli* that are 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. 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.

And 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 the 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 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 COG numbers.

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%. So 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.

The table below 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.

Table 3.1: Example of anomalous COG identified in the preliminary analysis.

COG Number	Bacillus subtilis	Escherichia coli	Comparison	Bacillus subtilis Normalized ↓	Escherichia coli Normalized	Comparison Normalized
454	12.09	35.7	9.71	0.34	1	0.27

3.2 Automation

With rapidly increasing number of organisms being studied by researchers and more number of proteins being crystallized in organisms, it would be a good idea to automate the process of identifying HGT.

The dataset containing all the protein structures in all the bacteria from the two phyla *Firmicutes* and *Proteobacteria* were downloaded from the PROFESS database. The following query was used.

```
SELECT link_cog_pdb.cog_number, link_pdb_taxon.pdb_id,
name,  lineage
FROM taxonomy, link_pdb_taxon, pdb, link_cog_pdb
WHERE (lineage LIKE '%Proteobacteria%'
       OR lineage LIKE '%Firmicutes%')
      AND link_pdb_taxon.taxon = taxonomy.taxon
      AND link_pdb_taxon.pdb_id=pdb.pdb_id
      AND pdb.cog_number = link_cog_pdb. cog_number ;
```

This query can be run directly on the web interface for PROFESS and result downloaded as a CSV (Comma Separated Value) file. There are about 9949 unique PDB IDs for *Proteobacteria* and 4298 unique PDB IDs for *Firmicutes* in PROFESS. The output from the query would be of the following format.

Table 3.2: Sample result set from PROFESS.

COG Number	PDB ID	Bacteria Name	Lineage
276	*2hk6*	*Bacillus Subtilis*	*Firmicutes*
280	*1td9*	*Bacillus Subtilis*	*Firmicutes*
299	*1cdd*	*Escherichia Coli*	*Proteobacteria*
...	...	...	...

The automation process was twofold. Since we have large number of pair-wise structural comparisons to be done using the DaliLite user interface, it was more feasible to automate this process rather than entering the pairs into the web interface manually. This comparison was performed in Holland Computing Center at the University of Nebraska – Lincoln. The result of this automation process was Z-score measures for all combinations of proteins with in the same COG classification for all the bacteria listed in PROFESS database. Sample result set from the DaliLite is as follows:

Table 3.3: Sample result set from DaliLite.

COG Number	PDB ID - 1	PDB ID - 2	Z - Score
270	*10mh*	*8mht*	*52.4*
270	*10mh*	*9mht*	*51.7*
221	*117e*	*1e6a*	*50*
221	*117e*	*1e9g*	*50.1*
...	...	...	...

3.2.1 Data modeling:

A database is modeled with the dataset obtained from DaliLite and PROFESS, easing the process of writing custom queries for further analysis. The database consists of three primary tables. The Entity-Relationship diagram of the database is as follows:

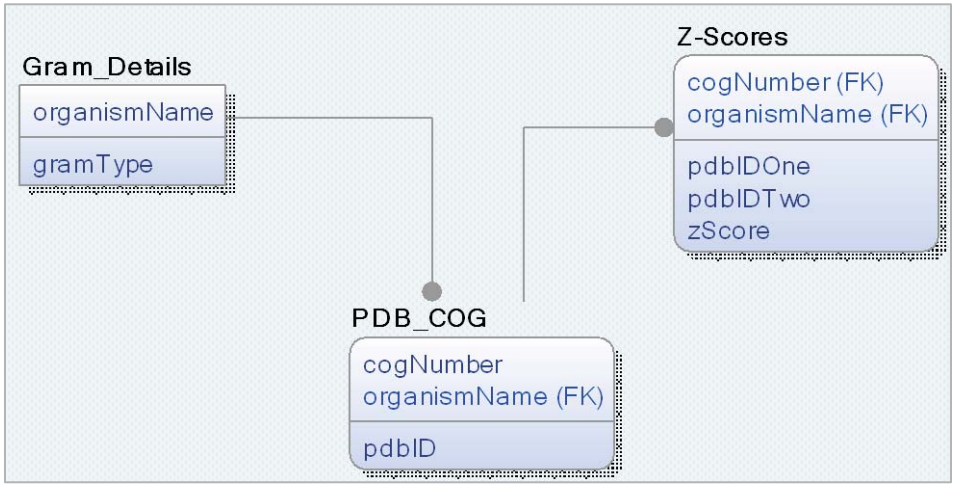

Figure 3.1: E-R Diagram of the database

The second module of automation includes a user interface in which opposite gram bacteria can be compared to get average of Z-score in each COG classification. The result set can then be exported to a spreadsheet on which further analysis is performed. The user interface looks like follows. Modules to add / modify organisms' data, add/modify PDB and COG data files, add/ modify Z-score data files has also been incorporated in the interface.

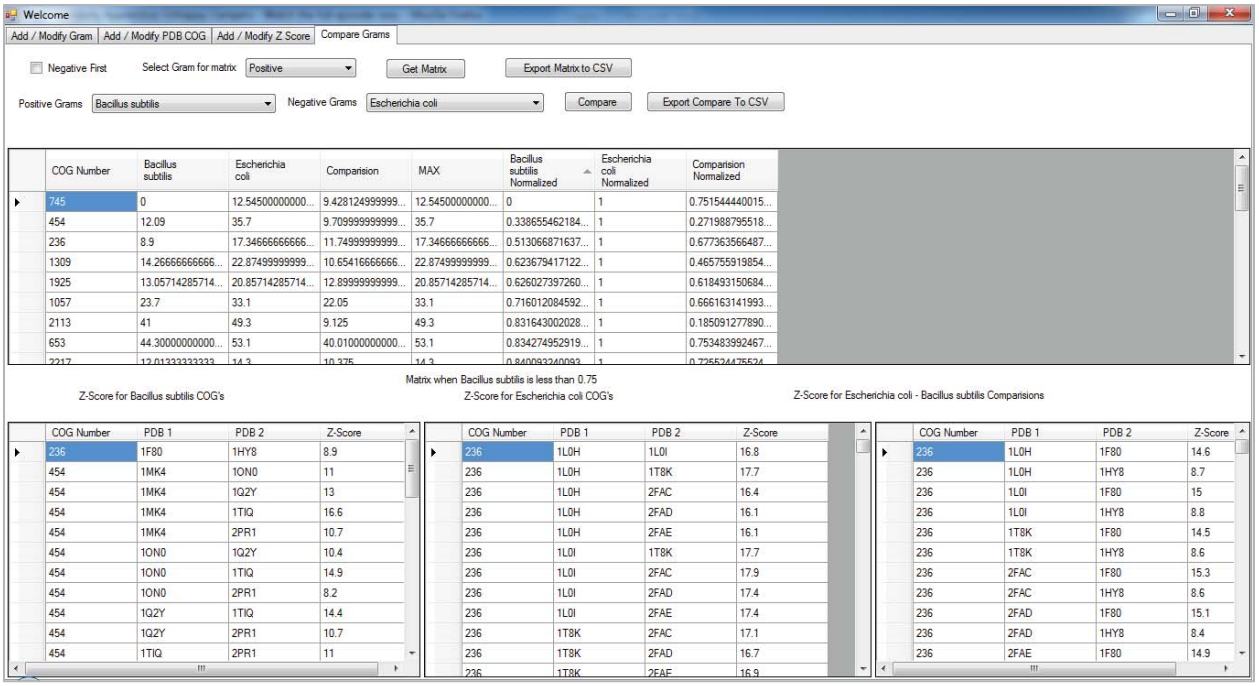

Figure 3.2: Module of structurally comparing opposite gram bacterial proteins.

This program has been used to compare *Bacillus subtilis* with all other bacteria belonging to *Proteobacteria*, to detect possible HGT in *Bacillus subtilis* from *Proteobacteria*.

Chapter 4

Analysis and 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 19 gram negative organisms only 5 organisms had matching COG numbers with the ones in *Bacillus subtilis*.

The Gram negative organisms compared with *Bacillus subtilis* are:

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 5 gram negative organisms and tabulated in tables 4.1 – 4.5. This comparison is performed only for the common COGs among the two different classes of bacteria i.e., 1 Gram positive organism and 5 Gram negative organisms.

In the tables 4.1 - 4.5 the COG numbers which have average Z-score values less than or equal to 75% of the average Z-score values in the other organism within the same COG are highlighted. This can be identified from the row labeled '*Bacillus subtilis Normalised*' in each of the tables. The value .75 is chosen as threshold value to identify anomalous COGs.

For example in Table 4.1 for COG 454 which is common in *Bacillus subtilis* and *Escherichia coli*, the average Z-score of *Bacillus subtilis* is *12.09* and the average Z-score of *Escherichia coli* is *35.7*. The normalized Z-score for *Bacillus subtilis* is *0.34* which is less than the chosen threshold value of 75%, which is because of the significant difference in average Z-scores of the two bacteria. Hence further analysis of this particular COG is required since it might provide evidence as to why the average Z-score of *Bacillus subtilis* is very less compared to *Escherichia coli,* which might be attributed to the fact of HGT occurrence in *Bacillus subtilis.* It is to be noted that we are concerned only with the normalized Z-score values of *Bacillus subtilis* and not *Escherichia coli* because we are trying find evidence of HGT in *Bacillus subtilis* from other bacteria.

The above mentioned procedure is a preliminary step to identify anomalous COGs. We focus our interest on the highlighted COGs in the tables and perform further analysis.

Table 4.1: Z-score structural comparison between *Bacillus subtilis* and *Escherichia coli*.

COG Number	Bacillus subtilis	Escherichia coli	Comparison	Bacillus subtilis Normalized ↓	Escherichia coli Normalized	Comparison Normalized
745	0	12.55	9.43	0	1	0.75
454	12.09	35.7	9.71	0.34	1	0.27
236	8.9	17.35	11.75	0.51	1	0.68
1309	14.27	22.88	10.65	0.62	1	0.47
1925	13.06	20.86	12.9	0.63	1	0.62
1057	23.7	33.1	22.05	0.72	1	0.67
2113	41	49.3	9.13	0.83	1	0.19
653	44.3	53.1	40.01	0.83	1	0.75
2217	12.01	14.3	10.38	0.84	1	0.73
4948	44.77	52.3	37.68	0.86	1	0.72
834	23.2	25.73	22.28	0.9	1	0.87
784	21.48	23.7	16.56	0.91	1	0.7
2050	21.3	22.59	15.12	0.94	1	0.67
2132	67.53	71.01	44.35	0.95	1	0.62
2351	23.93	24.6	17.32	0.97	1	0.7
207	42.13	43.21	33.62	0.98	1	0.78
34	62.2	60.31	44.92	1	0.97	0.72
171	42.83	40.23	36.31	1	0.94	0.85
363	45.7	45.07	37.39	1	0.99	0.82
500	39.5	12	16.55	1	0.3	0.42
503	27.63	20.67	11.8	1	0.75	0.43
511	16.27	12.9	7.93	1	0.79	0.49
526	19.2	13.38	10.2	1	0.7	0.53
563	37.1	31.67	29.25	1	0.85	0.79
596	47.6	24.8	28.13	1	0.52	0.59
604	55	36.77	41.1	1	0.67	0.75
789	15.2	10.35	8.33	1	0.68	0.55
840	29.6	5.6	1.63	1	0.19	0.05
1278	13.2	5.1	7.21	1	0.39	0.55
1609	41.6	28.96	28.87	1	0.7	0.69
1985	53.1	51.97	38.17	1	0.98	0.72
2141	63.7	53.3	32.5	1	0.84	0.51
2202	23.5	22.58	10.37	1	0.96	0.44

Table 4.2: Z-score structural comparison between *Bacillus subtilis* and *Pseudomonas aeruginosa.*

COG Number	Bacillus subtilis	Pseudomonas aeruginosa	Comparison	Bacillus subtilis Normalized ↓	Pseudomonas aeruginosa Normalized	Comparison Normalized
1057	23.7	36.63	21.58	0.65	1	0.59
689	36.27	42.5	33.68	0.85	1	0.79
454	12.09	13.3	13.28	0.91	1	1
1309	14.27	11.67	12.78	1	0.82	0.9
1846	14.6	14.11	13.96	1	0.97	0.96

Table 4.3: Z-score structural comparison between *Bacillus subtilis* and *Pseudomonas putida.*

COG Number	Bacillus subtilis	Pseudomonas putida	Comparison	Bacillus subtilis Normalized ↓	Pseudomonas putida Normalized	Comparison Normalized
1309	14.27	32.4	15.03	0.44	1	0.46
4948	44.77	64.16	43.68	0.7	1	0.68
1304	52.8	62.51	25.62	0.84	1	0.41
1902	62.15	52.47	48.28	1	0.84	0.78

Table 4.4: Z-score structural comparison between *Bacillus subtilis* and *Haemophilus influenzae.*

COG Number	Bacillus subtilis	Haemophilus influenzae	Comparison	Bacillus subtilis Normalized ↓	Haemophilus influenzae Normalized	Comparison Normalized
2050	21.3	28.8	14.55	0.74	1	0.51
822	21	15.4	6.7	1	0.73	0.32
1854	30.02	27	21.35	1	0.9	0.71

Table 4.5: Z-score structural comparison between *Bacillus subtilis* and *Helicobacter pylori*.

COG Number	Bacillus subtilis	Helicobacter pylori	Comparison	Bacillus subtilis Normalized ↓	Helicobacter pylori Normalized	Comparison Normalized
745	0	11.92	8.58	0	1	0.72
171	42.83	39.8	27.96	1	0.93	0.65

The following table gives the summary of the proteins structure comparisons performed in our preliminary analysis.

Table 4.6: Summary of candidates for HGT among the compared protein structures.

COG	Bacterial Pairs		Findings
	Number of Structures in Bacillus	*Number of Structures in 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
	Number of Structures in Bacillus subtilis	*Number of Structures in Pseudomonas aeruginosa*	
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

COG	Bacterial Pairs		Findings
	Number of Structures in Bacillus subtilis	*Number of Structures in Pseudomonas putida*	
1309	3	5	Most likely a good example of HGT
4948	3	5	Most likely a good example of HGT
	Number of Structures in Bacillus subtilis	*Number of Structures in Haemophilus influenzae*	
2050	2	2	The two Gram-positive protein structures are of the same protein and one of the protein structures of the Gram-negative organism is a protein fragment.
	Number of Structures in Bacillus subtilis	*Number of structures in Helicobacter pylori.*	
745	2	4	The two Gram-positive protein structures are completely dissimilar. Two of the Gram negative structures are same with different conformations, one is a protein fragment.

4.1 Summary of Suspected HGT

A further detailed analysis of the proteins in these candidate HGTs resulted in identification of the proteins 1VI0 in COG-1309 and 2GGE in COG-4948 as possible HGT to *Bacillus subtilis*.

Table 4.7: Summary of Proteins suspected as HGT.

PDB-ID	COG	ΔZ-score	Receiving Bacteria	Donor Bacteria
1VI0	1309	3.49	Bacillus subtilis	Pseudomonas putida
2GGE	4948	8.49	Bacillus subtilis	Unknown

*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.

4.2 Detailed Analysis of COG-1309

Table 4.8: COG-1309 in Comparison between *Bacillus subtilis* and *Pseudomonas putida.*

Bacillus subtilis versus each other			
	1RKT	1SGM	1VI0
1RKT		12.3	15.5
1SGM			15
1VI0			

Pseudomonas putida proteins versus each other					
	2UXH	2UXI	2UXO	2UXP	2UXU
2UXH		32	32.4	32.5	32.3
2UXI			32.3	32.4	32.3
2UXO				32.8	32.5
2UXP					32.5
2UXU					

Bacillus subtilis versus Pseudomonas putida proteins					
	2UXH	2UXI	2UXO	2UXP	2UXU
1RKT	15.8	15.8	15.9	15.8	15.9
1SGM	11.9	11.9	11.9	11.9	11.9
1VI0	17.3	17.2	17.3	17.5	17.5

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. This is done using the Jmol tool.

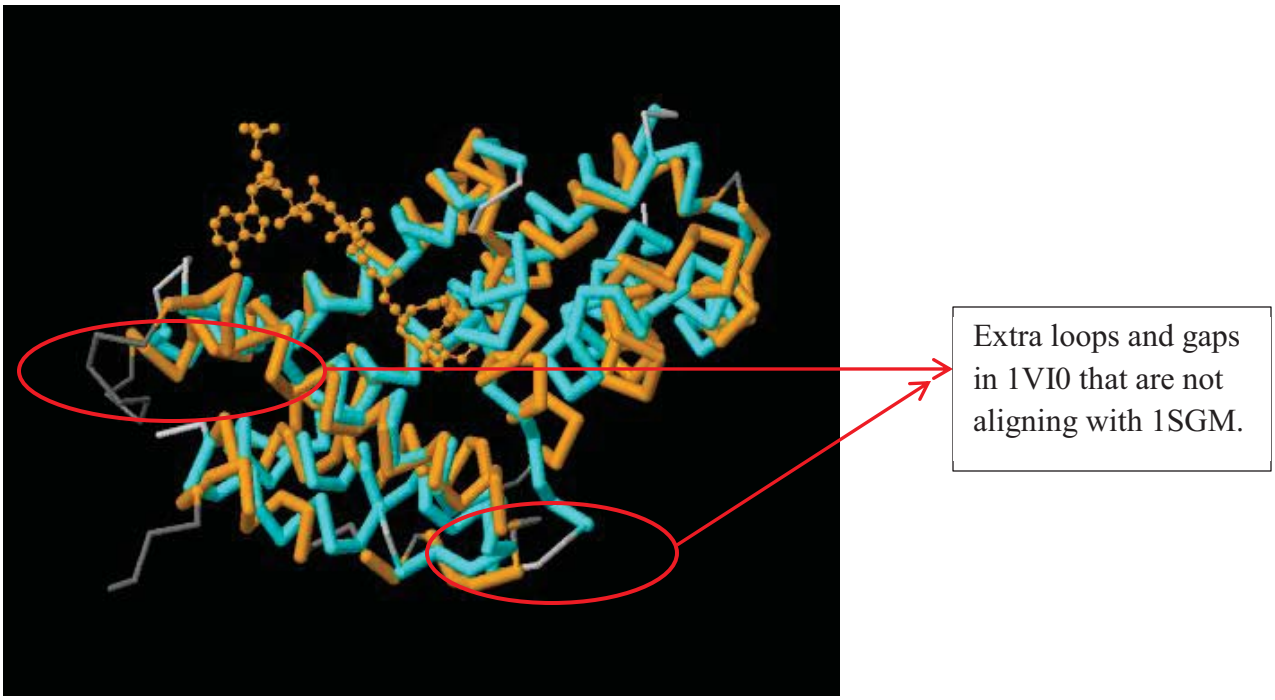

Figure 4.1: Pre-calculated jFATCAT-rigid structure alignment results 1VI0 (*Bacillus subtilis*) vs. 1SGM (*Bacillus Subtilis*).

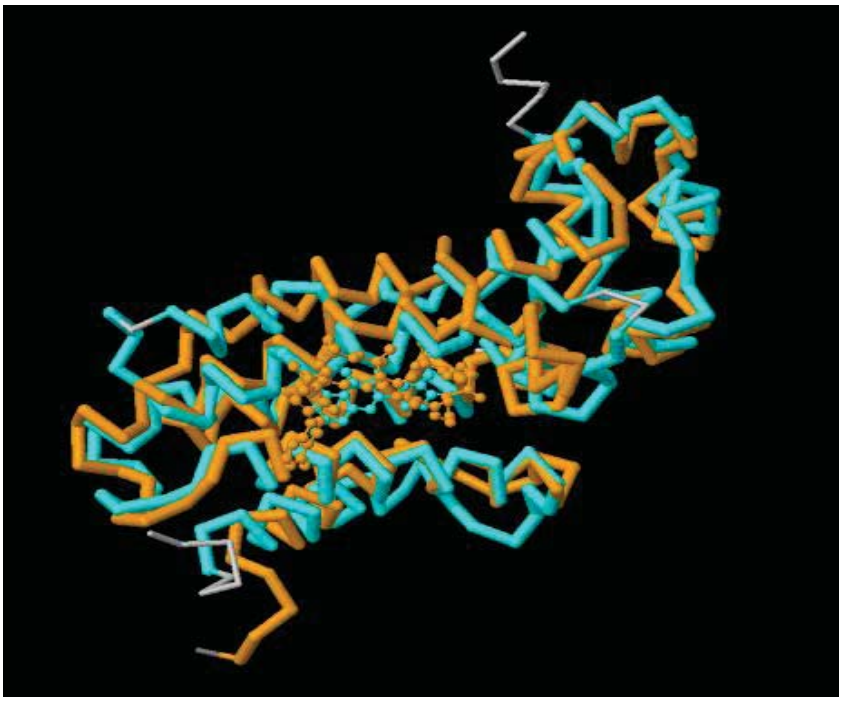

Figure 4.2: Pre-calculated jFATCAT-rigid structure alignment results 1VI0 (*Bacillus subtilis*) vs. 2UXH (*Pseudomonas putida).*

```
6:A          20:A      30:A      40:A        50:A      60:A      70:A
|   |    .   |    .   |    .   |    .      |    .   |    .   |
PKYMQIIDAAVEVIAENGYHQSQVSKIAKQAGVADGTIYLYFK--NKEDILISLFKEKMGQFIERMEEDI
.....|:..|.......|||......|.|:.|... ....:|.  .||::.|...............::..
DSREKILHTASRLSQLQGYHATGLNQIVKESGAPK-GSLYHFFPNGKEELAIEAVTYTGKIVEHLIQQSM
|   |    .   |    .   |    .   |     .    |    .   |    .   |
6:A          20:A      30:A      40:A       50:A      60:A      70:A

      80:A      90:A       100:A     110:A        120:A     130:A
 .    |    .   |    .     |    .   |       .    |    .   |    .
KEKATAKEKLALVISKHFSLLAG-DHNLAIVTQLELRQS---NLELRQKINEILKGYLNILDGILTEGIQ
.|.....|.:.|.|.|..|...  :....|...|...:.   ...||.....:.|.:..:....|.|...
DESSDPVEAIQLFIKKTASQFDNTESIKGIPVGLLASETALISEPLRTVCMKVFKSWEAVFARKLMENGF
.    |    .   |    .   |    .   |    .    |    .   |    .   |
     80:A      90:A      100:A     110:A     120:A     130:A     140:A

140:A     150:A     160:A     170:A     180:A     190:A
|    .   |    .   |    .   |    .   |    .   |
SGEIKEGLDVRLARQMIFGTIDETVTTWVMNDQKYDLVALSNSVLELLVSG
        :.......:|...|:..:.....|  |... .|  .:.|.:...
AEE-----EANQLGTLINSMIEGGIMLSLTN--KDKTPLL--LIAEQIPVL
.         |    .   |    .   |      .      |    .
          150:A     160:A     170:A         180:A
```

Clearly shows more gaps in the alignment

Figure 4.3: Sequence alignment results 1VI0 (*Bacillus subtilis*) vs. 1SGM (*Bacillus subtilis*).

```
6:A           20:A      30:A      40:A      50:A      60:A      70:A
|   |    .    |    .    |    .    |    .    |    .    |    .    |
PKYMQIIDAAVEVIAENGYHQSQVSKIAKQAGVADGTIYLYFKNKEDILISLFKEKMGQFIERMEEDIK-
....|||:||.......|........||..|||..|.||.:|.||.::...|...................
ETRAQIIEAAERAFYKRGVARTTLADIAELAGVTRGAIYWHFNNKAELVQALLDSLHETHDHLARASESE
|        |    .    |    .    |    .    |    .    |    .    |    .    |
11:A      20:A      30:A      40:A      50:A      60:A      70:A      80:A


     80:A      90:A         100:A     110:A       120:A     130:A
.    |    .    |    .       |    .    |      .    |    .    |    .
EKATAKEKLALVISKHFSLLAG---DHNLAIVTQLELRQSN--LELRQKINEILKGYLNILDGILTEGIQ
:.......:...:...|..|..   ...:..:........  .|:||.............:...|...:.
DEVDPLGCMRKLLLQVFNELVLDARTRRINEILHHKCEFTDDMCEIRQQRQSAVLDCHKGITLALANAVR
    .    |    .    |    .    |    .    |    .    |    .    |    .    |
          90:A      100:A     110:A     120:A     130:A     140:A     150:A


140:A     150:A     160:A     170:A     180:A     190:A
|    .    |    .    |    .    |    .    |    .    |
SGEIKEGLDVRLARQMIFGTIDETVTTWVMNDQKYDLVALSNSVLELLVSGIHN
.|::...||...|....|..:|..:..|.:.....||........:.....:..
RGQLPGELDAERAAVAMFAYVDGLIRRWLLLPDSVDLLGDVEKWVDTGLDMLRL
    .    |    .    |    .    |    .    |    .    |
          160:A     170:A     180:A     190:A     200:A


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

Figure 4.4: Sequence alignment results 1VI0 (*Bacillus subtilis*) vs. 2UXH (*Pseudomonas putida*).

From the sequence alignment above we can see that apart from couple of gaps most part of the two sequences align seamlessly.

4.3 Detailed Analysis of COG-4948

Table 4.9: COG-4948 in Comparison between *Bacillus subtilis* and *Pseudomonas putida.*

Bacillus subtilis versus each other			
	1JPM	1TKK	2GGE
1JPM		59.2	36.5
1TKK			38.6
2GGE			

Pseudomonas putida proteins versus each other					
	1BKH	1F9C	1MUC	2MUC	3MUC
1BKH		63	64.6	64.4	64.3
1F9C			63.4	63.4	62.5
1MUC				65.3	65.2
2MUC					65.5
3MUC					

Bacillus subtilis versus Pseudomonas putida proteins					
	1BKH	1F9C	1MUC	2MUC	3MUC
1JPM	46.7	46.5	46.7	46.6	46.5
1TKK	46.3	46.5	46.5	46.5	46.3
2GGE	38.1	37.8	38	38.1	38.1

The protein 2GGE is observed to have lesser Z-scores when compared to the other proteins in *Bacillus subtilis* (1JPM and 1TKK). Hence we can say that this protein probably has been horizontally transferred from other organism. However we can cannot be sure that it has been transferred form *Pseudomonas putida.* Using our method with other classifications of bacterial phyla might help us identify the organism from which the protein has been transferred.

4.4 False Positives

A situation where erroneously a positive result is observed is termed as false positive. During our analysis we noticed many situations that might cause false positives. They are listed as follows:

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-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.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

A protein structure based method to detect horizontal gene transfer has been devised. 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, as in it uses protein structures rather than 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.

5.2 Future Work

The false positives discussed in chapter 4 can cause erroneous results. This method can be improved by eliminating the cases for false positives.

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. Many underlying problems exist with this database, some of which are as follows:

1. Like any other biological database, PDB is incomplete, as in it does not contain complete protein structure information for all the organisms. It's a constant growing collection of sets of protein structure data. So there is limited flexibility when 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 [21]. 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 [16].

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:

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.

References

[1] K. Ochiai, T. Yamanaka, K. Kimura, O. Sawada, "Inheritance of drug resistance (and its transfer) between Shigella strains and between Shigella and E. coli strains," 1959.

[2] Susanna KA, den Hengst CD, Hamoen LW, Kuipers OP., "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".

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

[4] Thomas C. Butler, "Horizontal Gene Transfer and the Emergence of Darwinian Evolution," 2006.

[5] Maxim D. Frank-Kamenetskii, Unraveling DNA: The Most Important Molecule Of Life., 1997.

[6] Li SONG, Yi-bao NING, Qi-jing ZHANG, Cheng-huai YANG, Guang GAO and Jian-feng 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.

[7] Julia Goodrich, "Phylogenetic Pipeline for the Detection of Horizontal Gene Transfer," p. 1.

[8] Samuel Baron, Medical Microbiology. Texas: The University of Texas Medical Branch at Galveston, 1996.

[9] Koonin EV, Senkevich TG, Dolja VV., "The ancient Virus World and evolution of cells," National Center for Biotechnology Information, pp. 1-29, 2006.

[10] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE., "The Protein Data Bank," Nucleic Acids Resource, pp. 235-42, 2000.

[11] Philip E Bourne, Kenneth J Addess, Wolfgang F Bluhm, Li Chen, Nita Deshpande, Zukang Feng, Ward Fleri, Rachel Green, Jeffrey C Merino-Ott, Wayne Townsend-Merino, Helge Weissig, John Westbrook, Helen M Berman, "The distribution and query systems of the RCSB Protein Data Bank," Nucleic Acids Research, no. 32, pp. 223-225, 2004.

[12] Tatusov RL, Galperin MY, Natale DA, Koonin EV., "The COG database: a tool for genome-scale analysis of protein functions and evolution," Nucleic Acids Research, pp. 33-6, 2000.

[13] Roman L. Tatusov ,Darren A. Natale, Igor V. Garkavtsev, Tatiana A. Tatusova, Uma T. Shankavaram, Bachoti S. Rao, Boris Kiryutin, Michael Y. Galperin, Natalie D. Fedorova and Eugene V. Koonin, "The COG database: new

developments in phylogenetic classification of proteins from complete genomes," Nucleic Acids Research, pp. 22-28, 2001.

[14] "The Gene Ontology (GO) project in 2006," Nucleic Acids Research, pp. 322-326.

[15] T. Triplet, M. Shortridge, M. Griep, J. Stark, R. Powers, and P. Revesz., "PROFESS: a PROtein Function, Evolution, Structure and Sequence database," Database : the journal of biological databases and curation, 2010, p. baq011.

[16] Robert Powers, Jennifer C. Copeland, Katherine Germer, Kelly A. Mercier, Viswanathan Ramanathan, Peter Revesz, "Comparison of Protein Active Site Structures for Functional Annotation of Proteins and Drug Design," PROTEINS: Structure, Function, and Bioinformatics, vol. 65, no. 1, 2006.

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

[18] Lawrence JG, Ochman H., "Molecular archaeology of the Escherichia coli genome," Proc Natl Acad Sci U S A., pp. 9413-7, 1998.

[19] Aristotelis Tsirigos and Isidore Rigoutsos, "A new computational method for the detection of horizontal gene transfer events," Nucleic Acids Research, pp. 922–933, 2005.

[20] Matthew D. Shortridge, Thomas Triplet, Peter Z. Revesz, Mark A. Griep, Robert Powers, "Bacterial protein structures reveal phylum dependent divergence,"

Computational Biology and Chemistry, 2011.

[21] Lars Juhl Jensen, Philippe Julien, Michael Kuhn, Christian von Mering, Jean Muller, Tobias Doerks and Peer Bork, "eggNOG: automated construction and annotation of orthologous groups of genes," Nucleic Acids Res., 2008.

[22] Pere Puigbò, Antoni Romeu and Santiago Garcia-Vallvé, "Horizontal gene transfer in bacterial and archaeal complete genomes," Genome Res, vol. 10, pp. 1719-1725 , 2000.

[23] Carl R. Woese, "Interpreting the universal phylogenetic tree," vol. 97, pp. 8392-8396 , 2000.

[24] Santosh V.R.,Griep M., Revesz P., "Identifying Horizontal Gene Transfer Using Anomalies In Protein Structures And Sequences. C* Conference on ComputerScience & Software Engineering," , 2011.

[25] Grace Yim, "Attack of the Superbugs: Antibiotic Resistance," The Science Creative Quarterly, 2007.

[26] T. Triplet , M. Shortridge, M. Griep, R. Powers, and P. Revesz, "PROFESS: PROtein Functions, Evolution, Structures and Sequences," in 11th International Congress on Amino Acids, Peptides and Proteins, Vienna, Austria, 2009, p. 95.

[27] Karlin S., "Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes," Trends in Microbiology, pp. 335-43, 2001.

[28] Putonti C, Luo Y, Katili C, Chumakov S, Fox GE, Graur D, Fofanov Y., "A computational tool for the genomic identification of regions of unusual compositional properties and its utilization in the detection of horizontally transferred sequences," Molecular Biology and Evolution, pp. 1863-8, 2006.

[29] MWJ van Passel, A Bart, HH Thygesen, ACM Luyf, AHC van Kampen, and A van der Ende, "An acquisition account of genomic islands based on genome signature comparisons," BMC Genomics, 2005.