2013 International Conference on Advances in ICT for Emerging Regions (ICTer): 118 - 124

Prediction of Horizontal Gene Transfer in *Escherichia coli* using Machine Learning

P. G Sudasinghe[*1], C. R Wijesinghe[*2], A. R. Weerasinghe [*3]

[*] *University of Colombo School of Computing*
35, Reid Avenue, Colombo 7, Sri Lanka
[1]pgsudasinghe@gmail.com
[2]crw@ucsc.cmb.ac.lk
[3]arw@ucsc.cmb.ac.lk

Abstract— **Horizontal Gene Transfer (HGT), also known as Lateral Gene Transfer is a process where an organism acquires genetic material from another organism without being a descendant of that organism. Horizontal gene transfer is said to be the predominant method of evolution in prokaryotic organisms. This study is focused on constructing a method that employs genome comparison and semi supervised learning to identify genes that are horizontally transferred to *Escherichia coli O157:H7* and attempting to find a link between these genes and other organisms that display pathogenic behaviour. *E.coli O157:H7* is compared to *E.coli K-12* which is a harmless strain of the same organism. This comparison yields the set of genes that has not originated from the same ancestor (non-homologous) and is the possible cause of its pathogenic properties. A supervised self-organizing map was constructed to classify the non-homologous genes as either horizontally or vertically transferred. Most of the obtained horizontally transferred genes have shown a striking similarity to other pathological bacteria and Achaea. The results have indicated that, while it is possible to discern the mode of transfer of a gene based on compositional feature to a certain degree, it is better to combine several other features to further refine the findings.**

Keywords— **Horizontal gene transfer, Lateral Gene Transfer, Escherichia coli, Unsupervised learning, Self- organizing map, Supervised Self-organizing map, Codon adaptation index, GC content.**

I. INTRODUCTION

A Gene is defined as a unit of hereditary which usually resides in the nucleic acid called Deoxyribonucleic acid (DNA). Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring.DNA is made up of four chemical bases called Adenine, Thymine, Cytosine and Guanine. They are depicted as a string of four letter alphabet (A, T, G, and C) each corresponding to one of the above mentioned chemical bases. The most common form of DNA is a double helix structure, where two individual DNA strands twist around each other in a right-handed spiral [1].

Genes are transferred from organism to organism via two main methods; Vertical Transfer and Horizontal Transfer. Vertical transfer occurs when an organism receives genetic material from its ancestor. This is the predominant mode of gene transfer among members of species and among species.

Horizontal Gene Transfer also known as Lateral Gene Transfer is the process where an organism acquires genetic material from another organism without being a descendant of that organism (see Figure 1).

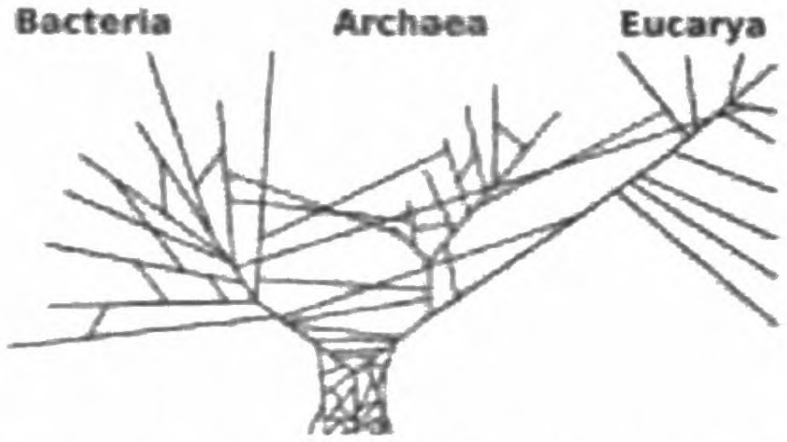


Fig. 1 Horizontal gene transfer between Bacteria, Archaea and Eucarya

This study was conducted to explore new avenues in identification of Horizontal Gene Transfer in *Escherichia coli str. O157:H7*genome and to identify if HGT has any effect in the pathogenic properties it displays. For this purpose, *E.coli O157:H7* was compared with a non-pathogenic strain of the same organism *E.coli K-12*. The dissimilar genes were then examined for signs of Horizontal transfer and origin.

Genomes of several organisms have now been fully sequenced and that data are available in public domain. The availability of sufficient test data has promoted computer scientists to search for methods to automate the process of gene analysing rather than resorting to time consuming and often tedious laboratory experiments. This data could also be used in this research to predict the genes that are transferred horizontally. Horizontal gene transfer is said to be very common among bacteria, and several studies have been carried out with respect to HGT in archaeal and bacterial organisms.

Horizontal gene transfer has been identified as the leading cause of bacteria acquiring new pathogenic and drug resistant genes. *Escherichia coli O157:H7* possess potentially lethal toxins and is the cause of food- borne illness in humans. While some strains of *E.coli* are harmless; there exist some strains like the aforementioned that are harmful to humans.

One example for illnesses caused by this bacterium is the *E.coli O104:H4* outbreak that started in Germany in May, 2011. At first several people in Germany were infected with bacteria leading to hemolytic-uremic syndrome (HUS), a serious medical emergency that requires urgent treatment. By 8th June, 26 people had died and around 500 had been hospitalized with HUS due to the outbreak.

Genomic sequencing by Beijing Genomics Institute has confirmed that the O104:H4 serotype has some enteroaggregative *E. coli* (EAEC or EAggEC) properties, presumably acquired by horizontal gene transfer [2].

978-1-4799-1276-6/13/$31.00 ©2013 IEEE

Due to these reasons, it is important to obtain an understanding of how the potentially harmful bacteria evolve. This study would provide an insight to this matter as well.

II. Literature Survey and Background

Charles Darwin described the evolution of species as The Tree of Life, which is a tree like representation of all living and extinct organisms. This became the standard pictorial depiction of evolution, implying the notion of a Common ancestor of all organisms and the bifurcating evolutionary process [3].Due to HGT, information move across normal mating barriers, between more or less distantly related organisms. This has resulted in what is now known as a "Web of Life" instead of the traditional "Tree of Life".

HGT has largely affected the increase of certain bacteria's ability to resist drugs. When one bacterial organism acquires the resistance to a certain drug, the responsible gene could be transferred to other species quickly.

This phenomenon also raised concern among the scientific community regarding the possibility of dangerous transgenic DNA spreading from species to species. The claim that there were over 100 bacterial genes in the genome fuelled this concern further; this has been proven to be exaggerated since the initial discovery.

Traditionally, phylogenetic methods have been used to prove that a gene has been horizontally transferred. The high similarity between different species was also an initial step in inferring such genes. The evolutionary trees of certain individual genes of a particular genome may conflict with the trees of majority of the genes in that genome. This is a strong indication that the conflicting tree belongs to a gene that has been horizontally transferred.

If majority of the genes in a particular genome share the same evolutionary history, these data are used to construct a consensus tree. Genes that conflict with this consensus tree are candidates of horizontal transfers. Identification of these genes is compulsory as it is apparent that organismal evolution cannot be inferred by studying just a several genes of the genome in question [4].

One of the earliest methods in the prediction of horizontally transferred genes utilized the C+G content, codon usage, and amino acid usage and gene position of genes of the genomes. A gene was considered to be acquired via HGT if they were extraneous from G+C content and codon usage, if they were longer than 300 bp and deviated from amino acid composition, or if they were included in blocks of genes acquired via horizontal transfer which are known as alien genomic strips.

An algorithm called Wn which produces a list of horizontally transferred genes when given a genome has been introduced in [5]. A 'typicality' score is assigned to each gene of a given genome. The higher this score is, the more probable that this gene is not horizontally transferred. This Wn method was then expanded to facilitate the identification of clusters of putative gene transfers.

In more recent researches several scientists have approached the detection of horizontally transferred genes using machine learning. The work described in [6] examines how the horizontal gene transfer inference by phylogenetic approach is affected by the properties of multiple sequence alignment. A support vector machine has been used as the classifier.

Another approach for detection of horizontal gene transfer employs unsupervised learning along with bipartition spectral analysis [4]. This involves the construction of a consensus tree from a subset of phylogenetic information overcoming the high computational demand that arises otherwise.

As described above, scientists have first started detecting horizontal gene transfer with statistical procedures. However with time the number of sequenced genomes has multiplied and the size of these genomes are large. Therefore they have moved towards using various machine learning techniques for this. The features used in both statistical approaches and machine learning approaches are similar to some extent.

Most researchers have agreed that C+G content and codon usage bias are characteristics that could be used in the detection of horizontally transferred genes [8]. Most of the above mentioned researches have used these in some way. All most all the researches carried out with respect to horizontal gene transfer concern those of bacterial and archaeal genomes. The most significant hindrance in analysing a multicellular genome according to these scientists is the sheer size of them

III. Design and Methodology

The following section contains the proposed design and the methodologies employed in the prediction of genes that are horizontally transferred into the selected organism.

The key steps of this research are listed below:

1. Compare the complete genomes of two pathogenic and non-pathogenic strains of the same organism (*Escherichia coli O157:H7* and *Escherichia coli str. K-12*) to obtain homologous genes.
2. The resulting data set contains homologous genes between the above mentioned genomes that are employed to construct the set of non-homologous genes.
3. Calculate necessary feature values of the non-homologous genes.
4. Construct a technique to predict which genes are horizontally transferred with the data obtained from the previous steps.

A. Genome Comparison

The genome that is focused on in this research is of the organism *Escherichia coli O157:H7* str. *EDL933*. This (shown in Figure 2) is a rod-shaped bacterium that is commonly found in the lower intestine of warm-blooded organisms. Most strains of *E. coli* are known to be harmless, however some strains such as the one focused in this research, the O157:H7 are the cause of food borne illnesses and could lead to hemorrhagic diarrhoea, and occasionally to kidney failure, especially in young children and elderly.

TABLE I
SCIENTIFIC CLASSIFICATION OF E. COLI

Domain	Bacteria
Phylum	Proteobacteria
class	Gammaproteobacteria
Order	Enterobacteriales
Family	Enterobacteriaceae
Genus	Escherichia

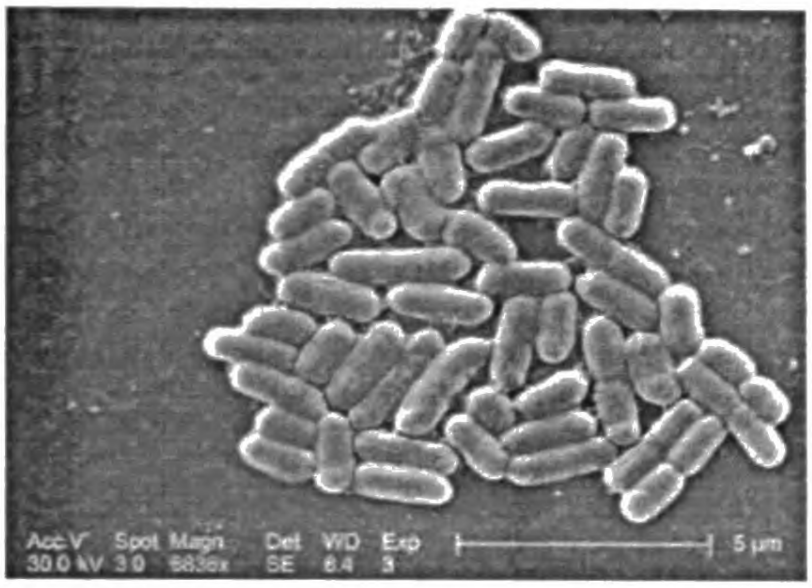


Fig. 2 Colorized scanning electron micrograph depicting a number of Gram-negative Escherichia coli bacteria of the strain O157:H7, magnification 6,836× [17].

E. coli K-12 was chosen as the non-pathogenic organism which *E. coli O157:H7* would be compared against. There are several completely sequenced genomes available for *E. coli K-12* strains as well as two such genomes for *O157:H7* strain in NCBI database.

TABLE II
COMPARISON OF E. COLI O157:H7 AND E. COLI K-12

Feature	O157:H7	K-12
Ref. Sequence	NC_002655	NC_000913
Length	5,528,445 nt	4,639,675 nt
GC content	50%	50%
%coding	87%	85%
Topology	Circular	Circular
No. of genes	5427	4494
Protein coding seq.	5298	4145
Structural RNA	128	175

B. Obtaining Non-homologous genes between E.coli O157:H7 and E.coli K-12

The two genomes in question were compared against each other to identify homologous genes and genetic regions between them. Homologs are genomic regions that share a common ancestry. The non-homologous genes of *E.coli* O157:H7 can be extracted from this.

C. Calculating gene features

Identify suitable features that would be used to mark a certain gene as horizontally transferred from another organism. The type of such features used so far in similar researches include the GC content, Codon Usage bias and gene trees.

D. Identify horizontally transferred genes.

The data obtained from the above phase will be then employed in the construction of the prediction mechanism. The sort of approach that would be successful depends on the data acquired from the previous phase.

IV. IMPLEMENTATION

The section is divided into two main parts each detailing the data acquisition process and machine learning process.

A. Data acquisition

The First step of acquiring necessary data is to perform a genome comparison between *E.coli O157:H7* and *E.coli K-12*. This is done with the aid of Synmap tool on CoGe[8]. Synmap enables the users to compare two complete genomes and identify syntenic regions (regions that are inherited from a common ancestor). First, the necessary genomes are selected in Synmap and necessary parameters are set.

This function employs a variation of BLAST algorithm called (B)LastZ for the nucleotide-nucleotide search. This LastZ[8] implementation has been parallelized to break up the query sequences into multiple pieces for searching. This is said to be the best algorithm to pick in terms of sensitivity and speed. The general principle is that the most likely and parsimonious way two genomes have a collinear series of homologous genes is when those genomic regions in each organism are derived from a common ancestral genomic region. Therefore, genomic synteny is inferred by a collinear arrangement of putatively homologous genes in two or more genomic regions.

After the execution of the program, Synmap produces several outputs. One is the Syntenic dotplot which depicts the syntenic regions of the two selected organisms in green dots and homologous genes in grey dots. The generated syntenic dotplot for *E.coli O157:H7* and *E.coli K-12* is depicted in Figure 3.

The syntenic dotplot is generated thus:

1. All protein coding regions (CDS)) are extracted from each genome
2. These sequences are blasted against each other to identify putative homologous gene pairs.
3. Putative homologous gene pairs are analysed to determine if they share a collinear order between the genomes.

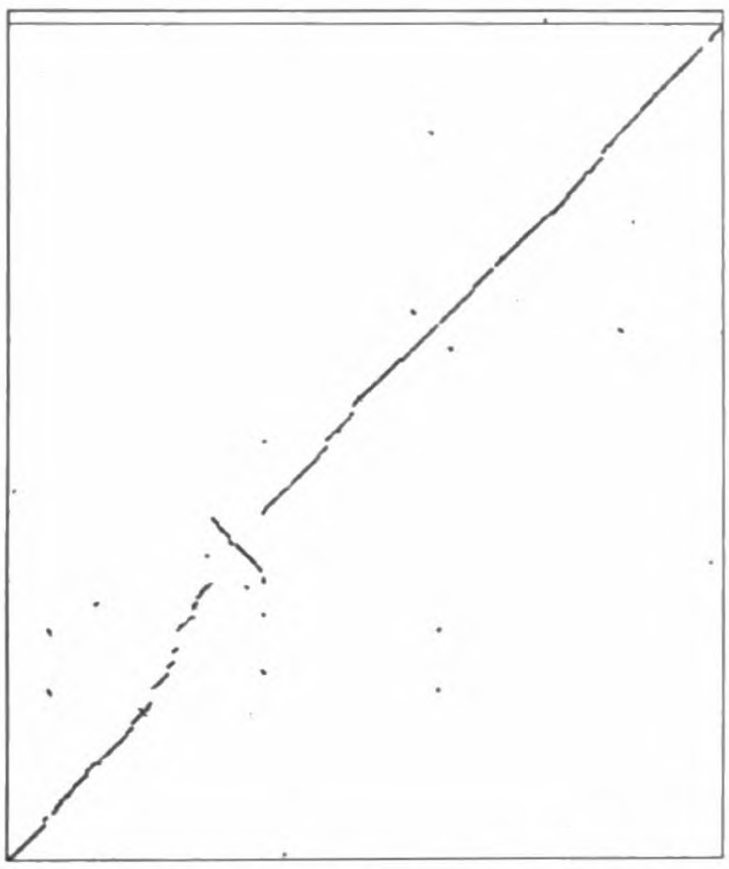


Fig. 3 Synmap: Syntenic dotplot for *E.coli O157:H7* and *E.coli K-12*

The bold lines in the map display syntenic regions while the grey dots represent homologs.

Synmap also makes available several of the text files that were generated in creating the above mentioned Syntenic dotplot. Out of these, the files that contain the entire coding sequences of *E.coli O157:H7* and the homologous genes between *E.coli O157:H7* and *E.coli K-12* are required to

to obtain the list of non-homologous genes between the two genomes.

The entire coding region on *E.coli O157:H7* contains 5543 records including the plasmids. The number of genes was 5445. Each record consists of two lines; the first givingthe necessary information about the gene and the second containing the gene sequence itself.

The file that contains the homologous genes has one record for each matching pair. Each record contains the gene identification details of the two genes and severalattributes that were calculated in creating the dotplot. The partition of the record that corresponds to the *E.coli O157:H7* genome was filtered from the record initially. As one gene can be homologous to one or more genes of the other genome, this set of genes were further filtered to remove the repeated genes.The number of homologous genes is 4008.

By comparing these two datasets, the set of genes that were in the coding sequencesof *E.coli O157:H7* but not in the homologous gene set were acquired. This is essentially theset of 1535 genes that are non-homologous to *E.coli K-12* genome.

After the set of non-homologous genes are obtained the gene sequences are yet again processed to identify if they are of the correct format. Each gene should only consistof letter A, T, C and G and should be divisible by 3. A considerable number of genes had to be then discarded because they failed to meet with this criterion. The main reason for this issue could possibly be an error in the conversion process in Synmap.

For this process MySQL Server version: 5.1.49-1ubuntu8.1 was used. The text files were read, parsed and repeated genes eliminated with the use of a Java program written for this purpose. After this processing, the total number of genes of *E.coli O157:H7* that were non-homologous to *E.coli K-12* were 1199.

B. Calculating gene features

Each gene has various features that could be used in identifying if it is indeed transferred from other organisms. As mentioned in the earlier section, the most commonly used such features were G+C content and patterns in codon usage. For this research, we decided to use six such features.

1. Codon Adaptation Index
2. G+C content of the entire gene.
3. G+C content of the first position in each codon of a given gene.
4. G+C content of the second position in each codon of a given gene.
5. G+C content of the last position in each codon of a given gene.
6. Percentage of Alanine, Cysteine, Glycine and Threonine amino acids composition

Codon Adaptation Index[9]: Codon adaptation index is a measurement of the relative adaptiveness of the codon usage of a gene towards the codon usage of highly expressed genes. Calculation of CAI needs a reference table [14] of Relative Synonymous Codon Usage (RSCU). The CAI is then calculated as the geometric mean of the RSCU values corresponding to each of the codons used in the gene in question, divided by the maximum possible CAI value for a gene of the same amino acid composition.

$$CAI = \frac{CAI_{obs}}{CAI_{max}}$$

Where;

$$CAI_{obs} = \left(\prod_{k=1}^{L} RSCUk\right)^{1/L}$$

And

$$CAI_{obs} = \left(\prod_{k=1}^{L} RSCUkmax\right)^{1/L}$$

RSCUobs = RSCU value of the k^{th} codon in the gene;
RSCUkmax = RSCU value for the amino acid encoded by the kth codon in the gene and
L = number of codons in the gene.

G+C Content: GC-content (or guanine-cytosine content) is the percentage of nitrogenous bases on a DNA molecule that are either guanine or cytosine.

$$GC = \frac{G + C}{A + C + G + T} \times 100$$

$$GC(1) = \frac{G + C \text{ of the first position of the codon}}{No. of Codons} \times 100$$

$$GC(2) = \frac{G + C \text{ of the 2nd position of the codon}}{No. of Codons} \times 100$$

$$GC(3) = \frac{G + C \text{ of the 3rd position of the codon}}{No. of Codons} \times 100$$

Amino Acid composition: calculated with the aid of functions provided in Biopython Protein Analysis module. Here, the amino acid composition is obtained as a percentage value. All 20 amino acid compositions could be calculated; however for the purpose of this project only four were taken into account. They are Alanine, Cysteine,Glycine and Threonine.

Several machine learning techniques were tested with the calculated data. They are K-Means, Self Organizing Map[11] and Supervised Self Organizing Map.

For supervised self organizing map, a separate training set was needed. This was constructed with the aid of the homologous genes separated in the earlier process. If the standard deviation of the GC content is σ, a set of genes with GC content between 1.5σ – mean GC and 1.5σ +mean GC where randomly selected. This was labelled as vertically transferred. It was assumed that since the genes belong to the same organism *E.coli O157:H7*, the features of vertically transferred genes should be similar despite being homologous to another organism.

Another set of genes were selected from various organisms (mainly from bacteria and archaea), for which the same feature values were calculated. As they belong to different organisms they were expected to have distinct features[13]. These features were selected from bacteria and archaea specifically because the most likely horizontal transfers occur within bacteria- bacteria and archaea to bacteria. Therefore it is fair to assume that the genes selected from different

organisms could be used and labelled as horizontally transferred in the training set. The compositional values of these genomes were also taken into account when selecting this data set. Genomes with mean total GC content which were both higher and lower than that of *E.coli* were selected and labelled as HGT2and HGT1 accordingly. The entire dataset contained 200 samples of both types of data.

The above methods were implemented in MATLAB which is a numerical computing environment and fourth-generation programming language developed by MathWorksInc. The version used here is R2008a and the execution environment was Windows XP operation system. For the Self organizing map algorithms, the SOMToolbox[12] was used.

The quality of the Self organizing maps was discerned from quantization and topographic errors. The former is measured using average quantization error between data vectors and their Best Matching Units on the map. The latter is the percentage of data vectors for which the first and second Best matching units are not adjacent units and measures topology preservation.

V. Results and Discussion

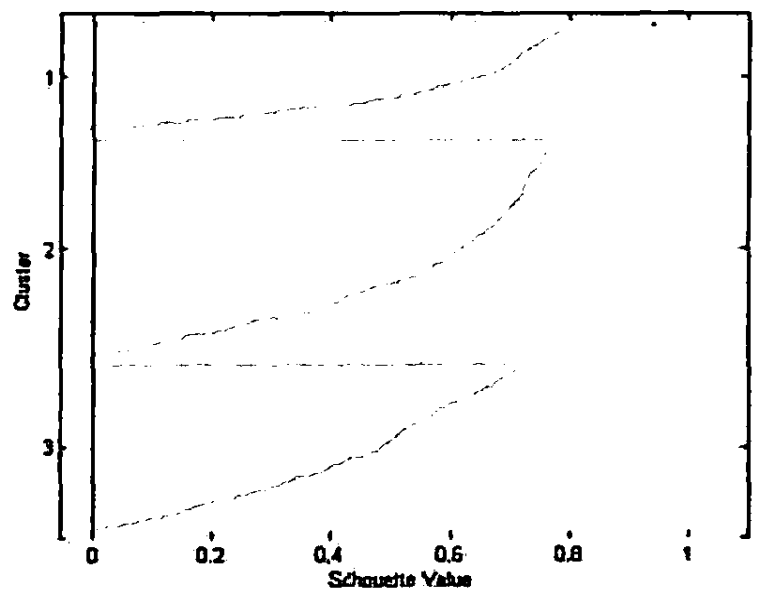
A. K- Means Clustering

Silhouette plots were constructed to discover the optimal number of clusters available in the data. As it can be seen in Figure 4, the clusters are not completely separated. However, this appears to be the best possible result with K-Means method due the clusters being increasingly hard to separate when the number of clusters increases.

The generic K-means algorithm where k=3 yields three clusters, which seemingly just divides the data set in to three portions.

In the next step initial cluster centres were given when clustering the entire data set. This produces three clusters with 411, 265, 523 genes. Again, the clustering does not provide us with a meaningful division of the data.

With these results, it is evident that the simple K-Means algorithm is not sensitive enough to produce an accurate clustering to the *E.coli* data set.



Fig,.1 Silhouette plot for three clusters using Euclidean distance

B. Self Organizing Map

The self organizing map was examined next. SOM is usually the initial step in a data-mining process as it gives the users a general idea on the structures available in the data. The local topology of the map constructed here is hexagonal and the global topology is a rectangular sheet. The availability of clusters should be depicted in the Unified distance matrix with blue areas separated by shades of red and yellow. The SOM contained 24 x7 map units and used Gaussian neighbourhood function to calculate the neighbourhoods of each cell.

However the SOM constructed for this dataset does not indicate any availability of clusters (Figure 5). Instead, the lower right area of the SOM displays cells coloured in shades of red and yellow. After the map has folded to cover all the data, it could be seen that a significant amount of data vectors share similar characteristics thus being mapped into the area coloured in shades of blue. The area coloured in shades of red in the above figure contains the data vectors that differ from each other exceedingly. However the quantization and topographic errors of o.o79 and o.o5 indicate a high accuracy in the map.

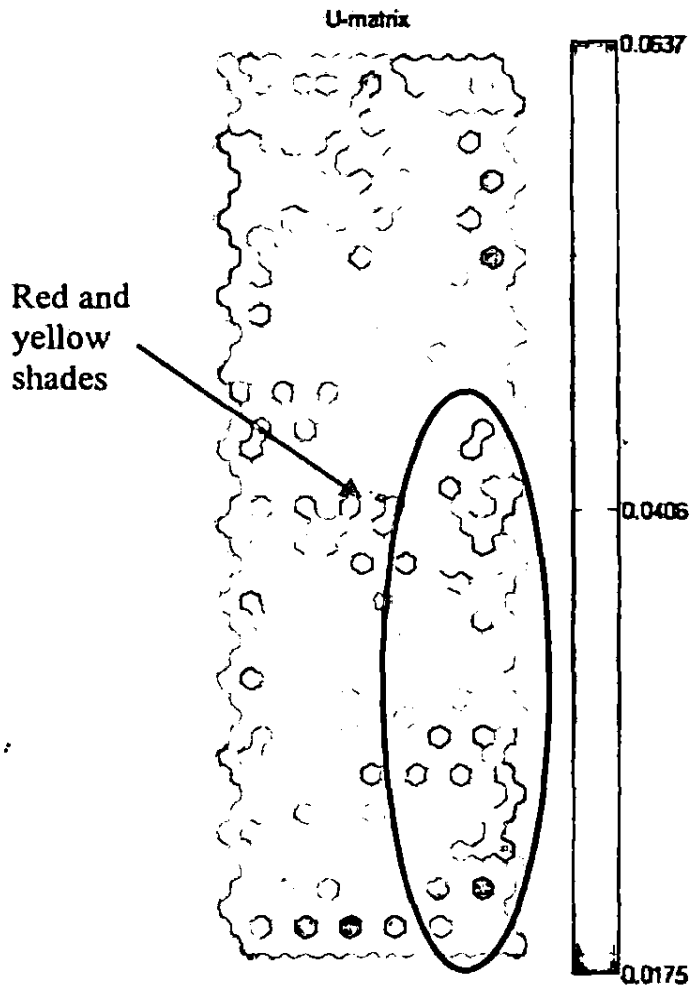


Fig.2 Unified Distance Matrix

With these results it became apparent that a completely unsupervised method is not successful. We then employed a supervised self-organizing map that makes use of labelled data to construct a better SOM.

C. Supervised Self-Organizing Map

The final quantization error and final topographic error of the supervised self organized map that was produced (Figure 6) was 0.079 and 0.020 respectively. This indicates that the data have been mapped into the cells in a significantly accurate manner. The cluster of vertically transferred genes is represented in a fairly consistent blue colour.

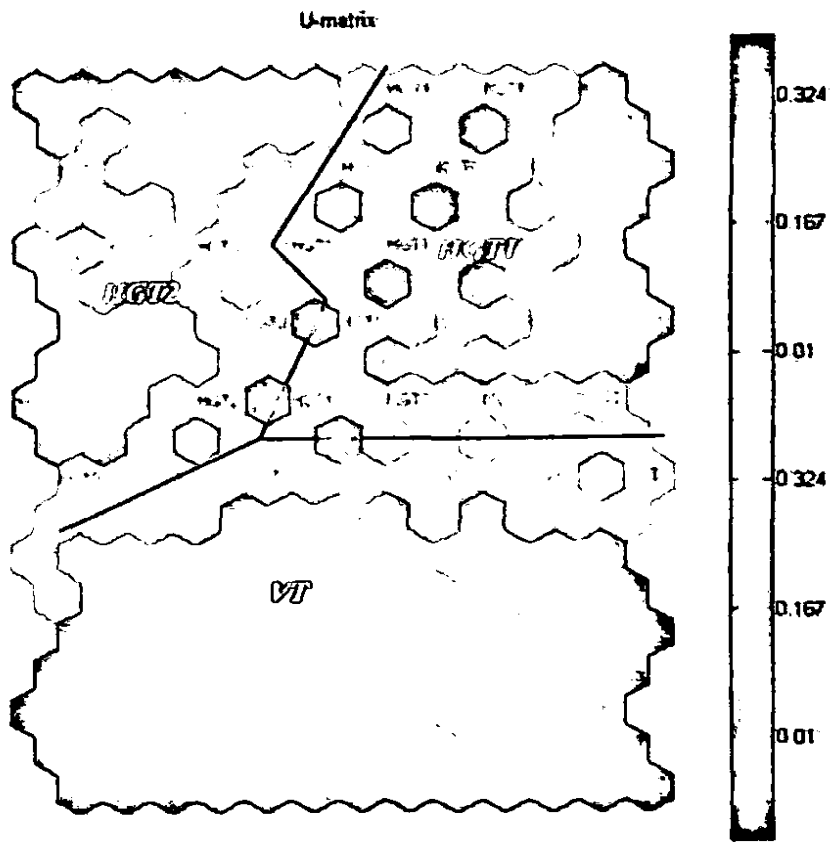


Fig.3. Unified Distance Matrix of Supervised SOM

The above SOM was then used to obtain the best matching unit of the testing set. From this 719 genes were classified as vertically transferred and the remaining 523 as horizontally transferred (HGT1 and HGT2). However the likelihood of such a large portion of genes being horizontally transferred is low. Previous studies have said that the total number of horizontally transferred genes in E. coli O157:H7 amounts to 11% of the total number of genes. To gain an insight to the genes that have been classified as horizontally transferred in this method, we compared our results with the earlier studies carried out and described in [7] using it as a benchmark. The results of the above mentioned study is freely available in [16].

From this comparison it was discerned that the accuracy of this method is 59.96%.This method identified vertically transferred genes with an accuracy of 63% but misclassified the horizontally transferred genes. To discover the reasons behind this phenomenon, the genes that were classified as HGT were closely examined. Here, we discovered that the data classified as horizontally transferred in this method, did indeed contain extraneous compositional features. Genes with total GC contents as low as 35.616% and as high 63% were classified as vertically transferred in the earlier study which were classified as horizontally transferred in our study. The rest of the calculated features also depicted a significant variation. The authors of [16] mention that their approach is fine-tuned to reduce the number of false positives. They have further mentioned that there could be a significant amount of false negatives as confirmed by the above mentioned comparison.

Therefore above 59.96% accuracy depends on the assumption that the classification on the earlier study is accurate.

VI. Evaluation

A. Accuracy

To validate the supervised self-organizing map discussed in the earlier, another dataset constructed in the same manner of the training set was used. This dataset included 30 genes from the homologous genes between *E.coli O157:H7* and *E.coliK-12* which were classified as vertically transferred. This was also cross referenced with the results obtained from [8] to make sure none of the horizontally transferred genes they have found was in is in this dataset.

As horizontally transferred genes in the validation set, data from several bacterial and archaeal genomes were acquired. The mean compositional features of these ranged either higher or lower than that of *E.coli O157:H7*. However these genes contained some feature values that might fall in the range of vertically transferred genes while the others remained distinct. The entire dataset contained 60 genes and was then used to evaluate the SOM.

$$Sensitivity = \frac{No.\ of\ True\ Positives}{No.\ of\ True\ Positives + No.\ of\ False\ Negatives} = 0.6206$$

$$Specificity = \frac{No.\ of\ True\ Negatives}{No.\ of\ True\ Negatives + No.\ of\ False\ Positives} = 0.837$$

This indicates that the SOM has a reasonable accuracy in identifying extraneous values that will correspond to horizontally transferred genes, but will have less accuracy in identifying vertically transferred genes. This occurs due to the highly variable feature values in *E.coli O157:H7*. Therefore to obtain more reliable results, the compositional feature values of a genome should be combined with positional features and phylogenetic data.

B. Possible Donors of Horizontally Transferred Genes

A subset of genes that were marked as horizontally transferred out of the 1199 non-homologous set of genes by this process was compared against the NCBI databases[15] to attempt and find a possible donor organism.

Majority of these genes (e.g. genes bearing NCBI Reference numbers; NP_285711.1, NP_ 290966.1, NP_ 285711.1, and NP_285715.1) showed a striking similarity to *Shigellaboydii*, *Shigellaflexeneri* and *Salmonella entrica*. All these organisms are responsible for food borne illnesses in humans in much the same was as *E. coli O157:H7*.

Another important finding is that, genes classified as horizontally transferred in our method but not in the previous study have a high similarity to pathogenic bacteria and other virulent organisms. Therefore it could be concluded that the accuracy of our method is higher than 56%.

VII. Conclusion and Future Work

Horizontal gene transfer is an important phenomenon in nature. Considered the primary mode of evolution in prokaryotes, the consequences of this has influenced all living organisms. The factors that promote or discourage horizontal transfer of genes are yet unknown, adding to the concern of scientists. Bacteria's ability to quickly acquire drug resistant genes and other harmful virulent genes has promoted the scientific community to focus their attention on decrypting the mystery of HGT.

This research was focused in identifying the connection between *Escherichia coli O157:H7* and other donor organisms that leads to its pathogenic properties. This organism was compared with another genome of a different strain (*E. coli K-12*) which is considered harmless. This was

carried out with the aid of a web based system called CoGe which implements LastZ algorithm to compare entire genomes. The genes that were not common to the two genomes (non homologous) were then examined to identify horizontally transferred genes.

We have employed several machine learning (semi-supervised and unsupervised) techniques to obtain the necessary information. Compositional feature of the selected genes were employed in this process as most researchers have recommended features such as Guanine and Cytosine content to differentiate native genes from horizontally transferred genes. We have used positional GC contents as well as the total GC content of a gene, Codon Adaptation Index which indicates the codon usage of a gene and amino acid compositional values of Alanine, Cysteine, Glycine and Threonine. The calculated feature values were then manipulated using K-Means, Self organizing map and Supervised Self Organizing map algorithms.

K-Means algorithm failed to produce any useful divisions in the data that could be used to identify horizontally transferred genes. Self Organizing Map, which can visualize the structure of a given dataset indicated that the dataset contains a large amount of data with similar properties which could be assume to correspond to vertically transferred genes, and a significant amount of genes that contains compositional values that differ significantly from each other. This algorithm too, failed to give sufficient information that could aid in identifying horizontally transferred genes. The final method examined in this research is a semi supervised algorithm; the Supervised self organizing map which uses labelled data to construct the self organizing map utilizing the classes of each data vector. This was trained using a dataset constructed with feature values of homologous genes whose features were sufficiently closer to those values of the entire genome labelled as vertically transferred genes and genes obtained from different bacteria and archaea labelled as horizontally transferred. The SOM trained with these data were then employed in classifying the non-homologous genes. The results of this classifier were then compared with the results of a previously carried out research to obtain an idea of the accuracy of this method.

The final results lead to the conclusion that, while compositional values such as the Guanine and Cytosine content, codon usage and amino acid composition of a gene will give an indication as to the nature of the transfer of the gene, no decision should be made depending solely on these data. As mentioned earlier, the factors causing horizontal transfer of genes between organisms are yet to be fully discovered. Which horizontally transferred genes would be discovered also depend on the sort of method used. Statistical methods and other methods that use the compositional features of a gene will indicate certain genes as horizontally transferred while phylogenetic methods could indicate genes that were not discovered in the statistical methods. Therefore a combination of these methods should be used to obtain maximum accurate results.

The process detailed here could be adapted to be used with other genomes. As an example, if the requirement is to identify how one strain of *Oryza Stliva (*rice*)* differs from another strain and if horizontal gene transfer has had any effect on this, the entire process described here could be carried out. If one merely needs to acquire a list of potential horizontally transferred genes, the genome comparison step is not required; the clustering and classification step could be carried out at once.

The biggest challenge in adapting this process to other organisms would be dealing with the sheer amount of data. The genome of Rice (*Oryza Sativa*) contains 12 chromosomes and each chromosome contains thousands of genes. Calculating feature values therefore becomes a problem. The solution would be to distribute this workload among several computers and then combine each output to construct datasets.

The next step in this research is to identify non-compositional features of genes and combining them with the outcome described in this report for more accurate results.

If the above mentioned shortcomings were addressed, we are certain that the horizontally transferred genes of any given genome could be identified with high accuracy.

REFERENCES

[1] N. C. Jones and P. A. Pevzner, An Introduction to Bioinformatics Algorithms (Computational Molecular Biology), ch. 3. The MIT Press, 1 ed., 2004.

[2] Robert Koch-Institut. (2011, June). EHEC/HUS O104:H4 Outbreak, Germany, May/June, 30 June 2011. Retrieved September 2011, from Rober Koch Institut. [Online]. Available : www.rki.de/EN/Home/EHEC_Report.html

[3] J. P. Gogarten and J. P. Townsend, "*Horizontal gene transfer, genome innovation and evolution*" Nature reviews. Microbiology, vol. 3, pp. 679-87, 2005

[4] L. Hamel, N. Nahar, M. S. Poptsova, O. Zhaxybayeva, and J. P. Gogarten, "Unsupervised learning in detection of gene transfer," Journal of biomedicine & biotechnology, January 2008

[5] A. Tsirigos and I. Rigoutsos, "A new computational method for the detection of horizontal gene transfer events," Nucleic acids research, vol. 33, pp. 922{33, January 2005

[6] M. Roettger, W. Martin, and T. Dagan, \A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies.," Molecular biology and evolution, vol. 26, pp. 1931-9, 2009.

[7] S. Garcia-Vallve,"Horizontal gene transfer in bacterial and archaeal complete genomes," Genome Research, vol. 10, pp. 1719{1725, 2000

[8] "Synmap: Whole genome synteny." [Online] Available : http://synteny.cnr.berkeley.edu/CoGe/ SynMap.pl

[9] R. S. Harris, "Improved pairwise alignment of genomic DNA. PhD thesis, The Pennsylvania State University, 2007

[10] P. M. Sharp and W.-H. Li, "The codon adaptation index: A measure of directional synonymous codon usage bias,and it's potential applications,", Nucleic Acid Research, vol. 15, no. 3, pp. 1281-1295, 1987

[11] T. Kohonen, "The self-organizing map," in Proceedings of the IEEE, vol. 78, pp. 1464-1480, September 1990

[12] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, "Som toolbox for matlab 5," tech. rep., Helsinki University of Technology, P.O. Box 5400, FIN-02015 HUT, Finland, January 2000

[13] A. Tsirigos and I. Rigoutsos, "A new computational method for the detection of horizontal gene transfer events," Nucleic acids research, vol. 33, pp. 922-33, January 2005

[14] Y. Nakamura, "Codon usage database." [Online]. Available: http://www.kazusa.or.jp/codon/

[15] "National center for biotechnology information." [Online]. Available: http://www.ncbi.nlm.nih.gov/

[16] Garcia-Vallve et al. (2003). Horizontal gene transfer database (HGT-DB) [Online]. Available: http://genomes.urv.es/HGT-DB/

[17] Public Health Image Library. "*Centers for Disease Control and Prevention,*" [Online]. Available: http://phil.cdc.gov/phil/details.asp?pid=10068