Chapter 1

IDENTIFICATION OF LOW NUTRIENT RESPONSE GENES IN THE BACTERIUM *PSEUDOMONAS AERUGINOSA* WITH HIERARCHICAL CLUSTERING

Bertrand Sodjahin[1], Shauna Reckseidler-Zenteno[1,2], Shawn Lewenza[1,2], Vive S. Kumar[1], Junye Wang[1]
[1] *Faculty of Science and Technology, Athabasca University, Canada*
[2] *Department of Microbiology, Immunology & Infectious Diseases, University of Calgary, Canada*

Abstract: *Pseudomonas aeruginosa* is a bacterial organism known for its ubiquity in the ecosystem and for its ability to resist antibiotics. It can survive at length in any environment it reaches, in particular hospital surfaces and is deemed to cause various diseases, in humans, animals, and plants. It is a common cause of nosocomial, hospital-acquired infections. It has been shown that this organism can be isolated from water in a number of intensive care units. The hypothesis is that *P. aeruginosa* is capable of long-term survival in water due to the presence of particular genes which encode for proteins that facilitate persistence. The objective of our research is then to identify genes involved in the survival of *P. aeruginosa* in water by looking at genes responsive to a low nutrient environment. We conducted on a gene expression data a hierarchical clustering analysis in Weka, which is a collection of machine learning algorithms for data mining tasks. The results appear to be interesting, yielding a list of 91 distinct genes accounting for approximately 8% of the genome and identified as potentially responsible for the survival in water of the bacterium.

Key words: Machine Learning, Hierarchical Clustering, genes expression, *Pseudomonas aeruginosa*, Weka.

1. INTRODUCTION

Pseudomonas aeruginosa is a gram-negative bacterium that is ubiquitous in the environment and is known for its ability to inhabit a number of environments, causing disease in plants, animals, and humans (Jørgensen et al., 1999). This diverse organism is also a common cause of hospital-acquired infections, mostly causing skin infections in burn patients, infections of indwelling devices such as catheters, and fatal lung infections in patients with cystic fibrosis (Driscoll, Brody, & Kollef, 2007). Studies have shown that *P. aeruginosa* may survive for months on hospital surfaces (Kramer, Schwebke, & Kampf, 2006). Infection by this bacterium is very difficult to treat because of its resistance to a number of antibiotics (Driscoll et al., 2007). It utilizes certain mechanisms to resist the effects of antibiotics including efflux pumps, modification of the outer membrane to reduce permeability, and inactivation of drugs through the production of enzymes (Driscoll et al., 2007). One of the most effective ways of combating the effects of antibiotics is for the organism to exist as a biofilm (Driscoll et al., 2007; Harrison, Turner, & Ceri, 2005; Ryder, Byrd, & Wozniak, 2007) which is the result of a complex aggregation of microorganisms surrounded by a protective and adhesive matrix. These biofilms are dramatically more resistant, up to 1000 fold, to antibiotic treatment due to the protection provided by the surrounding matrix polymers (DNA, protein, polysaccharides), the slow growth rates of nutrient limited cells within a biofilm and the presence of multidrug tolerant persister cells (Harrison et al., 2005). We are interested in understanding how *P. aeruginosa* is able to survive in the environment, particularly in water. The ability of the organism to persist at length in water without any nutrients may be responsible for its introduction in hospital environments, leading to patients infections. Not only is *P. aeruginosa* an important opportunistic pathogen and causative agent of nosocomial infections, it can also be considered a model organism for the study of diverse bacterial mechanisms that contribute to bacterial persistence. One of the reasons for its ability to survive in a number of conditions may be due to the large genome it possesses (Stover et al., 2000; Wolfgang et al.,, 2003). The presence of a large number of genes, 50% more genes than *E. coli* permits diversity and adaptability by the organism. Our research goal is to identify these genes involved in the survival of *P. aeruginosa* through the genes' response to low nutrient water. The paper is organized as follows. Methodology and underlying literature in section 2 first describes the data and the analysis environment. Then it discusses the data pre-processing methods as well as the analysis configuration and

implementations in Weka. In Section 3, we present our results and discussions. Section 4 is consecrated to the conclusion of our current work and to some details concerning our future works.

2. METHODOLOGY AND UNDERLYING LITERATURE

2.1 Data description and analysis environment

Fundamental study interests in genetics and microbiology mostly concern functional genomics, genes sequencing, gene profiling or genes expression level for the identification of genes associated for instance with a certain phenotype manifested by an organism. Molla, Waddell, Page, & Shavlik, (2004) introduce genes as components of DNA which encodes for protein and define gene expression as the sequential steps of the transcription of the DNA, which it is part of, into RNA and the translation of this latter into associated protein. In other words, the expression level of a gene is measured with as proxy, the observation of protein fabrication rate in an organism which in response to its environment switches on or off its protein production. Though measuring the expression of an individual gene has been previously achieved, it wasn't until the advent of microarray technology that simultaneous expression measurements of thousands of organism's genes are made possible. Babu (2004) described microarray as a glass slides assembly in which DNA molecules are orderly fixed at particular places referred to as spots or features, each containing millions identical copies of DNA molecules corresponding uniquely to a gene. This microarray technology makes it easy to capture at once integral biological activities and therefore conducive to obtaining high-throughput data, useful for example in the inference of cells regulatory pathways. One of the predominant uses of microarray in gene expression is in the comparison of expression measure of a set of genes originally maintained under a certain condition, with the same set under different other conditions. This permits the study of the impact of these conditions on gene expression.

In our current research study, to identify the genes involved in the survival of *P. aeruginosa* without nutrients, an existing transposon library of *P. aeruginosa* mutants was utilized. This mini-Tn5-*luxCDABE* transposon mutant library of *P. aeruginosa* PAO1 is a collection of random transposon mutants, each containing a mutation in a different gene. This is the result of insertion of a mini-Tn5 transposon into the gene, which prevents effective transcription and eventually translation of the gene into a functional protein.

Each insertion of the mini-Tn5 transposon contains the *luxCDABE* operon, which results in light production as the gene is being transcribed. This allows for determination of gene expression under a variety of conditions. The *luxCDABE* operon is derived from the bacteria *Photorhabdus luminescens*, which is a luminescent marine bacterium (Winson et al., 1998). The mini-Tn5-*luxCDABE* library in PAO1 contains 9,000 mutants, 2,500 of which have been mapped and characterized. Of the 2,500 characterized mutants, 1,384 of these were determined to produce light. This collection of 1,384 mutants was screened for gene expression in water and the gene expression data (Lewenza, Kobryn, de la Fuente-Nunez, & Reckseidler-Zenteno, 2015) has therefore approximately 15,000 data points to be analyzed. It has overall 15 columns. The first of which is the well ID that in fact represents the array identification of wells in which the mutants have been inoculated. The second column is the gene, the third is the PA number, the fourth is the product name, the fifth is the original well ID before the transfer, and columns 6 to 15 (T_4 – T_{672}) represent the ten different time points of the gene expression which represents the ratio of the actual measurement (absolute value) at time at time T_i ($i>0$) by the value for the same gene at time zero (T_0). This ratio establishing procedure is known as normalization (from the absolute value to a relative value). Its use is justified by the fact that, accurately estimating the absolute expression level of certain genes is challenging (Molla et al., 2004). So normalization is a way of canceling systematic variations that are induced by various sources such as different amount of starting mRNA material in two examples (Babau, 2004). Because gene expression matrix may be made up of absolute value or relative value, in order to prevent erroneous analysis, one must always first identify the type of values (absolute or relative) contained in a gene expression matrix, before undertaking any processing step. At this stage of our work, we focus mostly on the identification of the genes responsible for the persistence of the bacterium in low nutrient water.

For this analysis, among software and platforms we've explored are SPSS, a software package used for statistical analysis; AMOS, a statistical software package for structural equation modeling (produced by SPSS); and R, a software environment for statistical computing and graphics. One that appears to be the most practical and suitable is Weka, for several reasons. Weka environment, endowed with friendly usability, is a collection of machine learning algorithms (which include classification, regression, clustering and association) used to mine data. These embedded algorithms can either be applied directly to a dataset or called from one own Java code. In addition to the algorithms, Weka contains data pre-processing and visualization tools. Though we are not using its development components at this stage, it is actually well-suited for devising machine learning schemes. These make Weka of first choice for our future development given its compatibility with

the two main stream Operation Systems of reference: Mac OS and Windows. The latter is the one we use here.

2.2 Pre-processing methods and literature

The data raw described above (Lewenza et al., 2015) is typically a microarray data that includes more information columns than we need. The first step of our work consists in trimming it down to only the columns of genes and those of the 10 time point gene expression measures (table 1). This brings our dataset to the form that Babu (2004) classified as gene expression matrix's relative measurement. Discussing the various representations of gene expression data, Babu (2004) actually discussed discretization of the time point measurement. Intrinsically, most physical measurements are continuous and discretizing them is often needed, not only based on applications requirements but also as a mean of noise cancellation. Whether gene expression matrix is in absolute or relative representation, it can be discretized.

In our present study case, the values of the gene expression measurement are all numerical (Real numbers). As in gene expression analysis studies in general, discretization is required here as well. We base it on the significance of expression level measure as threshold according to which we qualify a gene as expressed or repressed. We define a variable *Gene_Express_Val* as the gene expression measurement value. We discretize the values by representing every value equal or greater than 2 (2-fold) as "YES" meaning expressed or up-regulated, and "NO" for those that are below i.e., down-regulated. Though this pre-processing can be partially done in Weka which offers such tools, we used excel. One main reason for this is that our original data is in Excel format and it is easier to remove in Excel columns that are not relevant to our work at this time. Those that we kept include genes column and all the 10 time point gene expression measurement columns which we discretize in excel into binary nominal attributes (YES, NO) through Algorithm 1.

Algorithm 1: Our Pseudo code for the gene expression discretization

```
1. If (Gene_Express_Val >= 2) {
2. Gene_Express_Val := YES;
3. } Else if (Gene_Express_Val < 2){
4. Gene_Express_Val := NO;
5. } Else {
6. Do nothing; // this is for missing values with "?"
7. }
```

Table 1. Sample of the trimmed down microarray gene expression matrix.

Genes	**T4**	**T8**	…	**T672**
PA5398	1.926365	1.427299	…	0.030316
PA5400	2.138769	1.51678	…	0.048796
…	…	…	…	…
?	1.066667	1.390476	…	0.819048
…	…	…	…	
Tgt	1.06	0.64	…	0.73

Table 2. Sample of the pre-processed microarray gene expression matrix.

Genes	**T4**	**T8**	…	**T672**
PA5398	NO	NO	…	NO
PA5400	YES	NO	…	NO
…	…	…	…	…
?	NO	NO	…	NO
…	…	…	…	…
Tgt	NO	NO	…	NO

There are missing data in some cells of the original dataset. Though we could use k-nearest neighbour (kNN) method for data imputation as in Low et al., (2014, October), due to the minimalistic number of missing data, we just managed this by filling in within Excel the empty cells with "?" to denote unknown value more specifically non-identified gene.

Babu (2004) has highlighted some key terminologies that are very useful in referencing portions of a data in the gene expression matrix. The first is *gene expression profile* which corresponds to the cumulative expression levels for a gene across all the experimental conditions. The second, *sample expression profile* alludes to the cumulative expression levels for all the genes in a single experimental condition. *Vectors space* is another representation alternative of gene expression data. It is in fact a mathematical concept domain in which gene expression profiles and sample expression profiles are represented respectively as horizontal and vertical vectors. Such domain is very useful in applications involving matrix operations in certain data processing procedures such as the rotation of an image by an angle α given its data matrix represented as vectors space. This representation is actually what we have in table 2 and because our study

concerns with genes pattern identification, we concentrate on *gene expression profile*.

Because our data file is in Excel (sample illustrated in table 2), and Weka requires input format of *.ARFF* file, we first converted the Excel file into *.CSV* format and then use an online CSV to ARFF converter called **csv2arff** to format our file into *.ARFF*. The resulting file is then used as the main data input for our analysis.

2.3 Analysis configuration in Weka: methods selection and literature background

A rich review of algorithmic techniques for gene expression data analysis is conducted by Kerr, Ruskin, Crane, and www.wcci2016.orgDoolan, (2008). Interestingly enough, they noted that selecting a method that best fits an experimental dataset is not without challenge. In other words there is no panacea method for all data. So this selection process has to be carefully carried out to obtain a technique that yields optimized results.

Molla et al. (2004) in their work on using machine learning to interpret gene-expression microarray in biological applications distinguished supervised learning and unsupervised learning. In our dataset, the genes are not already categorized or labelled and our research goal is to isolate group of genes that are responsible for the persistence of *Pseudomonas aeruginosa* bacterium. Therefore unsupervised learning is suitable for our analysis. Under the unsupervised learning Molla et al. (2004) discussed two main groups of learning algorithms: Clustering and Bayesian Networks. Both of these groups of learning algorithms are of key interest to our research work. First, clustering methods consist in grouping or clustering examples provided through a dataset, from which it learns by evaluating the similarity of their feature values, notably gene-expression values in our case here. According to Molla et al. (2004), the flexibility and intuitiveness of clustering make it widely adopted by biologist researchers and is well used in the domain of bioinformatics. For instance, Do and Choi (2008) surveyed the basic principles of clustering DNA microarray from various clustering algorithms. Babu (2004) broadly divided clustering methods in two majors groups: Hierarchical and non-hierarchical, though they are much more complex in their categorization as shown in table 3, which is excerpted from Han, Kamber, and Pei (2011a). Bayesian Network, the second type of learning algorithms, is to be considered in the next stage of our research and we will briefly discuss it in our future work section.

As far as method and process that we proposed here are concerned, after the pre-processing that resulted in the data format observed in table 2, we

need to proceed to its clustering. One of the primordial requirements for clustering analysis is the determination of the number of clusters which is generally challenging and requires domain knowledge. Based on our dataset, the goal of our research (isolate genes responsible for *P. aeruginosa* survival under low nutrient environment) and our background knowledge, we intuitively posit we need two clusters. This implies that genes that are not part of this group of interest will form a second group. Therefore through our analysis, the genes are to be partitioned in two groups of respectively similar gene expression pattern. Among the various clustering methods as shown in figure 1 that is excerpted from Chaudhari and Parikh, (2012), hierarchical appears to be the most suitable at this stage of our study work. First because by visually and "humanly" looking at the data, we forebode a certain hierarchy and interaction as we observe that genes which were initially down-regulated become up-regulated at a later time and vice versa, across time. Of previous researchers that have abounded in the same direction of hierarchical clustering, are Eisen, Spellman, Brown, and Botstein (1998) who used this technique to repeatedly pair two most similar examples, for the grouping of genes based on similarity in their expression pattern.

Indeed, as part of the gene expression data analysis is the distance measure which is the quantification of similarity between the sample objects under consideration, here genes. For the computation of similarity or dissimilarity, under hierarchical clustering in Weka, we have options between Euclidian, Manhattan, Minkowski, and Chebyshev distances. Though Euclidian distance is known to be of the most popular usage in clustering algorithm, we choose Chebyshev distance. It is a generalization of Minkowski distance which is itself a generalized form of Euclidian and Manhattan distances respectively (Han et al., 2011b) and it attributes equal distance to all its height neighbours as per its grid representation (Figure 1). So we choose the most generalized form, **Chebyshev distance**. Integral part of this clustering algorithm are also the distance approach considerations: *single linkage* (smallest distance), *complete linkage* (longest distance), *average linkage* (average distance) and *centroid linkage* (center distance). These are to be selected based on the clustering objectives and the domain. In fact computing Chebyshev distance is finding attribute f=f_{max} (among the P attributes in the dataset) for which the distance between two objects (here genes) is maximum. This definition makes our distance approach selection to be the **Complete Linkage**. Given two genes *Gi* and *Gj*, and based on Chebyshev distance formula from Han et al. (2011b), we write the equation (1) below that computes the Chebyshev distance between these genes.

$$d(Gi, Gj) = max_f^P \left| f_{Gi} - f_{Gj} \right| \quad (1)$$

The hierarchical methods, as observed in figure 2, is subdivided into agglomerative (bottom up) and divisive (top down) clustering. While the first proceeds through an iterative agglomeration of individual genes till all genes form a single cluster, the second uses an iterative division till each gene forms a group of its own. In Weka, as noted in the synopsis, the class *weka.clusters.HierarchicalCluster* is agglomeration based and is therefore what we use. Its general algorithm is presented in Algorithm 2.

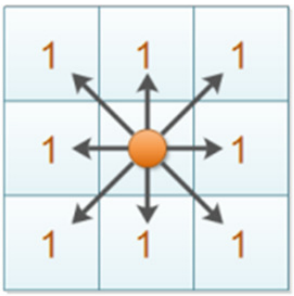

Figure 1. Chebyshev distance grid

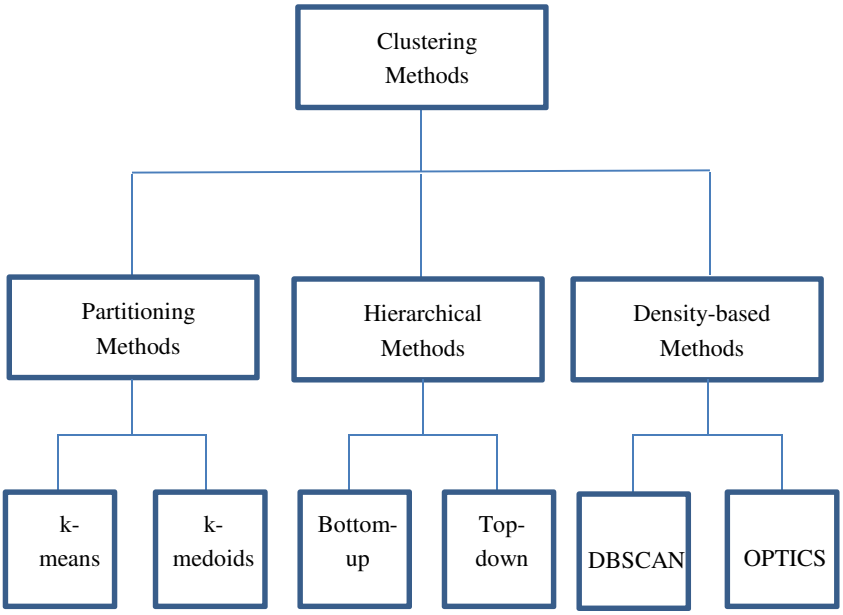

Figure 2. Graphical Overview of clustering methods

Algorithm 2: Basic agglomeration clustering algorithm

1. Compute the distance matrix, here **Chebyshev**
2. **Repeat**
3. Merge the closest two clusters
4. Update the distance matrix to reflect distance between new clusters
5. Until only one cluster remains

Table 3. Description of clustering methods.

Methods	**General Characteristics**
Partitioning Methods (e.g., K-Mean)	- Find mutually exclusive clusters of spherical shape - Distance-based - May use mean or medoid (etc.) to represent cluster center - Effective for small- to medium-size data sets
Hierarchical Methods	- Clustering is a hierarchical decomposition (i.e., multiple levels) - Cannot correct erroneous merges or splits - May incorporate other techniques like microclustering or consider object "linkages"
Density-based Methods	- Can find arbitrarily shaped clusters - Clusters are dense regions of objects in space that are separated by low-density regions - Cluster density: Each point must have a minimum number of points within its "neighborhood" - May filter out outliers
Grid-based Methods	- Use a multiresolution grid data structure - Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

3. RESULTS AND DISCUSSIONS

In our work while conducting various analysis methods experiments in Weka on the same dataset, we realized as discussed in Do and Choi, (2008) that with different clustering algorithms, similarity metrics and number of clusters, results vary substantially. Our approach had then been to first have a theoretical review of the methods and with our background knowledge to

select the suitable parameters and practically test them to see the one that provides the best results and this is what we did in section 2.3.

We have in figure 3 the results of the clustering as well as the parameters used. For instance it can be observed that **Hierarchical clustering** is used with **Complete Linkage** distance calculation. In fact based on our experiment, Complete Linkage provides the best clustering compared to the other distance approaches discussed in section 2.3 which were unable generate any result. We see in the same figure 3 that the percentage repartition shows 92% in cluster 0 and 8% in cluster 1. In other words, this means that 92% on one hand have similar expression profile while 8% on the other hand share same expression pattern. Mesquita, Soares-Castro, & Santos, (2013) noted that a comparative genomic analysis of *Pseudomonas aeruginosa* revealed it could be considered as a mosaic of two components and Kung, Ozer, & Hauser, (2010) evaluated the core at approximately 90% of the total genome and by implication the accessory nears 10%. Therefore our result of 92% for cluster 0 and 8% for cluster 1 not only allows us to draw the conclusion that the smallest percentage of our results contains the group of genes that have similar expression pattern responsible for the survival in low nutrient environment. For this group, our percentage obtained is approximately equal to those of Kung et al., (2010) in which it was further pointed that "the accessory genome may encode gene products that contribute to the niche-based adaptation of the bacterium, such as increase in host range, survival in new environment and utilization of new nutrients". This comes in a good alignment with the outcome of our results which makes us conclude that the 8% of the overall genes contains the genes we purposed to identify and which are potentially responsible for the persistence of the of *Pseudomonas aeruginosa* bacterium. And we should further zone in into this restricted group using additional biology information genes of this group that might be outliers

We visualized the clusters and obtained figure 3 showing red cluster elements as belonging to cluster1 and blue ones as belonging to cluster 0. We notice that the line is very fine between the two clusters and not easily separable when we observe the graphic. So we actually saved the result which generated another .ARFF file. This file comes out as a modified version of our input file that we loaded in Weka. It now includes an additional attribute which is named Cluster (see table 4) and is of nominal type (cluster0, cluster1). For each data tuple in this file, cluster 0 or cluster 1 has been added to classify each gene as result of our analysis. We then converted this file into CSV and opened it in Excel where we applied A-Z sorting by cluster column in order to group tuples with cluster 0 in the first rows and cluster 1 in the last rows. This allows us to copy at once the genes

labelled cluster 1 as result of our analysis. We also removed duplicates as well as non-gene rows. This as result allows us to isolate list of 91 genes.

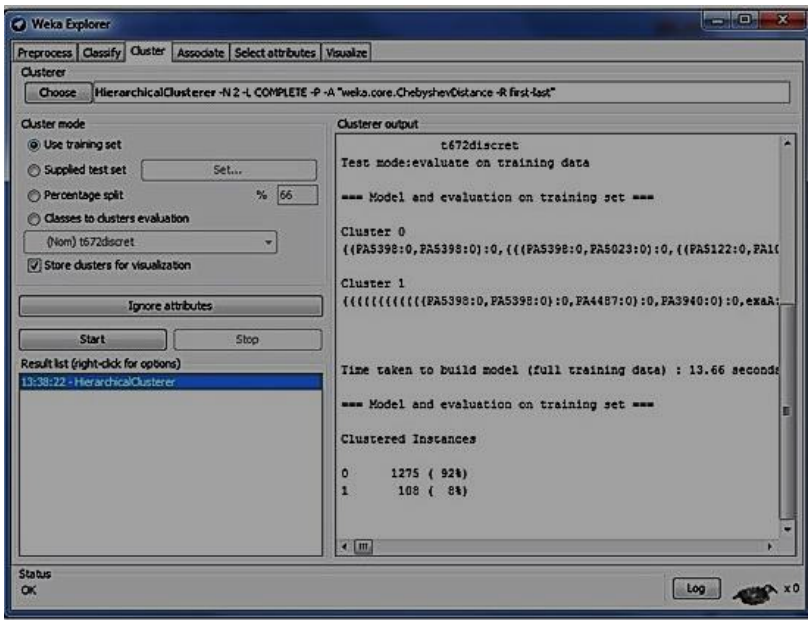

Figure 3. Parameters used and clustering results

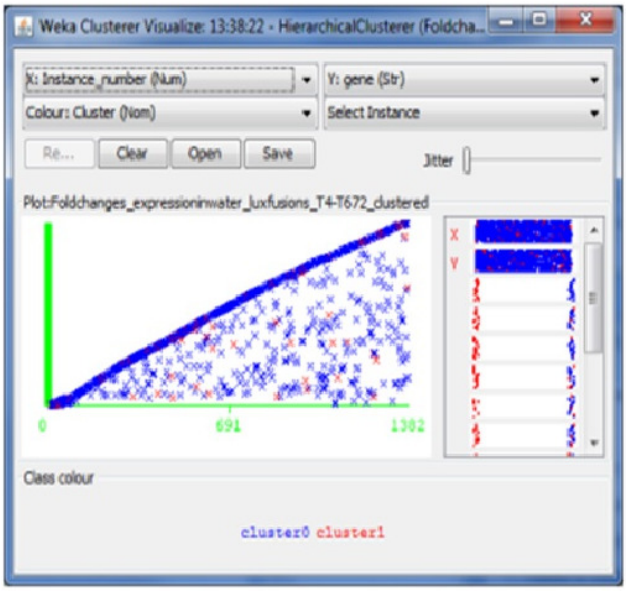

Figure 4. Clusters visualization

Table 4. Sample of the analysis result matrix with cluster labels.

Genes	**T4**	**T8**	…	**T672**	**Cluster**
PA5398	NO	NO	…	NO	cluster1
PA5400	YES	NO	…	NO	Cluster1
…	…	…	…	…	…
?	NO	NO	…	NO	…
…	…	…	…	…	…
Tgt	NO	NO	…	NO	Cluster0

Not only are other types of clustering found in the literature such as such as K-means (Yeung, Fraley, Murua, Raftery, & Ruzzo, 2001) that have proven successful though not suitable in all cases, there are also non-clustering methods such anti-clustering filtering (Raza, & Mishra, 2012) and time series based methods such as dynamic Bayesian Network (Low et al., 2014) that is the next target in our future work.

4. CONCLUSION AND FUTURE WORK

4.1 Conclusion

In this work whose objective is to identify gene or group of genes responsible for the survival of *Pseudomonas aeruginosa* bacterium in low nutrients water. We used Machine Learning, especially hierarchical clustering with Chebyshev and complete linkage as similarity distance calculation metric in Weka environment. We identified a group of 91 distinct genes, representing about 8% of the overall genome of the *Pseudomonas aeruginosa* in the dataset studied. As discussed in our previous section (section 3), these interesting results of the 8% i.e., 91 genes isolated as potential source of survival in low nutrient water would be verified in our future work with more advance methods and infer the underlying survival mechanism. Because additional work are still to be carried out, we do not published here the full list of the 91 genes identified.

4.2 Future work

The above proposed method is the first stage of works yet to be pursued. We will use Bayesian network as a second level of analysis. Bayesian network because like clustering it is an unsupervised learning algorithm that is suitable to our dataset. Its probabilistic approach of processing is of interest to our analysis given the uncertainty involved in the genes interaction. Also according to Molla et al., (2014) the application of learning Bayes' nets to gene expression microarray data has drawn much attention due to the insight it provides pertaining to the interaction networks within cells that regulates gene expression. Within the Bayesian network methods, we find that Dynamic Bayesian network would be a best fit for our research. First because it is time series based and our dataset is a record of gene expression measurement at different time points. So we would use our data to infer a temporal direction for the interaction among the genes and would therefore highlight causal relations. Additionally we will investigate possible additional data source which would provide background knowledge for our analysis as it is known that prior or background knowledge is quite useful in improving learning. We will not only compare the performance of Clustering and Bayesian network for this dataset (as well as other existing methods) but we will also evaluate which one is the best in learning and accurately identifying the genes responsible for the persistence. Another important task for our research is to further investigate (beyond the identification of genes responsible for the persistence of *Pseudomonas aeruginosa*) what new knowledge can be acquired from the data. Our approach here is to use the power of Induction Logic Programming (ILP). The rules of ILP are easily interpreted by human and this makes it popular and well accepted in domains other than computer science. King et al. (2009) is a good example of the power of ILP which they used to include experiment design to devise an autonomous scientist which discovered new knowledge about functional genomic of yeast.

REFERENCES

Babu, M. M. (2004). Introduction to microarray data analysis. Computational Genomics: Theory and Application, 225-249.

Chaudhari, B., & Parikh, M. (2012). A Comparative Study of clustering algorithms Using weka tools. *International Journal of Application or Innovation in Engineering & Management (IJAIEM), 1*(2).

Driscoll, J. A., Brody, S. L., & Kollef, M. H. (2007). The epidemiology, pathogenesis and treatment of Pseudomonas aeruginosa infections. *Drugs*, 67(3), 351-368.

Do, J. H., & Choi, D. K. (2008). Clustering approaches to identifying gene expression patterns from DNA microarray data. *Molecules and cells*, (25), 279-88.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25), 14863-14868.

Han, J., Kamber, M., & Pei, J. (2011). Cluster Analysis: Basic Concepts. *Data mining: concepts and techniques: concepts and techniques* (pp. 443-495). Elsevier

Han, J., Kamber, M., & Pei, J. (2011). Getting to Know Your Data. *Data mining: concepts and techniques: concepts and techniques* (pp. 39-82). Elsevier.

Harrison, J. J., Turner, R. J., & Ceri, H. (2005). Persister cells, the biofilm matrix and tolerance to metal cations in biofilm and planktonic Pseudomonas aeruginosa. *Environmental Microbiology*, 7(7), 981-994.

Jørgensen, F., Bally, M., Chapon-Herve, V., Michel, G., Lazdunski, A., Williams, P., & Stewart, G. S. A. B. (1999). RpoS-dependent stress tolerance in Pseudomonas aeruginosa. *Microbiology*, 145(4), 835-844.

Kerr, G., Ruskin, H. J., Crane, M., & Doolan, P. (2008). Techniques for clustering gene expression data. *Computers in biology and medicine*, 38(3), 283-293.

Kramer, A., Schwebke, I., & Kampf, G. (2006). How long do nosocomial pathogens persist on inanimate surfaces? A systematic review. BMC infectious diseases, 6(1), 130.

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., ... & Clare, A. (2009). The automation of science. Science, 324(5923), 85-89.

Kung, V. L., Ozer, E. A., & Hauser, A. R. (2010). The accessory genome of Pseudomonas aeruginosa. *Microbiology and Molecular Biology Reviews*, 74(4), 621-641.

Lewenza S., Kobryn M., de la Fuente-Nunez C., & Reckseidler-Zenteno S. L., In preparation.

Low, S. T., Mohamad, M. S., Omatu, S., Chai, L. E., Deris, S., & Yoshioka, M. (2014, October). Inferring gene regulatory networks from perturbed gene expression data using a dynamic Bayesian network with a Markov Chain Monte Carlo algorithm. *In Granular Computing (GrC), 2014 IEEE International Conference on* (pp. 179-184). IEEE.

Mesquita, C. S., Soares-Castro, P., & Santos, P. M. (2013). Pseudomonas aeruginosa: phenotypic flexibility and antimicrobial resistance.

Molla, M., Waddell, M., Page, D., & Shavlik, J. (2004). Using machine learning to design and interpret gene-expression microarrays. *AI Magazine*, 25(1), 23.

Raza, K., & Mishra, A. (2012). A novel anticlustering filtering algorithm for the prediction of genes as a drug target. arXiv preprint arXiv:1211.2194.

Ryder, C., Byrd, M., & Wozniak, D. J. (2007). Role of polysaccharides in Pseudomonas aeruginosa biofilm development. *Current opinion in microbiology*, 10(6), 644-648.

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., ... & Olson, M. V. (2000). Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature*, 406(6799), 959-964.

Winson, M. K., Swift, S., Hill, P. J., Sims, C. M., Griesmayr, G., Bycroft, B. W., ... & Stewart, G. S. (1998). Engineering the luxCDABE genes from Photorhabdus luminescens to provide a bioluminescent reporter for constitutive and promoter probe plasmids and mini-Tn5 constructs. *FEMS microbiology letters*, 163(2), 193-202.

Wolfgang, M. C., Kulasekara, B. R., Liang, X., Boyd, D., Wu, K., Yang, Q., ... & Lory, S. (2003). Conservation of genome content and virulence determinants among clinical and environmental isolates of Pseudomonas aeruginosa. Proceedings of the National Academy of Sciences, 100(14), 8484-8489.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., & Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. Bioinformatics, 17(10), 977-987.