INSILICO MICROARRAY ANALYSIS FOR *HELICOBACTER PYLORI* USING GENE EXPRESSION DATA

Ashish Gupta[1*], Santosh Kumar[2], Ashish Trivedi[3], Samantha Vaishnavi[1], Smita Rastogi[5], Prachi Srivastava[3] Ashish Runthala[4]

[1]College of Engg. , School of Biotechnology, Shri Mata Vaishno Devi University, Katra (J&K), India
[2]BCS Insilico Biology, Lucknow (U.P.), India
[3]Amity Institute of Biotechnology, Amity University, Lucknow (U.P.), India
[4]Dept of Biological Sciences, BITS Pilani, Pilani, Rajasthan, India
[5]Department of Biotechnology, Integral University, Lucknow (U.P.), India.

ABSTRACT: *Helicobacter pylori* is a Gram-negative, microaerophilic bacterium that inhibits various areas of the stomach and duodenum. It causes a chronic low-level inflammation of the stomach lining and is strongly linked to the development of duodenal and gastric ulcers and stomach cancer. To better understand adaptive mechanisms utilized by *H.pylori* within the context of the host environment, spotted-DNA microarrays was utilized to characterize in a temporal manner, the global changes in gene expression in response to low pH in the pathogenic *H. pylori* strain G27. Raw data of this microarray work was available in Stanford Microarray Database. Co-regulated genes may share similar expression profiles, may be involved in related functions or regulated by common regulatory elements. There are different approaches to analyse the large-scale gene expression data in which the essence is to identify gene clusters. This approach has allowed us to (i) determine expression profiles of previously described developmentally regulated genes, (ii) identify novel developmentally regulated genes. The *Helicobacter pylori* is an important human and veterinary pathogen. In this work raw data of *Helicobacter pylori* is used as a sample to find out the coexpressed gene.

Keywords: Micro array, hierarchical clustering, Kmeans clustering, co expression, pathogen, ulcer, cancer, Normalization, G27, Tree view, gene expression, Stanford Microarry Database

INTRODUCTION

In recent year there has been an exponential growth in molecular genetics technologies, such as DNA microarrays, allow us for the first time to obtain a “Whole” view of the cell. At the same time development in high throughput computing and bioinformatics provide various computational algorithm and tool to define functional, structural and regulatory behavior of such genes which are analyzed by DNA microarray technology.

At present there are various different technologies for measuring gene (mRNA) expression levels are present but cDNA microarrays are preferably used in scientific community. This allows for a quantitative readout of gene expression on a gene-bygene basis (Brown and Botestein, 1999; Duggan et al., 1999). Microarrays have opened the possibility of creating the possibility of creating data sets of molecular information to represent many systems of biological or clinical interest. Gene expression profiles can be used as input to large-scale data analysis such as – to discover regulatory genomics, to discover taxonomy, to discover new gene of drug importance, and to increase our understanding of normal and disease states (Debouck and Goodfellow, 1999; Alizadeh et al., 2000).

A first step to analyze this all type of information is to examine the extremes, i.e. genes with significant differential expression in two individual samples or in a time series after a given treatment. This simple technique can be extremely efficient. For example, in screens for potential tumor markers or drug targets (Debouck and Goodfellow, 1999). However, such analyses do not address the full potential of genome-scale experiments to alter our understanding of cellular biology by providing, through an inclusive analysis of the entire repertoire of transcripts, a continuing comprehensive window into the state of a cell as it goes through a biological process. What is needed instead is a holistic approach to analysis of genomic data that focuses on illuminating order in the entire set of observations, allowing biologists to develop an integrated understanding of the process being studied.

A natural basis for organizing gene expression data is to Clustering genes with similar patterns of expression (Eisen et al., 1998; Alizadeh et al., 2000). The first step for it is to adopt a mathematical description of similarity. For any series of measurements, a number of sensible measures of similarity in the behavior of two genes can be used, such as the Euclidean distance, angle, or dot products of the two n-dimensional vectors representing a series of n measurements. We have found that the standard correlation coefficient (i.e., the dot product of two normalized vectors) conforms well to the intuitive biological notion of what it means for two genes to be "co-expressed;" this may be because this statistic captures similarity in "shape" but places no emphasis on the magnitude of the two series of measurements.

Conserved DNA sequences are present in all types of organism as Motif or in any other form, which correspond to transcriptional regulatory motifs in upstream regions of genes (McGurie et al., 2001). These conserved regions are often binding sites for DNA-binding proteins and some time also work as TFBS for transcription factor and known as gene regulatory elements. In bacteria it is difficult to locate the regulatory region for a gene found within an operon (Jacob and Monad, 1961), since the promoter for that operon can lay several genes upstream, and it is difficult to predict which gene is at the head of the operon (Price et al., 2005). In addition, there are fewer instances of most regulatory motifs in a bacterial genome than in the yeast or any other Eukaryotes, as there is usually only one instance of a regulatory motif per operon instead of one instance per gene. It is easier to discover a motif that is found in more copies in the genome. However, one can increase the number of instances of a conserved regulatory motif by pooling together upstream sequence from co-express (co-regulated) or orthologous genes in closely related organisms, assuming the motif is conserved across these organisms. Now days in this bioinformatics era, various statistical based software are available which are based on different algorithm, for analysis of such conserved motif in given upstream sequences which work as regulatory elements, but main thing is to get biological relevance from them (Lescot et al., 2002; McGurie et al., 2001; Hughes et al., 2000; Hu et al., 2004; Fogel et al., 2004; Fogel et al., 2008).

Considering the above assumption, we have designed an analysis of gene expression pattern and regulatory genomics of some medicinal important genes of helicobacter pylori (small gramnegative bacteria), which is a pathogenic bacteria and infect over 50% of the global population by causing disease like ulcer (Matysiak-Budnik and Megraud, 1994) and gastric cancer (Graham and Yamaoka, 1999; Aguilar et al., 2001; Parkin, 2001; Ohata et al., 2004). Here we try to find out the consensus nucleotide elements (motifs) and there pattern, which are responsible for expression of gene and work as regulatory sequence in helicobacter pylori.

Helicobacter pylori is a very significant organism for medical purpose, the whole genome sequence of its widely available strain (26695 and J99) are completely sequenced, they contain a single circular genome of 1.7 million base pairs and around 1,500 predicted coding sequences (Tomb et al., 1997; Alm et al., 1999). The objective of present work is to find out co expression pattern of genes. This study is based on well established concept that genes with similar gene expression patterns are most probably share common regulatory machinery (Altman and Raychaudhuri et al, 2001; Allocco et al., 2004).

Material and Methods

Retreival of Raw data

From the Stanford Microarray Database, *Helicobacter pylori* was selected, data was downloaded by the Stanford microarray database, for " pH regulated gene expression of the gastric pathogen *Helicobacter pylori*" (Merrell DS et al, 2003) Different excel sheets 21168, 21169, 21170, 21171, 21340, 21341, 21342, 21343 obtained. In these files, first four files of experiment 1, and last four of experiment 2 at 30, 60, 90,120 minutes respectively.

Export data in Genesis

On SMD this raw data are provided in 6 excel file sets (each file have equivalent weight) and each file contain expression data of different time and different pH. It consists of a set of 4623 gene expression data. The raw data files are sorted and scaled by taking logarithm at base 2 of R/G normalized (mean) ratio. All excel files were merged in one excel file. Then data is normalzed by the rule if missing value in a row more than 80% then deletes that row. After normalization we got 4607 genes in excel file. Finally this file imported in the genesis (Sturn A, et al, 2002).

Clustering for data:

Here we have obtained a Hierarchical cluster as output by using microarray gene expression data in cluster which can visualize in tree-view as a hierarchical tree. We have found that this tree contain all given data in a hierarchical form. According to gene expression value, closely related (co-express) gene would in same cluster. By using different correlation type we also found that the centered correlation is better and suitable for hierarchical clustering and gives more appropriate output for further process (Eisen et al., 1998; Wen et al., 1998), Then through manual subclustering we got 10 clusters. For the k means clustering numbers of clusters of hcl is used. After that k means clustering was done, parameter number of cluster was 10 and maximum iteration of 50 was selected.

Comparision of clusters

Each and every cluster of hcl is compared with the k means clusters (10 clusters). It was done manually. For clusters which were having same expression patterns (by comparing hcl and k means clusters), genelist were obtained.

Result and Discussion

Prediction of coexpressed genes

From the manual analysis of k means clusters and HCl clusters, 4 clusters were found which were showing same expression pattern, which is shown in table1.

Table1: cluster of HCl matching with K means cluster and their expression pattern

Cluster no. of HCl	Cluster no of k means	Shown in figure
Hcl1	K4	Fig.1
Hcl4	K5	Fig.2
Hcl5	K7	Fig.3
Hcl3	K8	Fig.4

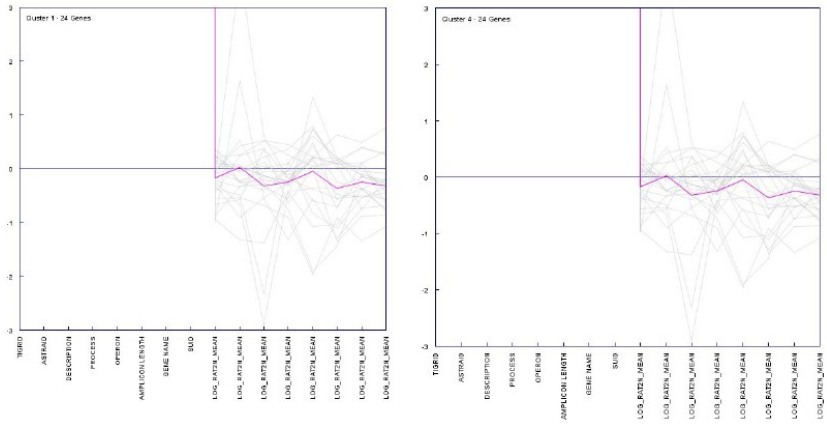

Figure.1 Expression pattern of HC1 & k4

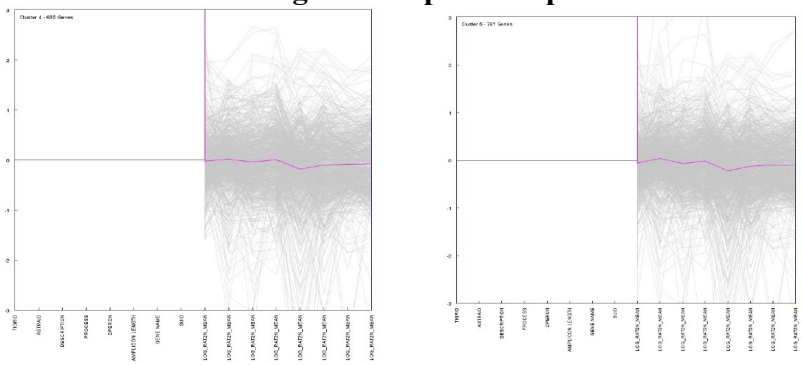

Figure. 2 Expression pattern of HCl4 & k5

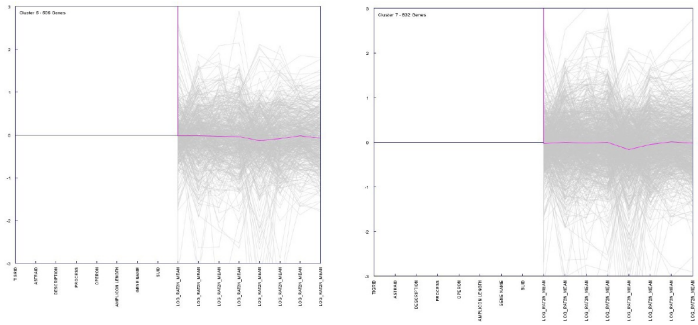

Figure.3 Expression pattern of HCl5 & k7

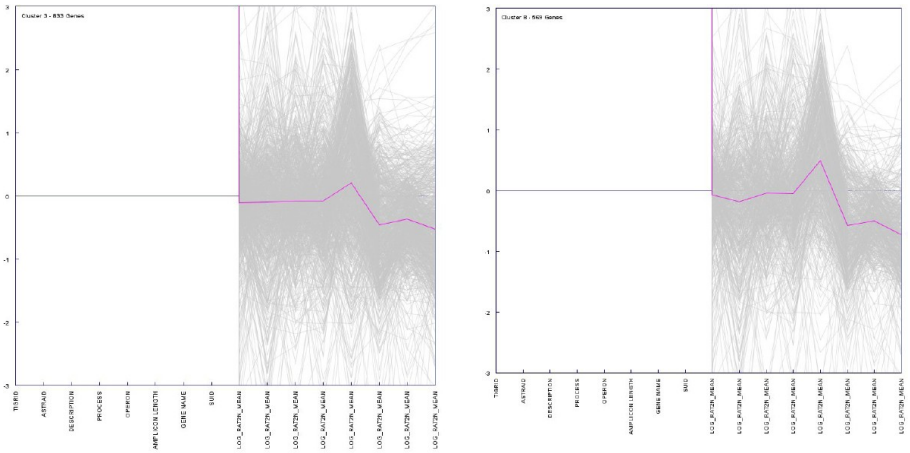

Figure.4 expression pattern of HCl3 & K8

Common genes in clusters

The genes of these selected clusters 'seed cluster' can be used for further analysis and rest of the cluster (Hcl 3 and K8) were discarded because for those process in which they involved is not availabale. This both cluster contains different number of gene which shows same gene expression.

By applying clustered gene in genesis and some other plotting option for gene expression pattern analysis, plot by them shows that according to time period and provided environment condition, the expression of gene become changed, and this change is observable. This fluctuation seems same for most of genes which are in same cluster Figure 1, 2 3 & 4.

Gene expression pattern shows that, when we going to calculate the variance for all gene at every given condition, we found that at some point it is very high for some gene. It is shows that after clustering there is a chance of getting some false positive gene in cluster. For those genes whose function is tilled not known but are came in these four clusters, we can assume that they are somehow related for pH regulation of *H pylori.*

Conclusion

Clustering result from both the methods (HCL clustering and K-means clustering) shows that genes which are common in specific clusters of Hierarchical Clustering and cluster of k-means clustering(k4 HCl1 , k5 HCl4, k7 hcl5, k8 hcl3) have similar expressions pattern of the respective clusters (which are present in both type of clustering) are also same.

Finally it was concluded that common genes of both clustering methods, *viz-aviz* the different clusters, obtained by matching of images of clusters, k4 hcl1, k5 hcl4, k7 hcl5, k8 hcl3 differentially coexpressed. Thus on the basis of comparative analysis this can be concluded that the coexpression is present within the genes of the same clusters.

Acknowledgement

AG acknowledges BCS insilico Biology, Lucknow, for providing essential facilities for completion of this research work.

REFERENCES

1. Aguilar GR, Ayala G, Zarate FG (2001) Helicobacter pylori recent advances in the study of its pathogencity and prevention. Salud Publica Mex 43: 237-47.
2. Alexander Sturn, John Quackenbush, Zlatko Trajanoski. (2002). Genesis: cluster analysis of microarray data, 18.1. 207-208.
3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS,et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-511.
4. Allocco DJ, Kohane IS, Butte AJ (2004) Quantifying the relationship between co-expression, coregulation and gene function. BMC Bioinformatics 5: 18.
5. Alm RA, Ling LS, Moir DT, King BL, Brown ED, et al. (1999) Genomicsequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. Nature 397: 176-80.
6. Altman RB, Raychaudhuri S (2001) Whole-genome expression analysis: challenges beyond clustering. Curr Opin Struct Biol 11: 340-7.
7. Brown PO, Botestein D, (1999) Exploring the new world of the genome with microarrays. Nat Genet 21: 33-7.
8. Debouck C, Goodfellow PN (1999) DNA microarrays in drug discovery and development. Nat Genet 21: 48-50.
9. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM (1999) Expression profiling using cDNA microarrays. Nat Genet 21: 10-4.
10. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci 95: 14863-8.
11. Fogel GB, Porto VW, Varga G, Dow ER, Craven AM, et al. (2008) Evolutionary computation for discovery of composite transcription factor binding sites. Nucleic Acids Res 36: e142.
12. Fogel GB, Weekes DG, Varga G, Dow ER, Harlow HB, et al. (2004) Discovery of sequence motifs related to coexpression of genes using evolutionary computation. Nucleic Acids Res 32: 3826-3835.
13. Graham DY, Yamaoka YH (1999) Pylori and cagA: relationships with gastric cancer, duodenal ulcer, and reflux esophagitis and its complications. Gut 44: 336-41.
14. Hu Z, Fu Y, Halees AS, Kielbasa SM, Weng Z (2004) SeqVISTA: a new module of integrated computational tools for studying transcriptional regulation. Nucleic Acids Res 32: W235-W241.
15. Hughes JD, Estep, Preston W, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccaromyces cerevisiae. J Mol Biol 296: 1205-14.
16. Jacob F, Monad J (1961) Genetic regulatory mechanism in the synthesis of proteins. J Mol Biol 3: 318-56.
17. Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, et al. (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res 30: 325.
18. Matysiak-Budnik T, Megraud F (1994) Helicobacter pylori in eastern European countries: what is the current status? Gut 35: 1683-6.
19. McGurie AM, Hughes JD, Church GM (2001) conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. Genome Research 10: 744-757.
20. Merrell DS, Goodrich ML, Otto G, Tompkins LS, Falkow S. (2003) pH-Regulated Gene Expression of the Gastric Pathogen *Helicobacter pylori*. INFECTION AND IMMUNITY. 71(6): 3529–3539
21. Ohata H, Kitauchi S, Yoshimura N, Mugitani K, Iwane M, et al. (2004) Progression of chronic atrophic gastritis associated with Helicobacter pylori infection increases risk of gastric cancer. Int J Cancer 109: 138-43.
22. Parkin DM (2001) Global cancer statistics in the year 2000. Lancet Oncol 2: 533-43.
23. Price MN, Huang KH, Alm EJ, Arkin AP (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res 33: 880-92.
24. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, et al. (1997) The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature 388: 539-547.
25. Wen X, Furham S, Michaels GS, Carr DB, Smith S, et al. (1998) Largescale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci 95: 334-339.