Journal of Computational Intelligence in Bioinformatics.
ISSN: 0973-385X Volume 4 Number 1 (2011) pp. 55—61

http://www.ripublication.com/jcib.htm

Cluster Analysis and Coexpression Study of *Helicobacter pylori* for Genome Wide Microarray Expression Data

Ashish Chandra Trivedi[1,2], Anshul Tiwari[1,3], Smita Rastogi[2] and Prachi Srivastava[1*]

[1]*Amity Institute of Biotechnology, Amity University Lucknow Campus, UP, India*
[2]*Department of Biotechnology, Integral University, Lucknow, UP, India*
[3]*Department of Ophthalmology C.S.M. Medical University, UP, India*

Abstract

Helicobacter pylori is one of the most common Gram-negative, microaerophilic bacterium bacterial pathogen in human that inhibits various areas of the stomach and duodenum. It is the causative agent of chronic superficial gastritis in humans, long persistence of this organism leads increases risk of development of gastric ulcer disease and adenocarcinoma and mucosa-associated lymphoid tissue lymphoma of the distal stomach. Till the mechanism of pathogenesis in humans remains largely unknown. Microarray technology has been widely applied in biological and clinical studies for simultaneous monitoring of gene expression in thousands of genes. Numerous microarray studies of understanding the mechanism of pathogenesis for *H.pylori* have been conducted. Gene clustering analysis is found useful for discovering groups of correlated or coexpressed genes potentially co-regulated or associated to the disease or conditions under investigation. In this current study we have conducted cluster analysis for *H.pylori* for identification of coexpressed gene by using Microarray raw data sets available at Stanford Microarray Database. 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 determine expression profiles of novel developmentally regulated genes. Finally we get the six genes which highly coexpressed and may be involved in pathogenesis. These genes are also novel developmentally regulated genes and used as Drug target.

Keywords: Microarray, Hierarchical clustering, K-means clustering, co-expression, pathogen, ulcer, gene expression, Stanford Microarray Database, Genesis Software.

Introduction

Helicobacter pylori is a common bacterial pathogen of the human stomach and infect an estimated 50% of the population worldwide. *H. pylori* causes gastritis infection initially and if allowed to Persist, can induce a range of pathologies (1). It is the initially causative agent of most peptic ulcers and many serious outcomes such as atrophic gastritis, intestinal metaplasia, and gastric cancer are correlated with long-term infections. Currently there is no report whether these outcomes are due to specific factors produced by the organism or whether they result from chronic inflammation due to efficient and persistent colonization of the gastric mucosa. Thus, colonization and persistence factors may themselves constitute virulence factors for this organism (2). It has already been reported during different microarray studies that several genes are responsible for virulence in *H.pylori*. But one of the cag pathogenicity island (PAI) play a crucial role in development of gastric ulcer and composed of 27 genes. The presence of the *cag* PAI correlates with more-serious disease outcomes, implying that when functioning, it is important in pathogenesis. (1)

CagA protein is also playing a major virulence factor of *H.pylori*, which is delivered into gastric epithelial cells and elicits growth factor–like responses. Once within the cells, CagA is tyrosine phosphorylated by Src family kinases and targets host proteins required to induce the cell responses. (3) CagA protein is secreted into gastric epithelial cells via the type IV secretion system of *H. pylori* and plays a pivotal role in the etiology of *H. Pylori* –associated gastric diseases. (4,5). Different genomic studies have been performed for *H.pylori* to analyze the differential expression of the genes involved in the causing ulceratis and related pathways, by adopting this advance microarray technology.(1)

Microarray gene expression data allow us to quantitatively and simultaneously monitor the expression of thousands of genes under different condition. Genes with similar expression pattern under various conditions or time course may imply coregulation or relation in functional pathways. Identification of such groups of genes with similar expression patterns is usually achieved by exploratory techniques such as cluster analysis. (6) By, clustering we can group the genes from multiple experiments into groups with similar expression patterns. Similar type groups of genes which have same type expression pattern are coexpressed or coregulated. Most popular traditional gene clustering algorithms are available like Hierarchical clustering, K-means, partitioning around medoids, self-organizing maps (SOM) (7, 8). In current times as it is most popular and authentic technique, hence several microarray expression profiles studies have been performed for *H.pylori* and enormous amount of published data sets are now available at Microarray databases(14). Exponential growth of research in microarray has given explosion in the volume of raw data available for analysis; there is a widening gap between statistically compel results and their biological interpretation. Research oftentimes becomes bogged down in an analytical maze of spreadsheets and arbitrary statistical significance thresholds. Tools are needed that can efficiently summarize huge amounts of information within a biological context in well relevant manner. (9).

To overcome such complexities for interpretation of experimental data, among different data analysis techniques clustering is one of the most important and well

defined data analysis technique which plays crucial role in grouping similar type of objects. (11)

The objective of present work is to find out coexpressed genes on the basis of their expression pattern. For this we performed the coexpression study of *H.pylori* virulence genes involved in Pathogenesis and find out some novel coexpressed or co-regulatory genes on the basis of comparative analysis between under expressed and over expressed gene cluster(s). (12) Identified genes are found to have some important function and can be used as therapeutic potential drug targets for *Human* pathogen *i.e H.pylori*.

Material and Methods

From the Stanford Microarray Database, *Helicobacter pylori* was selected, raw data sets was downloaded by the Stanford microarray database (13, 14). On SMD this raw data are provided in 6 arrays sets (each array have equivalent weight) and each array contain expression data of different time and different pH. The Raw expression data consists of a set of 4623 gene. Raw data files are sorted and scaled by taking gene expression ratio Log (base2) of R/G Normalized Ratio (Mean) from the downloaded array(s). Then array data was normalized for missing values, by following the neutral method, which was proposed by Alizadeh *et al.* for the analysis of diffused large B-cell lymphoma [10] and for that we replaced the missing (empty) values with zero or averaging by rows, if a row having more than 80% missing value then deletes that row. After pre-processing and normalization we got 2081 genes in expression array data file. Finally this expression array data file imported in the genesis for clustering (15).

Genesis tool was used for clustering of data. Here Hierarchical clustering and k-means clustering were done. In Hierarchical clustering (HCL) unweighted average linkage was used because it gives acceptable results. According to gene expression value, closely related (coexpress or similar) gene would in same cluster. By using different correlation type we also found that the centred correlation is better and suitable for hierarchical clustering and gives more appropriate output for further process. Tree was cutted by deciding the level and got 10 sub clusters in HCL. For the k-means clustering numbers of hcl cluster (10) and maximum iteration of 50 was selected.

Two sets of Cluster are selected for Comparison, First gene are under expressed and over expressed in HCL and K-means cluster(s) by viewing their respective expression pattern. Over expressed gene clusters are selected for coexpression studies which having similar type of mean expression patterns in HCL and K-means cluster(s) (16).

Result and Discussion

Prediction of coexpressed genes

The heat map is generated in the terms of “differential experimental condition”, along with Differential expression along with expression pattern of genes. The expression

pattern was selected for 2081 genes in six different experiments conducting results from microarray gene expression data of *H.pylori.* The color scale ranges from saturated green for log ratios -3.0 and below to saturated red for log ratios 3.0 and above. In our results expression in form of heat map are not showing, only expression pattern of gene clusters are demonstrated. Four clusters were found which were having same/similar type of expression pattern by comparing Hcl and K-means clusre(s), which is demonstrated in table1 and their respective expression pattern are showing in fig. 1(a,b,c,d) and fig. 2(a,b).

Table 1: Same expression pattern of HCL and K- means clusters.

Cluster no. of Hcl	Cluster no of k means	Clusters Over expressed/Under expressed	Shown in figure
Hcl1, Hcl4	K4, K5	Under expressed	Fig. 1
Hcl5	K7	Over expressed	Fig. 2

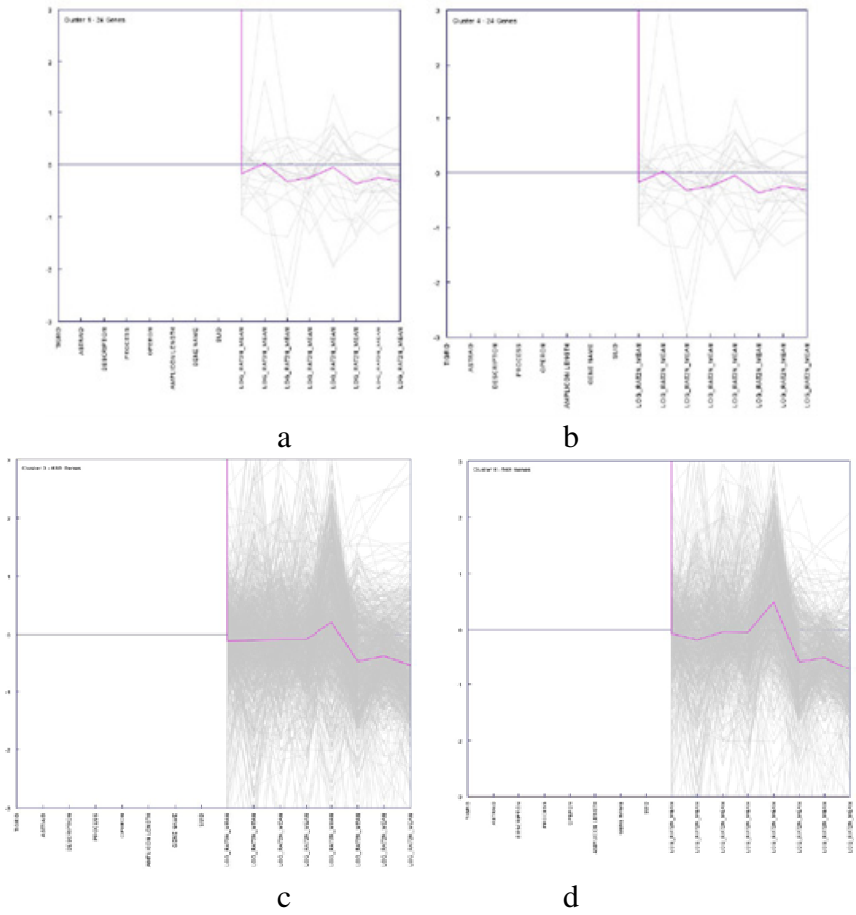
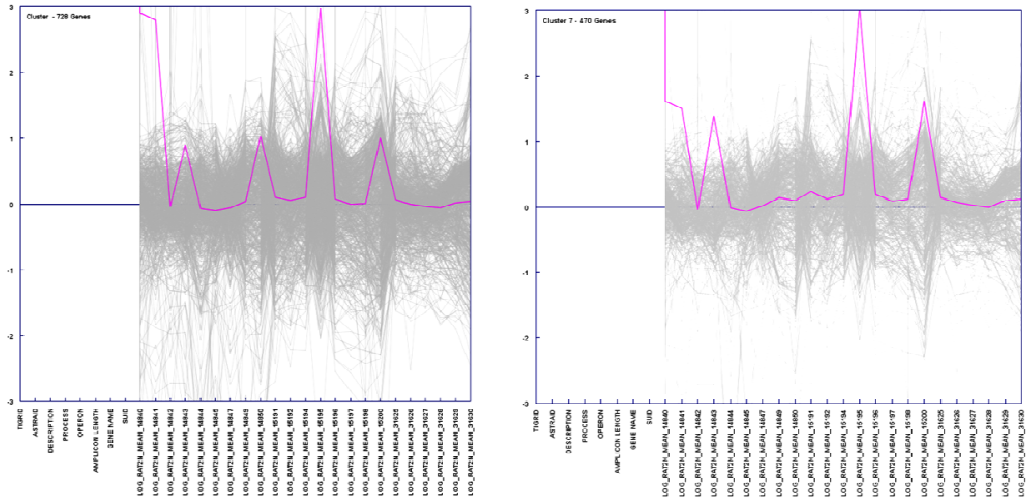


Figure 1: comparison of clusters with under expressed genes. (a & c) expression pattern Hcl Cluster 1, 4. (b & d) expression pattern of K-means cluster 4, 5.



a b

Figure 2: comparison of clusters with over expressed genes. (a) Expression pattern Hcl Cluster 5 (b) expression pattern of K-means cluster 7.

Fig.1 (a, b, c, d) exhibits cluster genes (Hcl & K-means) which are under expressed because their mean of expression is <0 and also having same mean expression pattern. The clusters of k-means 7 (fig.2 b) are very similar to the clusters found with hierarchical cluster 5 (fig.2 a) and genes comes in this cluster also highly over expressed (mean is >0).

Common genes in clusters

The genes of these selected clusters (hcl5 & K-means7) 'seed cluster' proceed for further analysis because this cluster genes are highly overexprssed and very important for this work, because, this cluster containing the Cag family genes like CagA protein. Which is responsible for Virulence factor in *H.pylroi*. The cluster size of both cluster also same contain approximately same gene size and having similar gene expression pattern. 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 fig 1 and fig 2.

Gene expression pattern also 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.

We screened out those genes are present in both cluster(s) (Hcl-5 & k-7). Total twenty genes screened out. Further validation of result BLAST was performed for these twenty gens against Database for Essential genes (DEG) for *H.pylori*. Out of twenty genes fourteen genes are Non Essential genes for this Pathogen and six genes are Essential for survival of this pathogen. These genes are VacA, rpoB, cag7, ruvA,

rps6. VacA may play an important role in the pathogenicity of *H. pylori* and also VacA is a pore-forming toxin with unique structural properties *H. pylori.* VacA gene was detected in all primary liver cancer specimens and in 71% (5/7) of control liver specimens. rpoB, cag7 No Function available at NCBI gene Database Gene References Into function but similar to GP: 1800165 percent identity: 94.57; identified by sequence similarity.

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(Hcl1- k4, Hcl- k5, Hcl5- k7) have similar expressions pattern of the respective clusters (which are present in both type of clustering) are also same.

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. These genes are essential for pathogen and also involve in different process like vacuolating cytotoxin, DNA-directed RNA polymerase subunit beta/beta, Holliday junction DNA helicase RuvA (Homologous recombination), Translation, and Ribosomal proteins: synthesis and modification. final we target total six genes. Further study carried out the product of these genes like enzymes and role of these genes product in available metabolic pathway of *H.pylori.*

Thus on the basis of comparative analysis this can be concluded that the coexpression is present within the genes of the same clusters.

Reference

[1] David N. Baldwin, Benjamin Shepherd, Petra Kraemer, Michael K. Hall, Laura K. Sycuro,1,2 Delia M. Pinto-Santini and Nina R. Salama (2007). Identification of *Helicobacter pylori* Genes That Contribute to Stomach Colonization . Infect. Immun.Vol. 75, No. 2p. 1005–1016

[2] 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.

[3] Asahi, M., T. Azuma, S. Ito, Y. Ito, H. Suto, Y. Nagai, M. Tsubokawa,Y. Tohyama, S. Maeda, M. Omata, et al. (2000). *Helicobacter pylori* CagA protein can be tyrosine phosphorylated in gastric epithelial cells. *J. Exp. Med.* 191:593–602. 5.

[4] Blaser, M.J., G.I. Perez-Perez, H. Kleanthous, T.L. Cover, R.M. Peek, P.H. Chyou, G.N. Stemmermann, and A. Nomura. (1995). Infection with *Helicobacter pylori* strains possessing *cagA* is associated with an increased risk of developing adenocarcinoma of the stomach. *Cancer Res.* 55:2111–2115.

[5] Masato Suzuki,Hitomi Mimuro, Toshihiko Suzuki (2005).Interaction of CagA with Crk plays an important role in *Helicobacter pylori* induced loss of gastric

epithelial cell adhesion. The Journal of Experimental Medicine. Vol. 202, No. 9, 1235–1247

[6] Morag Park,Tadashi Yamamoto, and Chihiro SasakawBrown,P.O. and Botstein,D. (1999) . Exploring the new world of the genome with DNA microarrays. Nature Genetics, 21, 33–37.

[7] 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.

[8] Anbupalam Thalamuthu, Indranil Mukhopadhyay, Xiaojing Zheng and George C. Tseng(2006).Evaluation and comparison of gene clustering methods in microarray analysis. Vol. 22 no. 19, pages 2405–2412

[9] Pramod Katara, Neeru Sharma, Sugandha Sharma, Indu Khatri, Akansha Kaushik, Lalima Kaushal, Vinay Sharma(2010).Comparative microarray data analysis for the expression of genes in the pathway of glioma, Bioinformation Vol. 5: Issue 1

[10] 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.

[11] Anbupalam Thalamuthu, Indranil Mukhopadhyay, Xiaojing Zheng1 and George C. Tseng(). Evaluation and comparison of gene clustering methods in microarray analysis

[12] Heyer, L. J., Kruglyak, S. and Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. Genome Research.Nov; 9(11):1106-15.

[13] 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.

[14] Stanford Microarray Database [http://genome-www5.stanford.edu/MicroArray/SMD]

[15] A. Sturn, J. Quackenbush, and Z. Trajanoski (2002). Genesis: cluster analysis of microarray data. Bioinformatics. 18: 207-208.

[16] Allocco DJ, Kohane IS, Butte AJ (2004). Quantifying the relationship between co-expression, coregulation and gene function. BMC Bioinformatics 5: 18. 1471-2105.