Available online at www.elixirpublishers.com (Elixir International Journal)

Applied Biology

Elixir Appl. Biology 48 (2012) 9546-9550

ISSN: 2229-712X

Analyzing time course microarray data of *Toxoplasma gondii* asexual development and identification of developmentally regulated genes using bioconductor

Budhayash Gautam[1,*], Gurmit Singh[2] and Satendra Singh[1]

[1]Department of Computational Biology and Bioinformatics, Jacob School of Biotechnology and Bio-engineering, Sam Higginbottom Institute of Agricultural, Technology and Sciences, Allahabad 211007, U.P, India.

[2]Department of Computer Science and Information Technology, Shepherd School of Engineering and Technology, Sam Higginbottom Institute of Agricultural, Technology and Sciences, Allahabad 211007, U.P, India.

ARTICLE INFO

Article history:

Received: 4 April 2012;

Received in revised form:

28 June 2012;

Accepted: 19 July 2012;

Keywords

Toxoplasma gondii,

Bioconductor,

Differential gene expression,

Time course microarray,

T- tests,

Cluster analysis.

ABSTRACT

Toxoplasma gondii is an obligate intracellular apicomplexan parasite that can infect a wide range of warm-blooded animals including humans. In humans and other intermediate hosts, *toxoplasma* develops into chronic infection that cannot be eliminated by host's immune response or by currently used drugs. The ability of the parasite to convert to the bradyzoite stage and live inside slow-growing cysts that can go unnoticed by the host immune system allows for parasite persistence for the life of the infected host. Little is known, however, about how bradyzoites manipulate their host cell. Large scale microarray experiments are becoming increasingly routine, particularly those which track a number of different cell lines through time. This time course information provides valuable insight into dynamics of various biological processes. The proper statistical analysis, however, requires the use of more sophisticated tools and complex statistical models. In the current study, the open-source R programming environment in conjunction with the open-source Bioconductor software were used to analyze microarray data of *T. gondii*. Several statistical analysis procedures like (log) fold changes in conjunction with ordinary and moderated t-statistics have been used for determining differentially expressed genes. The differentially expressed genes were subjected to cluster analysis followed by the annotation of the up and down regulated genes based on the gene ontology. The findings in this study suggests the overall effect of the gene expression changes is to modulate the key metabolic pathways leading to compromised host immune response, enhancement in programmed cell death, depression in cell proliferation process and induction of various diseases

Introduction

Toxoplasma gondii, an intracellular pathogen, has the potential to infect nearly every warm-blooded animal but rarely causes morbidity. *Toxoplasma gondii* is an extremely common parasite in humans and animals. Although sexual reproduction of this intracellular protozoan takes place only within felines, the intermediate hosts (many species of mammals and birds) support asexual reproduction consisting of two stages: tachyzoites and bradyzoites. Tachyzoites replicate rapidly, disseminate through the host, and cause tissue damage. Most are then cleared by the host immune response but not before some have converted into the bradyzoite stage. Bradyzoites replicate slowly, form a cyst within the host cell, and sustain a chronic infection for the life of the mammalian host. These bradyzoites latently persist and cause little pathology in a healthy host but, in an immune-compromised animal, they can reconvert into the tachyzoite stage and cause potentially fatal encephalitis. This intracellular survival likely necessitates host cell modulation, and tachyzoites are known to modify a number of signaling cascades within the host to promote parasite survival. Little is known, however, about how bradyzoites manipulate their host cell (1, 2).

Toxoplasma has a variety of mechanisms to co-opt the host cell and evade host defenses, thereby promoting intracellular survival. In particular, a number of studies indicate that tachyzoites manipulate various signaling pathways within the host cell. For example, tachyzoite-infected cells have been shown to be resistant to the induction of apoptosis through the targeting of multiple, distinct steps (1, 2). *Toxoplasma tachyzoites* also manipulate host cell NF-B signaling (3, 4), as well as mitogen-activated protein kinase signaling based on the fact that tachyzoite-infected macrophages are refractory to additional stimulation by lipopolysaccharide (5). Recent research has also shown that tachyzoite proteins can be injected into the host cell upon invasion (6, 7) and that at least one of these, a protein kinase, can have major effects on host transcription (7). To better understand the interaction between parasite and host, microarray technology has been used by several groups for genome-wide analysis of the effects of the intracellular tachyzoite on the host cell transcriptome. Two groups have shown cell-specific responses to *Toxoplasma tachyzoites* in dendritic cells, macrophages, and retinal vascular endothelial cells (8). Another group compared host gene expression in human foreskin fibroblasts (HFFs) infected by *Toxoplasma tachyzoites* with infection by other pathogens and identified two genes specifically induced by *Toxoplasma* (MacMarcks and transferrin receptor) (9). Further studies have

Tele: +91-9451427316

E-mail address: budhayash.gautam@shiats.edu.in

confirmed that the parasite-induced increase in host transferring receptor aids parasite survival. In contrast to this large body of data on infection with tachyzoites, relatively little is known about changes mediated by intracellular bradyzoites. There are many biological differences between tachyzoites and bradyzoites that predict the host responses to these stages are probably very different. For example, a number of studies have revealed developmentally regulated *Toxoplasma* genes, including metabolic enzymes, secreted proteins, and surface proteins (10, 11). These differences in gene expression correspond with a much slower growth rate for bradyzoites and development of a cyst wall characteristic of this stage, both of which might contribute to bradyzoite persistence. Furthermore, unlike tachyzoites, which attract a strong proinflammatory response, bradyzoites often persist in the animal without attracting immune infiltrates (12). This led us to hypothesize that bradyzoites might produce a unique signature of changes in the host cell transcriptome.

Microarray analysis of the response of human foreskin fibroblasts to tachyzoite infection reveals an increase in abundance of transcripts encoding enzymes involved in cholesterol synthesis in infected cells as compared with non-infected host cells (12). Microarray analyses also led to the observation that the host transcription factor HIF1, which regulates the transcription of genes involved in cell growth, cell survival, iron metabolism, and glucose metabolism, is activated by the parasite and is necessary for parasite replication under physiologically relevant oxygen levels (13).

In the present study, *Toxoplasma gondii* cDNA microarrays were used to investigate whether and how the changes occur in gene expression during its asexual development (14). For this study microarrays data were analyzed to identify profile changes in *toxoplasma gondii* gene expression. Open-source R programming environment in conjunction with the open-source Bioconductor software were used to analyze microarray data of *T. gondii*. Several statistical analysis procedures like (log) fold changes in conjunction with ordinary and moderated t-statistics have been used for determining differentially expressed genes. The differentially expressed genes were subjected to cluster analysis followed by the annotation of the up and down regulated genes based on the gene ontology.

Materials and Methods

Importing and accessing probe – level data

The microarray data used in this present piece of work was obtained from the Array-Express, which is Database of gene expression and other microarray data at the European Bioinformatics Institute (EBI). (http://www.ebi.ac.uk/arrayexpress/). The microarray data was of bradyzoite cDNA library. There were 16 arrays of data which was extracted to the desired location. The data was Agilent and data source was smd.old. In the present study the microarray data was analyzed using the Bioconductor package. Bioconductor is an open source and open development software project to provide tools for the analysis and comprehension of genomic data. Bioconductor is based primarily on the R programming language, but does contain contributions in other programming languages.

Microarray data analysis

The analysis of the microarray data consisted of the following steps: 1) within-array and between-array normalizations; 2) fitting the data to a linear model; and 3) computing differential gene expression. For normalization purposes MA-plots were generated representing the (R, G) data (R = red for Cy5 and G = green for Cy3), in which the log ratio of R versus G (M value = log2 R/G) was plotted against the overall intensity of each spot (A value = log2 (R + G)/2. Within-array normalization was first applied and M-values were normalized within each array using the Global Loess Normalization method. Quantile normalization was then applied to the A-values as a method for between-array normalization, to assure that the intensities and log-ratios had similar distributions across arrays. To estimate the average M-value for each gene and assess differential gene expression, a simple linear model was fit to the data, and M-value averages and standard deviations for each gene were obtained. To find genes with significant expression changes between groups, empirical Bayes statistics were applied to the data by moderating the standard errors of the estimated M-values. P-values were obtained from the moderated t-statistic and corrected for multiple testing with the method. The null hypothesis, that there is no differential expression of genes between regeneration stages compared with normal tissues, was rejected for p-values lower than 0.00001 (15). Thus, the change in expression is given by the fold change while the believability of the change is given by the odds. Methods used to visualize the expression profiles included hierarchical and soft clustering such as partition around medoids (PAM).

Clustering

Clustering techniques have proven to be helpful to understand gene function, gene regulation, cellular processes, and sub-types of cells. Genes with similar expression patterns (co-expressed genes) can be clustered together with similar cellular functions. This approach may further understanding of the functions of many genes for which information has not been previously available (16, 21, 22) Furthermore, co-expressed genes in the same cluster are likely to be involved in the same cellular processes, and a strong correlation of expression patterns between those genes indicates co-regulation. One of the characteristics of gene expression data is that it is meaningful to cluster both genes and samples. On one hand, co-expressed genes can be grouped in clusters based on their expression patterns (16, 17, 21, and 22). In such *gene-based clustering*, the genes are treated as the objects, while the samples are the features. On the other hand, the samples can be partitioned into homogeneous groups. Each group may correspond to some particular macroscopic phenotype, such as clinical syndromes or cancer types. Such *sample-based clustering* regards the samples as the objects and the genes as the features. In the present study gene-based clustering is performed with all the clustering algorithms (hierarchical and partition around medoids) used, however heatmaps produced by the hierarchical clustering also depicts sample-based clustering as well. Library *amap* was used for calculating the hierarchical clustering for both genes and chips. Partitioning around medoids (PAM) was used for clustering the genes, a partitioning method which operates on a distance matrix, e.g., Euclidean distance matrix. Library *cluster* was used to perform PAM-clustering and also for the generation of heatmap.

Annotation based on Gene Ontology

Annotating the genes, or in other words, combining the gene expression data with other knowledge, is typically carried out after statistical testing. Bioconductor project produces annotation packages for many chiptypes, and these can be directly used for annotating the results. In this work annotation

was done manually from the supplemented data which was generated by M. D. Cleary., (2002) (14).

Results and Discussion:

Quality assessment before and after normalization

This step was used to determine any anomalies or defects in the probe level data before proceeding further for any analysis. In order to either correct the abnormalities or filter out the defected probe data, in the present study, the following steps were involved in quality assessment of the probe level data. Density plots and MA plots are the effective way to estimate errors in the probe level data. The MA plot gives a quick overview of the distribution of the data. Quality problems are most apparent from an MA plot in cases where the lowess smoother oscillates wildly or if the variability of the M values appears to be greater in one or more arrays relative to the others. However these anomalies did not occur in all the six arrays used in the present study, suggesting the good quality of the chips used (Figure 2 & 3) (18). It depicts the distribution of intensity ratio (M) of the genes plotted by the average intensity (A). Figure 1 show the density plot for the distribution of probe level intensities of 16 arrays. Again anomalies did not occur in all the six arrays used in the present study, suggesting the good quality of the chips used (19, 20).

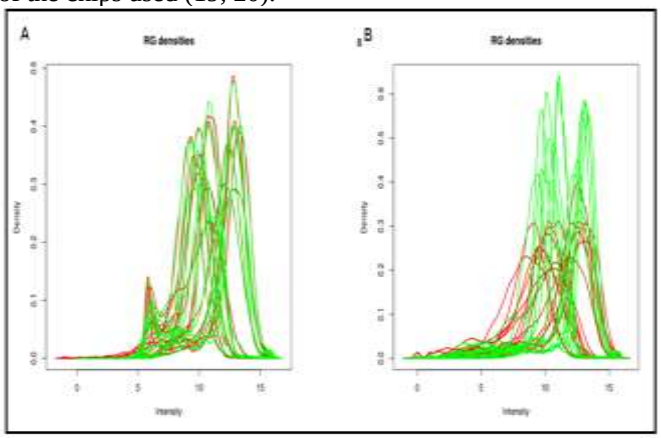

Figure 1: RG Density plot of 16 arrays of probe level data (A) after normalization and (B) before normalization.

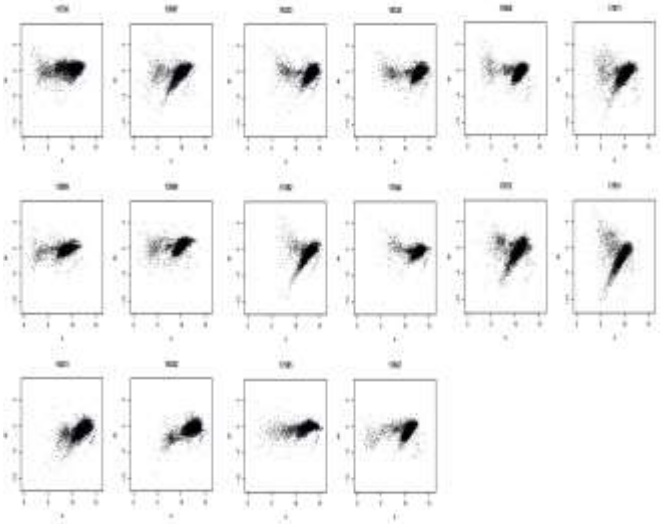

Figure 2: MA plots of 16 arrays of probe level data before normalization

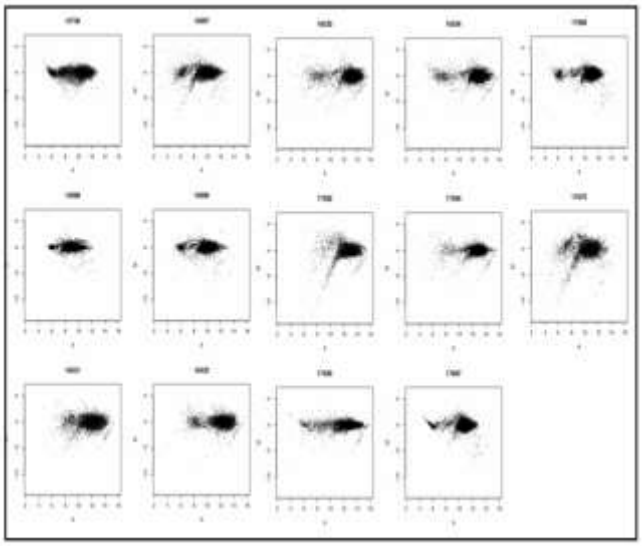

Figure 3: MA plots of 16 arrays of probe level data after normalization.

Statistical analysis

Empirical Bayes is a better analysis method than, say, traditional t-test for DNA microarray data, since it gives us more precise estimates of the statistical significance of the genes. The results of the statistical analysis was in the form of list of top genes which were expressed most (Table 1), based on the logFC, adj.P.Val, p-value and false discovery rate (19, 20). All the genes that have the unadjusted p-value at most 0.00001 were extracted (18, 19). The total number of differentially expressed genes was 648 in numbers. The volcano plots were plotted for these 648 genes. The plot was showing the number of up regulated and down regulated genes which were showing a twofold change in there expressions (Figure 4).

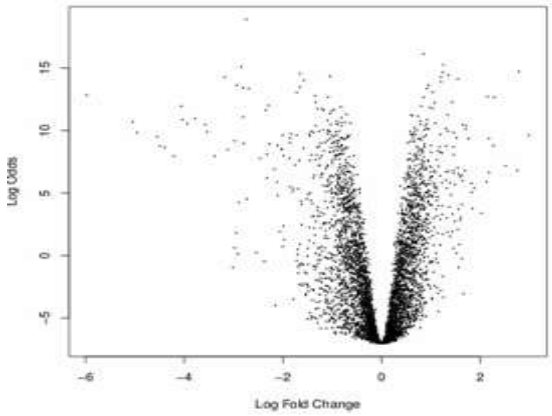

Figure 4: Volcano plot of all differentially expresses genes, which are changing their expression level twofold. Negative values shows under expression and positive values shows over expression.

Top 10 differentially expressed genes, along with their logFC, t, P-Value, adj.P.Val and B-values have been shown in table 1. There were 289 genes out of 648, which were up regulated and other 369 genes were down-regulated during the *T. gondii* asexual development. Out of 289 up-regulated genes only 7 genes were two folded up-regulated while out of 369 only 33 were two folded down-regulated (Table 2).

Table 1: Top 10 differentially expressed genes, along with their logFC, t, P-Value, adj.P.Val and B-values

	logFC	t	P.Value	adj.P.Val	B
4024	-2.745353	-17.33360	1.464955e-12	7.313056e-09	18.86940
4893	0.859428	14.55614	2.669199e-11	6.662322e-08	16.13553
3594	1.262280	13.78583	6.497866e-11	8.432470e-08	15.28079
3156	-2.848705	-13.63809	7.743754e-11	8.432470e-08	15.11147
2350	2.787771	13.30312	1.159456e-10	8.432470e-08	14.72087
238	1.236188	13.24849	1.239336e-10	8.432470e-08	14.65627
116	-1.649677	-13.16646	1.370331e-10	8.432470e-08	14.55879
152	1.357393	13.05186	1.578205e-10	8.432470e-08	14.42163
4395	-1.039419	-12.99997	1.682990e-10	8.432470e-08	14.35915
4191	-3.181548	-12.94962	1.791674e-10	8.432470e-08	14.29830

Clustering

Hierarchical clustering and soft clustering was performed using R of these top 648 genes which were showing two fold changes in their expression levels. The results of the clustering were as follows:

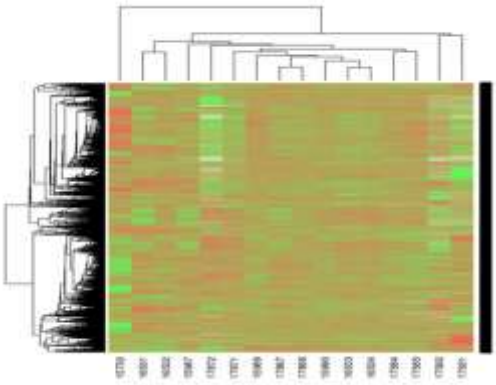

Figure 5: Heatmap of top 648 genes showing two major clusters of genes. The red spots show the up-regulated genes while the green spots indicate the down regulated genes.

The clustering by PAM also shows 2 clusters of genes (not shown here), which is almost same as the hierarchical clustering results, in which there are 2 main gene clusters (Figure 5). Thus classification and number of genes have been done properly by these two methods (18, 22).

Annotation

Annotating the genes is typically carried out after statistical testing. Bioconductor project produces annotation packages for many chip types, and these can be directly used for annotating the results. Table 2 was manually prepared from the list of top differentially expressed genes. Annotation of all the genes was prepared by the supplementary data provided by the SMD. There are some gene products which are necessary to be made by the gene but due to the *T. gondii* infection their normal level of production is decreased leading to the escape of the pathogen from the host immune system. This condition also helps the pathogen to reside in the host cell for a longer period of time and thus able to again infect the host or can transmitted from the host to another host.

Some of the genes of our interest out of top up regulated genes include the genes of following gene product – BAG1, DRPA-like, ENO1, LDH2, SAG2C/D, SAG4, GRA6-like, Toxo HSP70, HSP 90, HRA 5-like, Toxo actin, SRS9 etc. These gene products are of important enzymes and proteins as they are actively carrying out the cell metabolism and cycle (10).

The genes of our interest out of top down-regulated genes includes genes of the following proteins or gene products- ROP1, ROP4, GRA 1,GRA5, SRS2, NTP1,Fructosebiphosphate aldolase, BAG1,ATP synthase, Enolase, Heterotrimeric G protein, Apurinic endonuclease, Ubiquitin fusion protein etc. These gene products are down regulated so that the pathogen may evade the immune system of the host. These gene products play major roles in various immune processes (10, 11).

Table 2: List of top differentially expressed genes

S. No.	2-fold change up-regulated genes	Ctoxoqual contig no.	2-fold change down-regulated genes	Ctoxoqual contig no.
1.	BAG1	3906	ROP1	20
2.	DRPA-like	4436	ROP4	3826+2043
3.	ENO1	3908	GRA 1	619
4.	LDH2	4243+4054	GRA5	1406
5.	SAG2C/D	4135	SRS2	490
6.	SAG4	4196	NTP1	1589+4801
7.	GRA6-like	1344	Fructose-biphosphate aldolase	1665
8.	Toxo HSP70	1343	Ubiquitin fusion protein	2221
9.	HSP 90	1844	Apurinic endonuclease	2077
10.	GRA 5-like	4593	Heterotrimeric G protein	2378
11.	Toxo actin	244	ATP synthase	3071
12.	SRS9	4130	Enolase	2155
13.	Mucin-domain protein	3897	SAG1	3571
14.	VEE-repeat surface protein	4245+4131	G6PD	1694
15.	MET aminopeptidase	4080	ACT1	554
16.	Oligopeptidase	1284		

Conclusion

Thus from the findings reported in the present study it can be concluded that the overall effect of developmental changes accelerate the metabolism and genes which are responsible for living and surface attachment are more expressed then rest. The up and down regulation of the genes were showing important parts of the asexual development system leading to asexual growth of *Toxoplasma* which ultimately play important part in host compromised immunity, escape of the pathogen from the host immune system and provide tools to remain present in the host undetected for a longer period of time. The parasite up regulates a pathway in the host cell will identify drugs that down regulate that pathway as potential inhibitors of *Toxoplasma* growth. Ultimately, this information will have obvious benefit in opening up new potential avenues for intervention in disease. For instance, where the parasite and host cell each contribute to an essential process, combination therapy might be highly synergistic. Less obviously, we may also learn much about the basic biological processes of non-infected human or other mammalian host cells. That is, *Toxoplasma* is a powerful probe that, as it perturbs a human cell, reveals how such a host cell normally functions. Hence, as we observe the coordinated cascades of changes in gene expression in infected cells and find the host molecules that mediate these cascades, we are likely to find entirely new pathways that were previously 'known' only by the parasite.

Acknowledgment

The authors are grateful to the Sam Higginbottom Institute of Agriculture, Technology & Sciences, Deemed to be University, Allahabad for providing the facilities and support to complete the present research work.

References

1. Sinai A P, Payne T M, Carmen J C, Hardia L, Watson S J & Molestina R E, Mechanisms underlying the manipulation of host

apoptotic pathways by *Toxoplasma gondii*, *Int. J. Parasitol.*, 34 (2004) 381–391.
2. Goebel S, Gross U & Luder C G, Inhibition of host cell apoptosis by *Toxoplasma gondii* is accompanied by reduced activation of the caspase cascade and alterations of poly (ADP-ribose) polymerase expression, *J. Cell Sci.*, 114 (2001) 3495–3505.
3. Shapira S, Harb O S, Margarit J, Matrajt M, Han J, Hoffmann A, Freedman B, May M J, Roos D S & Hunter C A, Initiation and termination of NF-B signaling by the intracellular protozoan parasite *Toxoplasma gondii*. *J. Cell Sci.*, 118 (2005) 3501–3508.
4. Molestina R E & Sinai A P, Host and parasite-derived IKK activities direct distinct temporal phases of NF-kB activation and target gene expression following *Toxoplasma gondii* infection, *J. Cell Sci.*, 118 (2005) 5785–5796.
5. Lee C W, Bennouna S & Denkers E Y, Screening for *Toxoplasma gondii*-regulated transcriptional responses in lipopolysaccharide-activated macrophages, *Infect. Immun.*, 74 (2006) 1916–1923.
6. Gilbert L A, Ravindran S, Boothroyd J C & Bradley P J, *Toxoplasma gondii* targets a protein phosphatase 2C to the nucleus of infected cells, *Eukaryot. Cell*, **6 (2007) 73-83.**
7. Saeij J P J, Coller S, Boyle J P, Jerome M, White M W & Boothroyd J C, *Toxoplasma* co-opts host gene expression by injection of a polymorphic kinase homologue, *Nature*, 445 (2007) 324–327.
8. Knight B C, Brunton C L, Modi N C, Wallace G R & Stanford M R, The effect of *Toxoplasma gondii* infection on expression of chemokines by rat retinal vascular endothelial cells, *J. Neuroimmunol.*, 160 (2005) 41–47.
9. Gail M, Gross U & Bohne W, Transcriptional profile of *Toxoplasma gondii*-infected human fibroblasts as revealed by gene-array hybridization, *Mol. Genet. Genomics*, 265 (2001) 905–912.
10. Kim S K & Boothroyd J C, Stage-specific expression of surface antigens by *Toxoplasma gondii* as a mechanism to facilitate parasite persistence, *J. Immunol.*, 174 (2005) 8038–8048.
11. Schwarz J A, Fouts A E, Cummings C A, Ferguson D J & Boothroyd J C, A novel rhoptry protein in *Toxoplasma bradyzoites* and *merozoites*, *Mol. Biochem. Parasitol.*, 114 (2005) 159–166.
12. Blader I J, I. D. Manger I D & Boothroyd J C, Microarray analysis reveals previously unknown changes in *Toxoplasma gondii*-infected human cells, *J. Biol. Chem.*, 276 (2001) 24223–24231.
13. Spear W, Chan D, Coppens I, Johnson R S, Giaccia A & Blader I J, The host cell transcription factor hypoxia-inducible factor 1 is required for *Toxoplasma gondii* growth and survival at physiological oxygen levels, *Cell. Microbiol.*, 8 (2006) 339–352.
14. M. D. Cleary. "Toxoplasma gondii asexual development: identification of developmentally regulated genes and distinct patterns of gene expression". *Eukaryotic Cell.* 1(3):329–40, (2002).
15. Smyth G K, Limma: linear models for microarray data. In: 'Bioinformatics and Computational Biology Solutions using R and Bioconductor'. Gentleman, R., Carey, V., Dudoit, S`121tr, Irizarry, R., and Huber, W., (eds), Springer, New York, 2005, pp. 397--420.
16. Eisen M B, Spellman P T, Brown P O & Botstein D, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95 (1998)14863–14868
17. Ben-Dor A, Shamir R & Yakhini Z, Clustering gene expression patterns, *Journal of Computational Biology*, 6 (1999) 281–297.
18. Gregory A W, Roayaei J A, QuiÃ±nones O A, Schneider K T, A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R. Briefings in Bioinformatics 8 (2007) 415-431.
19. Dudoit S, Yang Y H, Speed T P & Callow M J, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, 12 (2002) 111–139.
20. Gillespie C S, Lei G, Boys R J, Greenall A & Wilkinson D J, Analysing time course microarray data using Bioconductor: a case study using yeast2 Affymetrix arrays, *BMC Research Notes*, 3 (2010) 81.
21. Wadhwa, Gulshan., Biochip and Biocomputers : The Future of Computing and Medicine, J. Sci. Ind. Res., Oct. 1990.
22. Budhayash Gautam, Pramod Katara, Satendra Singh and Rohit Farmer. "Drug target identification using gene expression microarray data of Toxoplasma gondii". "International Journal of Biometrics & Bioinformatics (IJBB)", Volume (4): Issue (3), (2010).