BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 6(3)

Microarray Data Analysis and Mining Tools

Saravanakumar Selvaraj, Jeyakumar Natarajan*

Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore - 641 046, India; Jeyakumar Natarajan – Email: n.jeyakumar@yahoo.co.in; *Corresponding author

Received February 02, 2011; Accepted February 03, 2011; Published April 22, 2011

Abstract:
Microarrays are one of the latest breakthroughs in experimental molecular biology that allow monitoring the expression levels of tens of thousands of genes simultaneously. Arrays have been applied to studies in gene expression, genome mapping, SNP discrimination, transcription factor activity, toxicity, pathogen identification and many other applications. In this paper we concentrate on discussing various bioinformatics tools used for microarray data mining tasks with its underlying algorithms, web resources and relevant reference. We emphasize this paper mainly for digital biologists to get an aware about the plethora of tools and programs available for microarray data analysis. First, we report the common data mining applications such as selecting differentially expressed genes, clustering, and classification. Next, we focused on gene expression based knowledge discovery studies such as transcription factor binding site analysis, pathway analysis, protein- protein interaction network analysis and gene enrichment analysis.

Keywords: Microarrays; Gene expression; Microarray data analysis; Bioinformatics tools.

Background:
Microarray is one such technology which enables the researchers to investigate and address issues which were once thought to be non traceable by facilitating the simultaneous measurement of the expression levels of thousands of genes **[1, 2]**. A microarray is simply a glass slide on which DNA molecules are fixed on an ordered manner at specific locations called spots or probes **[3]**. The spots are printed on the glass slide by different technologies such as photolithography to robot spotting. The DNA in a spot may either be complete copy of genomic DNA or short stretch of oligo-nucleotides that correspond to a gene. A typical microarray platform and its architecture and flow of experiential design and data analysis perspective are illustrated in **Figure 1**. Using microarrays one can analyze the expression of many genes in a single reaction quickly and in an efficient manner. It has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning of the human body. The core principle behind microarrays is hybridization between two DNA strands, the property of complementary nucleic acid sequences to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs. However, with the generation of large amounts of microarray data, it has become increasingly important to address the challenges of data quality and standardization related to this technology **[4]**. The recent advancement of the microarray technology has allowed for a very high resolution mapping of chromosomal aberrations with the use of their tiling array platform **[5]**. Computational data analysis tasks such as data mining which includes classification and clustering used to extract useful knowledge from microarray data. In addition, relating gene expression data with other biological information; it will provide kind of biological discoveries such as transcription factor biding site analysis, pathway analysis, and protein- protein interaction network analysis. In the present paper focus was given on biologist's perspective to get knowledge about the several tools and programs available for microarray data mining tasks. With this motivation at the end of each data mining task, we provided the list the commonly available tools with its underlying algorithms, web resources and relevant reference.

Microarray Data Analysis:
Microarray data sets are commonly very large, and analytical precision is influenced by a number of variables. So it is extremely useful to reduce the dataset to those genes that are best distinguished between the two cases or classes (e.g. normal vs. diseased). Such analyses produce a list of genes whose expression is considered to change and known as differentially expressed genes. Identification of differential gene expression is the first task of an in depth microarray analysis **[6]**. There are two common methods for in depth microarray data analysis, i.e. clustering and classification **[6]**. Clustering is one of the unsupervised approaches to classify data into groups of genes or samples with similar patterns that are characteristic to the group. Classification is supervised learning and also known as class prediction or discriminant analysis. Generally, classification is a process of learning-from-examples. Given a set of pre-classified examples, the classifier learns to assign an unseen test case to one of the classes.

Identification of Differentially Expressed Genes:
Differentially expressed genes are the genes whose expression levels are significantly different between two groups of experiments **[7]**. The genes are relevant for discovering potential drug targets and biomarkers. In the earlier stage, simple "fold change" approach was used to find differences under assumption that changes above some threshold, (For example, two-fold) were biologically significant. There are several univariate statistical methods were used later to determine either the expression or relative expression of a gene from normalized microarray data, including t tests **[8]**, modified t-test known as SAM **[9]**, two-sample t tests **[10]**, F-statistic **[11]** and Bayesian models **[12]**. For more complex datasets with multiple classes, Analysis of Variance (ANOVA) techniques were used **[13]**. Various software packages have been developed and available to identify changes in expression using the above statistical methods. The commonly used and freely available programs with its underlying algorithm are illustrated in **Table 1 (see Supplementary material)**.

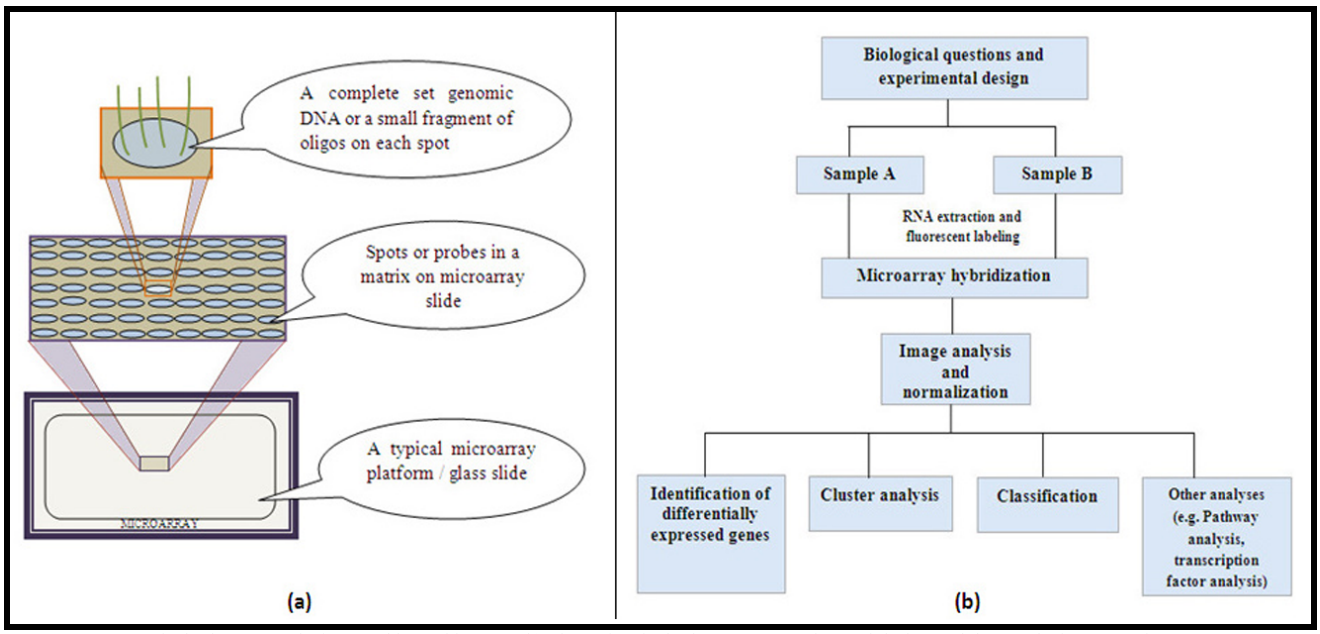

Figure 1: (a) A typical microarray platform and its architecture (b) Flow of typical microarray experimental design and data analysis perspectives

Cluster Analysis:
Clustering is the most popular method currently used in the first step of gene expression data matrix analysis. It is used for finding co-regulated and functionally related groups **[14]**. Clustering is particularly interesting in the cases when we have complete sets of an organism's genes. There are three common types of clustering methods (i.e.) hierarchical clustering, k-means clustering and self-organizing maps. Hierarchical clustering is a commonly used unsupervised technique that builds clusters of genes with similar patterns of expression **[15]**. This is done by iteratively grouping together genes that are highly correlated in terms of their expression measurements, then continuing the process on the groups themselves. It is a method of cluster analysis which seeks to build a hierarchy of clusters. A dendrogram represents all genes as leaves of a large, branching tree. The number and size of expression patterns within a data set can be estimated quickly, although the division of the tree into actual clusters is often performed visually. It generally falls into two categories (i.e.) agglomerative and divisive. Agglomerative is a bottom up approach where each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy. Divisive is a top down approach i.e., all observations start in one cluster and splits are performed recursively as one moves down the hierarchy.

K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships **[16]**. It is one of the simplest clustering techniques and it is commonly used in medical imaging and biometrics. The K-means clustering algorithm typically uses the Euclidean properties of the vector space. After the initial partitioning of the vector space into K parts, the algorithm calculates the center points in each subspace and adjusts the partition so that each vector is assigned to the cluster the center of which is the closest. This is repeated iteratively until either the partitioning stabilizes or the given number of iterations is exceeded **[17]**. A self-organizing map (SOM) is a neural network-based non-hierarchal clustering approach. (SOMs) work in a manner similar to K-means clustering **[18]**. The commonly used and freely available programs for clustering analysis are illustrated in **Table 2 (see Supplementary material)**.

Classification:
Classification is also known as class prediction, discriminant analysis, or supervised learning. Given a set of pre-classified examples, (for example, different types of cancer classes such as AML and ALL) a classifier will a find a rule that will allow to assign new samples to one of the above classes **[19]**. For classification task, one must have sufficient sample numbers to allow an algorithm to be trained known as training test and then to have it tested on an independent set of samples known as test set. Using normalized gene expression data as input vectors, classification rules can be built. There are a wide range of algorithms that can be used for classification, including k Nearest Neighbors (kNN), Artificial Neural Networks, weighted voting and support vector machines (SVM). The promising application of classification is in clinical diagnostics to find disease types and sub types. Popular examples includes finding classes of leukemia (ALL or AML) **[20]**, five classes of brain tumor (MD classis, MD desmoplastic, PNET, rhabdoide, glioblastoma) **[21]** and four classes of lymphoma **[22]**. The general data mining and machine learning application tools are used for classification tasks are illustrated in the **Table 3 (see Supplementary material)**.

Knowledge Discovery with Microarray Data:
Classification, clustering and identification of differential genes can be considered as basic microarray data analysis tasks with gene expression profiles alone. However, Gene expression profiles can be linked to other external resources to make new discoveries and knowledge. Some of the common applications that addressed with gene expression data with other biomedical information are discussed below.

Identification of transcription factor binding sites:
The identification of functional elements such as transcription-factor binding sites (TFBS) on a whole-genome level is the next challenge for genome sciences and gene-regulation studies. Transcription factors act as critical molecular switches in the gene expression profiling. Transcription factors play a prominent role in transcription regulation; identifying and characterizing their binding sites is central to annotating genomic regulatory regions and understanding gene-regulatory networks **[23]**. Various groups have exploited this problem and discovered putative binding sites in the promoter regions of genes that are co-expressed **[24]**. Some of common tools for transcription factor binding site prediction and underlying algorithm are illustrated in **Table 4 (see Supplementary material)**.

Protein-protein interaction network and pathway analysis:
Protein-protein interactions (PPI) are useful tools for investigating the cellular functions of genes. It is a core of the entire interactomics system of any living cell. PPI improves our understanding of diseases and can provide the basis for new therapeutic approaches **[25]**. Several databases that have been developed to store protein interactions such as the Biomolecular Interaction Database (BIND) **[26]**, Database of Interacting Proteins (DIP) **[27]**, IntAct **[28]**, STRING **[29]** and the Molecular Interaction Database (MINT) **[30]**. Combining co-expressed as well as interacting genes in the same cluster several meaningful predictions related to gene functions, evolutionary prelateships and pathways can be made **[31]**. Obviously, the next promising method for analyzing microarray data is pathway analysis as it involves the cascade of network interactions. Analyzing the microarray data in a pathway perspective could lead to a higher level of understanding of the system **[32]**. This integrates the

normalized array data and their annotations, such as metabolic pathways and gene ontology and functional classifications. Metabolic pathway analysis can identify more subtle changes in expression than the gene lists that result from univariate statistical analysis **[33]**. There are several web based tools and academic softwares are available to predict protein interactions and pathways from microarray data and are tabulated in **Table 5 (see Supplementary material)**.

Gene set enrichment analysis:
Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether a set of genes shows statistically significant and concordant differences between two biological states. The gene sets are defined based on prior biological knowledge, e.g., published information about biochemical pathways, located in the same cytogenetic band, sharing the same Gene Ontology category, or any user-defined set. The goal of GSEA is to determine whether members of a gene set tend to occur toward the top (or bottom) of the list, in which case the gene set is correlated with the phenotypic class distinction **[34]**. The freely available software packages for gene enrichment are illustrated in **Table 6 (see Supplementary material)**.

Conclusion:
DNA Microarray is a revolutionary technology and microarray experiments produce considerably more data than other techniques. Integrating gene expression data with other biomedical resources will provide new mechanistic or biological hypotheses. However, innovative statistical techniques and computing software are essential for the successful analysis of microarray data. This review shows the current bioinformatics tools and the promising applications for analyzing data from microarray experiments. The various data analysis perspectives and softwares mentioned in the paper will help the biological expertise as a good foundation for computational analysis of microarray data.

Acknowledgments:
We thank our lab members for valuable comments.

References:
[1] Schena M *et al*. *Science* 1995 **270**(5235): 467 [PMID: 7569999]
[2] DeRisi JL *et al*. *Science* 1997 **278**(5338): 680 [PMID: 9381177]
[3] Stears RL *et al*. *Nat Med*. 2003 **9**(1):140 [PMID: 12514728]
[4] Lockhart DJ & Winterer EA. *Nature* 2000 **405**(6788): 827 [PMID: 10866209]
[5] Lockwood WW *et al*. *Eur J Hum Genet*. 2006 **14**(2): 139 [PMID: 16288307]
[6] Mutch DM *et al*. *Genome Biol.* 2001 **2**(12): preprint0009 [PMID: 11790248]
[7] Wei C *et al*. *BMC Genomics.* 2004 **5**: 87 [PMID: 15533245]
[8] Troyanskaya OG *et al*. *Bioinformatics* 2002 **18**(11): 1454 [PMID: 12424116]
[9] Tusher VG *et al*. *Proc Natl Acad Sci U S A.* 2001 **98**(9): 5116 [PMID: 11309499]
[10] Fan J *et al*. *Proc Natl Acad Sci U S A*. 2005 **102**(49): 17751 [PMID: 16314559]
[11] Cui X *et al*. *Biostatistics* 2005 **6**(1): 59 [PMID: 15618528]
[12] Baldi P & Long AD. *Bioinformatics* 2001 **17**(6): 509 [PMID: 11395427]
[13] Kerr MK *et al*. *J Comput Biol*. 2000 **7**: 819 [PMID: 11382364]
[14] Svrakic NM *et al*. *Recent Prog Horm Res*. 2003 **58**: 75 [PMID: 12795415]
[15] Eisen MB *et al*. *Proc Natl Acad Sci U S A*. 1998 **95**: 14863 [PMID: 9843981]
[16] Tavazoie S *et al*. *Nat Genet.* 1999 **22**(3): 281 [PMID: 10391217]
[17] Brazma A & Vilo J. *FEBS Lett.* 2000 **480**(1): 17 [PMID: 10967323]
[18] Tamayo P *et al*. *Proc Natl Acad Sci U S A*. 1999 **96**(6): 2907 [PMID: 10077610]
[19] Quackenbush J. *Nat Rev Genet*. 2001 **2**(6): 418 [PMID: 11389458]
[20] Golub TR *et al*. *Science* 1999 **286**(5439): 531 [PMID: 10521349]
[21] Wang J *et al*. *BMC Bioinformatics* 2003 **4**:60 [PMID: 14651757]
[22] Kim JY *et al*. *Environ Mol Mutage*. 2005 **45**(1):80 [PMID: 15612046]
[23] Pritsker M *et al*. *Genome Res.* 2004 **14**(1): 99 [PMID: 14672978]
[24] Chowdhary R *et al*. *BMC Syst Biol*. (2010) **4**(Suppl 1):S4 [PMID: 20522254]
[25] Pellegrini M *et al*. *Expert Rev Proteomic.* 2004 **1**(2): 239 [PMID: 15966818]
[26] http://bond.unleashedinformatics.com/.
[27] http://dip.doe-mbi.ucla.edu/dip/Main.cgi.
[28] http://www.ebi.ac.uk/intact
[29] http://string.embl.de
[30] http://mint.bio.uniroma2.it/mint
[31] Guffanti A. *Genome Biol*. 2002 **3**(10): reports4031 [PMID: 12374572]
[32] Yue L & Reisdorf WC. *Curr Mol Med*. 2005 **5**: 11 [PMID: 15720266]
[33] Curtis RK *et al*. *TRENDS Biotechnol*. 2005 **23**(8): 429 [PMID: 15950303]
[34] Subramanian A *et al*. *Proc Natl Acad Sci U S A*. 2005 **102**(43): 15545 [PMID: 16199517]
[35] Zhang S. *BMC Bioinformatics*. 2007 **8**: 230 [PMID: 17603887]
[36] Saeed AI *et al*. *Biotechniques* 2003 **34**(2): 374 [PMID: 12613259]
[37] Pan F *et al*. *Bioinformatics* 2006 **22**(13): 1665 [PMID: 16672260]
[38] Leek JT *et al*. *Bioinformatics* 2006 **22**(4): 507 [PMID: 16357033]
[39] Lin M *et al*. *Bioinformatics* 2004 **20**(8): 1233 [PMID: 14871870]
[40] Heyer LJ *et al*. *Bioinformatics* 2005 **21**(9): 2114 [PMID: 15647303]
[41] Ramoni MF *et al*. *Proc Natl Acad Sci U S A*. 2002 **99**(14): 9121 [PMID: 12082179]
[42] http://www.cs.waikato.ac.nz/ml/weka/
[43] http://www.sas.com/technologies/analytics/datamining/miner/
[44] http://www.spss.com/software/modeling/modeler-pro/
[45] http://svmlight.joachims.org/
[46] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/.
[47] Ho Sui SJ *et al*. *Nucleic Acids Res*. 2005 **33**(10): 3154 [PMID: 15933209]
[48] Kel AE *et al*. *Nucleic Acids Res*. 2003 **31**(13): 3576 **[**PMID: 12824369]
[49] Hooghe B *et al*. *Nucleic Acids Res*. 2008 **36**: W128 [PMID: 18453628]
[50] Dubchak I & Ryaboy DV. *Methods Mol Bio*. 2006 **338**: 69 [PMID: 16888351]
[51] Faisst S & Meyer S. *Nucleic Acids Res*. 1992 **20**(1): 3 [PMID: 1738600]
[52] Heinemeyer T *et al*. *Nucleic Acids Res*. 1998 **26**(1): 364 [PMID: 9399875]
[53] Kiełbasa SM *et al*. *Nucleic Acids Res*. 2010 **38**: W275 [PMID: 20511592]
[54] Nikitin A *et al*. *Bioinformatics* 2003 **19**(16): 2155 [PMID: 14594725]
[55] Jiménez-Marín A *et al*. *BMC Proc.* 2009 **3**(Suppl 4): S6 [PMID: 19615119]
[56] Shannon P *et al*. *Genome Res*. 2003 **13**(11): 2498 [PMID: 14597658]
[57] http://www.springerlink.com/content/jfjpg0an9mm0g81d/
[58] Dahlquist KD *et al*. *Nat Genet.* 2002 **31** (1):19 [PMID: 11984561]
[59] Chung HJ *et al*. *Nucleic Acids Res.* 2004 **32**: W460 [PMID: 15215430]
[60] Goffard N & Weiller G, *Nucleic Acids Res.* 2007 **35**: W176 [PMID: 17586825]
[61] Wrobel G *et al*. *Bioinformatics* 2005 **21**(17): 3575 [PMID: 16020468]
[62] Shoop E *et al*. *Bioinformatics* 2004 **20** (18): 3442 [PMID: 15271779]
[63] Khatri P *et al*. *Nucleic Acids Res*. 2004 **32**: W449 [PMID: 15215428]
[64] Pandey R *et al. Bioinformatics* 2004 **20**(13): 2156 [PMID: 15145817]
[65] Zeeberg BR *et al*. *Genome Biol.* 2003 **4**(4): R28 [PMID: 12702209]
[66] Hu Z *et al*. *Nucleic Acids Res.* 2005 **33**: W352 [PMID: 15980487]
[67] Wu J *et al*. *Nucleic Acids Res*. 2006 **34**: W720 [PMID: 16845106]
[68] Backes C *et al*. *Nucleic Acids Res.* 2007 **35**: W186 [PMID: 17526521]
[69] Sartor MA *et al*. *Bioinformatics* 2010 **26**(4): 456 [PMID: 20007254]
[70] Kim SB *et al*. *Bioinformatics* 2007 **23**(13):1697 [PMID: 17468122]
[71] Paszkowski-Rogacz M *et al*. *BMC Bioinformatics* 2010 **11**: 254 [PMID: 20478033]

Edited by P Kangueane
Citation: Selvaraj & Natarajan. Bioinformation 6(3): 95-99 (2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: List of programs available for differential gene expression analysis.

S. No	Software	Algorithm/Method	URL/Reference
1.	SAM	Modified t-test known as SAM	*http://www-stat.stanford.edu/~tibs/SAM/* **[35]**
2.	MeV	non-parametric t-test/ ANOVA	http://www.tm4.org/mev/. **[36]**
3.	iArray	Student's t-test and Mann- Whitney test	http://zhoulab.usc.edu/iArrayAnalyzer.htm **[37]**
4.	EDGE	Optimal Discovery Procedure	http://www.genomine.org/edge/ **[38]**
5.	Cyber-T	Simple t-test or regularized t-tests	http://cybert.microarray.ics.uci.edu/ **[12]**

Table 2: List of programs available for cluster analysis.

S. No	Software	Algorithm/Method	URL/Reference
1.	Cluster and Treeview	Hierarchical clustering, K-means clustering self organizing maps etc.	http://rana.lbl.gov/EisenSoftware.htm **[15]**
2.	dChip	Hierarchical clustering, K-means clustering self organizing maps etc.	http://biosun1.harvard.edu/complab/dchip/ **[39]**
3.	MeV	Hierarchical clustering, K-means clustering, Tree EASE, self organizing maps, & QT-clustering etc.	http://www.tm4.org/mev/. **[36]**
4.	MAGIC Tools	Hierarchical clustering, K-means clustering, and QT-clustering	http://www.bio.davidson.edu/projects/magic/magic.html **[40]**
5.	CAGED	Bayesian clustering program on a-temporal expression data.	http://www.genomethods.org/caged. **[41]**

Table 3: List of programs available for classification.

S. No	Software	Algorithm/Method	URL/Reference
1.	weka	Artificial Neural Networks, Decision trees, k Nearest Neighbors, Support Vector Machines, and many	http://www.cs.waikato.ac.nz/ml/weka/ **[42]**
2.	SAS	Artificial Neural Networks, Decision trees, k Nearest Neighbors, Support Vector Machines, and many	http://www.sas.com/technologies/analytics/datamining/miner/ **[43]**
3.	IBM/SPSS Clementine	Artificial Neural Networks, Decision trees, k Nearest Neighbors, Support Vector Machines, and many	http://www.spss.com/software/modeling/modeler-pro/ **[44]**
4.	SVMlight	Support Vector Machines	http://svmlight.joachims.org/ **[45]**
5.	LIBSVM	Support Vector Machines	http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/ **[46]**

Table 4: List of tools for transcription factor binding site.

S.No	Software / Tool	Algorithm/Method	URL/Reference
1	oPOSSUM	Position weight matrix, Fisher exact test	http://www.cisreg.ca/oPOSSUM/ **[47]**
2	MATCH	Position weight matrix	http://www.gene-regulation.com/pub/programs.html#match **[48]**
3	ConTra	Position weight matrix, HMM	http://bioit.dmbr.ugent.be/ConTra/index.php **[49]**
4	Whole Genome rVISTA	Position weight matrix	http://genome.lbl.gov/vista/index.shtml **[50]**
5	TFSCAN	Position weight matrix, word-matching	http://mobyle.pasteur.fr/cgi-bin/portal.py?form=tfscan **[51]**
6	TFSEARCH	Position weight matrix	http://www.cbrc.jp/research/db/TFSEARCH.html **[52]**
7	TransFind	Position weight matrix, Fisher exact test	http://transfind.sys-bio.net/ **[53]**

Table 5: List of softwares for PPI and Pathway analysis.

S.No	Software	Application	URL/Reference
1	Pathway Studio	Pathway network analysis, data mining, and visualization etc.	http://www.ariadnegenomics.com/products/pathway-studio/ **[54]**
2	Ingenuity Pathway Analysis	Cancer pathway network analysis.	http://www.ingenuity.com/ **[55]**
3	Cytoscape	PPI network analysis, gene annotation and pathway integration, etc.	http://www.cytoscape.org/ **[56]**
4	Pajek	Analysis and visualization of large networks, etc	http://vlado.fmf.uni-lj.si/pub/networks/pajek/ **[57]**
5	GenMAPP2	Gene expression, Pathway analysis, and GO analysis	www.genmapp.org **[58]**
6	ArrayXPath	Mapping, visualizing expression data, and pathway analysis	http://www.snubi.org/software/ArrayXPath/ **[59]**
7	PathExpress	Pathway analysis and visualization	http://bioinfoserver.rsbs.anu.edu.au/utils/PathExpress/ **[60]**
8	GO-cluster	GO based pathway analysis	http://www.mpibpc.mpg.de/go-cluster/ **[61]**
9	GO-view	GO based pathway analysis	http://db.math.macalester.edu/goproject **[62]**
10	Onto-Express	GO based pathway analysis	http://vortex.cs.wayne.edu/Projects.html **[63]**
11	Pathway Miner	cellular and regulatory pathway analysis	http://www.biorag.org/pathway.html **[64]**
12	Gominer	GO based pathway analysis	http://discover.nci.nih.gov/gominer/ **[65]**

13	visANT 3.86	Pathway and network analysis	http://visant.bu.edu/ **[66]**
14	KOBAS	KEGG Orthology-based pathway analysis	http://kobas.cbi.pku.edu.cn **[67]**

Table 6: List of available programs for gene set enrichment analysis.

S. No	Software/Tools	Algorithm/ Method	URL/Reference
1	GSEA	Null hypothesis	http://www.broadinstitute.org/gsea/ **[34]**
2	MeV	Null hypothesis, linear model	http://www.tm4.org/mev/. **[36]**
3	GeneTrail	Dynamic-programming	http://genetrail.bioinf.uni-sb.de/ **[68]**
4	ConceptGen	Parametric and non-parametric tests	http://conceptgen.ncibi.org/core/conceptGen/index.jsp **[69]**
5	GAzer	Z-test, Parametric and non-parametric tests	http://expressome.kobic.re.kr/GAzer/index.faces **[70]**
6	PhenoFam	Mann-Whitney U test, ρ - Herrnstein's ρ statistic	http://appserver.biotec.tu-dresden.de/phenofam/ **[71]**