ISSN No. 0976-5697

Volume 5, No. 7, September-October 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# A Survey on Pre-processing of Microarray Gene Expression Data

Ajoy ku. Mishra, Dr. Subhendu ku. Pani, Dr. Bikram Kesari Ratha
PhD Scholar, UU, , Bhubaneswar,India
Associate Prof.,Dept. Of CSE, OEC, , Bhubaneswar,India
Reader, Utkal University, Bhubaneswar,India

***Abstract:*** Microarray data are often extremely asymmetric in dimensionality, highly redundant and noisy. Most genes are believed to be uninformative with respect to studied classes. Extracting useful knowledge and information from the microarrays has attracted the attention of many biologists and computer scientists. This type of experiment allows to determine relative levels of mRNA abundance in a set of tissues or cell populations for thousands of genes simultaneously. Naturally, such an experiment requires computational and statistical analysis techniques. At the outset of the processing pipeline, the computational procedures are largely determined by the technology and experimental setup that are used. Subsequently, as more reliable intensity values for genes emerge, pattern discovery methods come into play. The most striking peculiarity of this kind of data is that one usually obtains measurements for thousands of genes for only a much smaller number of conditions.This article reviews the methods utilized in pre-processing of Microarray gene expression data.

***Keywords:*** mRNA, Microarray,Gene Expression.

## I. INTRODUCTION

The DNA microarray is a way to measure the expression level of thousands of genes at the same time in a cell mixture [1]. The advent of microarray technology has provided the ability to measure the expression levels of thousands of genes simultaneously in a single experiment and made it possible that providing diagnosis for disease, in molecular level [2]. However, classification based on microarray data is very different from previous classification problems in that the number of genes greatly exceeds the number of samples, which result in the known problem of 'curse of dimensionality' and over-fitting of the training data [3]. The classification of gene expression data samples involves feature selection and classifier design. Several methods have been used to perform feature selection on the training and testing data. The two broad categories of feature subset selection have been proposed: filter and wrapper [4,5]. Although wrapper approach is more effective than filter approach, in our work we have utilized the advantage of filter approach such as gene ranking.

## II. PREPROCESSING OF MICROARRAY DATA

Before any kind of microarray data can be analysed for differential expression several steps must be taken. Raw data must be quality assessed to ensure its integrity. Unprocessed raw data will always be subject to some form of technical variation and thus must be preprocessed to remove as many unwanted sources of variation as is possible, to ensure that results are of the highest attainable level of accuracy. Ideally, the data being assayed should be preprocessed using several different methods, the results of which should be compared to identify which method is of the highest level of suitability[6]. The most appropriate method should then be used to preprocess the raw data before differential expression analysis.

### *A. Preprocessing Methods Implemented for Affymetrix GeneChip Array:*

#### *a. MicroArray Suite 5.0 (MAS5):*

MAS5 is an algorithm developed by Affymetrix and is described in their white paper "Statistical Algorithms Description Document" . This algorithm background corrects both PM and MM probes; MMs are then converted into ideal mismatches, where their values are always smaller than their corresponding PM values. Remeber than approximently 30% of the time MM values are greater than their PMs. If MM < PM, then MM value is left unchanged. A robust mean over the log2 transformed differences between PMs and the already calculated ideal mismatch is computed. Expression values are normalized by setting the trimmed mean of the original signals of each chip to a prespecified value. Hence, MAS5 data is normalized after summarisation, not before, as in many other algorithms.

#### *b. Probe Logarithmic Intensity Error Estimation (PLIER):*

PLIER is the current recommended algorithm from Affymetrix. Affymetrix claim that the algorithm improves on MAS5 by introducing a higher reproducibility of signal (lower coefficient of variation) without loss of accuracy; higher sensitivity to changes in abundance for targets near background and dynamic weighting of the most informative probes in a dataset to determine signal . In this system the PLIER algorithm is modified to include quantile normalization as PLIER does not normalize data by default.

#### *c. Robust Multi-Array Analysis (RMA)*

RMA is an academic alternative to Affymetrix's algorithms for converting probe level data to gene expression measures. This method is distinct from Affymetrix's methods in that it completely ignores the MM probe readings; the inventors of the algorithm claim that the MM probes introduce more noise and that, while acknowledging that these probes do provide useful information, have not, at the time of publication of the method, found a productive way to use it .

### *d. GeneChip RMA (GCRMA):*

GCRMA is largely based on RMA and in fact only differs in the background correction step where it uses probe sequence information to help estimate the background. This leads to improved accuracy in fold changes but at the expense of marginally lower precision .

### *e. Other Methods:*

The system can also carry out a preprocessing method by which the user can manually create the algorithm used, by specifying explicitly which of a selection of available functions, should be applied at each of the various stages, the options available to the user are as follows.

a) Background Correction:
(a). Mas5
(b). RMA
(c). RMA2
b) Normalization:
(a). Constant

## III. CONCLUSION

The biologist who uses an existing clustering algorithm to solve an underlying biological problem. The challenge before the biologist is to make an appropriate choice of an algorithm since different algorithms will produce different results. The other is the developer of clustering algorithms, who consistently strives to improve existing algorithms, so that the underlying biological problems can be solved efficiently. Raw data must be quality assessed to ensure its integrity data.

## IV. REFERENCES

[1]. Chien-Pang Lee, Yungho Leu, January 2011 A novel hybrid feature selection method for microarray data analysis Applied Soft Computing, Volume 11, Issue 1, p. 208-213.

[2]. Chien-Pang Lee, Yungho Leu, Wei-Ning Yang, Constructing gene regulatory networks from microarray data using A/PSO with DTW Applied Soft Computing, Volume 12, Issue 3, p. 1115-1124.

[3]. Michael Hecker, Peter Lorenz, Felix Steinbeck, Li Hong, Gabriela Riemekasten, Yixue Li, Uwe K. Zettl, Hans-Jürgen Thiesen, January 2012, Computational analysis of high-density peptide microarray data with application from systemic sclerosis to multiple sclerosis .

[4]. R. Xu and D. Wunsch, "Survey of Clustering Algorithms," IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645-678, May 2005.

[5]. J McQueen, "Some Methods for Classifications and Analysis of Multivariate Observations", in the Proc of 5th Barkeley Symposium on Mathematics, Statistics and Probability, pp. 281-197, 1967.

[6]. T. Zhang, R. Ramakrishnan, and M. Linvy, "BIRCH: An Efficient data Clustering Method for Very Large Datasets," Data Mining and Knowledge Discovery, vol. 1, no. 2, pp. 141-182, 1997.

[7]. A. M. Yip, M. K. Ng, E. H. Wu and T. F. Chan, "Strategies for Identifying Statistically Significant Dense Regions in Microarray Data," IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 415-429, July-September, 2007