DNA PROBE SIGNAL PROCESSING FOR IDENTIFICATION OF ABNORMAL GENE REGULATION AND PATHOGENETIC UNDERSTANDING - A DATA MINING APPROACH

AKHILESH REDDY MALE, VAIBHAV REDDY G J N

[1,2]Pursuing I MTech CSE Department of Computer Science & Engineering GVR&S College of Engineering and Technology
Guntur, Andhra Pradesh
E-mail: [1]male.akhileshreddy@gmail.com, [2]vaibhavreddy98@gmail.com

Abstract- Gene expression microarray leverages DNA probes to acquire signal intensity in the hybridized biological samples, and has become a major source for producing high-throughput experiment data. The raw, probe-level signal leads to a compre- hensive understanding of the overall microarray data set, which is especially useful when the goals of the research are different from the original data producer or contributor. Dissecting the genetic basis of complex diseases and understanding their pathogenesis thereby hinges on the successful processing of the DNA probe- level signal. Moreover, starting exploration from raw probe- level signal ensures the integrity of original data from being compromized, thus usually yielding reasonable instinct towards choosing the precise algorithms or techniques for further analysis. In this paper, we present steps towards processing probe-level signal from the microarray. As case studies of our approach, two public data sets are then used, starting from scratch: one describes the gene expression in synchronous and metachronous liver metastatic lesions from colorectal cancer, the other one uses biopsies from patients with EBV-positive undifferentiated nasopharyngeal carcinoma and from cancer-free controls. Com- pared with previous work, our approach not only identifies up/ down-regulated genes, but discovers insightful pathogenesis as well.

Keywords- Microarray, DNA probes, Signal intensity, Gene reg- ulation, Quality assessment, Filtering, Multiple testing, Taxonomic clustering, Pathogenesis

I. INTRODUCTION

An emerging problem in bioinformatics and biomedical en- gineering is the characterization of gene expression levels that underline the differences in heterogeneous and homogeneous organism. An enormous amount of genetic variants have been identified and cataloged through exploratory signal analysis and processing in genomics and proteomics [1] [2], through dissecting the genetic basis of gene expression. This leads to the identification of abnormal gene regulation, and further more the pathogenetic understanding.

Many scientific findings in this area are achieved through high-throughput experiment signal acquisition, precessing, and analysis. Gene expression microarray has become one major source for producing high-throughput experiment signal [3] [4] [5]. Microarray technology allows for simultaneous mea- suring of numerous genes for the expression-level quantities. Affymetrix [6] chips arrays has established itself as de facto standard for producing DNA microarrays. Semiconductor man- ufacturing techniques are used to produce GeneChips (5-by-5 micrometers or even smaller). Affymetrix GeneChips leverage single-stranded DNA as probes to match target samples labeled with fluorescent dye. Each cell of the chip grid holds DNA probes (DNA, complementary DNA, or Oligonucleotides), which can be configured as probe pairs and probe sets. When washed with biological samples in the form of a solution of target mRNA, the DNA probes hybridize with the target mRNA. After hybridization, the remaining unbound mRNAs are washed off, leaving the target mRNA sticked to the mi- croarray. Since the target is labeled with a particular fluorescent dye, the fluorescent signal intensity emitted by a probe marks the expression level of its corresponding gene. Hence, the low- level, probe intensity readings becomes the raw, unprocessed gene expression microarray data.

In the recent decade, enormous quantity of genomic and proteomic data sources are becoming available [7] [8]. Some gene expression data sets are extremely huge, such as the The Human Genome Project [9], one of the pioneering Big Data project. Given the abundance of gene data, making use of them is of particular significance for advancing the progress of biomedical engineering and bioinformatics.

Applying data mining techniques has been proved to be an effective approach towards knowledge discovery based on DNA probe-level signals consisting of millions of variables and often only several or dozens of observations [10] [11]. However, since practitioners and researchers engaged in this domain stem from various background, one cannot be expected to be extremely familiar with the techniques and algorithms in data mining. This phenomenon is fair and the trend that more people become interested in this interdisciplinary area is even encouraging, because every one perceives and pro- cesses information from his or her own academic or practice background, which actually fosters the development of genetic engineering. Nevertheless, one should be aware of the nature of the data handed in front: it may have been selected, screened, and adjusted. The likelihood is high for a number of reasons. First, intellectual property prevents the communication and spread of the details of data generating procedures. Second, competitions, regardless of their justice or purposes, oppose the disclose of experiment specifics, including environment (temperature and moisture), biological material, procedure configurations, experiment design, etc. Last but not least, people tend to choose a higher starting point in research, choosing the data presented on a silver plate rather than starting from scratch - the probe intensity. However, any rigorous conclusion in biomedicine and bioinformatics relies on a solid, comprehensive understanding of the entire research details. A good knowledge of data set source, such as its origin, the methods for converting it from its basic raw form to its present state in spreadsheet, and any related artificial modifications applied towards data, is not only required for the researchers, but also expected by the readers of the final report. Despite the fact that powerful algorithms and methods have been proposed in a large body of papers from data mining society, not every gene expression data may lead to significant finding, no matter what sophisticated algorithm is employed. Exquisite adjustment of a data mining techniques makes sense only when the data used is deemed worthy of further study. Moreover, a huge effort could be wasted if the raw, probe-level data is improperly handled and its integrity is compromised. Discern the data at hand for its authentic from contamination is inevitable and essential [12] [13]. Hence, the techniques of mining raw expression data is necessary.

In this work, we provide a systematic approach for pro- cessing DNA probe-level signals, using data mining methods. Major steps are summarized as follows for mining the raw, probe-level gene expression microarray data.

- Low-level preprocessing
- Additional preprocessing
- Quality assessment and filtering
- Hypothesis test
- Taxonomic clustering

The approach is applied to two real world microarray data sets, Series GSE10961 and GSE13597, on Gene Expres- sion Omnibus [14]. This public repository holds an abundant amount of high-throughput genomic and proteomic data, which is frequently visited and intensively used by the statisticians, computer scientists, and bioinformatic scholars. Meanwhile, other popular sources, such as ArrayExpress [15], have also been established to contribute raw microarray data to the bioin- formatic community. Data set GSE10961 was contributed to gene expression analysis of liver metastases of metachronous and synchronous single metastatic lesions of colorectal cancer [16]. Data set GSE13597 was the probe-level intensity signals of snap frozen biopsies from nasopharyngeal patients, with controls obtained from patients with no evidence of malignancy [17]. We verify our proposed data mining approach through these two case studies, and demonstrate that it can be competitive to the complicated analysis in [16] [17], yet with more insightful findings discovered.

The rest of the paper is organized as follows. Section II presents the proposed data mining procedures for DNA probe- level signal processing. Section III describes the experiment results from GSE10961 and GSE13597. More analysis and further discussion are provided in Section IV. Section V summarizes our work.

II. METHOD

This section outlines the stepwise approach for mining probe-level, gene expression microarray data. When a number of variants of the techniques exist for some of the following steps, the most popular one is selected. We also resist the temptation to torture or squeeze the data towards any particular object. Obtaining a general comprehension of any possibly intrinsic information, and establishing structure to unstructured data is the goal, instead.

A. Low-level Preprocessing

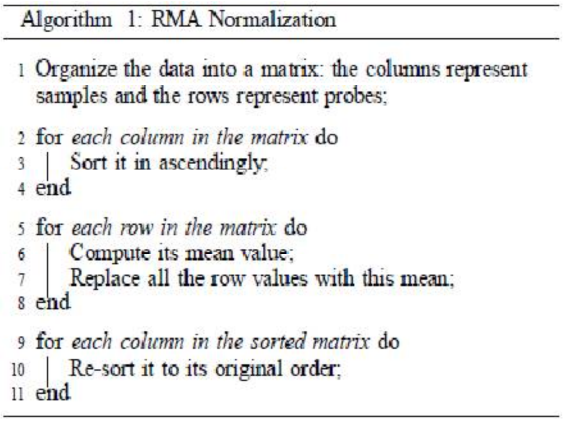
Low-level preprocessing aims to obtain the probe-level expression measurements. It consists of three procedures: background adjustment, normalization, and summarization at the probe set level. The most prevailing algorithms are MAS5 [18], RMA [19], and GCRMA [20], and PLIER [21]. We use RMA as an example, since it is recommended by [2] [16].

Background adjustment in RMA reduces background noises. The measured probe signal intensity Y is regarded as the actual signal S with the additive background noise B.

$$Y = S + B,$$

where S is assumed to be exponentially distributed as $S \sim \exp(\alpha)$, and B is normally distributed as $B \sim N(\mu, \sigma2)$. These three model parameters are thereby estimated during the RMA background adjustment for computing the actual probe signal S. This step can be seen as extracting the true signal S from the observed raw signal Y.

RMA normalization is given in Algorithm 1, combining all the previous probe-level signals into probe-set signals.



```
Algorithm 1: RMA Normalization
Organize the data into a matrix: the columns represent
samples and the rows represent probes;
for each column in the matrix do
    Sort it in ascendingly;
end
for each row in the matrix do
    Compute its mean value;
    Replace all the row values with this mean;
end
for each column in the sorted matrix do
    Re-sort it to its original order;
end
```

B. Additional Preprocessing

The term additional already indicates that this steps is not always necessary. Additional preprocessing essentially consists of further normalizing and global standardization for multiple microarrays. If the variance and co-variance of the probe signal intensities need to be suppressed, one can resort to logarithmic transformation [22], since a log function maps the original intensity measurements into a comparatively much smaller region, reducing the signal variation. This approach, however, has not been unanimously accepted by the bioinformatic society.

C. Quality Assessment & Filtering

Quality assessment gauge the data quality after (this in- dicates the previous steps may need be performed multiple times until satisfactory). Statistical plots, such as box plots,

histograms, and MA plots provide an effective approach to- wards visualizing the results.In particular, a MA plot can be used to compare the the log2 scale related to the mean log2 scale [23] for two microarray i and j:

$$M_k = log_2(\frac{X^i_k}{X^j_k}), \quad A_k = log_2 \quad X^i_k X^j_k),$$

where X I and X j stand for the intensity for probe set k of the component analysis family (PCA, LDA, CCA, etc.) is not the microarray i and j, respectively.

In this era of Big Data, not every bit and byte is deemed useful. Filtering is thereby employed to eliminate gene ex- pression measurements that are either contaminated by noise (which cannot be fixed by background adjustment) or irregular outliers whose intensities are too strong or too dim. Several criteria exist for filtering, for example, filtering by average ex- pression levels in a class, filtering by the number or proportions of present calls, by range values, etc.

D. Hypothesis Tests

At this stage, preliminary analysis of any statistical rela- tionships among genes can be assumed via hypothesis tests. Among the many choices of tests, we explain only two due to limited space: t test and multiple tests.

t test bridges the disciplines of signal processing and statistics. The test statistic is

$$t = \frac{x_1 - x_2}{s_e \quad \frac{1}{n_1} + \frac{1}{n_2}},$$

where x¯1 and x¯2 represent the average probe signals from two classes (case and control); n1 and n2 denote the number of biological samples in the case and control group, respectively; Se is the estimate of the standard deviation of the signals. In this way, t statistic can be viewed as the signal-to-noise ratio, since numerator computes the signal difference, while the denominator computes the signal variances caused by noise.

Multiple tests deserve special attention in our approach. They manifest themselves as a sequence of applying a certain hypothesis test multiple times, with each on a different data object. It is generally accepted that the odds of rejecting the null hypothesis is the significant level α (often set as 0.05), but the chance of an actual occurrence of only one false rejection is almost doomed [24]. This is consistant with the fundamental concept in probability theory that a rare event cannot happen in one trial, but is bound to happen in many trials. For this reason, compared with t Test and ANOVA F Test, the significant level α should be adjusted into α in Multiple Tests, where m denotes the number of hypothesis tests.

Among the multiple test family, the step-up Benjamini and Hochberg procedure [25] is especially useful for the high- dimentional gene microarray data.

E. Taxonomic Clustering

Clustering is a major topic in both data mining and signal processing. Grouping genes or samples through clustering analysis according to their underlying similarity is meaningful in all aspects. Clustering also provides an organised structure to the unstructured matrix data - the gene microarray.

It has been claimed that component analysis brings about similar results compared with clustering [2]. Clustering and component analysis are somehow closely related, especially when the data set under investigation follows gaussian distribution [26]. For the spirit of keeping our approach succinct, the component analysis family (PCA, LDA, CCA, etc.) is not adopted. Instead, we employ a plot-based strategy, such as scatter plots and heatmaps, for straitforward visualization and effective examination of potential clusters.

III. EXPERIMENTS

A. Data and Setup

Data set GSE10961 and GSE13597 were downloaded from Gene Expression Omnibus [14]. GSE10961 has 1164x1164 features (22 kb), with 54675 genes recorded for a number of 18 samples (10 synchronous and 8 metachronous lesions). GSE13597 has 712x712 features (16 kb), with 22283 genes recorded for a number of 28 samples (25 nasopharyngeal carcinoma and 3 cancer-free controls). The platforms for GSE10961 and GSE13597 were Affymetrix HG-U133 Plus 2.0 and Affymetrix HG-U133A, respectively.

The probe level, raw data sets from Gene Expression Om- nibus repository are usually presented as CEL files. A generic CEL file consists of the following sections: (1) a version number; (2) header description; (3) intensity set; (4)masks; (5) outliers; (6) modified description. Header description and intensity set are the two most informative parts for probe-signal precessing. The former contains the size (i.e., the number of columns and rows of the microarray), the source platform, the parameters configured for the operating program, etc. The latter contains cell number, 2-D coordinates for the cells, means and standard deviations for the intensity, etc. Minor changes exist among the raw microarray data for different types of gene chips or operating programs.

All the analyses were performed with R version 3.2.1 and Bioconductor packages. The platform is x86 64-apple- darwin13.4.0 (64-bit).

B. Results for Up Regulation of Genes

This section shows the results of using our proposed approach for DNA probe signal processing of the up gene regulation in synchronous and metachronous liver metastatic lesions from colorectal cancer. RMA is employed for low-level preprocessing.We present the box plot in Fig 1 and histogram in Fig 2 after this step. From Fig 1, the maximum values of the probe intensity obviously need further adjustment, though the IQR (interquartile range) containing the middle 50% of the data are acceptable. Therefore, additional preprocessing is deemed necessary.

The MA plot in Fig 3 shows the results after low-level preprocessing and additional preprocessing. From Fig 3, we observe that most of the clouds are centered around the M = 0 horizontal lines,suggesting that only a small number of probe sets have significantly different expressions (over- and under expressed).This conclusion, drawn from graphical plots, needs to be further justified, however.

Next, applying quality assessment and filtering by choosing IQR = 0.5, 25650 genes remained. Note this is less than half of the original 54675 genes, indicating that 6.2% of the genes are removed as outliers, and 46.9% of the gene expressions spread over the interval defined by the upper limit and Quartile 3, and the interval defined by lower limit and Quartile1.

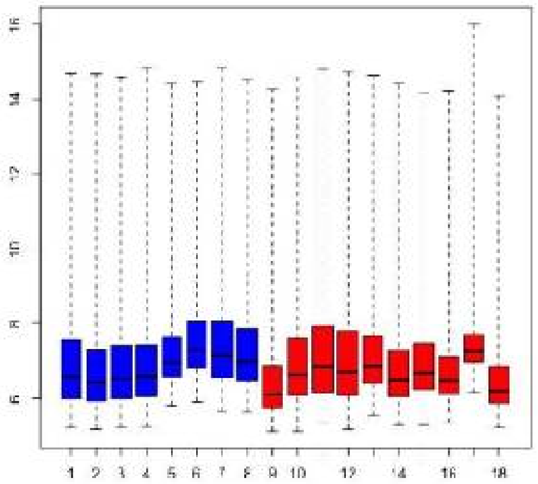


Fig. 1: Box plot for GSE10961

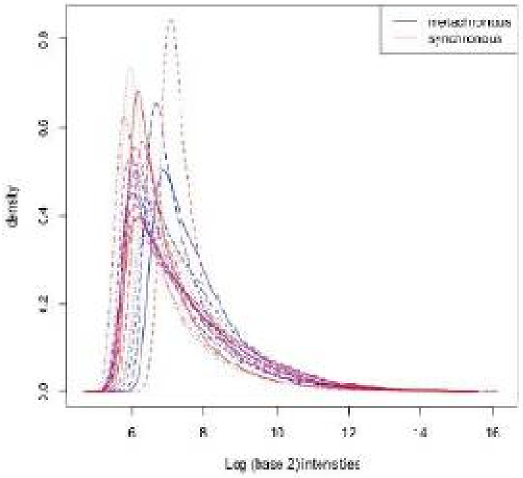


Fig. 2: Histogram for GSE10961

t Test, Bonferroni procedure, Benjamini and Hochberg procedure are the hypothesis tests employed testing between synchronous and metachronous liver metastases. The results are collectively presented in Table I. The iteration ends after six steps.

The heat map plot in Fig 4 shows the result of taxonomic clustering. In a heat map of the microarray data, the expression levels are represented as color intensities. In Fig 4, given most of the probe set color are near yellow to orange, we conclude that most genes are expressed moderately, and the intensity reading are valid. For 18 samples, the dendrogram on the top the plot indicates an even grouping, meaning each of the two major cluster consists of 9 samples.

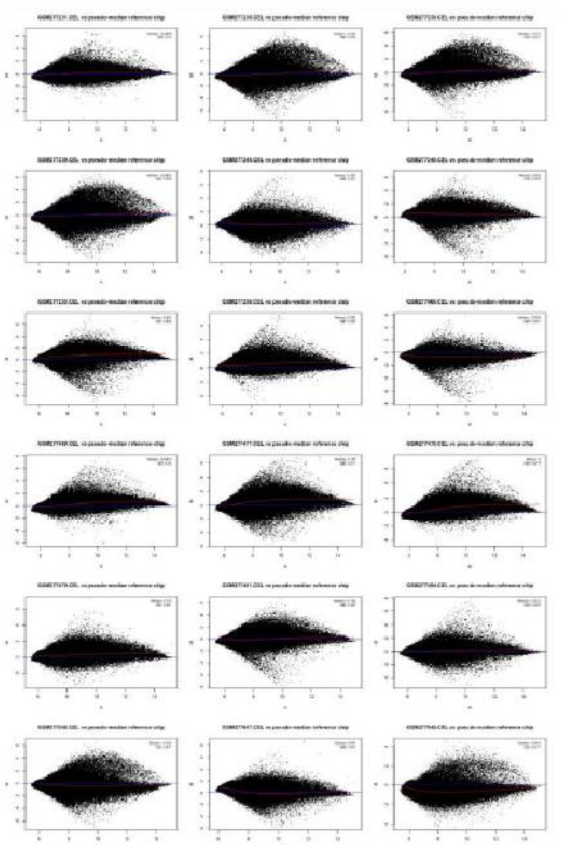


Fig. 3: MA plots for data set GSE10961.

C. Results for Down Regulation of Genes

This section shows the results of DNA probe signal processing of the down gene regulation in nasopharyngeal carcinoma. The box plot and histogram produced after RMA are presented in Fig 5 and Fig 6, respectively. Fig 7 demonstrates MA plots for each of the biopsies (25 cases + 3 controls). Since much of the step by step procedures resemble the ones in the previous section, similar description is omitted. By choosing IQR = 0.5, 5880 genes are remained after filtering. The results for employing hypothesis tests are collectively presented in Table II. The heat map plot in Fig 8 shows the result of taxonomic clustering. Note, instead of using the yellow-red ”heat” color scheme as in 4, here the topographical colors (yellow-green-blue) are used. There is no distinctly unique advantage of choosing one scheme over the other, but the latter offers a bit higher of resolution.

TABLE I: Results of hypothesis testing on GSE10961.

steps	raw p	Bonferroni method	Benjamini & Hochberg procedure
[1]	1.183462e-05	0.3035580	0.2264337
[2]	2.792442e-05	0.7162615	0.2264337
[3]	3.767055e-05	0.9662496	0.2264337
[4]	4.893162e-05	1.0000000	0.2264337
[5]	5.689843e-05	1.0000000	0.2264337
[6]	5.967680e-05	1.0000000	0.2264337

TABLE II: Results of hypothesis testing on GSE13597.

steps	raw p	Bonferroni method	Benjamini & Hochberg procedure
[1]	1.780806e-05	0.1047114	0.09521299
[2]	4.951403e 05	0.2911425	0.09521299
[3]	6.302289e-05	0.3705746	0.09521299
[4]	6.477074e-05	0.3808519	0.09521299
[5]	1.086529e-04	0.6388790	0.10957783
[6]	1.118141e-04	0.6574670	0.10957783

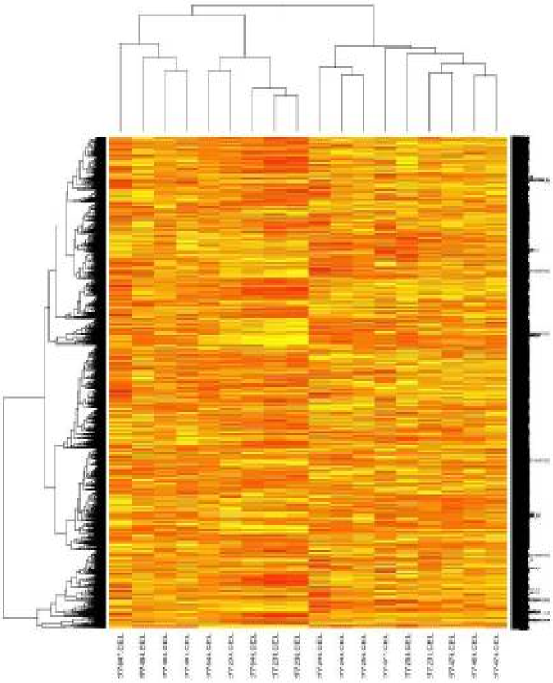


Fig. 4: Heat map for data set GSE10961 after processing.

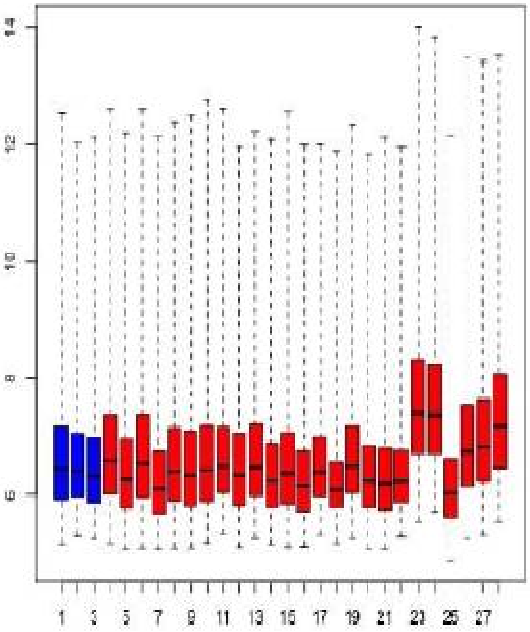


Fig. 5: Box plot for GSE13597

IV. DISCUSSION

One of the major concerns pathogenesis deals with is about whether patients of a certain disease at different stages can be treated differently, or whether the molecular portraits are able to differentiate between carcinomas, etc. Studies of this aim often employ permutation tests to verify the results, given that a very limited number of biological samples are available in experiment setting. The result of the permutation-base t tests used in [16] suggests that the molecular background of liver metastases may be different between the metachronous and the synchronous. This conclusion is verified in our work as well, as shown in Table I, where the raw p-values are at the order of 1e-5 level. Thus the null hypothesis that the difference between two classes of liver metastases is not significant should be rejected.

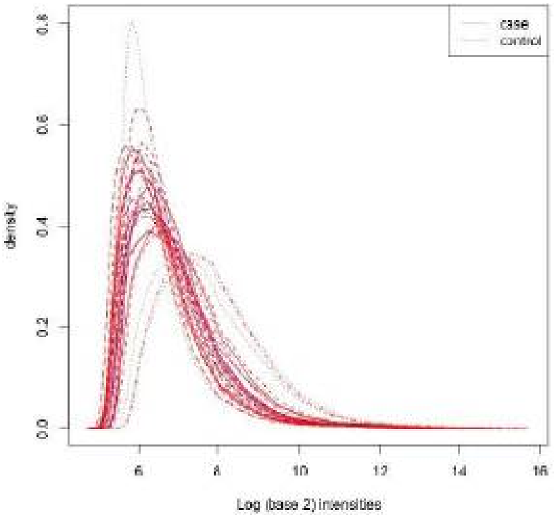


Fig. 6: Histogram for GSE13597

Similarly, tumour suppressing pathway are down-regulated in EBV-associated nasopharyngeal carcinoma, according to the t-tests in Table II, which substantiates the report in [17], using two-tailed t-test and Fishers test.

Further examination reveals that, according to Table I, although raw p-values reach 1e-5 level, the corrected p- values are not significant at all. Therefore, there is no sig- nificant differences in single genes between synchronous and metachronous liver metastases. We arrived this conclusion by conducting multiple tests, in which the significant level is adjusted (divided by the number of actual tests) since each single gene is tested. The two genes reported in [16], cyclooxygenase-2 (COX-2) and epidermal growth factor re- ceptor, are not particularly significant for regulating the path- ways in metachronous and synchronous metastases. Further statistical analysis on more clinical trials is needed to identify specific genes related to different pathways. This is also true according to the results of Bonferroni method and Benjamini and Hochberg procedure reported in Table II.

Identifying the up-regulation or down-regulation of genetic markers is therefore different from detecting the gene expres- sion signatures. The latter is a collective behavior of multiple genes, usually related to a certain pathway, on different biolog- ical samples [27] [28]. The former confines itself to a gene-to-gene comparison scenario. Considering the fact that numerous gene expression intensities are simultaneously obtained from gene microarray, the significant level must be adjusted to lower the threshold for rejecting the null hypnosis.

The heatmap plot in Fig 4 again proves the above conclu- sion that no significant differences exist in single gene expres- sion between synchronous and metachronous liver metastases, since there is no cluster of high expression levels (particularly bright spots). Similarly, in Fig 8, no rows of genes whose tran- scription are statistically and significantly different between 25 nasopharyngeal carcinoma samples and 3 cancer-free controls.

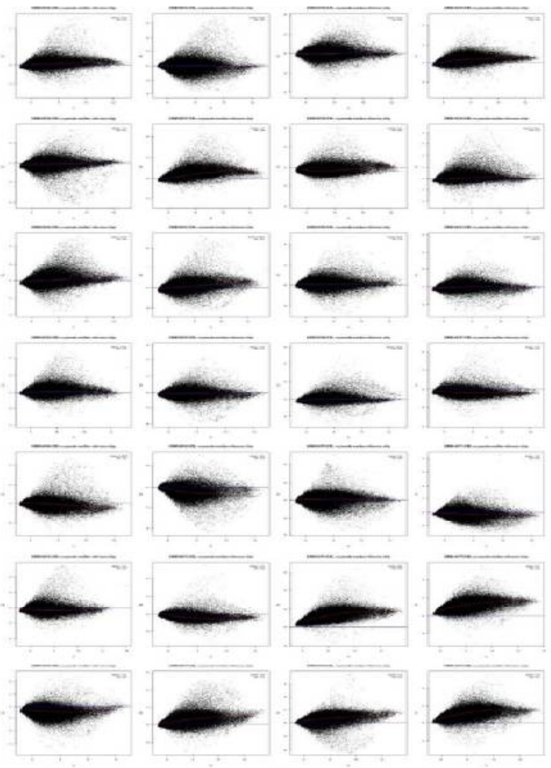


Fig. 7: MA plots for data set GSE13597.

To justify our claim, we leverage a scatter plot for com- paring probe signal intensities of a nasopharyngeal carci-noma sample (GSM342152) and a cancer-free control sample (GSM342155), as in Fig 9. The scatter plot is smoothed by kernel density estimation, as shown by the gradually changed blue band. Only the black dots that are distant from blue band represent the genes abnormally regulated between the case and control, since the x-y values for the dots in blue band are close. Pair-wised comparisons of all 25 cases and 3 controls, by using scatter plots of this kind and an union operation of all the black dots that are distant from blue band, produce a set of genes (represented by the distinctively distributed black dots). Unfortunately, none of the suspect genes in this set are consistently down regulated.

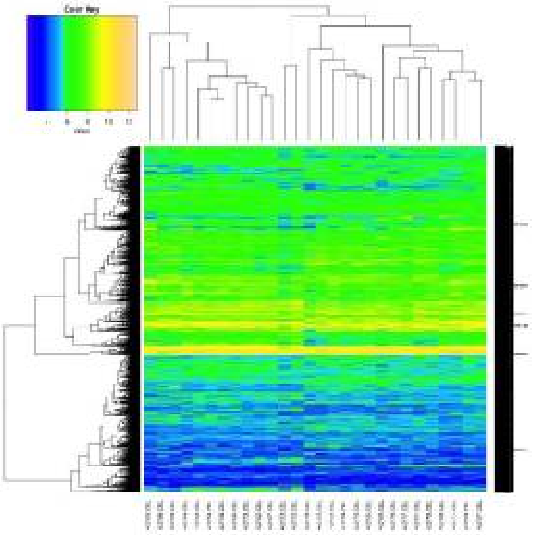


Fig. 8: Heat map for data set GSE13597.

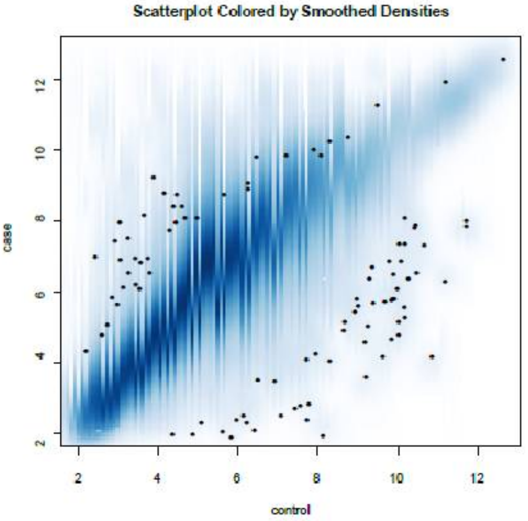


Fig. 9: Scatter plot of probe signal intensity for gene expression profile GSM342152 and GSM342155.

CONCLUSION

This paper presents a data-mining based approach to pro- cess the DNA probe signals from gene microarray data. This approach is validated through cases study of synchronous and metachronous lver metastatic lesions from colorectal cancer, and nasopharyngeal carcinoma with cancer-free controls. The motivation for mining raw microarray data is mainly to obtain a general comprehension of any possibly intrinsic information, to establish structure to unstructured data, and to be able to conduct serious research starting from scratch, employing more straitforward and suggestive figures. Besides, compared with previous work, our data mining approach leads to comprehen- sive pathogenetic understanding.

ACKNOWLEDGMENT

This work was supported by the Natural Science Founda- tion of Hubei Province of China (Grant No. 2016CFB611).

REFERENCES

[1] The 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," Nature, vol. 467, pp.1061–1073, 2010.

[2] D. M. Dziuda, Data Mining for Genomics and Proteomics. John Wiley & Sons, Inc, 2010.

[3] N. Rodriguez-Ezpeleta, M. Hackenberg, and A. M. Aransay, Bioinfor- matics for High Throughput Sequencing. Springer, 2012

[4] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Class prediction by nearest shrunken centroids with applications to DNA microarrays," Statistical Science, vol. 18, pp. 104–117, 2003.

[5] F. Chu and L. Wang, "Applications of support vector machines to cancer classification with microarray data," International Journal of Neural Systems, vol. 15(6), pp. 475–484, 2005.

[6] "Affymetrix, Inc." http://www.affymetrix.com.

[7] [7] C. Seidel, "Introduction to dna microarray," Developments in Biologi- cals, vol. 126, pp. 17–21, 2006.

[8] T. K. Karakach, R. M. Flight, S. E. Douglas, and P. D. Wentzell, "An introduction to dna microarrays for gene expression analysis," Chemometrics & Intelligent Laboratory Systems, vol. 104(1), pp. 28–52, 2010.

[9] "The human genome project race, UC Santa Cruz Center for Biomolec- ular Science and Engineering." www.cbse.ucsc.edu/research/hgp race.

[10] J. D. Hadfield and S. Nakagawa, "General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait mod- els for continuous and categorical characters," Journal of Evolutionary Biology, vol. 23(3), pp. 494–508, 2010.

[11] T. M. Przytycka, M. Singh, and D. K. Slonim, "Toward the dynamic interactome: it's about time." Briefings in Bioinformatics, vol. 11(1), pp. 15–29, 2010.

[12] J. B. Sass and J. P. Devine, "The center for regulatory effectiveness invokes the data quality act to reject published studies on atrazine toxicity," Environmental Health Perspectives, vol. 112:A18, 2004.

[13] J. J. Tozzi, W. G. Kelly, and S. Slaughter, "Correspondence: data quality act: response from the center for regulatory effectiveness," Environmental Health Perspectives, vol. 112:A18, 2004.

[14] "Gene Expression Omnibus," http://www.ncbi.nlm.nih.gov/geo/.

[15] "ArrayExpress," www.ebi.ac.uk/microarray-as/ae/.

[16] M. A. Pantaleo et al., "Gene expression profiling of liver metastases from colorectal cancer as potential basis for treatment choice." British Journal of Cancer, vol. 99(10), pp. 1729–1734, 2008.

[17] S. Bose et al., "The ATM tumour suppressor gene is down-regulated in EBV-associated nasopharyngeal carcinoma," Journal of Pathology, vol.99(10), pp. 1729–1734, 2009.

[18] Affymetrix Expression Console Software Version 1.0 User Guide, Affymetrix, Inc., Santa Clara, CA.

[19] B. M. Bolstad, "Preprocessing and normalization for affymetrix genechip expression microarrays." Methods in microarray normaliza- tion, pp. 41– 59, 2008.

[20] Z. Wu and R. A. Irizarry, "Stochastic models inspired by hybridization theory for short oligonucleotide arrays,"

Journal of Computational Biology, vol. 12(6), pp. 882–893, 2005.

[21] T. M. Therneau and K. V. Ballman, “What does PLIER really do?”Cancer Informatics, vol. 6, pp. 423– 431, 2008.

[22] S. M. Lin, P. Du, W. Huber, and W. A. Kibbe, “Model-based variance- stabilizing transformation for illumina microarray data,” Nucleic Acids Research, vol. 36(2):e11, 2008.

[23] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke, “A variance-stabilizing transformation for gene-expression microarray data.” Bioinformatics, vol. 18(Supplement 1), pp. S105–S110, 2002.

[24] L. Wasserman, All of Statistics. Springer, 2004.

[25] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: A practical and powerful approach to multiple testing.” Journal of the Royal Statistical Society, vol. Series B (Methodological) 57(1), pp. 289– 300, 1995.

[26] K. P. Murphy, Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.

[27] M. E. Garber et al., “Diversity of gene expression in adenocarcinomas of the lung.” Proceedings of the National Academy of Sciences, vol.98(24), pp. 13 784–9, 2001.

[28] M. H. Jones et al., “Two prognostically significant subtypes of high- grade lung neuroendocrine tumours independent of small-cell and large- cell neuroendocrine carcinomas identified by gene expression profiles.” Lancet, vol. 363, pp.775–781, 2005.

★★★