CLASSIFICATION OF COLORECTAL CARCINOGENIC TISSUE WITH DIFFERENT DNA CHIP TECHNOLOGIES

Ana Madevska Bogdanova, Monika Simjanoska and Zaneta Popeska

Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering
Skopje, Republic of Macedonia
ana.madevska.bogdanova@finki.ukim.mk, m.simjanoska@gmail.com, zaneta.popeska@finki.ukim.mk

Abstract

We explore increased or decreased colorectal gene expression levels since they are the reason for improper work of the cells in the colorectal region, i.e. the processes they are associated with are disrupted. In the previous work, we have unveiled the genes responsible for the colorectal cancer occurrence (the biomarkers), and made a model for classification which determines whether one patient is carcinogenic. The model uses a developed methodology that calculates the Bayesian posterior probability for classification. The gene expression profiling was done by using the DNA microarray technology from the Illumina microarray technology.

The motivation of this research is the comparation between the two different DNA chip technologies, Illumina and Affymetrix, which misses in the literature, especially for the problem of colorectal cancer classification. We examined the gene expression data obtained from the Affymetrix, in order to analyze the differences in the classification process.

Keywords - DNA microarray, Illumina, Affymetrix, machine learning, colorectal cancer, Bayes' theorem, posterior probability, Support Vector Machines.

1 INTRODUCTION

According to the World Health Organization and the GLOBOCAN project the colorectal cancer causes 8% of total cancer deaths. This fact makes the colorectal cancer the fourth most common cause of death from cancer [1].

In this paper, the colorectal cancer is considered as a problem of particular genes which have increased or decreased expression levels in the colorectal region. In the previous research, the gene expression profiling was done by using the Illumina HumanRef-8 v3.0 Expression BeadChip microarray technology [2].

In this paper we explore the problem of colorectal cancer analysis deeper, to point the problems when using the expressions of the same genes, but with the Affymetrix Human Genome U133 Plus 2.0 Array. We intent to examine if there is platform independence between this two technologies, i.e. whether we can use the gene expressions independently in the classification process of the colorectal carcinogenic tissue.

The paper is organized as follows: 2. Related work, 3. Methods and methodology, 4. Experiments and results and 5. Summary and conclusions.

2 RELATED WORK

In this section we briefly review some of the research literature related to both Affymetrix and Illumina DNA chips used for colorectal cancer gene expression analysis. We also present some of the literature related to the used data for our research.

The authors in [3] show comparison between the Affymetrix and Illumina platforms which indicate very high agreement, particularly for genes which are predicted to be differentially expressed between the two tissues. They assume that the agreement is strongly correlated with the level of expression of a gene, which is very useful statement for our research. Another research of this kind is presented in [4] where the authors performed series of analysis to compare different platforms and confirmed an intraplatform consistency across test sites as well as a high level of interplatform concordance in terms of genes identified as differentially expressed.

When studying Illumina and Affymetrix gene expression experiments, the authors in [5] state that systematic processing noise is very common in microarray experiments but is often ignored despite its potential to confound or compromise experimental results. They conclude that careful experimental design is important to protect against noise, detailed meta-data should always be provided, and diagnostic procedures should be routinely performed prior to downstream analyses for the detection of bias in microarray studies.

Another research [6] aims to directly combine appropriate Affymetrix and Illumina datasets for reanalysis and finds out that despite fundamental differences in the technology, data from these platforms can legitimately be combined at the normalized and corrected intensity level.

Unlike the research that claims high level correlation between Affymetrix and Illumina platforms, the authors in [7] exhibit cross-platform comparisons which unfortunately showed a disappointingly low concordance between lists of regulated genes between the platforms; therefore they conclude that each platform requires different statistical treatment.

Gene expression data sets used in our paper have also been used in other scientific researches. The authors in [8] aimed to find a metastasis-prone signature for early stage mismatch-repair proficient sporadic colorectal cancer (CRC) patients for better prognosis and informed use of adjuvant chemotherapy. A transcriptome profile of human colorectal adenomas is given in [9] where they characterize the molecular processes underlying the transformation of normal colonic epithelium. One of the data sets has been used in [10] to clarify the difference between MSI and microsatellite stability (MSS) cancers and, furthermore, to determine distinct characteristics of proximal and distal MSI cancers. A similar research is presented in [11] where the scientists showed cross-study consistency of MSI-associated gene expression changes in colorectal cancers.

3 METHODS AND METHODOLOGY

In our previous research [12], we used Bayes' theorem to classify the colorectal carcinogenic tissue using the gene expression analysis. In order to achieve realistic results, we developed an original methodology that includes several steps – data preprocessing, statistical analysis, modeling the a priori probability for all significant genes and the classification process itself.

In this paper we analyse the classification of the gene expression with the Affymetrix Human Genome U133 Plus 2.0 Array, which contains 54675 probes, but distinctive genes are 21050. We preprocessed the data in the same way as we did with the Illumina chip, in order to extract the biomarkers for the collorectal cancer from the Affymetrix one. We used gene expression profiling of 32 colorectal tumors and matched adjacent 32 non-tumor colorectal tissues. With the presented methodology we obtained 818 biomarkers (577 unique genes). At the Volcano Plot filtering, instead of fold change with value 1.2, we used value 4, because there were too many genes. We wanted to explore the ones whose expressed value is of greater difference compared to its starting value, to emphasize its biological significance, i.e. the biomarkers.

We used the following procedure developed in [12].

A. Preprocessing

1) Normalization – Quantile Normalization in order to make the distribution of the gene expression as similar as possible across all samples [13].
2) Low entropy filter – Higher entropy of a gene means that its expression levels are more randomly distributed [14], while low entropy of a gene reveals that there is low variability [15] in its expression levels across the samples. Therefore, we used low entropy filter to remove the genes with almost ordered expression levels.
3) T-test – it is most commonly used method for finding marker genes that distinguishes carcinogenic from healthy tissue. But, using the t-test only, we confronted with the problem of false positives, i.e. the genes which are considered statistically significant when in reality differential expression doesn't exist. To remove such genes from further analysis, we used False Discovery Rate (FDR) method.

4) False Discovery Rate – The significance in terms of false discovery rate is measured as a q-value. It can be described as a proportion of significant genes that turn out to be false positives [16]. The t-test and the FDR method identified differential expression in accordance with statistical significance values.

5) Volcano Plot – in order to consider biological significance [17], we used the volcano plot visual tool to display both statistically and biologically significant genes using a p-value threshold of 0.01 and a fold change threshold of 1.2. The genes that lie in the area cut off by the horizontal threshold, which implicates statistical significance, and the vertical thresholds, which implicate biological significance, are the genes that are up or down regulated depending on the right and the left corner of the plot respectively [12].

B. *Modeling the a priori probability*

In order to represent the differences in using the Illumina and Affymetrix DNA chips, we also modeled the a priory probability of the latter. Using the histogram visual tool, we represented gene expressions at carcinogenic and healthy tissues. Fig. 1 and Fig. 2, present the Illumina chip gene expression distribution in a carcinogenic and healthy tissue respectively, and Fig.3 and Fig.4 show the corresponding histograms of the Affymetrix chip.

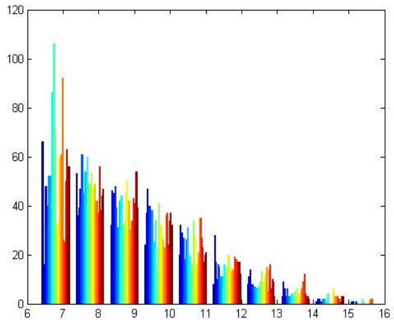


Fig. 1. Gene expression distribution of the carcinogenic tissue samples, Illumina

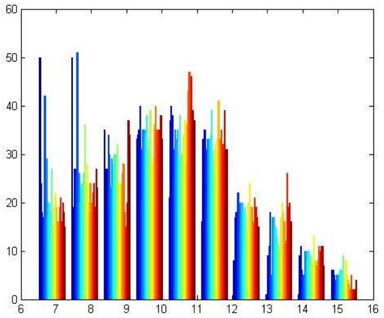


Fig. 2. Gene expression distribution of the healthy tissue samples, Illumina

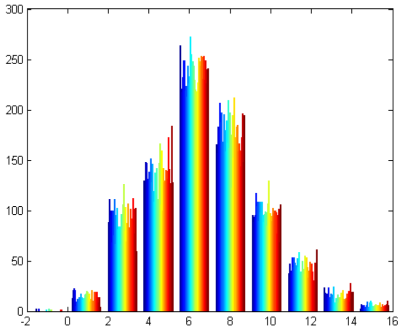


Fig. 3 Gene expression distribution of the carcinogenic tissue samples, Affymetrix

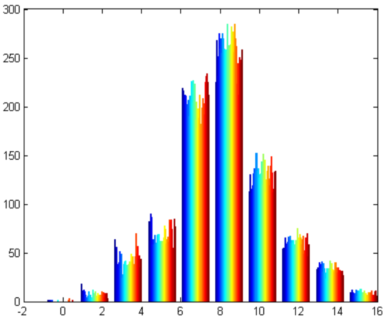


Fig. 4. Gene expression distribution of the healthy tissue samples, Affymetrix

One can notice that corresponding probability distributions (histograms) obtained from the Affymetrix chip are different from the ones obtained from the Illumina chip [12]. We can also notice that there are genes that have negative expression level, which may be explained that Affymetrix does not produce ratios, but each probe produces only an absolute intensity. Also, systematic processing noise is very common in microarray experiments [5].

It is obvious that the range of the gene expression values are stretched between different values in the observed DNA chips – the Illumina is in the range of (6, 16), and the Affymetrix is in the range (0, 16). We extract a set of mutual genes, the overlapping biomarkers and continue the further analysis from there. We used the following procedure:

- The duplicate biomarkers from Ilumina are filtered, and there are 191 left (out of 215);

– Their names are searched through the earlier established Affymetrix biomarker base.

After this procedure, there are **80 overlapped biomarkers,** the same genes in the both DNA chips. This is the control set in the rest of the analysis.

Exploring the histograms of the overlapped biomarkers, we can see that the correlation between the Illumina and Affymetrix is 0.8 for the healthy tissues, and 0.7 for the cancerogenic tissues. The correlation is substantial and it gives the assurance that although the same genes have different expression levels in the two DNA chips, they would give similar results in the classification system that we built.

3.1 Classification Techniques

The used classification techniques are the same as in [12], in order to analyse the difference between the Illumina and Affymetrix DNA chip for the given problem.

We used supervised learning methods established in [12] to diagnose whether the tissue from a given patient is healthy or carcinogenic.

A. Bayesian classifiers

In order to use the Bayes' Theorem for classification, we tried to classify the Affymetrix tissues (patients) on the modeled Illumina class-conditional densities and the prior probabilities, but the results were poor. The explanation lies in the fact that the probability distributions of the overlapping biomarkers are different. Therefore, we proceeded with the idea to model new class-conditional densities with the overlapping biomarkers for the Affymetrix biomarkers, and proceeded to calculate the posterior probability and to classify the tissues using (1), by the rule

If $p\,(C_1 \,|\, \vec{x}\,) > p\,(C_2 \,|\, \vec{x}\,)$, then choose C_1 (1)

If $p\,(C_2 \,|\, \vec{x}\,) > p\,(C_1 \,|\, \vec{x}\,)$, then choose C_2.

In the figures 5-8 we give the histograms of the overlapping set of biomarkers for Illumina and Affymetrix DNA chip accordingly.

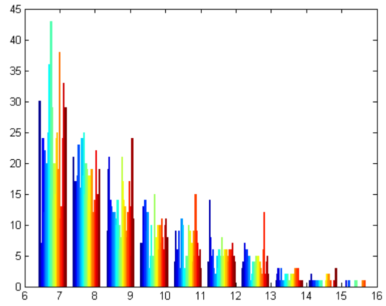


Fig. 5. Distribution of overlapped biomarkers – carcinogenic tissue, Illumina

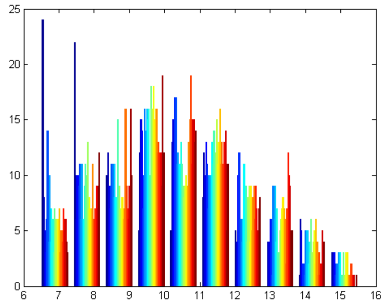
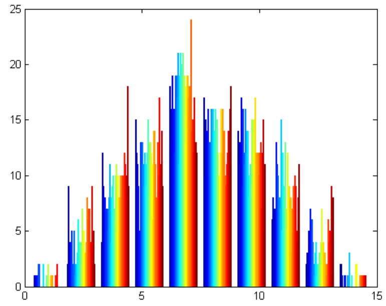


Fig. 6. Distribution of overlapped biomarkers – healthy tissue, Illumina



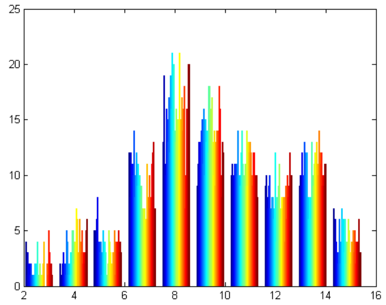


Fig. 7 Distribution

of overlapped biomarkers – carcinogenic tissue, Affymetrix

Fig. 8 Distribution of overlapped biomarkers- healthy tissue, Affymetrix

B. Support Vector Machines

The second classification method is SVM, since it was the second best method in the Illumina classification [12]. SVM is a method that easily classifies high-dimensional data. Given the overlapping biomarkers, we constructed tissue vector $\vec{x}$ for each patient. This binary classifier is supposed to choose the maximum margin separating hyperplane among the many [18] that separates the carcinogenic from healthy samples in the m-dimensional expression space, where m is the number of overlapping biomarkers. In order to investigate the expression data separability, we trained the classifier using three types of kernels: linear kernel, quadratic kernel and radial basis function. In order to avoid over-fitting, we used hold-out cross-validation technique which avoids the duplication between training data and test data, providing a more accurate estimate for the generalization performance of the algorithm [19].

4 EXPERIMENTS AND RESULTS

The new database for the colorectal microarray data are retrieved from Gene Expression Omnibus functional genomics data repository using GEO accessions: GSE9348, GSE13294, GSE8671 and GSE4554 [19].

We performed a series of analyzes according to the methodology presented in 2.2 that led us to the following results.

Table 1. Statistical results before and after implementing normalization method

Tissue	**Statistics**	**Unnormalized**	**Normalized**
Cancer tissue	*Sample min.*	3,6812	1,3034
	1st Quartile	40,0109	4,7968
	Median	100,3441	6,3918
	2nd Quartile	475,5168	8,7972
	Sample max.	83533,025	16,3324
	Outliers	7678	254
Normal tissue	*Sample min.*	4,2862	1,4456
	1st Quartile	46,6735	5,0796
	Median	118,7797	6,6896
	2nd Quartile	487,25	8,8644
	Sample max.	88580,57	16,4226
	Outliers	7410	295

Table 2. Filtering homogeneus gene expression

Filter type	Biomarkers
Low entropy	49206

Table 3. Biomarkers determining methods

Methods	Biomarkers		
	up expressed	*down expressed*	*sum*
T-test	7227	10240	17467
FDR	7100	9923	17023
Volcano Plot	190	628	818

This preparation methodology has enabled us to explore the colorectal cancer problem. Since we have the a priori knowledge such as the gene expression levels and the two possible health conditions, we used two classificators, representing the two categories of classificators – generative (Bayesian classification) and discriminative (the SVM method) in order to compare the results with those chosen methods when worked with the Illumina chip [12].

Bayesian classification: first, we used generative approach - modeling the prior distributions by ourselves. We tried to model the prior distributions of the Affymetrix overlapping biomarkers (Fig. 7 and Fig. 8). The Kolmogorov-Smirnov testing has shown that cancer and healthy tissues have different probability distributions, but among the hypotessis about the distribution acceptance (Gamma, Extreme value, LogNormal and Normal), the only one that was accepted is that both of the tissues have Normal distribution. This modelling of the Affymetrix overlapping biomarkers has lead to ambivalence, because the probability distribution is very similar for the healthy and cancerogenic tissues. Since the Bayesian classification is distribution based, one can expect that it wouldn't make the distinction between the cancerogenic and healthy tissue during the classification process.

Therefore, we conclude that Bayesian classification is not suitable classification method when Affymetrix gene expression values are used.

Table 4. Bayesian posterior classification

Bayes' theorem		
Platform	*Sensitivity*	*Specificity*
Affymetrix	1	0.0625

SVM classification: table 5 represents the results obtained from the SVM classification for the overlapping biomarker pool. In the training process, we used three types of kernels. We used hold-out cross-validation technique which involved 10% of the samples in the training set and 90% in the testing set. The Table 6 contains the results of test set which involve patients from distinct data set, separate from the data set used for training and biomarkers revealing.

Table 5. SVM classification results

SVM	Affymetrix		Illumina	
	Sensitivity	*Specificity*	*Sensitivity*	*Specificity*
Linear Kernel	**1**	**1**	**0.9565**	**0.9565**
Quadratic Kernel	09285	1	0.6956	0.9130

RBF	1	0.6428	0.3043	1

Table 6. SVM results for new patients

SVM	**Affymetrix**	
	Sensitivity	*Specificity*
Linear Kernel	**1**	**1**
Quadratic Kernel	0.9832	1
RBF	1	0.0833

We conclude that the classification process for the Affymetrix data is best performed when we used the SVM Linear Kernel. This is confirmed in Table 6 when using it on completely unknown set of patients.

5 SUMMARY AND CONCLUSIONS

In this work we continued to explore the gene expression levels of the genes involved in the colorectal tissues – carcinogenic and healthy, to produce realistic classification system [12]. We analyzed the gene expressions from the Affymetrix DNA chip in order to see if the classification process is platform independent. The procedure developed for the Illumina chip was repeated in a sense of establishing the overlapping biomarker pool for both of the platforms (Illumina/Affymetrix). The overlap set contains 80 mutual genes with highly expressed values.

We concluded that the probability distributions of the biomarker expression levels are platform-dependent. In Illumina data set, biomarkers expression levels distribution is obviously and statistically confirmed to be distinct at carcinogenic and healthy tissues. In Affymetrix, this distribution is far more similar at both carcinogenic and healthy tissues. Thus when tested on Lognormal, Gamma, Extreme values, i.e. the distribution used for Illumina [12], and additionally for Normal distribution, most of the biomarkers show equal distribution for both of the tissues, which prevents Bayesian classifier from obtaining accurate results. This directly influences the classification process - the choice between the generative and discriminative type of classification. In the Illumina case, we have shown that the generative type of classifier performs better than a discriminative [12]; whereas in the Affymetrix case, using SVM is preferably, precisely the Linear kernel. Its Sensitivity and Specificity showed excellent classification capability for the given distinct test set, which is consisted of different tissues from the ones used for training and isolating the biomarkers.

Our future work will be to explore even further the relation of the gene expression levels between the Illumina and Affymetrix DNA chip technologies for the colorectal cancer classification.

References

[1] GLOBOCAN, 2008. [Online]. Available: http://globocan.iarc.fr/factsheets/cancers/colorectal.asp

[2] T. Hinoue, D. Weisenbuerger, D. Van Den Berg, and P. Laird,“Gene expression analysis of colorectal tumors and matched adjacent non-tumor colorectal tissues,” 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25070

[3] M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, and P. Pavlidis, “Experimental comparison and cross-validation of the affymetrix and illumina gene expression analysis platforms,” Nucleic acids research, vol. 33, no. 18, pp. 5914–5923, 2005.

[4] L. Shi, L. Reid, W. Jones, R. Shippy, J. Warrington, S. Baker, P. Collins, F. de Longueville, E. Kawasaki, K. Lee et al., “The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements,” Nature biotechnology, vol. 24, no. 9, pp. 1151–1161, 2006.

[5] R. Kitchen, V. Sabine, A. Simen, J. Dixon, J. Bartlett, and A. Sims, “Relative impact of key sources of systematic noise in affymetrix and illumina gene-expression microarray experiments,” BMC genomics, vol. 12, no. 1, p. 589, 2011.

[6] A. Turnbull, R. Kitchen, A. Larionov, L. Renshaw, J. Dixon, and A. Sims, “Direct integration of intensity-level data from affymetrix and illumina microarrays improves statistical power for robust reanalysis,” BMC Medical Genomics, vol. 5, no. 1, p. 35, 2012.

[7] W. Wong, M. Loh, and F. Eisenhaber, “On the necessity of different statistical treatment for illumina beadchip and affymetrix genechip data and its significance for biological interpretation,” Biology direct, vol. 3, no. 1, p. 23, 2008.

[8] Y. Hong, T. Downey, K. Eu, P. Koh, and P. Cheah, “A metastasispronesignature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics,” Clinical and Experimental Metastasis, vol. 27, no. 2, pp. 83–90, 2010.

[9] J. Sabates-Bellver, L. Van der Flier, M. de Palo, E. Cattaneo, C. Maake, H. Rehrauer, E. Laczko, M. Kurowski, J. Bujnicki, M. Menigatti et al., “Transcriptome profile of human colorectal adenomas,” Molecular Cancer Research, vol. 5, no. 12, pp. 1263–1275, 2007.

[10] T. Watanabe, T. Kobunai, E. Toda, Y. Yamamoto, T. Kanazawa, Y. Kazama, J. Tanaka, T. Tanaka, T. Konishi, Y. Okayama et al., “Distal colorectal cancers with microsatellite instability (msi) display distinct gene expression profiles that are different from proximal msi cancers,” Cancer research, vol. 66, no. 20, pp. 9804–9808, 2006.

[11] R. Jorissen, L. Lipton, P. Gibbs, M. Chapman, J. Desai, I. Jones, T. Yeatman, P. East, I. Tomlinson, H. Verspaget et al., “Dna copy-number alterations underlie gene expression differences between microsatellite stable and unstable colorectal cancers,” Clinical Cancer Research, vol. 14, no. 24, pp. 8061–8069, 2008.

[12] M. Simjanoska, A. M. Bogdanova, and Z. Popeska, “Recognition of colorectal carcinogenic tissue with gene expression analysis using bayesian probability,” in ICT Innovations 2012. Springer Berlin / Heidelberg, 2012, *in print*.

[13] Z. Wu and M. Aryee, “Subset quantile normalization using negative control features,” Journal of Computational Biology, vol. 17, no. 10, pp. 1385–1395, 2010.

[14] A. Butte and I. Kohane, “Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements,” in Pac Symp Biocomput, vol. 5, 2000, pp. 418–429.

[15] C. Needham, I. Manfield, A. Bulpitt, P. Gilmartin, and D. Westhead, “From gene expression to gene regulatory networks in arabidopsis thaliana,” BMC systems biology, vol. 3, no. 1, p. 85, 2009.

[16] J. Storey and R. Tibshirani, “Statistical significance for genomewide studies,” Proceedings of the National Academy of Sciences, vol. 100, no. 16, pp. 9440–9445, 2003.

[17] A. Tarca, R. Romero, and S. Draghici, “Analysis of microarray experiments of gene expression profiling,” American journal of obstetrics and gynecology, vol. 195, no. 2, pp. 373–388, 2006.

[18] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. Furey, M. Ares, and D. Haussler, “Knowledge-based analysis of microarray gene expression data by using support vector machines,” Proceedings of the National Academy of Sciences, vol. 97, no. 1, pp. 262–267, 2000.

[19] P. Refaeilzadeh, L. Tang, and H. Liu, “Cross-validation,” Encyclopedia of Database Systems, vol. 5, 2009.

[20] GEO, “Gene expression omnibus.” [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/

[21] A. M. Bogdanova and N. Ackovska, “New support vector machines based approach over dna chip data,” in Innovations in Information Technology. IEEE, 2008, pp. 16–19.

[22] A. Bogdanova and N. Ackovska, “Data driven intelligent systems,” in ICT Innovations 2010, 2010.