Recognition of Colorectal Carcinogenic Tissue with Gene Expression Analysis Using Bayesian Probability

Monika Simjanoska[1], Ana Madevska Bogdanova[2] and Zaneta Popeska[2]

Ss. Cyril and Methodius University in Skopje, Faculty of Computer Science and Engineering

[1]m_simjanoska@yahoo.com,
[2]{ana.madevska.bogdanova, zaneta.popeska}@finki.ukim.mk

Abstract. According to the WHO research in 2008, colorectal cancer caused approximately 8% of all cancer deaths worldwide. Only particular set of genes is responsible for its occurrence. Their increased or decreased expression levels cause the cells in the colorectal region not to work properly, i.e. the processes they are associated with are disrupted. This research aims to unveil those genes and make a model which is going to determine whether one patient is carcinogenic. We propose a realistic modeling of the gene expression probability distribution and use it to calculate the Bayesian posterior probability for classification. We developed a new methodology for obtaining the best classification results. The gene expression profiling is done by using the DNA microarray technology. In this research, 24,526 genes were being monitored at carcinogenic and healthy tissues equally. We also used SVMs and Binary Decision Trees which resulted in very satisfying correctness.

Keywords: DNA microarray, machine learning, colorectal cancer, Bayes' theorem, posterior probability, Support Vector Machines, Binary Decision Trees

1 Introduction

According to the World Health Organization and the GLOBOCAN project which provided research of the cancer incidence, mortality and prevalence worldwide in 2008, colorectal cancer is responsible for nearly 608,000 deaths, or, it 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. The gene expression profiling is done by using the Illumina HumanRef-8 v3.0 Expression BeadChip microarray technology. This whole-genome expression array allows 24,526 transcript probes.

Gene expression data used in this paper is downloaded from the ArrayExpress, EMBL-EBI biological database [2]. It is collected according to the MIAME standard and can be accessed using the unique identity number E-GEOD-25070.

The paper is organized as follows: 2. Methods and methodology, where we give an overview of the related work and the developed original procedure for using the Bayes' theorem, 3. Experiments and results and 4. Summary and conclusions.

2 Methods and Methodology

2.1 Related Work

In this section we briefly review some of the research literature related to colorectal cancer statistical and discriminant analysis.

L. C. LaPointe [3] in his Ph.D. thesis describes the discovery and validation of biomarker candidates for colorectal neoplasia. Some genes exhibit gene expression patterns which correlate with the neoplastic phenotype and these results enable investigation of the central practical aim: the identification from the pool of differentially expressed genes those candidate biomarkers which could serve as leads for clinical assay research and development in the future. He has given an overview to discriminant analysis, Bayes' theorem, and machine learning algorithms in candidate biomarkers identification.

Gene expression data set used in our paper has also been used in other scientific researches. The experiment authors in [1] together with Christopher P.E. Lange et al. [4] used this expression data to perform analysis of the aberrant DNA methylation in colorectal cancer.

Another paper that used the same gene expression data and that can be helpful in comparing different methods for identification of colorectal cancer genes is the research done by Bi-Qing Li, et al. [5]. In order to identify colorectal cancer genes, they used method based on gene expression profiles and shortest path analysis of functional protein association networks.

In the Shizuko Muro's, et al. [6] research, when making the classification model they assumed that the gene expression data is distributed according to a mixture of Gaussian distributions.

2.2 The methodology

In our research, 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. Furthermore, we used the Support Vector Machines and Decision trees to compare the obtained classification results.

2.2.1 Data Preprocessing

Gene expression profiling of 26 colorectal tumors and matched adjacent 26 non-tumor colorectal tissues is retrieved for further analysis. The gene expression data

consists of raw and processed data. Processed data is log2 transformed and normalized using Robust Spline Normalization (RSN) method.

Normalization methods. Our research aims to unveil the differential expression of the genes expression level. We assume that only a small set of genes are differentially expressed. In such cases Quantile normalization is a suitable normalization method. Quantile normalization (QN) makes the distribution of the gene expression as similar as possible across all samples [7]. However, Quantile normalization forces same distribution for intensity values across different samples which can cause small differences among intensity values to be lost [8]. Therefore, we also analyzed processed data which is normalized using the RSN method. RSN method combines the good features from the Quantile and the Loess normalization. Rather, it combines the strength of Quantile normalization and the curve fitting [9].

Filtering methods. Some genes may not be well distributed over their range of expression values, i.e. low expression values can be seen in all samples except one [10]. This can lead to incorrect conclusion about gene behavior. To remove such genes, we used an entropy filter. Entropy measures the amount of information (disorder) about the variable. Higher entropy for a gene means that its expression levels are more randomly distributed [11], while low entropy for a gene means that there is low variability [12] in its expression levels across the samples. Therefore, we used low entropy filter to remove the genes with almost ordered expression levels.

Statistical tests. This model assumes that whole-genome gene expression follows normal distribution [13]; therefore, we used unpaired two-sample t-test for differential expression. The t-test is most commonly used method for finding marker genes that discriminate carcinogenic from healthy tissue. Here we have two independent groups, cancer vs. normal tissue. We expect that most of the genes are not differentially expressed. Thus, the null hypothesis states that there is no statistical difference between the cancer and the normal samples. The rejection of the null hypothesis depends on the significance level which we determine. In this paper we consider the genes as statistically significant for a p-value less than 0.01, which means that the chances of wrong rejection of the null hypothesis is less than 1 in 100.

Using the t-test only, we confront with the problem of false positives. The term false positive refers to 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. FDR method is defined as a measure of the balance between the number of the true positives and false positives [14]. For a threshold of 0.01 we expect 10 genes to be false positive in a set of 1000 positive genes. 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 [14]. This method is supposed to reduce the number of significant genes supplied from the t-test.

The t-test and the FDR method identified differential expression in accordance with statistical significance values. However, these methods do not consider biological significance. The biological significance is measured as a fold change which describes how much the expression level changed starting from the initial value. Fold change is measured as ratio between the two expression intensities and does not take into account the variance of the expression levels. Because of its simplicity it is usually used in combination with another statistical method [15]. Therefore, 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.

2.2.2 Modeling the a priori probability

Using the histogram[1] visual tool, we represented gene expressions at carcinogenic and healthy tissues. Observing the Fig. 1 and Fig. 2, which show the gene expression distribution in a carcinogenic and healthy tissue respectively, we perceived that their distribution substantially differs one from another. In order to confirm the visual assumption of difference, we used the Kolmogorov-Smirnov test at which all genes rejected the null hypothesis of having the same distribution. Having the prior knowledge about gene expressions distribution, we can use the Bayes' theorem to compute the posterior probability $p\ (C_i \mid \vec{x})$, where $\vec{x} = \{e_1, e_2, \ldots, e_n\}$ is a tissue vector containing the expression values for all significant genes, and C_i is one of the classes – carcinogenic or healthy. The posterior probability expresses how probable the class C_i is for a given tissue $\vec{x}$. According to the Bayes' theorem [16], in order to obtain this probability, we must determine the class-conditional densities $p\ (\vec{x} \mid C_i)$ for each class C_i individually, and the class prior probabilities $p\ (C_i)$.

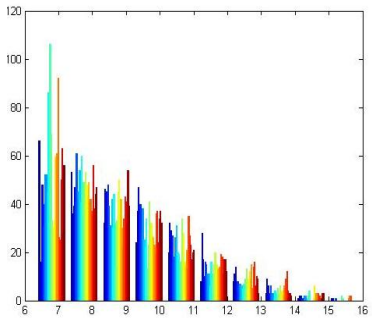

Fig. 1. Gene expression distribution of the carcinogenic tissue samples

[1] The histogram represents each sample (patient) with different color, putting its expression values on the x-axis and the number of genes on the y-axis.

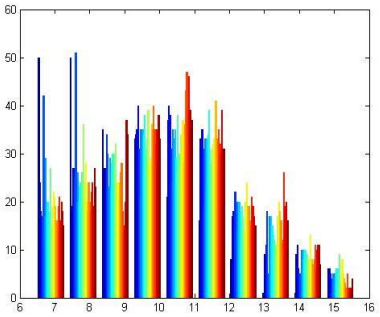

Fig. 2. Gene expression distribution of the healthy tissue samples

Since the probability $p(\vec{x})$ is calculated using the law of total probability and is the same for both cases it is usually ignored the Bayes' theorem takes the form

$$p(C_i \mid \vec{x}) = p(\vec{x} \mid C_i)\, p(C_i). \tag{1}$$

Estimating the class prior probabilities $p(C_i)$ is simple in this case, because we have equal number of samples into both of the classes - carcinogenic and healthy. Thus, the prior probabilities are $p(C_1) = p(C_2) = 0.5$ for the carcinogenic and the healthy class, respectively.

The class-conditional density $p(\vec{x} \mid C_i)$ is the probability density function for $\vec{x}$ given the particular class C_i. Unlike most of the models which assume Gaussian distribution, we followed generative approach and modeled class-conditional densities by ourselves. Thus, assuming independence of gene distribution we modified the class-conditional densities as

$$p(\vec{x} \mid C_i) = \Pi\, f_1 f_2 \ldots f_n, \tag{2}$$

where f_i is the continuous probability distribution of each gene distinctively.

In order to determine the distribution of each gene, we needed to observe a large quantity of data. Therefore, using the holdout cross-validation technique, we involved ¾ of the data in the training process. For each gene we performed statistical tests over the continuous and asymmetric Lognormal, Gamma and Extreme value distribution and we have chosen the one with the highest probability. Once we have modeled the class-conditional densities and the prior probabilities, we used the Bayes formulation to calculate the a posteriori probability to classify the tissues (1).

2.3 Classification techniques

As we revealed the genes whose differential expression is significant for the colorectal cancer in the data preprocessing part, we can use supervised learning methods to diagnose whether the tissue is healthy or carcinogenic and choose the one that recognizes the carcinogenic tissues with highest precision.

Bayes' Theorem. Once we have modeled the class-conditional densities and the prior probabilities, we proceeded to calculate the posterior probability and to classify the tissues using (1), by the rule

$$\text{If } p(C_1 \mid \vec{x}) > p(C_2 \mid \vec{x}), \text{ then choose } C_1 \qquad (3)$$

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

Support Vector Machines. SVM is a method that can classify high-dimensional data as are multiple genes' expression levels. Given significant genes expression levels, we constructed tissue vectors $\vec{x}$ for each patient. This binary classifier is supposed to choose the maximum margin separating hyperplane among the many [17] that separates the carcinogenic from healthy samples in the m-dimensional expression space, where m is the number of significant genes. 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. To avoid over-fitting, we used hold-out cross-validation technique which avoids the overlap between training data and test data, yielding a more accurate estimate for the generalization performance of the algorithm [18]. In addition, we also used bootstrapping method for accuracy improvement.

Decision Trees. Decision tree is a hierarchical data structure implementing the divide-and-conquer strategy. The tree codes directly the discriminants separating class instances without caring much for how those instances are distributed in the regions. The decision tree is a discriminant-based, whereas the statistical methods are likelihood-based in that they explicitly estimate the likelihood before using the Bayes' rule and calculating the discriminant. Discriminant-based methods directly estimate the discriminants, bypassing the estimation of class densities [19]. The reason for using this method is because it is easy to implement and it solves the classification problem using completely different approach from the SVM and Bayes' theorem, which gives useful insight for methods efficiency comparison.

3 Experiments and Results

We retrieved colorectal microarray data from the ArrayExpress biological database [2]. To obtain realistic modeling of the specific genes gene expression probability distribution, we performed a series of analyzes according the methodology presented in 2.2 that leaded to these results.

As far as we normalized gene expression levels (Table 1), we continued with genes reduction methods. Starting from the initial condition of 24,526 genes for 52 tissues we implemented a few statistical tests to separate the significant genes suitable for classification modeling, i.e., the data preprocessing. At first we removed the genes with low variability in their expression values using the low entropy filter (Table 2). Assuming the whole-genome distribution follows a normal distribution and most of

the genes are not differentially expressed, we performed t-test statistics to find marker genes that discriminate carcinogenic from healthy tissue. The number of genes significantly reduced to approximately 3500 for up expression and 2900 for down expression. To remove the false positives, we used the FDR method which eliminated nearly 400 genes at both up and down expression. Eventually, using the volcano plot visual tool, we cut off the genes considering both statistical and biological significance, which resulted in a set of nearly 200 genes, most of them down expressed. The results are given in Table 3.

Table 1. Normalization results for the gene expression levels

Tissue	Statistics	Before	QN	RSN
tumor tissue	Sample min.	6,3517	6,3884	6,5971
	1st Quartile	6,9229	6,9719	7,1066
	Median	7,6698	7,7381	7,7613
	2nd Quartile	9,4721	9,5295	9,3357
	Sample max.	13,2958	13,3551	12,6789
	Outliers	425	430	659
normal tissue	Sample min.	6,3624	6,4057	6,6123
	1st Quartile	6,9220	6,9618	7,0968
	Median	7,6770	7,7213	7,7439
	2nd Quartile	9,4879	9,5542	9,3498
	Sample max.	13,3289	13,4410	12,7262
	Outliers	460	417	676

Table 2. Removing homogenous gene expressions

Filter	QN	RSN
Low entropy	22073	22073

Table 3. Finding significant marker genes

Norm. / Methods	T-test		FDR		Volcano Plot		
	up	down	up	down	up	down	**sum**
QN	3515	2865	3108	2598	50	165	**215**
RSN	3729	2968	3410	2736	41	151	**192**

Once we discovered marker genes that discriminate carcinogenic from healthy tissue, we used them to make a model according to which we can diagnose the patients'

health condition. Since we have the a priori knowledge such as the gene expression levels and the two possible health conditions, we used few supervised learning methods in order to choose the one with best performance.

First, we used generative approach - modeling the prior distributions by ourselves. We modeled the prior distributions (Fig.1 and Fig.2) and used them in the Bayes' theorem to calculate the posterior probability. Thus, we maintained very high correct rate, especially for the carcinogenic samples, which is very important in the diagnosing process (Table 4).

Table 5 represents the results obtained from the SVM classification. When training the classifier we used three types of kernels. We used hold-out cross-validation technique which involved ⅔ of the samples in the training set and ⅓ in the testing set. In addition, we used bootstrapping method, but it gave very poor results. The SVM method produced good results, but they vary in every subsequent trial depending on the chosen training set.

Furthermore, we used Binary Decision Trees because of their simplicity and the different approach of discriminant calculation. The results in Table 6 show that it correctly classifies healthy tissues.

According to the overall results, the Bayes' theorem is the most accurate classification method in the problem of classifying colorectal carcinogenic tissue.

Table 4. Bayesian posterior probability classification

Bayes' theorem	**Cancer**		**Healthy**		**Total**	
	all	test	all	test	all	test
QN	100%	100%	92.30%	83.33%	96.15%	91.67%
RSN	96.15%	100%	96.15%	100%	96.15%	100%

Table 5. Support Vector Machines classification

SVM results	**Linear kernel**		**Quadratic kernel**		**GRB**	
	cancer	healthy	cancer	healthy	cancer	healthy
QN	100%	87.5%	75%	87.5%	0%	100%
RSN	87.5%	100%	87.5%	100%	100%	25%
	total		total		total	
QN	93.75%		81.25%		50%	
RSN	93.75%		93.75%		62.5%	
Bootstrapping	cancer	healthy	cancer	healthy	cancer	healthy
QN	20%	35%	10%	90%	70%	30%
RSN	20%	45%	45%	70%	45%	35%

Table 6. Binary Decision Tree classification

BDT results	Cancer	Healthy	Total
QN	75%	100%	87.5%
RSN	87.5%	100%	93.75%

The ability of the test to correctly classify positive and negative samples is measured as sensitivity and specificity respectively. Sensitivity refers to the true positive rate; whereas specificity takes into consideration the true negative rate. This analysis also indicates that the Bayes' theorem is the most suitable classifier in this case.

Table 7. Classifiers' Sensitivity and Specificity

Results	Bayes' theorem	Linear kernel	Quadratic kernel	GRB	BDT
Sensitivity	1	1	0.8750	1	0.75
Specificity	0.9231	0.8750	0.8750	0	1

4 Summary and Conclusions

As we are well introduced with the incidence and mortality caused by the colorectal cancer worldwide, we used DNA microarray data to observe its gene expression behavior. We assumed that the responsibility for its occurrence lies in the disrupted gene expression levels, and therefore, we performed different statistical tests to unveil those genes. Those tests discovered approximately 200 marker genes that discriminate carcinogenic from healthy tissue, which we used to build an accurate diagnostic system. Histogram representation confirmed different gene expression pattern at carcinogenic and healthy tissues distinctively. Subsequently, we used few different classification methods in order to choose the most accurate one. The best results were achieved using the Bayes' theorem - we obtained over 90% correctness when classifying the tissues. We can conclude that the reason the Bayes learning model was most accurate for this problem is in the realistic modeling of the a priori probability.

The results from this paper can be used for future research in upgrading the model in order to obtain even more accurate diagnostic system. Furthermore, the unveiled significant marker genes can be used in building ontology which can be very useful in developing new pharmaceutical molecules.

References

1. GLOBOCAN 2008, http://globocan.iarc.fr/factsheets/cancers/colorectal.asp
2. Weisenberger, D.J., Van Den Berg, D., Laird, P.W., Hinoue, T.: Gene Expression Analysis of Colorectal Tumors and Matched Adjacent Non-Tumor Colorectal Tissues. In: EMBL-EBI, ArrayExpress, Experiment: E-GEOD-25070 (2011)

3. LaPointe, Lawrence C.: Gene Expressions Biomarkers for Colorectal Neoplasia. Flinders University of South Australia, School of Medicine, Dept. of Medicine, http://theses.flinders.edu.au/public/adt-SFU20091011.090028/index.html (2008)
4. Hinoue, T., Weisenberger, D.J., Lange, C.P.E., Shen, H., Byun, H.M., Van Den Berg, D., Malik, S., Pan, F., Noushmehr, H., Van Dijk, C.M., Tollenaar, R.A.E.M., Laird, P.W.: Genome-scale Analysis of Aberrant DNA Methylation in Colorectal Cancer. In: Genome Res. February 2012 22, pp. 271-282 (2011)
5. Li, B.Q., Huang, T., Liu, L., Cai, Y.D., Chou, K.C.: Identification of Colorectal Cancer Related Genes with mRMR and Shortest Path in Protein-Protein Interaction Network. In: PLoS ONE, 7, e33393, doi:10.1371/journal.pone.0033393 (2012)
6. Muro, S., Takemasa, I., Oba, S., Matoba, R., Ueno, N., Maruyama, C., Yamashita, R., Sekimoto, M., Yamamoto, H., Nakamori, S., Monden, M., Ishii, S., Kato, K.: Identification of Expressed Genes Linked to Malignancy of Human Colorectal Carcinoma by Parametric Clustering of Quantitative Expression Data. In: Genome Biol., 4:R21 (2003)
7. Wu, Z., Aryee, M.J.: Subset Quantile Normalization Using Negative Control Features. In: Journal of Computational Biology, vol. 17, nu. 10, pp. 1385-1395 (2010)
8. Du, P., Feng, G., Kibbe, W.A., Lin, S.: Using Lumi, a Package Processing Illumina Microarray (2012)
9. Du, P., Lin, S.: Towards an Optimized Illumina Microarray Data Analysis Pipeline. Midwest Symposium on Computational Biology & Bioinformatics (2007)
10. Kohane, I.S., Kho, A.T., Butte, A.J.: Microarrays for an Integrative Genomics. MIT (2003)
11. Butte, A.J., Kohane, I.S.: Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements. Pacific Symposium on Biocomputing 5, pp. 415-426 (2000)
12. Needham, C.J., Manfield, I.W., Bulpitt, A.J., Gilmartin, P.M., Westhead, D.R.: From Gene Expression to Gene Regulatory Networks in Arabidopsis Thaliana. BMC Systems Biology 3:85 (2009)
13. Yu, H., Tu, K., Xie, L., Li, Y.Y.: Digout: Viewing Differential Expression Genes as Outliers. In: Journal of Bioinform. and Comput. Biol., vol. 8, suppl. 1, pp. 161-175 (2010)
14. Storey, J.D., Tibshirani, R.: Statistical Significance for Genomewide Studies. In: Proceedings of the National Academy of Sciences of the United States of America, vol. 100, no. 16, pp. 9440-9445 (2003)
15. Tarca, A.L., Romero, R., Draghici, S.: Analysis of Microarray Experiments of Gene Expression Profiling. In: American Journal of Obstetrics and Gynecology, 195(2), pp. 373-88 (2006)
16. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
17. Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M.Jr., Haussler, D.: Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. In: Proceedings of the National Academy of Sciences of the United States of America, 97(1), pp. 262-7 (2000)
18. Refaeilzadeh, P., Tang, L., Liu, H.: Cross-Validation. In: Enc. of Database Systems (2009)
19. Alpaydin, E.: Introduction to Machine Learning. MIT Press, Cambridge, MA (2010)
20. Bogdanova, A.M., Ackovska, N.: New Support Vector Machines-Based Approach over DNA Chip Data. In: Innovations in Information Technology, IEEE, pp. 16-19, Al Ain, UAE, 978-1-4244-3397-1/08, Dec. 16-18 (2008)
21. Bogdanova, A.M.: DNA Chips in Bioinformatics. In: Computational Intelligence and Information Technologies, CIIT'07, Molika, Macedonia, Jan. 21-25 (2007)
22. Bogdanova, A.M., Ackovska, N.: Data Driven Intelligent Systems. In: ICT Innovations, Proceedings ISSN 1857-7288 (2010)