Bayesian Posterior Probability Classification of Colorectal Cancer Probed with Affymetrix Microarray Technology

Monika Simjanoska, Ana Madevska Bogdanova and Zaneta Popeska
Ss. Cyril and Methodius University,
Faculty of Information Sciences and Computer Engineering,
Skopje, Macedonia
Email: m.simjanoska@gmail.com, {ana.madevska.bogdanova, zaneta.popeska}@finki.ukim.mk

***Abstract*—Colorectal cancer is one of the most common types of cancer worldwide. Assuming increased or decreased gene expression is the reason for abnormal cells work and processes interference in the colorectal region, in our previous work we used data from Illumina microarray technology to analyse gene expression values. Once we have unveiled biomarker genes and developed methodology for Bayesian posterior probability classification, we proceeded with implementing the same methodology on data obtained from Affymetrix microarray technology. However, our research results showed that different microarray technologies require different statistical approach for classification analyses. In this paper we use colorectal data probed with Affymetrix microarray technology, and propose a new methodology that intends to eliminate the noise and produce more robust preprocessed data appropriate for prior distribution modelling. This allows us to construct an efficient Bayesian a posteriori classificator. In order to test the procedure reliability we used different set of carcinogenic and healthy patients.**

***Index Terms*—Colorectal Cancer, Bayesian Classification, Affymetrix, Illumina, Microarray Technology, Machine Learning**

I. Introduction

The World Health Organization provided a research of the cancer incidence, mortality and prevalence worldwide in 2008. According to the GLOBOCAN project results, colorectal cancer is the third most common cancer in men, and the second in women with a total incidence of 1,234,000 cases, of which, 608,000 deaths. The mortality results make this type of cancer to be the fourth most common cause of death from cancer. Prevalence results showed that almost 60% of the cases occur in developed regions [1].

In this paper we consider the colorectal cancer problem as tightly connected to the gene expression phenomena. We believe that the reason for its occurrence lies in the increased or decreased level of expression of particular genes which intent to disrupt the biological processes they are associated with. Gene expression profiling by microarrays should advance the progress of personalized cancer treatment based on the molecular classification of subtypes [2]. Therefore, in our previous work [3] we showed that using the Illumina microarray technology data in combination with appropriate statistical analysis can lead to unveiling a set of particular genes, referred to as biomarkers, that can be used in building an accurate diagnostic system based on Bayesian a posteriori probability computation.

Assuming platform independence as presented in [4], [5], we proceeded our research in [6] using the same methodology as in [3] on colorectal data probed with the Affymetrix Human Genome U133 Plus 2.0 Array. However, the results showed poor distinctive capability of the biomarker genes. In this paper, we embraced another conclusion derived in [7] which states that each platform requires different statistical treatment.

Hence, we confronted the challenge of inventing a new methodology that leads to unveiling the Affymetrix colorectal cancer biomarkers, eliminate the noise and produce more robust data appropriate for prior distribution modelling that can be used, and therefore, is appropriate for as a part of Bayesian a posteriori classification. In order to obtain reliable results, we used data from patients with different geographical distribution, probed with Affymetrix Human Genome U133 Plus 2.0 Array. Data sets are retrieved from the Gene Expression Omnibus biological database, [8], [9], [10], [11].

The rest of the paper is organized as follows. In Section II we give a brief preview of the related work and the latest results. The methodology used in this paper is described in Section III. In Section IV we exhibit the experiments and the results according to which we derive conclusion in Section V.

II. Related Work

In this section we present some of the research related to the problem of statistical preprocessing methods and classification.

Considering the preprocessing is crucial in pointing out statistical and biological significance, the authors in [12] present a comprehensive study of the effect that normalization, gene selection, the number of selected genes and machine learning method have on the predictive performance of resulting models. The best machine learning methods in this study were Support Vector Machines with the three basic kernel configurations in comparison to Artificial Neural Networks and Decision Trees.

The authors in [13] state that at the moment there is no commonly agreed gold standard pre-processing method and

each researcher has the responsibility to choose one method, incurring the risk of false positive and false negative features arising from the particular method chosen. Therefore, they aim at providing a method of analysis of a gene-expression experiment that combines and synthesises the information from several pre-processing methods, to obtain a better calibrated estimate of differential expression. We agree with the necessity of combining several methods to confirm different gene expression, but we additionally evaluate the efficiency of the methods by including the differently expressed genes in the process of patients classification.

Since we are interested in Affymetrix microarray experiment noise, we present two studies of this kind. The authors in [14] assume that a typical microarray experiment has many sources of variation which can be attributed to biological and technical causes. Their analysis showed that the greatest source of variation at Affymetrix is biological variation, and the variation due to labelling. Similarly, a research on the of systematic noise in Affymetrix and Illumina gene-expression microarray experiments is presented in [15]. The authors suggest that it is not recommended to analyse individual test samples (e.g. to try and classify), but instead to run several experiments at the same time to get a better estimate of the experimental variation.

Gene expression data sets used in this research have also been used in other scientific researches. The authors in [16] 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 [17] where they characterize the molecular processes underlying the transformation of normal colonic epithelium. One of the data sets has been used in [18] 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 [19] where the scientists showed cross-study consistency of MSI-associated gene expression changes in colorectal cancers.

III. The Methodology

In this section we present the original methodology which includes several steps to prepare the data for the classification process.

A. Preprocessing

We used gene expression profiling of 32 colorectal tumors and matched adjacent 32 non-tumor (healthy) colorectal tissues probed with Affymetrix Human Genome U133 Plus 2.0 Array, which contains 54675 probes, but the unique genes observed are 21050. All the statistical analysis presented below aim to reduce the number of genes in order to distinguish the genes with most biological information, the colorectal cancer biomarkers.

1) Normalization: Since our aim is to unveil the difference in gene expression levels between the carcinogenic and healthy tissues, choosing the appropriate normalization method is essential. Our assumption states that only a small set of genes are differently expressed compared to the biomarker genes, i.e., most of the genes are not correlated to the colorectal cancer. In such cases Quantile normalization (QN) is a suitable normalization method, because it makes the distribution of the gene expressions as similar as possible among each other across all samples [20].

2) 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 [21]. This can lead to incorrect conclusion about gene behaviour. To remove the genes of this kind, 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 [22], while low entropy for a gene means that there is a low variability in its expression levels across the samples [23]. Therefore, we used low entropy filter to remove the genes with almost ordered expression levels.

3) Paired-sample T-test: Knowing the facts that both the carcinogenic and healthy tissues are taken from the same patients, and that the whole-genome gene expression follows normal distribution [24], we used a paired-sample t-test. Assuming that the most of the genes do not have different expressions, the null hypothesis states that there is no statistical difference between the carcinogenic and the healthy 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.

4) False Discovery Rate: False Discovery Rate (FDR) is a reduction method that usually follows the t-test. FDR solves the problem of false positives, i.e., the genes which are considered statistically significant when in reality there is not any difference in their expression levels. 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 FDR is measured as a q-value. It is described as a proportion of significant genes that turn out to be false positives [25].

5) Volcano Plot: Both the t-test and the FDR method identify different expressions in accordance with statistical significance values, and do not consider biological significance. The biological significance is measured as a fold change [26] 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. In order to display both statistically and biologically significant genes we used volcano plot visual tool. The genes that lie in the area cut off by the horizontal threshold, i.e. the p-value of 0.01 which implicates statistical significance, and the vertical thresholds, i.e. the fold change of 1.68 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. When plotting the genes a better transformation procedure is to take the logarithm base 2 value of the expression. This has the major advantage that it treats different up-regulation and down-regulation equally, and also has a continuous mapping space [27].

B. Modelling the a Priori Distribution

Once we have unveiled biomarker genes that discriminate carcinogenic tissue from the healthy one, we proceeded with another series of methods to model the a priori distribution of these biomarkers. In our previous research [6] we showed that the distribution of biomarker genes highly overlaps at both tissues. Therefore, in this paper we used the modified methods in III-A, and obtained another set of biomarker genes that showed different statistical distribution, even visually. However, few more steps are needed to generate more reliable prior distributions.

1) Round-up threshold method: When observing gene expression values, we notice that a large percentage of the data set values are negative. The authors in [28] explain this phenomena within a few processing steps. At the beginning, the background signal is estimated and then it is subtracted from all expression values. The algorithms used in Affymetrix, calculate gene expression values by comparing the signals obtained with perfect-match and one-base-mismatch hybridization oligonucleotides on the microarrays. Sometimes the one-base-mismatch oligonucleotides hybridize to other mRNAs, which does not always gives a good representation for non-specific background signal, thus, negative expression values for genes can result. Negative values might also be results because of noise in the data. One way to remove these genes is to transform all gene expression values below some threshold cut-off value to that threshold value [29]. This method is known as Round-up threshold method. In this paper we used 0 as threshold cut-off value, and any expression value below this threshold is mapped into the interval [0,2]. In order to sustain the prior distribution shape, we want to avoid eventual gene accumulation at one point. Thus, we chose a whole interval instead of particular value.

2) Appropriate tissue selection: Instead of using the cross-validation method for determining the training and the testing set as we did in our previous work [3], [6], here we propose different approach for choosing the training set. Considering the visual representation of the distribution of both tissues, we realized that the data have differently skewed distributions. Therefore, in order to eliminate the possibility of randomly picking up the tissues whose distributions overlap at some genes, we choose the training set according to the skewness factor. If the skewness factors are with opposite signs, then these tissues are involved into the training process.

3) Hypothesis testing: Our generative model fits four types of distributions: Normal, Lognormal, Gamma, and Extreme Value. At first, the two sets of tissues are confirmed to be differently distributed using the Kolmogorov-Smirnov test. After we have fulfilled our assumption for different distribution, we proceed with appropriate hypothesis testing for all the distribution mentioned above. The distribution parameters are estimated using the Maximum Likelihood Estimation (MLE) method, with a confidence level of $\alpha = 0.01$. Then we perform the Chi-square goodness-of-fit test of the default null hypothesis that the data in the tissue (vector) comes from the particular distribution with mean and variance estimated from the MLE method, again with significance level of 0.01. Once we have obtained the probabilities from the testing for each gene distinctively, we choose the distribution whose probability is the highest and we assign it to the particular gene.

Once we modelled the prior distributions, we can use them to calculate the Bayesian a posteriori probability and correctly classify the tissues.

C. Bayesian a Posteriori Classification

Using all the methods explained above, we eventually reached our purpose of building an accurate Bayesian classifier based on posterior probability computation. Modelling the prior distributions, we are now able to compute the class-conditional densities, $p(\vec{x}|C_i)$, which we calculate as the product of the continuous probability distributions of each gene distinctively:

$$p(\vec{x}|C_i) = \prod f_1 f_2 ... f_n \tag{1}$$

For the prior probabilities $P(C_i)$, we defined two test cases:

- Test Case 1: Since we have equal number of tissues into both of the classes, the prior probabilities are also equal $P(C_1) = P(C_2) = 0.5$;
- Test Case 2: The prior probabilities are estimated according to the statistics in [1], and $P(C_1) = 0.0002$ and $P(C_2) = 0.9998$, where C_1 denotes carcinogenic class, and C_2 denotes healthy class.

Therefore, we calculate the posterior probability $P(C_i|\vec{x})$, as:

$$p(C_i|\vec{x}) = \frac{p(\vec{x}|C_i) * P(C_i)}{\sum_{1}^{2} p(\vec{x}|C_i) * P(C_i)} \tag{2}$$

The tissue $\vec{x}$ is classified according to the rule of maximizing the a posteriori probability (MAP):

$$C_i = \max p(C_i|\vec{x}) \tag{3}$$

IV. Experiments and Results

In this section we present the experiments and the obtained results.

Our main goal is to construct an efficient Bayesian classifier that can distinguish healthy from carcinogenic tissue when gene expression levels from Affymetrix DNA chip are used. For our research purpose we retrieved data sets from GEO functional genomics data repository. The 32 colorectal tumors and matched adjacent 32 non-tumor colorectal tissues have

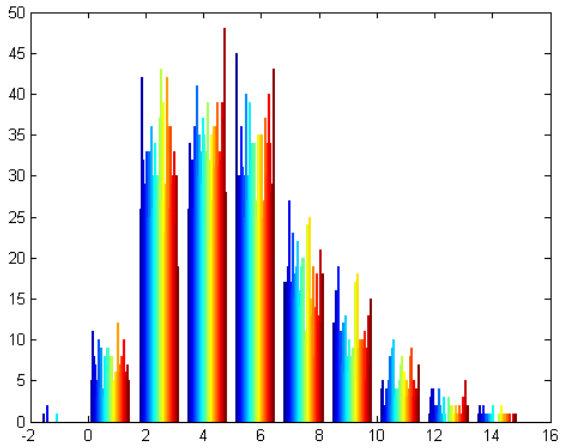

Fig. 1. Biomarkers distribution at carcinogenic tissues

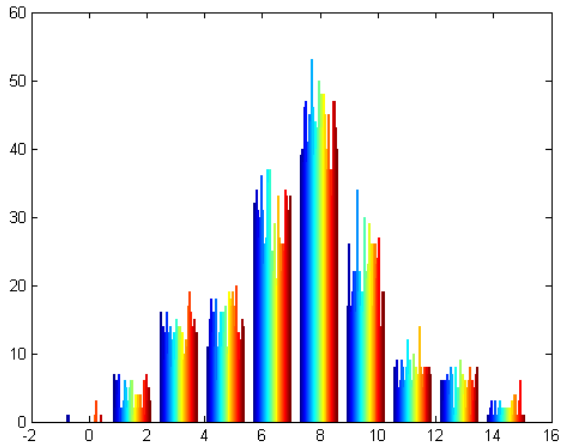

Fig. 2. Biomarkers distribution at healthy tissues

been probed with Affymetrix Human Genome U133 Plus 2.0 Array. This type of technology allows 54675 probes.

We intent to use as biomarkers only the genes which are considered to be statistically and biologically significant. The developed methodology presented in III-A will preprocess the data in order to make the best model of the prior distributions.

After normalizing the gene expression values, we used the low-entropy filter which reduced the number of genes to 49,206. However, this number of gene is still inappropriate to model the carcinogenic and the healthy distribution. Furthermore, we proceed with the paired-sample t-test, the FDR method, and eventually we applied the volcano plot. The results are given in Table I. The final number of revealed biomarkers is 138.

TABLE I
BIOMARKERS REVEALING

Statistical methods	Up expression	Down expression	Sum
Paired-sample t-test	7630	10691	18321
FDR	7581	10539	18120
Volcano plot	29	109	138

The number of genes left to be biomarkers is appropriate for modelling the prior distribution. Using a histogram tool we presented the frequency of the genes within particular range for each of the carcinogenic and healthy tissues distinctively. Figure 1 and Figure 2 depict the prior distributions of the carcinogenic and healthy tissues, respectively. Obviously, the distributions differ from each other. Before we proceed with hypothesis testing, we must cut-off the negative values, since we assume they occurred as a result of a noise.

After the process of rounding up the negative values using a predefined threshold value, both of the distributions depicted in Figure 3 and Figure 4, suffered minor reshape.

Observing the two distributions, we can easily assume that both of them are with opposite skewness. Therefore, to avoid

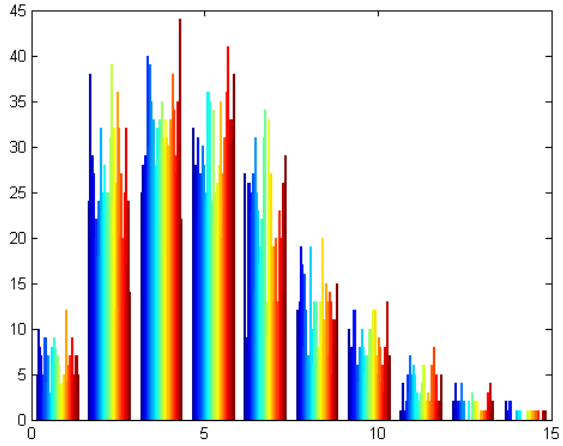

Fig. 3. Biomarkers distribution at carcinogenic tissues after Round-up

the possible selection of tissues with the same skewness by coincidence, we used the method of appropriate tissues selection. This method outcome determined 15 appropriate tissues to be involved in the training process. The distribution of those tissues is depicted in Figure 5 and Figure 6, respectively.

The testing of the biomarkers extracted from the carcinogenic and healthy tissues respectively for equality in their probability distribution, is done by the Kolmogorov-Smirnov test. It confirmed that 135 out of 138 has declined the null hypothesis ”belong to the same distribution”. Hereupon, we performed generative probability distribution fitting by using appropriate hypothesis testing over the Normal, Lognormal, Gamma and Extreme Value distribution. Once we assigned the distribution which showed to be the most probable to each gene, we proceeded to calculate the Bayesian posterior probability as defined in III-C.

Moreover, to confirm the procedure reliability, we addition-

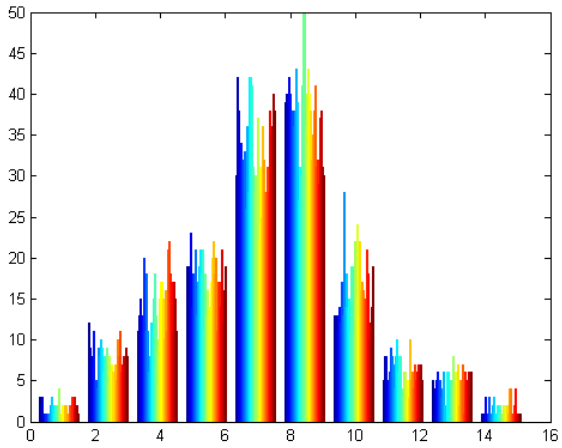

Fig. 4. Biomarkers distribution at healthy tissues after Round-up

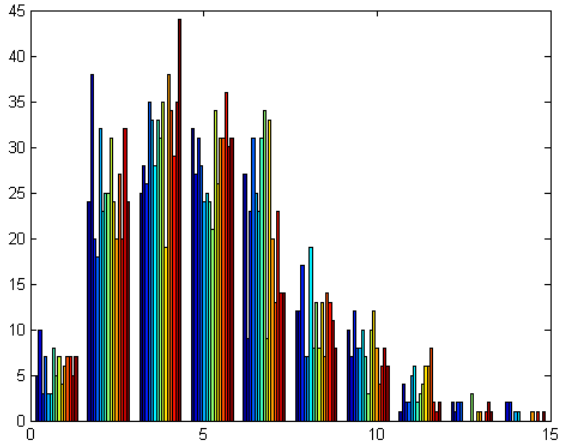

Fig. 5. Biomarkers distribution of selected tissues at carcinogenic case

ally tested our trained classificator with new 239 carcinogenic patients, as well as, with new 12 healthy patients. In order to avoid any negative values, the new patients are also processed with the Round-up threshold method.

The prior probabilities we used are those we defined earlier as the two test cases in III-C. The results presented in Table II are derived according to all three equations specified in III-C.

TABLE II
BAYESIAN A POSTERIORI CLASSIFICATION

Bayesian classification	Test Case 1	Test Case 2
32 carcinogenic tissues	100%	93.75%
32 healthy tissues	84.38%	100%
239 carcinogenic patients	97.5%	89.5%
12 healthy patients	91 %	100%

We evaluated the classificator performance through relative trade-off between the true positives and the false positives.

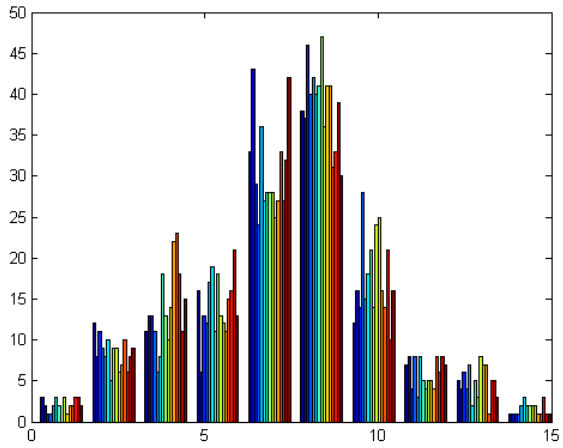

Fig. 6. Biomarkers distribution of selected tissues at healthy case

True positive rate (TPR), known as recall or sensitivity, is a term that refers to the ability of the classificator to correctly classify carcinogenic tissues. The ability of the classificator to correctly classify healthy tissues is measured as specificity. The results given in Table III show the ability of the classificator to correctly classify carcinogenic and healthy conditions with high accuracy at both tissues and patients which we used as a test sets.

TABLE III
SENSITIVITY AND SPECIFICITY

Performance	Sensitivity	Specificity	Test Cases
Tissues	1	0.84	Test case 1
	0.94	1	Test case 2
Patients	0.98	0.91	Test case 1
	0.90	1	Test case 2

V. CONCLUSION

In this paper we showed how to construct an efficient Bayesian classifier that distinguishes healthy from carcinogenic tissue when gene expression levels from Affymetrix DNA chip are used.

In order to make the best model of the prior distributions, which is essential for proper Bayesian classification, we developed an original methodology presented in III-A that preprocess the input data.

We focused on the genes whose expression clearly differs in the carcinogenic opposed to the healthy tissue. As we discovered a set of biomarkers, in the next few steps we provided series of statistical analysis in order to produce a well shaped distribution for both classes of tissues. Once we applied our generative approach, we classified the tissues calculating the Bayesian a posteriori probability.

This research follows two previous papers where we examined colorectal cancer using gene expression values obtained

from Illumina microarray BeadChip [3] and performed platform comparison between Illumina and Affymetrix [6]. Comparing the two platforms we realized that gene expressions obtained from Affymetrix require different statistical approach.

The obtained results in this paper, confirmed that the generative Bayesian approach gives excellent performance when appropriate preprocessing methodology is used. In our future research we will focus on distinguishing the different colorectal cancer stages in the carcinogenic tissues.

REFERENCES

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

[2] K. Jain, "Applications of biochips: From diagnostics to personalized medicine." *Current opinion in drug discovery & development*, vol. 7, no. 3, pp. 285–289, 2004.

[3] M. Simjanoska, A. M. Bogdanova, and Z. Popeska, "Recognition of colorectal carcinogenic tissue with gene expression analysis using bayesian probability," in *to be published in ICT Innovations 2012*. Springer Berlin / Heidelberg, 2012.

[4] 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.

[5] 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.

[6] A. M. Bogdanova, M. Simjanoska, and Z. Popeska, "Classification of colorectal carcinogenic tissue with different dna chip technologies," Tech. Rep., 2013.

[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] G. Marra, J. Sabates, and H. Rehrauer, "Transcriptome profile of human colorectal adenomas," 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8671

[9] Y. Hong, "Expression data from healthy controls and early stage crc patient's tumor," 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9348

[10] T. Orntoft and C. Andersen, "Expression data from primary colorectal cancers," 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13294

[11] T. Watanabe, T. Kobunai, E. Toda, Y. Okayama, Y. Sugimoto, and T. Oka, "Gene expression signature of colorectal cancer with microsatellite instability," 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4554

[12] J. Önskog, E. Freyhult, M. Landfors, P. Rydén, and T. Hvidsten, "Classification of microarrays; synergistic effects between normalization, gene selection and machine learning," *BMC bioinformatics*, vol. 12, no. 1, p. 390, 2011.

[13] M. Blangiardo and S. Richardson, "A bayesian calibration model for combining different pre-processing methods in affymetrix chips," *BMC bioinformatics*, vol. 9, no. 1, p. 512, 2008.

[14] S. Zakharkin, K. Kim, T. Mehta, L. Chen, S. Barnes, K. Scheirer, R. Parrish, D. Allison, and G. Page, "Sources of variation in affymetrix microarray experiments," *BMC bioinformatics*, vol. 6, no. 1, p. 214, 2005.

[15] 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.

[16] Y. Hong, T. Downey, K. Eu, P. Koh, and P. Cheah, "A metastasis-pronesignature 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] I. Kohane, A. Butte, and A. Kho, *Microarrays for an integrative genomics*. MIT press, 2002.

[22] 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.

[23] 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.

[24] Y. Hui, T. Kang, L. Xie, and L. Yuan-Yuan, "Digout: Viewing differential expression genes as outliers," *Journal of Bioinformatics and Computational Biology*, vol. 8, no. supp01, pp. 161–175, 2010.

[25] 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.

[26] K. HC and S. AA, "Gene expression profile analysis by dna microarrays: Promise and pitfalls," *JAMA*, vol. 286, no. 18, pp. 2280–2288, 2001.

[27] M. B. M., "An introduction to microarray data analysis." [Online]. Available: http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/chapter-final.pdf

[28] M. Weir and M. Rice, "Microarray clustering analysis." [Online]. Available: https://wesfiles.wesleyan.edu/courses/biol265/microarray_lab.htm

[29] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences*, vol. 96, no. 6, pp. 2907–2912, 1999.