Novel Methodology for CRC Biomarkers Detection with Leave-one-out Bayesian Classification

Monika Simjanoska and Ana Madevska Bogdanova

Ss. Cyril and Methodious University, Faculty of Computer Science and Engineering,
Rugjer Boshkovikj 16, 1000 Skopje, Macedonia
m.simjanoska@gmail.com, ana.madevska.bogdanova@finki.ukim.mk

Abstract. In our previous research we developed a methodology for extracting significant genes that indicate colorectal cancer (CRC). By using those biomarker genes we proposed an intelligent modelling of their gene expression distributions and used them in the Bayes' theorem in order to achieve highly precise classification of patients in one of the classes carcinogenic, or healthy. The main objective of our new research is to subside the biomarkers set without degrading the sensitivity and specificity of the classifier. We want to eliminate the biomarkers that do not play an important role in the classification process. To achieve this goal, we propose a novel approach for biomarkers detection based on iterative Bayesian classification. The new Leave-one-out method aims to extract the biomarkers essential for the classification process, i.e. if they are left-out, the classification shows remarkably degraded results. Taking into account only the reduced set of biomarkers, we produced an improved version of our Bayesian classifier when classifying new patients. Another advantage of our approach is using the new biomarkers set in the Gene Ontology (GO) analysis in order to get more precise information on the colorectal cancer's biomarkers' biological and molecular functions.

Keywords: Colorectal Cancer, Biomarkers Detection, Bayesian Classification, Gene Ontology

1 Introduction

Recently, the scientists provide intensive gene expression profiling experiments in order to compare the malignant to the healthy cells in a particular tissue. The advantage of the microarray technologies enables simultaneous observation of thousands of genes and allows the researchers to derive conclusions whether the disorder is a result of the abnormal expression of a subset of genes.

In our previous work we have used gene expression data from Affymetrix Human Genome U133 Plus 2.0 Array to perform analysis of CRC and healthy tissues [1]. During the research we developed a methodology for biomarkers detection based on the two types of tissues, carcinogenic and healthy. The obtained

set of biomarkers was then used to build a machine learning (ML) based classifier capable of distinguishing between carcinogenic and healthy patients.

Since the classification analysis resulted in very high accuracy when classifying both CRC and healthy patients, we proceeded to inspect whether the biomarkers we discovered play important biological role in the CRC development [2]. For that purpose, we provided GO analysis and inspected the molecular functions and the biological processes of a particular set of genes that showed to be overrepresented among all biomarkers. Considering the colorectal cancer significance of the biomarker genes, we confirmed few biomarkers to be tightly related to the disease: $CHGA$, $GUCA2B$, $MMP7$, $CDH3$ and PYY.

In this paper we address another issue - reducing the biomarkers set in order to improve the classification reliability and to extract new information from the Gene Ontology analysis. We developed a new methodology for subsiding the biomarkers set without degrading the sensitivity and specificity of the built classifier. We want to eliminate the biomarkers that do not play an important role in the classification process. To achieve this goal, we propose a new approach, based on iterative Bayesian classification. In order to eliminate the non-informative genes, we use a Leave-one-out method. Taking into account only the reduced set of biomarkers (the subsided gene set), we produced an improved version of our Bayesian classifier when classifying new patients.

The GO helps us to find the biological meaning of the gene data and their role in the functions connected to the CRC. The GO analysis [2] has showed the possible biological and molecular functions connected to the CRC. The novel proposed methodology, gives us an advantage in the GO analysis, because we can obtain more precise knowledge about the expressed genes in the CRC disease.

The rest of paper is organized as follows. In Section 2 we present the literature related to CRC and GO analysis. The analysis flow is presented in Section 3, whereas the GO analysis are described in Section 4. In Section 5 we discuss the experiments from both the classification and the GO analysis. In Section 6 we present the conclusion from our work and our plans for further research.

2 Related Work

In this section we give a review of the recent literature that relates to CRC and GO analysis.

Authors in [3] sum up the biomarkers results from 23 different researches on CRC and GO analyses. Even though most of them show diversity in the significant genes revealed, the authors in their research take into account the unique biomarkers, which are nearly 1000, and perform GO analysis by using few different tools. They mainly hold on to the ontology results of the enriched set of genes, rather than verifying the biomarkers with classification methods so that we can compare our results.

Similarly, in [4] the researchers use Affymetrix microarray data from 20 patients to reveal significant gene expression, which resulted in 1469 biomarkers.

From the GO results they ranked top 10 most important pathways. Comparing our results to theirs, we realized that there is no overlap between ours and their biomarkers sets. Even though they lack a classification analysis, we may include their biomarkers in our future work and test the ability of the Bayesian approach to make an appropriate modelling using different biomarkers revealing procedure.

Since the non overlapping between the biomarkers sets discovered in different scientific papers is very common, a new meta-analysis model of CRC gene expression profiling studies is proposed in [5]. As the authors ranked the biomarker genes according to various parameters, the gene CDH3 which we found to play role in the CRC [2] is also found by their meta-analysis model.

Another interesting approach maintained with classification analysis is presented in [6], where the authors constructed disease-specific gene networks and used them to identify significantly expressed genes. A particular attention is given to five biomarkers, from which one of them, IL8 was also detected by the GO enrichment analysis of our new subsided set of biomarkers. In order to test the power of the colon cancer-specific gene network biomarkers revealing ability, they use five different classifiers: Diagonal linear discriminate analysis, 3 Nearest neighbours, Nearest centroid, Support Vector Machines and Bayesian compound covariate.

3 The Methodology

In this section we present the new approach for biomarkers detection which is an extension of a previously defined methodology [1].

3.1 The Previous Methodology

The biomarkers which we use in this paper were detected by using the following methods:

- *Quantile normalization.* Since our aim is to unveil the difference in gene expression levels between the carcinogenic and healthy tissues, we proposed the Quantile normalization (QN) as a suitable normalization method [7].
- *Low entropy filter.* We used low entropy filter to remove the genes with almost ordered expression levels [8], since they lead to wrong conclusions about the genes behaviour.
- *Paired-sample t-test.* Knowing the facts that both carcinogenic and healthy tissues are taken from the same patients, and that the whole-genome gene expression follows normal distribution [9], we used a paired-sample t-test.
- *FDR method.* 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.

– *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. In order to display both statistically and biologically significant genes we used volcano plot visual tool.

3.2 The Novel Approach

After we discovered the significant genes, we proposed a generative approach by building the prior distributions of the two classes (CRC and healthy) which we used in the Bayes' Theorem to classify new patients.

Given the classes C_i for $i = 1, 2$ and a vector $\boldsymbol{x}$ of biomarkers gene expression values, we calculated the prior distributions, $p(\boldsymbol{x}|C_i)$, as a product of the continuous probability distributions of each biomarker distinctively:

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

Once we determined the class-conditional densities, we applied them in the Bayes' theorem to obtain the a posteriori probability $P(C_i|\boldsymbol{x})$:

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

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 [10]. Therefore, $P(C_1) = 0.0002$ and $P(C_2) = 0.9998$, where C_1 denotes the carcinogenic class, and C_2 denotes the healthy class.

A new set of biomarkers gene expression, $\boldsymbol{x}$, is classified according to the rule of maximizing the a posteriori probability (MAP):

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

By using this methodology we achieved very high classification accuracy whose results are presented in Section 5.

The high sensitivity and specificity results from the Bayesian classifier intrigued us to go into more detail and make the prior distributions even more precise. In order to achieve our aim, we need to reduce the number of genes whose prior distribution varies when compared to the prior distributions of the majority of the genes.

That is the origin of the idea to use the generative Bayesian model as an additional method for biomarkers detection. As described in Figure 1, the method in based on iterative leave-one-out classification until we reach a set of genes whose classification power is higher than the initial set of biomarkers. Therefore the method was applied as follows.

Let's have n number of biomarkers. Since we wanted to reduce the biomarkers set as well as to sustain the good features of Distribution models for Bayes classification process, we performed $n * \frac{3}{4}$ iterations. We chose this number of iterations to be fixed since the analysis reported in our previous research [2] showed that this number of biomarkers is approximately optimal for Bayesian classification.

In each iteration $i = 0, ..., n * \frac{3}{4}$ we perform $n - i$ retrainings and classifications by cutting-off one biomarker in each one. Once we obtain the results from all the classifications in the particular iteration, we chose the biomarker that have degraded the classification results at most, and put it into the new set of biomarkers. That biomarker is excluded from the initial set of biomarkers. The initial set of biomarkers is now reduced by one biomarker and we can proceed to the next iteration. Eventually, after $n * \frac{3}{4}$ iterations we complete the new set of biomarkers which will consist of $n * \frac{3}{4}$ biomarkers - the subsided gene set.

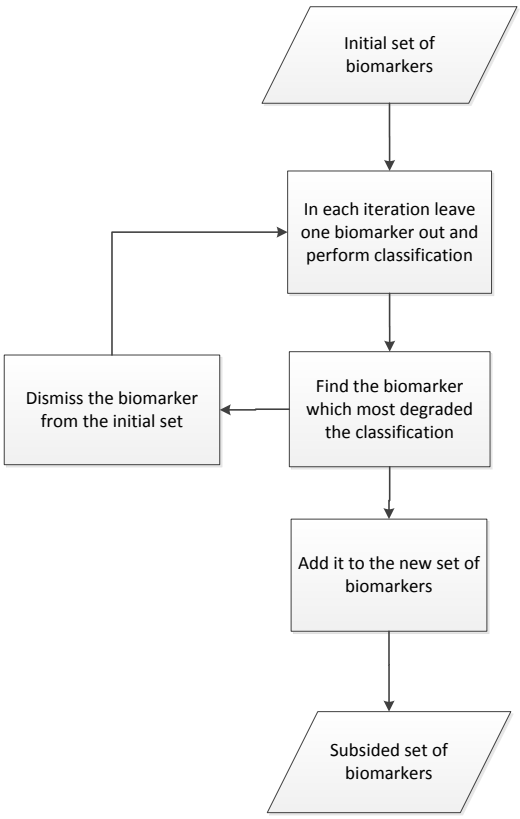

Fig. 1. The novel approach

4 GO Analysis

In the past the analyses of single markers have been in the focus of the genome-wide association studies. However, it often lacks the power to uncover the relatively small effect sizes conferred by most genetic variants. Therefore, using prior biological knowledge on gene function, pathway-based approaches have been developed with the aim to examine whether a group of related genes in the same functional pathway are jointly associated with a trait of interest [11].

The goal of the GO Consortium is to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing [12]. The GO project provides ontologies to describe attributes of gene products in three non-overlapping domains of Molecular Biology [13]:

1. **Molecular Function** describes activities, such as catalytic or binding activities, at the molecular level. GO molecular function terms represent activities rather than the entities that perform the actions, and do not specify where, when or in what context the action takes place.
2. **Biological Process** describes biological goals accomplished by one or more ordered assemblies of molecular functions.
3. **Cellular Component** describes locations, at the levels of subcellular structures and macromolecular complexes.

There are many tools based on GO resource; however, in this research we use the freely accessible Gene Ontology Enrichment Analysis Software Toolkit (GOEAST). It is a web based tool which applies appropriate statistical methods to identify significantly enriched GO terms among a given list of genes. Beside the other functions, GOEAST supports analysis of probe set IDs from Affymetrix microarrays. It provides graphical outputs of enriched GO terms to demonstrate their relationships in the three ontology categories. In order to compare GO enrichment status of multiple experiments, GOEAST supports cross comparisons to identify the correlations and differences among them [14].

We use cross comparisons of the old and new GO analyses to derive conclusions of the three Molecular Biology domains acquired from the subsided set of biomarkers.

5 Experiments and Results

In this section we present the experiments and the results obtained from the previously defined methodologies.

5.1 Microarray Data Analysis

In order to extract significant colorectal cancer genes we used gene expression profiling of 32 colorectal tumors, adenomas, and matched adjacent 32 non-tumor

Table 1. Old Sensitivity and Specificity

Chip	Performance	Sensitivity	Specificity	Test Cases
Affymetrix	Tissues	1	0.84	Test case 1
		0.94	1	Test case 2
	Patients	0.98	0.92	Test case 1
		0.90	1	Test case 2

Table 2. New classification results

Results	Test Case 1	Test Case 2
32 CRC Tissues	100%	96.87%
32 Healthy Tissues	81.25%	100%
239 CRC Patients	97.90%	94.14%
12 Healthy Patients	100%	100%

Table 3. New Sensitivity and specificity

Chip	Performance	Sensitivity	Specificity	Test Cases
Affymetrix	Tissues	1	0.81	Test case 1
		0.97	1	Test case 2
	Patients	0.98	1	Test case 1
		0.94	1	Test case 2

colorectal tissues probed with Affymetrix Human Genome U133 Plus 2.0 Array. It contains 54,675 probes, but the unique genes observed are 21,050.

The gene expression values were preprocessed according to the methodology described in Section 3.1. The methods produced a set of 138 biomarkers. The prior distributions of the biomarkers were modelled and a generative Bayesian classifier was produced whose results are reported in Table 1. In order to test the classifier on completely new patients, we used additional gene expression values from 239 CRC and 12 healthy patients.

The new methodology from Section 3.2 reduced the set of genes by retaining the most important ones - the subsided biomarkers set. In Table 2 we present the results from the classification procedure where we performed classification analysis of 64 tissues from 32 patients, and additional 251 patients that weren't involved in the training process.

We evaluated the classificator performance through relative trade-off between the true positives and the false positives. True positive rate (TPR), the sensitivity, refers to the classifier's ability to correctly classify CRC tissues, whereas the ability of the classifier to correctly classify healthy tissues is measured in terms of specificity. The results from the new approach that improved the sensitivity and the specificity of the classifier are presented in Table 3.

Figure 2 depicts the comparison of the new sensitivity (green) and specificity (violet) results with the old sensitivity (blue) and specificity (red) results.

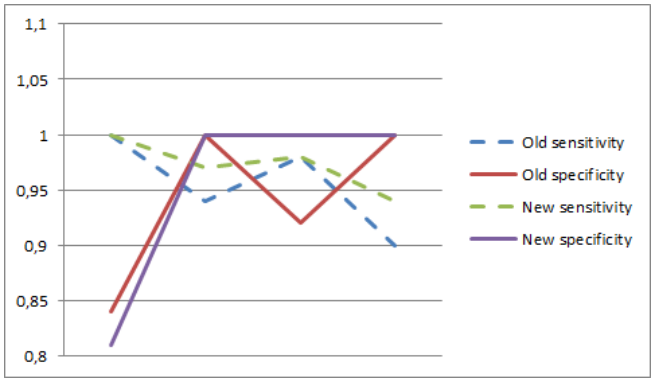

Fig. 2. Performance comparison

5.2 GO Results

In order to compare the GO results from the analysis of subsided set of biomarkers, we performed comparisons with the GO analysis of the old set of 138 biomarkers.

Biological Processes (BP). Figures 3 and 4 present the comparison of the biological processes (BP) of the two sets of biomarkers. Even though the subsided set of 100 genes is a subset of the old set of 138 genes, some of the processes that were not enriched in the previous analysis, now show significant enrichment. The results from the old analysis are marked with red, whereas the results from the new analysis are marked with green. All the common enriched terms are labelled with yellow. The newly enriched BP and their GO descriptions are as follows:

- **Negative regulation of cell proliferation** - Any process that stops, prevents or reduces the rate or extent of cell proliferation.
  Genes: SST, MSX2, CCL23, FABP6, IL8, SCG2.
- **Transmembrane receptor protein serine/threonine kinase signaling pathway** - A series of molecular signals initiated by the binding of an extracellular ligand to a receptor on the surface of the target cell where the receptor possesses serine/threonine kinase activity, and ending with regulation of a downstream cellular process, e.g. transcription.
  Genes: GREM2, MSX2, CHRDL1.
- **Indole-containing compound biosynthetic process** - The chemical reactions and pathways resulting in the formation of compounds that contain an indole (2,3-benzopyrrole) skeleton.
  Genes: TPH1

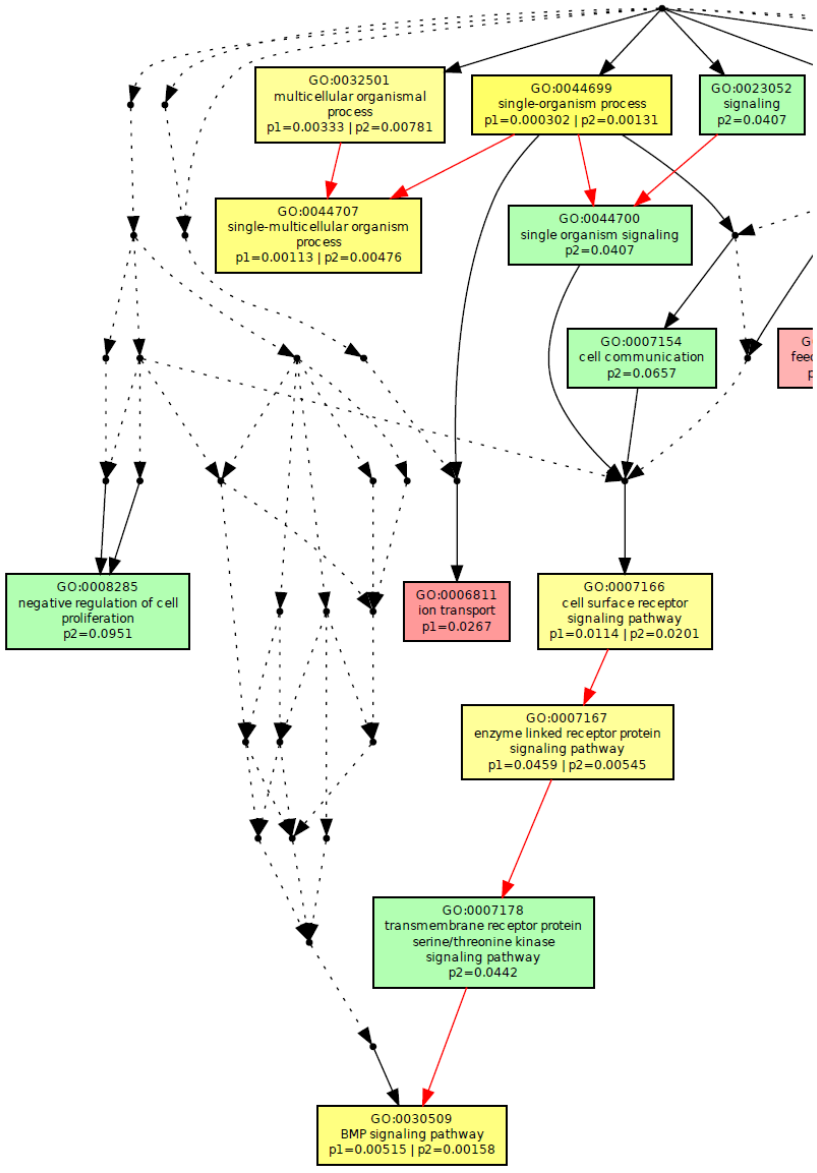

Fig. 3. Biological Processes Comparison (part 1)

Molecular Functions (MF). Considering the comparison of the molecular functions, we see the following enriched functions are no longer present in the new analysis:

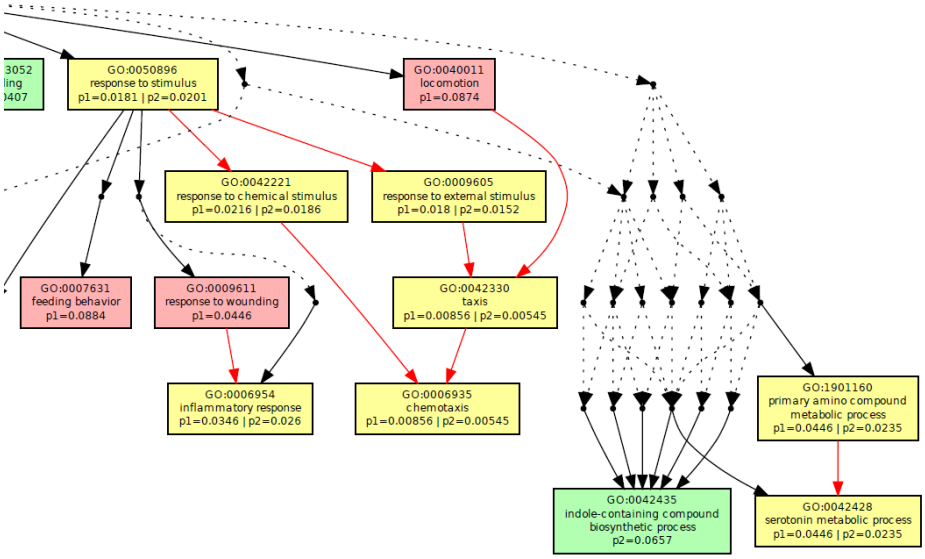

Fig. 4. Biological Processes Comparison (part 2)

- **G-protein coupled receptor binding** - Interacting selectively and non-covalently with a G-protein coupled receptor.
- **Hormone activity** - The action characteristic of a hormone, any substance formed in very small amounts in one specialized organ or group of cells and carried (sometimes in the bloodstream) to another organ or group of cells in the same organism, upon which it has a specific regulatory action.
- **Alcohol dehydrogenase activity, zinc-dependent** - Catalysis of the reaction: an alcohol + NAD+ = an aldehyde or ketone + NADH + H+, requiring the presence of zinc.
- **Sodium channel activity** - Catalysis of facilitated diffusion of a sodium ion (by an energy-independent process) involving passage through a transmembrane aqueous pore or channel without evidence for a carrier-mediated mechanism.

Cellular Components (CC). Eventually, we compared the cellular components results and we found the following enriched terms are excluded in the new results:

- **Apical part of cell** - The region of a polarized cell that forms a tip or is distal to a base. For example, in a polarized epithelial cell, the apical region has an exposed surface and lies opposite to the basal lamina that separates the epithelium from other tissue.
- **Apical plasma membrane** - The region of the plasma membrane located at the apical end of the cell.

– **Sodium channel complex** - An ion channel complex through which sodium ions pass.

Further analysis of the relation of the new GO results to CRC will be presented in our future work, where we will discuss the results from a biological point of view.

6 Conclusion

The aim of this paper was to enforce the classification system created for CRC diagnosis - the Bayes classification process that uses the chosen biomarker set [1]. We used the built generative Bayesian model as an additional method for meaningful reduction of the biomarkers set. We addressed this issue by choosing the biomarkers that contribute to the classification process the most. To achieve this goal, we proposed a new approach, based on iterative Bayesian classification. In order to eliminate the non-informative genes, we used a Leave-one-out method - we picked the ones that degrade the classification process when excluded from building the classification system. Taking into account only the reduced set of biomarkers (subsided set of biomarkers), we produced an improved version of our Bayesian classifier when classifying new patients and tissues.

We also engaged the GO analysis to understand the biological processes, the molecular functions and the cellular components when using the subsided biomarkers set. We compare the GO analysis of the initial set of biomarkers [2] with the analysis from the novel proposed methodology, with the subsided biomarkers set. The novel approach gives us an advantage in the GO analysis, because we can obtain more precise knowledge about the expressed genes and processes they are connected to in the CRC diagnosis. We obtained newly enriched BP and their GO descriptions and found out about enriched functions that are no longer present in the new analysis for the CC and MF.

Future work includes the investigation if the subsided biomarkers set can improve the methodology for CRC stages diagnostics [15], and a close collaboration with the Molecular Biology experts, that will validate our results for molecular diagnostics, evaluation and prognostic purposes in patients with colorectal cancer.

References

1. Simjanoska, M., Madevska Bogdanova, A., Popeska, Z.: Bayesian posterior probability classification of colorectal cancer probed with Affymetrix microarray technology. In: Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention on. pp. 959–964. IEEE (2013)
2. Simjanoska, M., Madevska Bogdanova, A., Panov, S.: Gene ontology analysis of colorectal cancer biomarkers probed with Affymetrix and Illumina microarrays. In: Proceedings of the 5th International Joint Conference on Computational Intelligence, IJCCI, 2013. pp. 396–406. IJCCI (2013)

3. Lascorz, J., Chen, B., Hemminki, K., Försti, A.: Consensus pathways implicated in prognosis of colorectal cancer identified through systematic enrichment analysis of gene expression profiling studies. PloS one 6(4), e18867 (2011)
4. Xu, Y., Xu, Q., Yang, L., Liu, F., Ye, X., Wu, F., Ni, S., Tan, C., Cai, G., Meng, X., et al.: Gene expression analysis of peripheral blood cells reveals toll-like receptor pathway deregulation in colorectal cancer. PloS one 8(5), e62870 (2013)
5. Chan, S.K., Griffith, O.L., Tai, I.T., Jones, S.J.: Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers. Cancer Epidemiology Biomarkers & Prevention 17(3), 543–552 (2008)
6. Jiang, W., Li, X., Rao, S., Wang, L., Du, L., Li, C., Wu, C., Wang, H., Wang, Y., Yang, B.: Constructing disease-specific gene networks using pairwise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements. BMC systems biology 2(1), 72 (2008)
7. Wu, Z., Aryee, M.: Subset quantile normalization using negative control features. Journal of Computational Biology 17(10), 1385–1395 (2010)
8. Needham, C., Manfield, I., Bulpitt, A., Gilmartin, P., Westhead, D.: From gene expression to gene regulatory networks in arabidopsis thaliana. BMC systems biology 3(1), 85 (2009)
9. Hui, Y., Kang, T., Xie, L., Yuan-Yuan, L.: Digout: Viewing differential expression genes as outliers. Journal of Bioinformatics and Computational Biology 8(supp01), 161–175 (2010)
10. GLOBOCAN: (2008), `http://globocan.iarc.fr/factsheets/cancers/colorectal.asp`
11. Wang, K., Li, M., Hakonarson, H.: Analysing biological pathways in genome-wide association studies. Nature Reviews Genetics 11(12), 843–854 (2010)
12. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature genetics 25(1), 25 (2000)
13. Harris, M., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., et al.: The gene ontology (go) database and informatics resource. Nucleic acids research 32(Database issue), D258 (2004)
14. Zheng, Q., Wang, X.J.: Goeast: a web-based software toolkit for gene ontology enrichment analysis. Nucleic acids research 36(suppl 2), W358–W363 (2008)
15. Simjanoska, M., Bogdanova, A.M., Popeska, Z.: Bayesian multiclass classification of gene expression colorectal cancer stages. In: ICT Innovations 2013, pp. 177–186. Springer (2014)