Gene Ontology Analysis on Behalf of Improved Classification of Different Colorectal Cancer Stages

Monika Simjanoska, Ana Madevska Bogdanova and Sasho Panov

Abstract The colorectal cancer is a serious cause of death worldwide. Diagnosing the current colorectal cancer stage is crucial for early prognosis and adequate treatment of the patients. Even though the scientists have developed various techniques, determining the real colorectal cancer stage is still critical. In this paper we utilize Gene Ontology analysis information to address this issue. We compose a set of special genes that are used to obtain two main results—we show the distinction between the carcinogenic and healthy tissue by difference in the range of their DNA gene expressions, and we propose a novel methodology that improves the colorectal cancer stages classification.

Keywords Gene ontology · Colorectal cancer stages · Gene expression · Bayes' theorem

1 Introduction

In 2008, the World Health Organization (WHO) conducted a research on the cancer's incidence, mortality and prevalence. The results showed that the colorectal cancer (CRC) deserves serious attention since it causes approximately 608,000, which is 8 % of total cancer deaths [1]. The incidence and prevalence results showed that 60 % of the 1,234,000 new cases occur in the developed regions, from which 663,000 at man and 571,000 at women.

M. Simjanoska (✉)
Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia
e-mail: m.simjanoska@gmail.com

A.M. Bogdanova
Natural Sciences and Mathematics, Ss. Cyril and Methodius University, Skopje, Macedonia
e-mail: ana.madevska.bogdanova@finki.ukim.mk

S. Panov
Institute of Biology, Skopje, Macedonia
e-mail: sasho@mt.net.mk

K. Madani et al. (eds.), *Computational Intelligence*,
Studies in Computational Intelligence 613,
DOI 10.1007/978-3-319-23392-5_27

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 used gene expression experiments from Affymetrix Human Genome U133 Plus 2.0 Array to perform analysis of colorectal carcinogenic and healthy tissues [2]. During the research we developed 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 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 colorectal cancer development [3]. For that purpose, we provided gene ontology (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.

Consequently, since gene expression profiling by microarrays is expected to advance the progress of personalized cancer treatment based on the molecular classification of subtypes [4], we used the same set of biomarkers to model the different CRC stages (I–IV) [5]. The modelling resulted in an accurate Bayesian classifier that showed satisfying results when diagnosing tissues in the critical stages, I and IV, and, II and III, which, as presented in Sect. 2, are often found to be problematic for prognosis.

Even though, we exceeded the problems of distinguishing between CRC stage I and IV, and, II and III, that remained common problem in the literature, we decided to go deeper in the problem in order to improve our classification results. In this paper we conduct a research that follows two threads of our previous work, the GO analysis of the biomarkers [3] and the classification of the different CRC stages [5]. In this research we preform GO analysis for each of the different CRC stages probed with the same Affymetrix platform. Our aim is to compare the stages that are critical for diagnosing and also the neighbouring stages, in order to derive conclusions on their common biological and molecular functions (enriched genes). Obtaining the enriched genes involved in the common GO functions and inspecting their range of DNA expression is very important for determining the distinguishing functions between the CRC stages. Once we discovered the enriched genes, we were able to remodel the prior probabilities of the different CRC stages and we got significantly improved classification results.

The rest of the paper is organized as follows. In Sect. 2 we briefly present the latest work related to our point of interest of this paper. In Sect. 3 we describe the methods for biomarkers selection and GO analysis. The results from the analysis are presented in Sect. 4 and eventually, we derive our conclusions and present our plans for future work in the final Sect. 5.

2 Related Work

In this section we present a work related to CRC stages analysis and GO appliance in the research of various diseases.

Recently, the classification of different CRC stages has been in the focus of many researches. Even though, the authors developed many procedures for diagnosis and survival prediction [6, 7], the analysis showed that an accurate classification of intermediate-stage cases, II and III, as well as stage I and stage IV, is problematic [8, 9].

The microarray data used in this paper, has also been used for distinguishing patterns in different CRC stages.

Laibe et al. [10] profiled both stage II and stage III carcinomas. They realized that expression profile of stage II colon carcinomas distinguishes two patterns, one pattern very similar to that of stage III tumors, based on a 7-gene signature. The function of the discriminating genes suggests that tumors have been classified according to their putative response to adjuvant targeted or classic therapies. Tsukamoto et al. [11] performed gene expression profiling and found that the over expression of OPG gene may be a predictive biomarker of CRC recurrence and a target for treatment of this disease. Hong et al. [12] aimed to find a metastasis-prone signature for early stage mismatch-repair proficient sporadic CRC patients for better prognosis. Their best classification model yielded a 54 gene-set with an estimated prediction accuracy of 71 %. Another problem of limited discrimination for Dukes stage B and C disease is presented by Jorissen et al. [13]. They conclude that metastasis-associated gene expression changes can be used to refine traditional outcome prediction, providing a rational approach for tailoring treatments to subsets of patients. Finally, three of the five microarray data sets used in this paper, have also been used by Schlicker et al. [14]. They model the heterogeneity of CRC by defining subtypes of patients with homogeneous biological and clinical characteristics and match these subtypes to cell lines for which extensive pharmacological data is available, thus linking targeted therapies to patients most likely to respond to treatment.

Regarding ontology and classification analysis related to colorectal cancer, authors in [15] sum up the biomarkers results from 23 different researches. 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 ontology analysis using various tools. Similarly, in [16] the researchers use Affymetrix microarray data from 20 patients to reveal significant gene expression, which resulted in 1469 biomarkers. From the ontology analysis they ranked top 10 most important pathways. Since the non overlapping between the biomarkers sets discovered in different scientific papers is very common, a new meta-analysis model of colorectal cancer gene expression profiling studies is proposed in [17]. As the authors ranked the biomarker genes according to various parameters, the gene CDH3 which we found to play role in the colorectal cancer [3], is also found by their meta-analysis model. Another interesting approach maintained with classification analysis is presented in [18], 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 our methodology [3], but it was not considered important in our research since no specific connection to the colorectal cancer was found in the literature.

3 Methods and Methodology

In this section we define the methodology that we developed to detect the genes whose expression is statistically and biologically markable among the different CRC stages and the healthy tissues. We also present the GO procedure that we used to obtain the genes involved in the common biological and molecular functions of the all four CRC stages. Eventually, we present the modified classification procedure that was used for obtaining the new improved results.

3.1 The CRC Stages

Colorectal stages systems are designed to enable physicians to stratify patients in terms of expected predicted survival, to help select the most effective treatments, to determine prognoses, and to evaluate cancer control measures [19]. All data is organized into four CRC stages [20]:

1. *Stage I*—In this stage cancer has grown through the superficial lining, i.e., mucosa of the colon or rectum, but has not spread beyond the colon wall or rectum.
2. *Stage II*—In this stage cancer has grown into or through the wall of the colon or rectum, but has not spread to nearby lymph nodes.
3. *Stage III*—In this stage cancer has invaded nearby lymph nodes, but is not affecting other parts of the body yet.
4. *Stage IV*—In this stage cancer has spread to distant organs.

3.2 Choosing the Biomarkers

In the process of CRC stages biomarkers selection, instead of using the whole genome data, we use the same set of biomarkers which ability to distinguish carcinogenic and healthy patients is previously confirmed by classification and GO analysis [2, 3].

Once we obtained the initial set of biomarkers, B, we repeated the procedure for biomarkers selection for each stage S_i, where i is the current CRC stage, versus $Healthy$ tissues, in order to produce subsets of biomarkers, B_i. The process for revealing the biomarkers consists of the following steps:

1. *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 [21].
2. *Low entropy filter.* We used low entropy filter to remove the genes with almost ordered expression levels [22], since they lead to wrong conclusions about the genes behaviour.
3. *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 [23], we used a paired-sample t-test.
4. *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.
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. In order to display both statistically and biologically significant genes we used volcano plot visual tool.

3.3 Gene Ontology Analysis

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 [24].

The goal of the Gene Ontology 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 [25]. The GO project provides ontologies to describe attributes of gene products in three non-overlapping domains of molecular biology [26]:

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 Gene Ontology 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 [27].

In this paper we define few test cases to compare the ontologies of the critical and the neighbouring stages:

1. *Test case 1*—Compare Stage I and Stage II
2. *Test case 2*—Compare Stage II and Stage III
3. *Test case 3*—Compare Stage III and Stage IV
4. *Test case 4*—Compare Stage I and Stage IV

3.4 Remodelling the Prior Distributions

Previously revealed biomarkers showed high precision while diagnosing both carcinogenic and healthy patients [2]. In order to produce improved CRC stage classification, we used the developed procedure [5], and introduced a powerful key subprocedure that enables reshaping the probability distributions of the training and test set:

1. *Round-up threshold method*
2. *Normalization*
3. *Smoothing method*
4. **Boosting the enriched biomarkers**: as we have analysed the common biological and molecular functions of all four CRC stages from the GO analysis, we introduce an additional method which as presented in Sect. 4, produced an improved prior distributions modelling of the CRC stages. We chose special biomarker genes that play role in the common biological functions among the CRC stages. In order to increase the importance of the special genes, we multiplied the set in ratio 3 : 1, so that the new set is now a leading factor in the distributions shape.
5. *Hypothesis testing*.

3.5 Multiclass Bayesian Classification

As we remodelled the prior distributions of all four CRC stages, we are now able to use them in the Bayes' theorem and to calculate the posterior probability for each patient to belong to each of the four classes. Given the prior distributions we can calculate the class conditional densities, $p(\boldsymbol{x}|C_i)$, as the product of the continuous probability distributions of each gene from $\boldsymbol{x}$ distinctively:

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

Since we have unequal number of patients in all four classes, considering the total number of 657 tissues, we defined the prior probabilities $P(C_i)$, to be $P(C_1) = 0.2085$, $P(C_2) = 0.3912$, $P(C_3) = 0.2770$ and $P(C_4) = 0.1233$. Therefore, we calculate the posterior probability $P(C_i|\boldsymbol{x})$, as:

$$p(C_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_i) * P(C_i)}{\sum_{1}^{4} p(\boldsymbol{x}|C_i) * P(C_i)} \tag{2}$$

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

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

4 Experiments and Results

4.1 Gene Expression Data

In order to unveil the biomarker genes in the initial biomarkers set B, discussed in Sect. 3.2, we used the microarray experiment retrieved from Gene Expression Omnibus database [28] with GEO accession ID $GSE8671$, where 32 carcinogenic and 32 adjacent normal tissues were probed with the $Affymetrix\ Human\ Genome\ U133\ Plus\ 2.0\ Array$.

The microarray experiments used for CRC stages biomarkers detection are retrieved by using the following GEO accession IDs: $GSE37892$, $GSE21510$, $GSE9348$, $GSE14333$ and $GSE35896$. The experiments have been performed using the same $Affymetrix$ platform. All data is organized into four CRC stages:

- *Stage I* contains gene expression from 137 patients.
- *Stage II* contains gene expression from 257 patients.
- *Stage III* contains gene expression from 182 patients.
- *Stage IV* contains gene expression from 81 patients.

4.2 Gene Ontology Results

According to the methodology we defined in Sect. 3.2, for each CRC stage we created new subset of biomarkers, B_i, for each stage $i = 1, .., 4$:

- $B_1 = 70$
- $B_2 = 72$
- $B_3 = 73$
- $B_4 = 66$

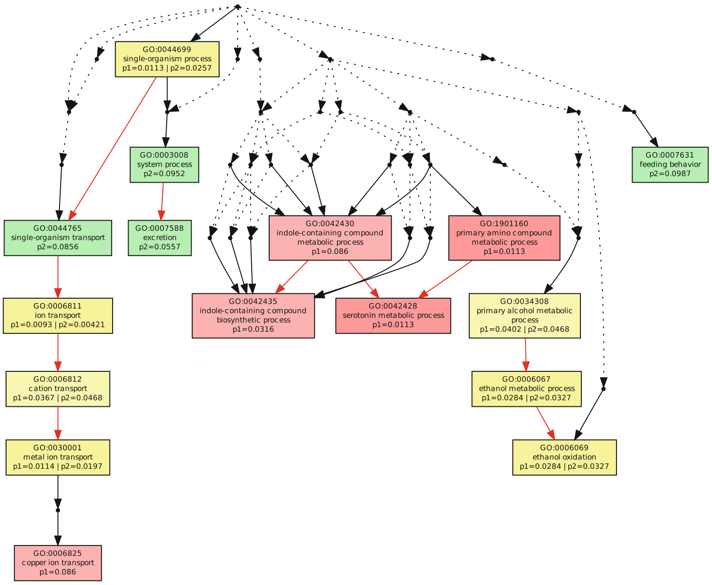

Fig. 1 Biological processes: stage I versus stage II

Thus, instead of analysis all the 138 initial biomarkers as we did in our previous work [3], we analyse the subsets of biomarkers that represent the current CRC stage.

For each subset B_i we performed GO-stage analysis using the GOEAST online tool previously discussed in Sect. 3.3. In order to compare multiple GO results, we used the Multi-GOEAST tool and produced three types of ontologies to describe: Biological processes, Molecular function and Cellular component. The different colour saturation degrees in the graphs present the enrichment significance of each GO term, defined by the p-value. In the graphical output of Multi-GOEAST results, each set is represented with different colour. Therefore, red and green boxes represent enriched GO terms only found in one of the biomarkers set, whereas yellow boxes represent commonly enriched GO terms in both experiments.

Since all ontologies refer to the same problem, in this paper we present only the *Biological processes* view.

Figure 1 presents the comparison of the ontology analysis between the neighbouring stage I and stage II. As we can see, those stages have 7 biological processes in common. Considering the critical stages II and III, Fig. 2 depicts their common biological processes, most of them overlapping with the common processes between stage I and stage II. Figure 3 presents the common biological processes in the neighbouring stages III and IV. They have nearly the same common biological processes as in the previous test cases.

Finally, we compared the critical stages I and IV (Fig. 4). As a result of these comparisons, we choose the following processes for further analysis, since they are common in the all four cases:

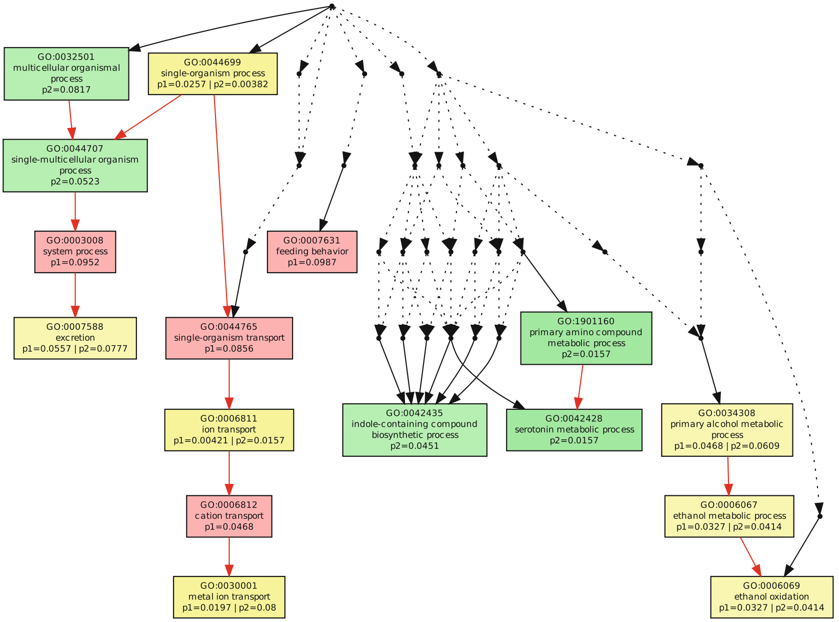

Fig. 2 Biological processes: stage II versus stage III

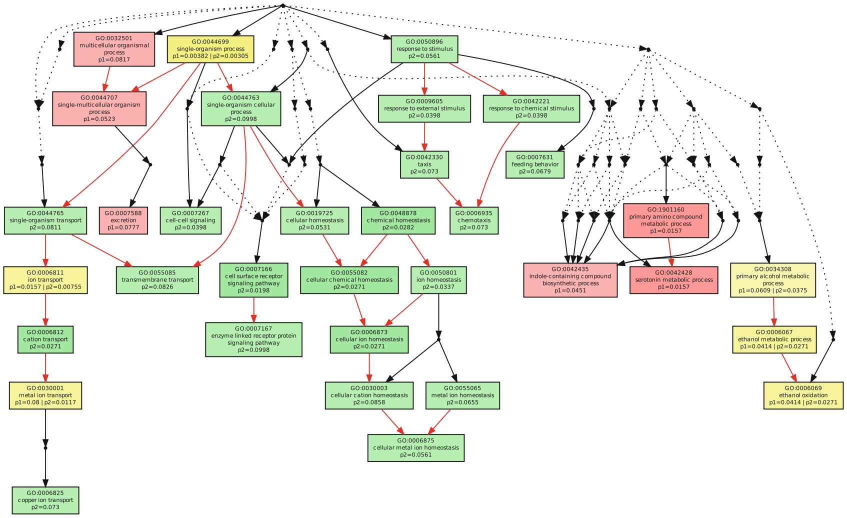

Fig. 3 Biological processes: stage III versus stage IV

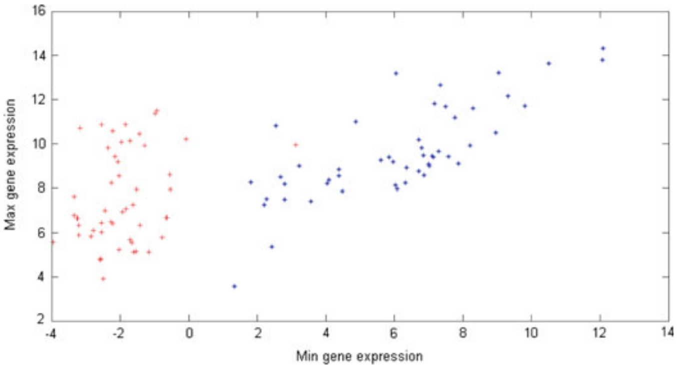

Fig. 4 Biological processes: stage I versus stage IV

- Single organism process;
- Ion transport;
- Metal ion transport;
- Primary alcohol metabolic process;
- Ethanol metabolic process and
- Ethanol oxidation.

Considering the common biological processes, we extracted 50 gene probes, which we refer to as $special\ genes$, that are directly involved in these processes. In Table 1 we present the gene symbols that are equal to the gene probes. An important fact is that four ($CHGA$, $GUCA2B$, $CDH3$ and PYY) of the five biomarkers we found to be highly correlated with the CRC phenomena, are found in the common biological processes.

Table 1 Biomarkers from common biological processes

Biomarker genes				
UNC5C	TRPM6	TPH1	SST	SLC6A19
SLC30A10	SLC26A3	SLC25A34	SCNN1B	SCN9A
SCN7A	RSPO2	PYY	PRKAA2	PLP1
NRXN1	NEUROD1	LGI1	INSM1	INSL5
GUCA2B	GREM2	GCNT2	GCG	FCRLA
FAM5C	CXCL13	CP	CLDN8	CHGB
CHGA	CHAD	CDH3	CDH19	CCL23
BEST4 ADH1B	BCHE	ASCL2	ANGPTL1	AFF3

4.3 Classification Results

As we finished the GO-stage analysis and obtained the enriched set of genes (the special genes), we inspected the ranges of the gene expression of the special genes at both carcinogenic and healthy tissues from the data in our disposition. Figure 5 presents the ranges of expression at test patients which were not involved in the biomarkers selection process and we can conclude that the special genes clearly distinguish the carcinogenic and the healthy patients. Following this first result, we proceeded with further experiments in order to improve the classification results in CRC stages classification.

Hereupon, we applied the methodology in Sect. 3.4 to remodel the gene expression distributions of each CRC stage. A key point in remodelling the probability distributions was to boost the enriched set of genes **3** times, thus the special genes are now a leading factor in the distributions shape. Using the boosting method, we additionally avoid the overlap between the probability distributions of the critical CRC stages. Therefore, as a results we got a set of 238 biomarkers which produced the distributions of the training sets depicted in Fig. 6.

As we remodelled the probability distributions, we used them in the Multiclass Bayesian classifier developed in Sect. 3.5 and achieved the improvements presented in Table 2. The improvement of recognition in the first three stages is significant, so we can decide to use this procedure, even the fourth-stage classification has decreased for few points. The *Old results* refer to the results published in [5].

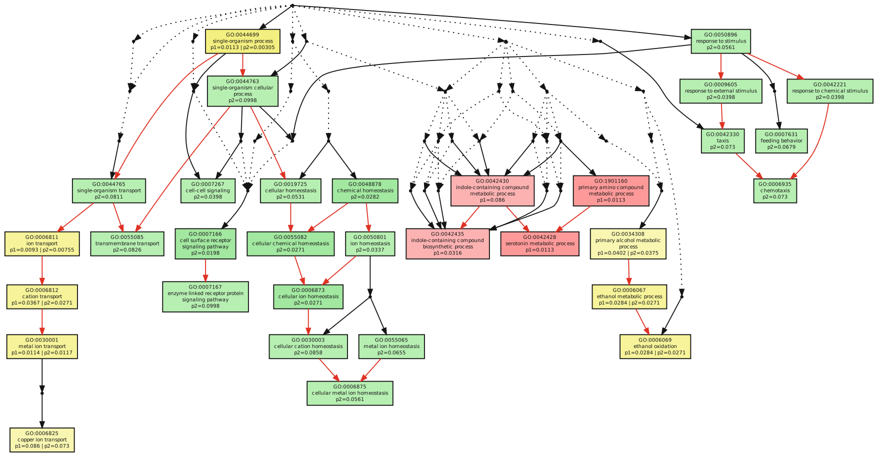

Fig. 5 Gene expression ranges of the special genes in carcinogenic and healthy testing tissues

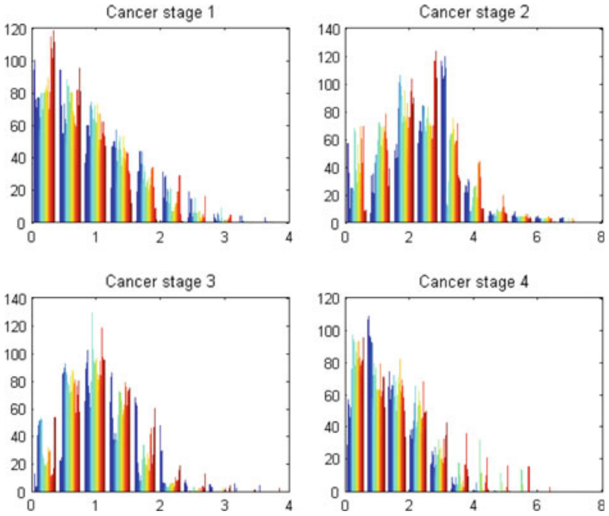

Fig. 6 Probability distributions of the training sets

Table 2 Classification results

Class	Old results (%)	New results (%)
Stage I	73.72	80.29
Stage II	53.69	63.42
Stage III	72.52	92.30
Stage IV	64.19	58.02

5 Conclusions and Future Work

The paper follows two threads of our previous work on CRC DNA chip gene expressions: the developed classificator of the four CRC stages [5] and GO CRC biomarkers analysis [3]. In this paper we developed GO-stages analysis (Biological processes) for each of the cancer stages (I–IV) using the 138 biomarkers. We have compared the neighbouring GO-stages as they are more difficult to distinguish one from another (GO_1–GO_2, GO_2–GO_3, GO_3–GO_4, GO_1–GO_4) and extracted several conclusions from this analysis. In this process we have differentiated 6 functions, common for the all 4 stages. We deducted the 50 gene probes directly included in the 6 common functions. The experiments has shown that the DNA chip expressions of the 50 special genes, clearly distinguish the carcinogenic versus healthy patients. Following this important result, we upgraded the existing methodology for preprocessing the multiclass Bayesian classification by adding the boosting method which multiplied the existing enriched biomarkers. As a result we achieved new probability distributions applicable for Bayesian modelling. The novel preprocessing procedure was confirmed by the experiments.

This result of distinguishing the genes responsible for the common functions for all 4 CRC stages, is very important for the further analysis of the colorectal cancer stages. In the future work, GO stages will help us to determine the distinguishing functions between the CRC stages, that will enable us to further improve the stages classification process and pinpoint the biomarkers that are responsible for the different CRC stages.

References

1. GLOBOCAN: (2008)
2. Simjanoska, M., Bogdanova, A.M., 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, IEEE, pp. 959–964 (2013)
3. Simjanoska, M., Bogdanova, A.M., 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, IJCCI, pp. 396–406 (2013)
4. Jain, K.: Applications of biochips: from diagnostics to personalized medicine. Curr. Opin. Drug Discov. Devel. **7**(3), 285–289 (2004)
5. 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)
6. Ahmed, F.E.: Artificial neural networks for diagnosis and survival prediction in colon cancer. Molecular cancer **4**(1), 29 (2005)
7. Frederiksen, C.M., Knudsen, S., Laurberg, S., Ørntoft, T.F.: Classification of dukes' b and c colorectal cancers using expression arrays. J. Cancer Res. Clin. Oncol. **129**(5), 263–271 (2003)
8. Eschrich, S., Yang, I., Bloom, G., Kwong, K.Y., Boulware, D., Cantor, A., Coppola, D., Kruhøffer, M., Aaltonen, L., Orntoft, T. F., et al.: Molecular staging for survival prediction of colorectal cancer patients. J. Clin. Oncol. **23**(15), 3526–3535 (2005)
9. Salazar, R., Roepman, P., Capella, G., Moreno, V., Simon, I., Dreezen, C., Lopez- Doriga, A., Santos, C., Marijnen, C., Westerga, J., et al.: Gene expression signature to improve prognosis prediction of stage ii and iii colorectal cancer. J. Clin. Oncol. **29**(1), 17–24 (2011)
10. Laibe, S., Lagarde, A., Ferrari, A., Monges, G., Birnbaum, D., Olschwang, S.: The COL2 Project: A seven-gene signature aggregates a subgroup of stage ii colon cancers with stage iii. OMICS: J. Integr. Biol. **16**(10), 560–565 (2012)
11. Tsukamoto, S., Ishikawa, T., Iida, S., Ishiguro, M., Mogushi, K., Mizushima, H., Uetake, H., Tanaka, H., Sugihara, K.: Clinical significance of osteoprotegerin expression in human colorectal cancer. Clin. Cancer Res. **17**(8), 2444–2450 (2011)
12. Hong, Y., Downey, T., Eu, K.W., Koh, P.K., Cheah, P.Y.: A metastasis-prone signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. Clin. Exp. Metastasis **27**(2), 83–90 (2010)
13. Jorissen, R.N., Gibbs, P., Christie, M., Prakash, S., Lipton, L., Desai, J., Kerr, D., Aaltonen, L.A., Arango, D., Kruhøffer, M., et al.: Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage b and c colorectal cancer. Clin. Cancer Res. **15**(24), 7642–7651 (2009)
14. Schlicker, A., Beran, G., Chresta, C.M., McWalter, G., Pritchard, A., Weston, S., Runswick, S., Davenport, S., Heathcote, K., Castro, D.A., et al.: Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. BMC Med. Genomics **5**(1), 66 (2012)

15. 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)
16. 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)
17. Chan, S.K., Griffth, O.L., Tai, I.T., Jones, S.J.: Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers. Cancer Epidemiol. Biomarkers Prev. **17**(3), 543–552 (2008)
18. 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 pair-wise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements. BMC Syst. Biol. **2**(1), 72 (2008)
19. O'Connell, J.B., Maggard, M.A., Ko, C.Y.: Colon cancer survival rates with the new American joint committee on cancer sixth edition staging. J. Natl. Cancer Inst. **96**(19), 1420–1425 (2004)
20. MayoClinic: Colon cancer (2013)
21. Wu, Z., Aryee, M.: Subset quantile normalization using negative control features. J. Comput. Biol. **17**(10), 1385–1395 (2010)
22. Needham, C., Manfield, I., Bulpitt, A., Gilmartin, P., Westhead, D.: From gene expression to gene regulatory networks in arabidopsis thaliana. BMC Syst. Biol. **3**(1), 85 (2009)
23. Hui, Y., Kang, T., Xie, L., Yuan-Yuan, L.: Digout: viewing differential expression genes as outliers. J. Bioinf. Comput. Biol. **8**(supp01), 161–175 (2010)
24. Wang, K., Li, M., Hakonarson, H.: Analysing biological pathways in genome-wide association studies. Nat. Rev. Genetics **11**(12), 843–854 (2010)
25. 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. Nat. genetics **25**(1), 25 (2000)
26. 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) (2004) D258
27. 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)
28. Gene Expression Omnibus: (2013)