A Combinatorial Approach to Construct Core and Generic Gene Co-Expression Networks of Colon Cancer

Mustafa Özgür Cingiz, Göksel Biricik, Banu Diri
Computer Engineering Department
Yildiz Technical University
Istanbul, Turkey
{mozgur, goksel, banu}@ce.yildiz.edu.tr

***Abstract*—Biological experiments can be set in order to detect the causes of diseases. However, they are expensive and time consuming. Recent developments in sequencing technologies help researchers to more easily reveal the underlying mechanisms of the diseases. In this study, we propose a combinatorial method to construct generic and core gene co-expression networks (GCNs) to discover the genes and their interactions related to colon cancer. We apply five gene network inference (GNI) algorithms and combine their estimations with Simple Majority Voting to specify the frequently inferred gene interactions and obtain the resulting GCNs on two different gene expression datasets. We then apply the intersection and union operators on these GCNS to derive the core and generic GCNs, respectively. The evaluation results of overlap analysis and topological features of GCNs for the colon cancer show that the networks produced with the proposed approach fit to the power-law degree distribution better.**

Keywords— gene co-expression network; gene network inference; ensemble based decision making; overlap analysis; topological features

I. INTRODUCTION

Protein synthesis alterations in organisms can generate different phenotypes, such as diseases. The reason behind protein level changes is generally based on chromosomal conditions as well as gene mutations that are either inherited from parents or related to influence of the environment. Changes in expression levels of genes, which are responsible for protein synthesis, induce alterations [1, 2]. For this reason, the discovery of the disease related genes and their interactions can reveal the underlying mechanisms of the diseases. In humans, there are approximately 20,000-25,000 genes that are responsible for protein synthesis. The relations between these genes can be derived with biological experiments. However, these tasks are expensive and time consuming. The developments in next generation sequencing technologies enable researchers to discover the disease related genes and their interactions more accurately. A variety of techniques in sequencing technologies help us to overview the whole genome or transcriptome more explicitly. Microarray gene expression, RNA-Seq, ChIP-seq, ChIP-chip, copy number variation, miRNA expression, DNA methylation are the examples of commonly used biological datasets to infer relations between molecules.

Microarray gene expression experiment is a widely used approach to infer interactions at gene level. Gene regulatory networks (GRNs) infer relations between genes and transcription factors (TFs) on microarray gene expression data. On the other hand, gene co-expression networks (GCNs) infer relations between genes on microarray gene expression data. GCNs reveal the relation between co-expressed genes that have tendency to involve in similar biological processes or have similar functions in the cell.

As we mentioned before, deriving these networks without biological experiments save both time and money. In order to achieve this, many gene network inference (GNI) algorithms are proposed in the literature. However, the impact of combining an ensemble of GNI algorithms is not yet fully explored. This state motivates us to investigate the network inference performance of a combined GNI algorithms ensemble.

GNI algorithms can use several measures to determine the interactions between genes, among which mutual information and correlation coefficient are known as the most popular ones. In network representation, nodes represent genes and edges represent interactions between genes. If two genes have a relation, these genes are named as co-expressed genes that show similar gene expression patterns in different conditions or different phenotypes. These genes form gene modules that have tendency to be involved in similar biological processes in the cell. In addition, detection of co-expressed genes is also important to understand the GRNs. Target genes of TFs are generally co-expressed genes that are derived in GCNs. There are many studies that aim to find disease related modules to discover the reasons of diseases [3, 4].

In the recent years, integration studies to obtain GCNs and GRNs of various diseases gained attraction and became an active research area. Researchers integrate multiple datasets to infer reliable gene networks. These datasets can be obtained with either same or different biological techniques. Using multiple microarray gene expression datasets to obtain extensive and reliable gene networks corresponds to the

978-1-5090-5795-5/17/$31.00 ©2017 IEEE

utilization of same biological technique [5-7]. On the other hand, extensive gene networks can be inferred with the combinatorial integration of GCNs that are obtained by using different biological techniques like RNA-Seq, miRNA expression and literature data [8-10].

In this study, we focus on deriving GCNs of colon cancer by using information based GNI algorithms. For this purpose we use two independent gene expression datasets related to colon cancer. The integrated GCNs of each dataset are derived using majority voting technique that utilizes all GCNs of GNI algorithms. After retrieval of GCNs from two datasets, we used union and intersection operators to aggregate two final GCNs. This aggregation builds extensive GCNs of colon cancer. In the light of our motivation, the combinatorial approach by using multiple GNI algorithms on independent datasets to construct the GCNs is our contribution. For performance analysis, we use overlap analysis and topological features of GCNs.

This paper is organized as follows. In section two we introduce the proposed GCN inference approach. Third section presents the performance results of GCNs. Finally we conclude by discussing the results and address future work.

II. Materials and Methods

We introduce the independent datasets related to colon cancer and our proposed combinatorial approach in the following subsections.

A. Datasets

The first dataset is *expO* gene expression dataset that contains 259 RAS signaling pathway related genes from 292 colon tumors. Microarray platform of *expO* is Affymetrix HG-U133PLUS2 and it can be accessed through GSE2109 in GEO repository [11].

The second dataset is *jorissen* gene expression dataset that contains 259 RAS signaling pathway related genes from 290 colon tumors. Microarray platform of *jorissen* is Affymetrix HG-U133PLUS2 and its accession number is GSE14333 in GEO repository [12].

Genes of both datasets are related to KRAS mutations. Gene expression datasets are retrieved from *predictionet* R package [13] and both of them contain normalized data.

B. Proposed Combinatorial GCN Inference Approach

Information based algorithms are known with their success on inferring genome-wide GCNs and GRNs on microarray data. It is shown that they are less complicated and faster than bayesian network based and regression based gene inference algorithms [14]. Information based GNI algorithms build GCNs by calculating the association scores between genes. We combined five information based GNI algorithms and applied them on *expO* and *jorissen* datasets. The selected algorithms are ARACNE [15], C3NET [16], CLR [17], MRNET [18] and MRNETB [19].

The selected algorithms calculate mutual information values of gene pairs to measure relevance of gene-gene interactions. ARACNE eliminates indirect relations by considering a triplet of genes. CLR eliminates weak gene-gene associations with converting mutual information value to z-scores of mutual information values. This step applies background correction via distribution of mutual information values. MRNET utilize maximum relevance/minimum redundancy feature selection method to prune interactions at second step. MRNETB improves the performance of MRNET using backward elimination and sequential replacement approach. C3NET selects interactions of a target gene considering only the highest mutual value of target gene.

We aim to integrate GCNs of two independent gene expression datasets to obtain core and generic GCNs of prostate cancer. Figure 1 illustrates the workflow and details of our proposed approach. At the first step, we separately apply the five selected GNI algorithms to *expO* and *jorissen* datasets. Each GNI algorithm infers GCNs on two datasets. Thus, we obtain five GCNs for each dataset. The ensemble of these GCNs increase the diversity of the gene-gene interaction predictions. As a support to this outcome, it is shown that the genes in common gene-gene interactions predicted by multiple GCNs are potentially related. [20]

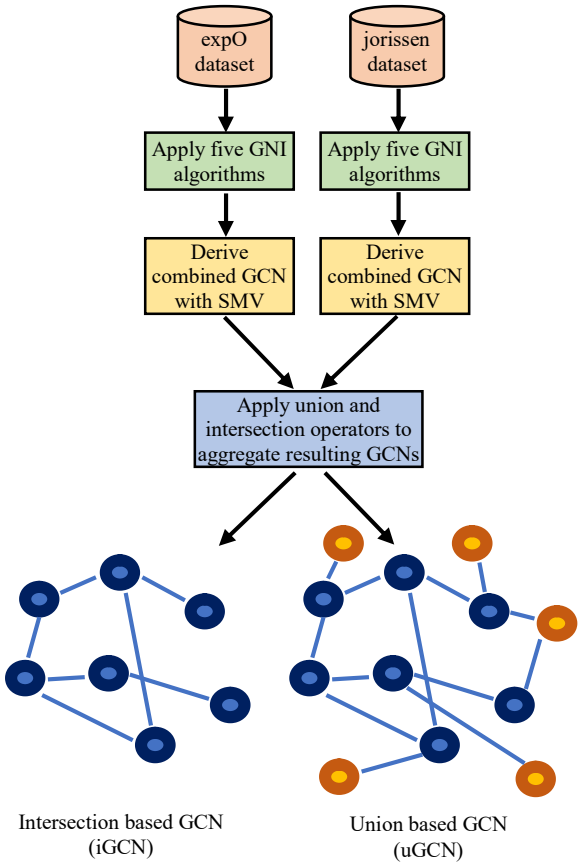

Fig. 1. Workflow of the proposed approach.

After the GCN inference step, we combine all GCNs on dataset basis with simple majority voting (SMV) algorithm in order to build more reliable GCNs. SMV provides an ensemble to determine the common gene-gene interactions from the resulting GCNs. The details of SMV algorithm are given in (1). In (1), every gene pair x, y is the member of the same gene set (N), A is the association (or interaction) between genes x and y, and $B \in \{0, 1\}$. The Kronecker delta function (δ) compares the values of its parameters and return either 0 or 1. In our case we

evaluate all possible gene pairs in the gene set. If the number of a gene pair is greater than 0, then SMV($A_{x,y}$) equals to 1, according to the sum of δ. If at least three GNI algorithms out of five detect $A_{x,y}$, then we assume there is an interaction between gene *x* and gene *y*. The application of SMV leads us to derive combined networks of *expO* and *jorissen* datasets. Although ensemble based decision making approaches do not guarantee the highest performance, the estimation is based on common gene-gene interactions that are derived from five GNI algorithms in our case, resulting in more generic GCNs. At the last step we use intersection and union operators to aggregate final GCNs of two datasets. This integration increases the diversity of interactions and enables us to obtain more generic and robust GCNs.

$$\mathrm{SMV}(A_{x,y}) = \max_{b \in B} \sum_{k=1}^{5} \delta\left(\mathrm{GCN}_k(A_{x,y}), b\right) \quad (1)$$

III. Performance Analysis of the Proposed Approach

We use overlap analysis and topological features of GCNs for the performance evaluation. In each of the following subsections we focus on these evaluation methods.

A. Overlap Analysis of GCNs

We use protein-protein interactions (PPI) in the overlap analysis. More than 1.5 million interactions result in a very large validation dataset. For this reason, we utilize precision and the number of true positives (TPs) in performance evaluation. An interaction derived with GNI algorithm that is also present in the PPI dataset indicates a TP. Otherwise, we label it as a false positive (FP).

In overlap analysis we have to be confident about whether the inferred gene-gene interactions are found due to the chance factor or not. For this purpose, we use Fisher's exact test. Null hypothesis of Fisher's exact test claims that there is no statistically significant difference between inferred interactions and random interactions. If the null hypothesis is valid, then these interactions become insignificant. We determine the threshold value as 0.05 for the Fisher's exact test. We discard any GCN with a p-value which is higher than the threshold. Thus, the p-values of all inferred GCNs are lower than threshold value.

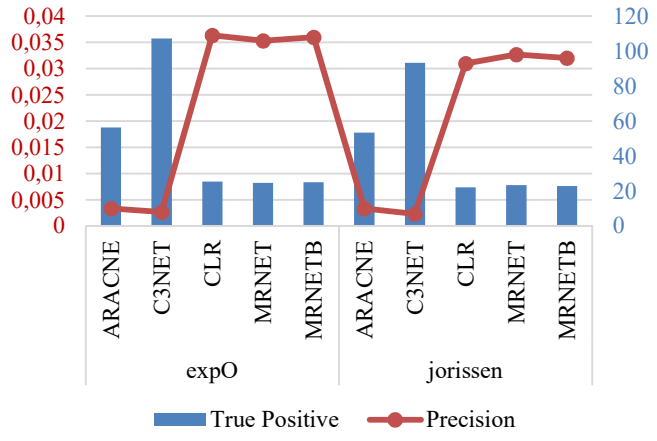

Fig. 2. Performance comparison of individual GNI algorithms.

Figure 2 shows the individual performance results of GCNs derived with the selected GNI algorithms on *expO* and *jorissen* datasets. C3NET and ARACNE predict lower number of gene-gene interactions on both datasets. However, the precision values of two algorithms are higher than CLR, MRNET and MRNETB. Although these three algorithms score lower precision values, their estimations are significantly higher than ARACNE and C3NET.

All precision values in Figure 2 vary between 0.003 and 0.035. These values are low but they are compatible to the previous results obtained in the literature [21, 22]. Figure 2 also shows that the results of GNI algorithms perform similar on both datasets. After inspecting the individual performances of GNI algorithms, we measure the performances of the integrated GCNs and their combinations. Figure 3 visualizes the results obtained with these GCNs.

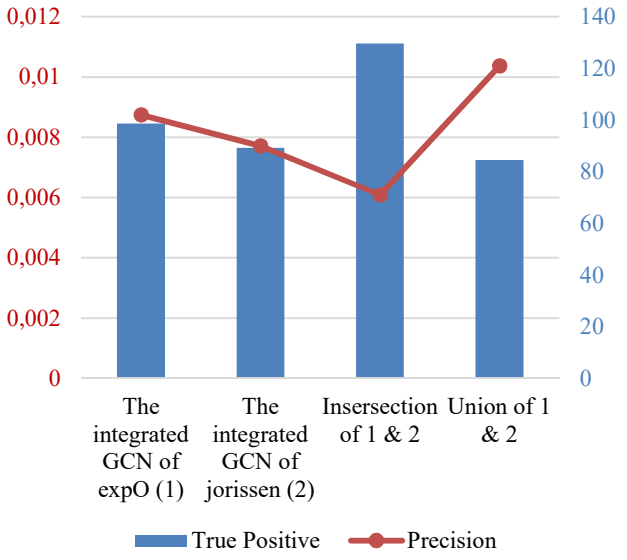

Fig. 3. Performance comparison of integrated GCNs and their combinations.

The results of integrated GCN on *expO* dataset are very close to the results of CLR, MRNET and MRNETB. Due to the lack of diversity in two groups (ARACNE-C3NET and CLR-MRNET-MRNET3) of five selected GNI algorithms, SMV cannot effectively improve the performance of the integrated GCN on *expO*. This situation is also valid for the integrated GCN on *jorissen* dataset. Nevertheless, the performance of the integrated GCN on *expO* is slightly higher than the performance of the integrated GCN on *jorissen*. Applying intersection operator on these GCNs increases the precision, but as expected, decreases the TP. The highest precision and the lowest TP are obtained with the intersection of the integrated GCNs. On the other hand, the union operator increase TP and decrease precision. Results show that C3NET and ARACNE predict scarce number of gene-gene interactions. If we focus on the average, our proposed approach increases the performances and prediction power of GNI algorithms. Besides, the diversity is increased with application of the union operation. Based on the interpreted results, we name the GCN obtained with the union of integrated GCNs as a generic GCN. Similarly, intersection of the integrated GCNs is a core GCN

whose interactions are commonly predicted by GNI algorithms on two different datasets.

B. Topological Features of GCNs

In gene networks, the degrees of nodes generally fit power-law degree distribution. In this type of distribution, genes in limited number have more connections than other genes in the gene network. These genes are named as the hub genes. Hub genes are important as they associate gene-disease relations. In this study, we extract hub genes in terms of degrees of nodes.

Other measures for the evaluation of GCNs are the average number of neighbors, network heterogeneity and clustering coefficient. Average number of neighbors indicates the average number of interactions for genes in a gene network. Network heterogeneity reflects the relevance to the power-law degree distribution. Clustering coefficient gives information about modularity of nodes in a gene network. All of these measures are obtained using NetworkAnalyzer tool [23]. Table 1 lists the topological features of integrated GCNs and their combinations.

TABLE I. TOPOLOGICAL FEATURES OF GCNS.

GCNs	Average # of Neighbors	Network Heterogeneity	Clustering Coefficient	Top 5 Hub Genes
CLR based GCN of *expO*	99.328	0.166	0.476	ACOT7 NOLC1 PLAUR OXSR1 PALMD
MRNET based GCN of *jorrisen*	96.803	0.165	0.444	PLAUR NIPAL1 CHST11 PTGS2 IL23A
Integrated GCN of *expO*	93.035	0.185	0.451	NOLC1 PALMD ACOT7 BCL6 ITPR3
Integrated GCN of *jorrisen*	90.757	0.180	0.442	PLAUR NIPAL1 CHST11 PTGS2 SLC2A3
Intersection of integrated GCNs	49.174	0.328	0.350	PLAUR SLC2A3 CDCP1 PHLDA2 NIPAL1
Union of integrated GCNs	134.618	0.119	0.560	HRAS NFKBIZ IL11 IL1A IL1B

In all GCNs listed in Table 1, the average number of neighbors is high. SMV determines gene interactions by using the GCNs of the five selected GNI algorithms. This leads to a decrease in modularity (clustering coefficient) and average number of neighbors. Intersection operator takes common interactions of the integrated GCNs into account. Thus, interaction decreases the average number of neighbors and clustering coefficient. On the contrary, the union operator aggregates all interactions of the integrated GCNs. This operation increases the average number of neighbors and clustering coefficient. The ensemble of the GNI algorithms and the intersection operator together decreases the number of interactions. This results in the elimination of the genes that have scarce number of interactions. Conversely, the union operator causes to retrieve higher network heterogeneity.

Hub genes are associated to cancer related biological processes. For this reason, revealing of hub genes is an important task to understand diseases. Studies show that PLAUR, NOLC1, PALMD, IL11, IL1A, IL1B, SLC2A3 (GLUT3) genes promote colon cancer formation and progression [24-27]. In this study, these genes are found as hub genes in the combinatorial GCNs. Thus, our predictions are compatible with the previous works in the literature.

IV. DISCUSSION AND CONCLUSION

Although the relations between the genes that are related to diseases can be derived with biological experiments, these tasks are expensive and time consuming. For this reason, researchers tend to reveal these relations by using the GNI algorithms in order to discover the underlying mechanisms in diseases. There are many GNI algorithms presented in the literature but the impact of an ensemble of GNI algorithms is not investigated in details. In this study, we focused on this issue and aimed to derive generic and core GCNs for colon cancer by using an ensemble of GNI algorithms. In order to achieve this goal, we use two small sized independent gene expression networks to construct GCNs. These datasets increase the diversities of the GCNs and enable to infer various GCNs of colon cancer. In addition, we increased the diversity by using an ensemble based decision making method. We combined the estimations of the selected GNI algorithms with SMV and derived the integrated GCNs. With this approach, we were able to obtain more robust gene- gene interactions. We applied intersection and union operators on these integrated GCNs to obtain the core and the generic GCNs, respectively.

The results show that the integrated GCNs derived from the independent datasets did not significantly improve the precision and TP, when compared with the selected GNI algorithms. This is due to the lack of diversities in GCNs of the selected GNI algorithms. However, we were able to derive core and generic GCNs of colon cancer, based on all GNI algorithms. In addition, intersection of the integrated GCNs decreased TP and increased precision. On the other hand, the union operator increased precision and decreased TP. Focusing on the average, our proposed approach increased the performances and prediction power of GNI algorithms.

We evaluated the results of our proposed method with the topological features of GCNs derived with the GNI algorithms. The topological features of the GCNs derived with SMV seemed similar to the features of the GCNs of the selected individual GNI algorithms. However, the integrated GCNs enabled us to obtain biological networks whose nodes fit better to the power-law degree distribution. In addition, all of the

colon cancer related hub genes that are inferred with the proposed approach are consistent with the hub genes reported in the literature.

The weaknesses of this study can be reported in two main titles. First, we used a limited number of gene expression datasets in the ensemble for the integration. This situation forwarded us to a lack in obtaining significant GCNs. Second, we utilized fundamental techniques such as the SMV, the union and the intersection operators for integration. We could consider weighting the inferred GCNs. However, the limited number of gene expression datasets used eliminated this decision.

On contrary to the weaknesses, the integration of biological datasets obtained from the selected techniques increases the diversity of interaction predictions. Thus, we can say that our approach is applicable.

For the future work, we scheduled to occupy miscellaneous cancer related biological datasets for the further improvement of our approach. We also planned to use various operators to increase the diversities of GCNs.

REFERENCES

[1] G.L. Semenza, "Targeting HIF-1 for cancer therapy", Nature reviews cancer, vol. 3(10), pp. 721-732, 2003.

[2] B. Levine, "Cell biology: autophagy and cancer", Nature, vol. 446(7137), pp. 745-747, 2007.

[3] J.A. Miller, M.C. Oldham and D.H. Geschwind, "A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging", Journal of neuroscience, vol. 28(6), pp. 1410-1420, 2008.

[4] H.N. Kadarmideen, N.S. Watson-Haigh and N.M. Andronicos, "Systems biology of ovine intestinal parasite resistance: disease gene modules and biomarkers", Molecular Biosystems, vol. 7(1), pp. 235-246, 2011.

[5] X.B. Huang and Zhao Tian, "Inferring Gene Regulatory Networks using Heterogeneous Microarray Data Sets", in 2nd International Conference on Bioinformatics and Biomedical Engineering, Shangai, 2008, pp.518-522.

[6] Y.K. Wang, D.G. Hurley, S. Schnell, C.G. Print and E.J. Crampin, "I ntegration of Steady-State and Temporal Gene Expression Data for the Inference of Gene Regulatory Networks", PloS ONE, vol.8(8), pp. e72103, 2013.

[7] A. Sirbu, M.Crane and H.J. Ruskin, "Data Integration for Microarrays: Enhanced Inference for Gene Regulatory Networks", Microarrays, nol.4(2), pp. 255-269, 2015.

[8] K. Lemmens, et al., "Inferring transcriptional modules from ChIP-chip, motif and microarray data", Genome Biology, vol.7(5), pp. R37, 2006.

[9] T.T. Vu and J. Vohradsky, "Inference of active transcriptional networks by integration of gene expression kinetics modeling and multisource data", Genomics, vol. 93(5), pp.426-433, 2009.

[10] J.K. Rhee, et al., "Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer", Nucleic Acids Research, vol. 41(18), pp. 8464-8474, 2013.

[11] A.H. Bild, et al., "Oncogenic pathway signatures in human cancers as a guide to targeted therapies", Nature, vol. 439(7074), pp. 353-357, 2006.

[12] R.N. Jorissen, et al., "Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer", Clinical Cancer Research, vol. 15(24), pp. 7642-7651, 2009.

[13] O. Catharina, et al., "Inference and validation of predictive gene networks from biomedical literature and gene expression data", Genomics, vol. 103(5), pp. 329-336, 2014.

[14] W.C. Young, A.E. Raftery and K.Y. Yeung, "Fast Bayesian inference for gene regulatory networks using ScanBMA", BMC Systems Biology, vol. 8(1), pp. 47, 2014.

[15] A.A. Margolin, et al., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context", BMC Bioinformatics, vol. 7(1), pp. S7, 2006.

[16] G. Altay and F. Emmert-Streib, "Inferring the conservative causal core of gene regulatory networks", BMC Systems Biology, vol. 4(1), pp. 132, 2010.

[17] J.J. Faith, et al., "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles", PLoS Biol, vol. 5(1), pp. e8, 2007.

[18] P.E. Meyer, K. Kontos, F. Lafitte and G. Bontempi, "Information-theoretic inference of large transcriptional regulatory networks", EURASIP J Bioinform Syst Biol, vol. (1), pp. 79879, 2007.

[19] P. Meyer, D. Marbach, S. Roy and M. Kellis, "Information-Theoretic Inference of Gene Networks Using Backward Elimination", Biocomp, 2010.

[20] M.M.Babu, "Introduction to microarray data analysis", Computationla Genomics: Theory and Application, vol. 17(6), pp. 225-249, 2004.

[21] F.M. Giorgi, C. Del Fabbro and F. Licausi, "Comparative study of RNA-seq-and microarray-derived coexpression networks in Arabidopsis thaliana", Bioinformatics, vol. 29(6), pp. 717-724, 2013.

[22] R. de Matos Simoes, S. Dalleau, K.E. Williamson and F. Emmert-Streib, "Urothelial cancer gene regulatory networks inferred from large-scale RNAseq, Bead and Oligo gene expression data", BMC Systems Biology, vol. 9(1), pp. 21, 2015.

[23] P. Shannon, et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks", Genome Res, vol. 13(11), pp. 2498-2504, 2003.

[24] Y. Zheng, J. Zhou and Y. Tong, "Gene signatures of drug resistance predict patient survival in colorectal cancer", The pharmacogenomics journal, vol 15(2), pp. 135-143, 2015.

[25] V. Malysheva, M.A. Mendoza-Parra, M.A. Saleem and H. Gronemeyer, "Reconstruction of gene regulatory networks reveals chromatin remodelers and key transcription factors in tumorigenesis", Genome medicine, vol. 8(1), pp. 57, 2016.

[26] W.T. Wei, S.Z. Lin, D.L. Liu and Z.H. Wang, "The distinct mechanisms of the antitumor activity of emodin in different types of cancer (Review)", Oncology reports, vol. 30(6), pp. 2555-2562, 2013.

[27] K.L. Bondurant, A. Lundgreen, J.S. Herrick, S. Kadlubar, R.K. Wolff and M.L. Slattery, "Interleukin genes and associations with colon and rectal cancer risk and overall survival", International journal of cancer, vol. 132(4), pp. 905-915, 2013.