International Journal on Advanced Science Engineering Information Technology

Vol.7 (2017) No. 4-2
ISSN: 2088-5334

Pathway-based Analysis with Support Vector Machine (SVM-LASSO) for Gene Selection and Classification

Nurul Athirah Nasrudin[#], Weng Howe Chan[#], Mohd Saberi Mohamad[*,+,^], Safaai Deris[$], Suhaimi Napis[%,@], Shahreen Kasim[&]

[#]*Artificial Intelligence and Bioinformatics Research Group, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia*

[*]*Faculty of Creative Technology and Heritage, Universiti Malaysia Kelantan, Karung Berkunci 01, 16300, Bachok, Kelantan, Malaysia*
E-mail: saberi@umk.edu.my

[+]*Center for Computing and Informatics, Universiti Malaysia Kelantan, City Campus, Pengkalan Chepa, 16100, Kota Bharu, Kelantan, Malaysia*

[^]*Artificial Intelligence and Big Data Institute, Universiti Malaysia Kelantan, City Campus, Pengkalan Chepa, 16100, Kota Bharu, Kelantan, Malaysia*

[$]*Soft Computing & Intelligent System Research Group, Faculty of Computer Systems & Software Engineering,Universiti Malaysia Pahang, 26300, Kuantan, Pahang, Malaysia*
E-mail: kohbalan@ump.edu.my

[%]*Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia*

[@]*Department of Cell and Molecular Biology, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400, Serdang, Selangor, Malaysia*

[&]*Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Johor, Malaysia*

Abstract— **Genomic knowledge has become a popular research field in bioinformatics biological process that providing further biological process information. Many methods have been done to address the issues of high data throughput due to increased use of microarray technology. However, it is still not able to determine the appropriate diseases accurately. This is because of existing non-informative genes that could be included in the analysis of context-specific data like cancer gene expression data, which affect the classification performance. This study proposed a pathway-based analysis for gene classification. Pathway-based analysis enables handling microarray data in order to improve biological interpretation of the analysis outcome. Secondly, Support Vector Machine with Least Absolute Shrinkage and Selection Operator algorithm (SVM-LASSO) is proposed, which to find informative genes for each pathway to ensure efficient gene selection and classification in every pathway. Experiments are done using lung cancer dataset and breast cancer dataset that widely used in cancer classification area. A stratified 10-fold cross validation is implemented to evaluate the performance of the proposed method in terms of accuracy, specificity, and sensitivity. Moreover, biological validation has been done on the selected genes based on biological literature and biological databases. Next, the results from the proposed methods are compared with the previous study throughout all the data sets in terms of performance. As a conclusion, this research finding can contribute in biology area especially in cancer classification area.**

Keywords— **genomic knowledge; gene analysis; microarray technology; pathway-based analysis; support vector machine; LASSO; 10-fold cross-validation; cancer classification**

I. INTRODUCTION

In Genome database, there are many genes stored with different types, features, and functions example gene expression data. Some of the biological knowledge need to identify such as gene structure, gene function, and gene expression. From the features, the genes can determine whether active, hyperactive or silent in normal or cancerous

tissues. Also, the genes need to be classified whether it is a normal gene or abnormal gene that can cause diseases such as cancer diseases. Gene selection process is involved during pathway-based analysis in order to identify the subset of informative genes. This process eliminates irrelevant and redundant genes that can lead to misleading of classification. It is important to overcome the high-throughput data problems, which contain several noise genes that hold less information related cancer diseases.

Microarray is one of the lab approaches to measure gene expression. There are another lab approaches such as Serial Analysis of Gene Expression (SAGE) that are based on isolation of unique sequence tag of mRNAs with a combination of long sequence serially for sequencing [1]. Microarray technology capable in analyse several thousand of genes simultaneously compared to SAGE. In microarray area, there are many types of microarray such as DNA microarray, tissue microarray, protein microarray, and etc. In a study of complex diseases and their causes, DNA microarray data had been used widely. Therefore, pathway knowledge into microarray data has been favoured by researchers in the area of bioinformatics owing to the improved biological interpretation of the analysis outcome Thus, the pathway-based analysis for gene classification is generally explained.

Recently, pathway-based analysis got a lot of attention in the genomic research area. This is because pathway-based analysis method capable in detect subtle compared to single-gene analysis [2-3]. In fact, the pathway analysis or pathway-based analysis had been widely used in research areas such as analysis of gene sets, metabolic pathway analysis and etc. [4]. In past few years, the focus has shifted to the pathway-based analysis that capable used large-scale omics data in order to improve the biological interpretations, which is important to the identification of complex diseases like cancers [5]. Commonly, pathway analysis has always been used to know signalling or metabolic pathways are activated under certain experimental condition. This method can be used to classify gene in via networks, pathways or paths where many linked components have been induced.

In this research, Support Vector Machine (SVM) had been implemented. Machine learning is an artificial intelligence technique that concern about design and development of an algorithm that allows the computer to learn behaviours based on the empirical data. SVM is one of the machine learning that widely used and a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. Recently, the previous researcher stated a hybrid variable selection or classification approach that is based on linear combinations of the gene expression profiles could maximize an accuracy measure is shown by receiver operating characteristics curve [6]. There are previous methods that have been used to select and classify the gene, but there still have limitation.

Generally, Least Absolute Shrinkage and Selection Operator (LASSO) is a technique for variable selection, especially when dealing with high dimensional data. It involves penalizing the absolute size of the regression coefficients. LASSO technique has been investigating for computing efficient model descriptions of nonlinear systems. It minimizes the residual sum of squares by the addition of a penalty term on the parameter vector of the traditional minimization problem. LASSO structure perform detection method was evaluated by using it to estimate the structure of a nonlinear polynomial model.

II. Materials And Methods

Microarray is one of the lab approaches to measure gene expression. There are another lab approaches such as Serial Analysis of Gene Expression (SAGE) that are based on isolation of unique sequence tag of mRNAs with a combination of long sequence serially for sequencing. Microarray technology capable in analyse several thousand of genes simultaneously compared to SAGE. In microarray area, there are many types of microarray such as DNA microarray, tissue microarray, protein microarray and etc. In a study of complex diseases and their causes, DNA microarray data had been used widely.

LASSO is one of the linear regression models that used for variable selection when dealing with high dimensional data. It can identify the important variables by adding an L_1 penalty to unusual regression objective function. LASSO also the best techniques in gene classification because it can improve the accuracy when selecting the variable. Because the data are complex and multivariate, it is important to develop a technique to detect systematic signals in gene expression patterns. Based on [6], LASSO is the suitable technique to solve the problem of variable selection and classification. Some of the variables will be estimated to be exactly zero to present genes that have no discriminatory power between two classes, while those genes with non-zero coefficients will present genes that can separate classes of tumour successfully. LASSO is used to greatly reduce the number of candidate pattern. The computational algorithm can handle an extremely large number of unknowns simultaneously.

LASSO is known to have automatic variable selection ability in linear regression analysis. LASSO is used to select the most informative genes for representing the probability of an example being positive as a linear function of the gene expression data. LASSO is more capable of selecting the important genes classification. Besides, LASSO also used in linear regression and also one of the most widely used techniques for robust regression and sparse estimation. Linear regression is a commonly used approach in bioinformatics. There are some challenges of linear regression, that is the number of regression weights needed to be determined is often at least one order of magnitude larger than a number of data points.

The data are (Yi, Xi), where Yi $(i=1,\ldots,n)$ is continuous variable. LASSO solution is to optimizing problem of minimizing

$$\sum_{i=1}^{n}(Y_i - \beta^T X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (1)$$

where $\beta = (\beta_1, \cdots, \beta_p)$ and $\boldsymbol{\lambda} \geq 0$ is a penalty term. This is the way LASSO estimate to utilized is an L_1constraint. An alternative way of formulating the above formulae is to minimizing. LASSO based classifiers have better performance in terms of average testing error. This shows

LASSO is more efficient at picking informative genes for linear classifiers.

An alternative way of formulating the above formulae is by minimization. LASSO based classifiers have better performance in terms of average testing error. This shows LASSO is more efficient at picking informative genes for linear classifiers. The pathway-based analysis had been proposed with a combination of SVM-LASSO to improve the gene selection and classification performance in complex gene dataset. The concept of pathway-based analysis is analysing the gene expression in a group that related to the pathway information. Thus, the gene selection and classification has been improved due to the concept. This is because the size of gene expression become smaller. Only the informative genes that relate with the pathway are selected. Fig. 1 shows illustrate of pathway-based analysis approach.

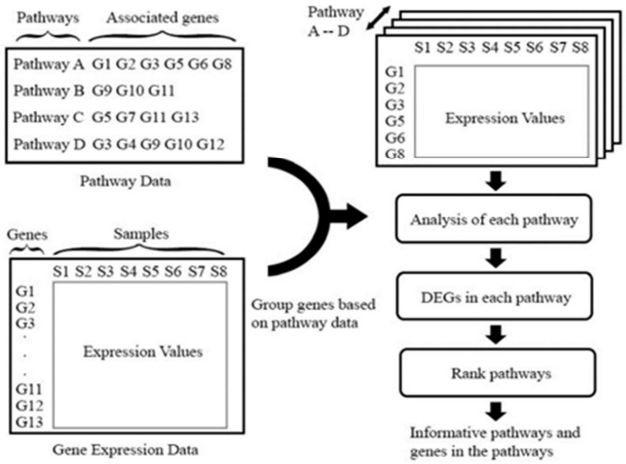

Fig. 1 Illustrate of pathway-based analysis

Next, each group of pathway will be executed to stratified 10-fold cross validation for accuracy performance validation. In this process, the data will divide into two groups of data which is training data set and testing data set. This two type of data set is used in support vector machine for the classification process. SVM will classify the classes by maximizing the margins, and this method is used for classification. For improving the selection process, an algorithm is implemented in SVM which is Least Absolute Shrinkage and Selection Operator. LASSO algorithm can control the complexity the size of the gene due to its capability in select informative gene automatically. SVM classified the gene expressions data in the pathway into their classes (e.g., tumor or normal). Next, the activities of this research are explained in Table 1 and Fig. 2 shows the flowchart of the proposed method.

In this research, gene expression microarray technologies and pathway data are used. The pathway data acts as prior biological knowledge where the gene expression data is grouped into corresponding gene sets based on the pathway data. The development of a computational system for classification gene based on cancer diseases creation based on Microarray data. Microarray dataset contains thousands of spots, and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene.

TABLE I
EXPLANATION OF ACTIVITIES

Activities	Explanation
Pre-processing data	First, the data set is obtained from microarray analysis and undergoes pre-processing data process. Next, the output form pre-processing data are defined as gene expression data. Then, gene expression data is divided into corresponding pathways according to the pathway data. Each pathway consists of a subset of gene expression data for the genes that associated to the pathway.
Selection by LASSO	The data are selected using a linear regression model to identify the informative information by reducing the number of gene expression. Thus, the noise data like uninformative genes are removed in order to reduce the complexity of experiment and in same time increase the accuracy of pathway performance.
Classification	In this research, support vector machine is used in the classification process. The data are separated based on decision plane where the sample has a different class which is a tumour and normal in cancer diseases case.
Rank the pathway	Finally, the pathway is rank based on accuracy value from higher to lower value. From that, the higher accuracy value shows that the pathway is more significant to targeted phenotype. As this research uses stratified 10-fold cross-validation (CV) for performance evaluation, after the data grouping process, samples from each pathway are partitioned based on the 10-fold CV requirement.

This experiment was done by using the datasets from GSEA and NCBI GEO which consists of lung gene expression and breast gene expression based on the previous researcher. Most of the researcher using this data to perform pathway-based analysis in order to classify the cancer genes. Available gene expression data includes Expressed Sequence Tags (ESTs), microarray files, and quantitative mass spectrometry protein measurements. An estimation of the error rate using 10-fold cross-validation was performed.

This generally gave an error rate between 15-20% for various choices, and there are some pre-processing steps for the dataset. The remained gene expression is for analysis. In addition, the pathway dataset is used to group the gene for the classification process. This data will match with pathway dataset and placed in one table or file. Genes have different name and associate number depend on the chips that used. Two gene expression datasets are used in this research. Both datasets can be download at http://www.broadinstitute.org/gsea/datasets.jsp and NCBI GEO database. Table 2 shows the gene expression data used in this research.

TABLE II
SUMMARY OF GENE EXPRESSION DATA SET

Dataset	No. of Sample	No. of Genes	Class	Ref.
Lung A	86	7129	2(normal/ tumour)	[6]
Breast	49	22218	3 tumour types	[7]

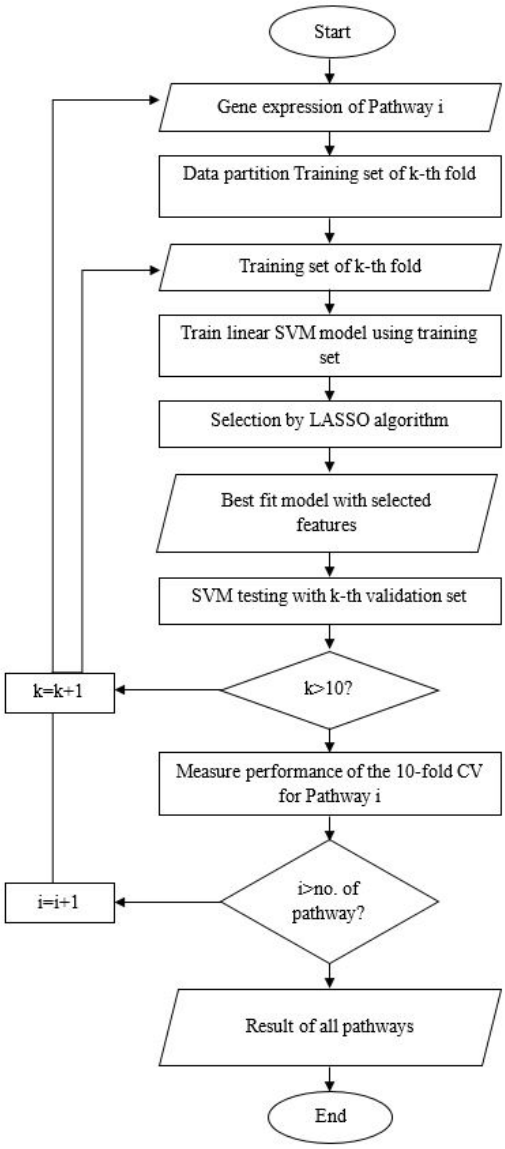

Fig. 1 Flowchart of proposed method

III. Result and Discussion

The result obtain from this study will be compared with previous methods and will be discuss in term of average accuracy of top 10 pathway and also the relational of gene with diseases.

In this research, gene selection and classification is implement by Pathway-based analysis with SVM-LASSO. The prediction performance of this functional analysis is measured by 10-fold cross validation accuracy and test set of correctness. The number of gene selected for each dataset is corresponding to the highest value of test set of correctness. The result of proposed method for both data sets are shows in Table 3.

According to the findings, Pathway-based analysis with SVM-LASSO performs well in prediction since the value accuracy of 10-fold cross validation is high and same goes to value of test correctness in overall. The differences between this two data in term of accuracy is 12%. Same goes to value of test correctness that is 1.91%. The result of this research are the pathway-based analysis of pathways and gene expression data with the application of SVM-LASSO. The data used in this re-search include Lung A [7] and Breast [8]. The different dataset are used to evaluate the performance of proposed method. These dataset are being analyse in term of test set correctness, 10-fold cross validation accuracy and average genes selected. The result is more significant with the implementation of pathway-based analysis as the gene expression being grouped followed the pathway dataset. Next, the data are analysed by the pathway that stored the gene expression value. The following section discusses and analysed the result obtained in this research. The comparison method between previous researches is presented in Table 4. The reading is recorded for five minutes.

TABLE III
Result of 10-Fold Cross-Validation of Two Data Set

Dataset	**Number of Genes**	**Number of Sample**	**Number of pathway**	**10-fold Cross Validation Accuracy (%)**
Lung A	7129	86	434	83.21%
Breast	22218	49	435	60%

TABLE IV
Comparison Performance Between Previous Researchers

Method	**10-fold CV accuracy (%)**	
	Lung A	Breast
Pathway-based with SVM-LASSO	83.21%	60%
Pathway-based with Random Forest (Chin *et al.*, 2012)	42%	60%

Based on the table above, the accuracy of the proposed method is higher than the previous method by using lung cancer data. However, the accuracies of breast cancer data for both methods are same. This shows the proposed method is not performed well in breast cancer data. Thus, the type of data used influences the accuracy of classification. Next, the selected pathway and genes are continues to biological validation.

In this section, the selected genes in the pathway produced by the proposed method will be validated using Genecards. For this process need to be done by manually. The pathway is selected and validated, for lung data in Table 5 and for breast cancer in Table 6.

In Table 5 above, shows the pathways with highest average 10-fold CV accuracy in this research. The top-ranked pathway is a CDK regulation of DNA replication, known as a cyclin-dependent kinase. Although there is no proof for the direct relationship of this pathway to lung cancer, CDK-containing fractions activated the replication of the SV40 virus in cellular extracts [22]. The CDK regulation pathway achieves 88.23% accuracy in the 10-fold CV. Next, from the proposed method has selected two informative genes for the classification.

Second-ranking goes to Nucleotide sugar metabolism pathway. This pathway is related to *Galactosemia* disease or known as an autosomal recessive disorder caused by a defect in one of the three enzyme genes for galactose metabolism [23]. It also can cause severe growth problems including

cataracts. Even though the pathway has higher accuracy in 10-fold CV, the pathway does not have any proof that related to cancer. But, the genes that selected are one of the markers in cancer disease.

TABLE V
PATHWAY FROM SVM-LASSO RANKED BY 10-FOLD CV CLASSIFICATION ACCURACY IN LUNG CANCER DATA

Pathway	10-fold CV accuracy (%)	Selected genes	Ref.
CDK Regulation of DNA Replication	88.23	MCM5 MCM3	[9]-[10]
Nucleotide sugar metabolism	85.71	GALE GALT UGP2	[11]-[13]
Regulators of Bone Mineralization	82.35	FURIN	[14]
Activation of Csk by cAMP-dependent Protein Kinase Inhibits Signaling through the T Cell Receptor	82.35	ALPL COL4A1 COL4A2 IBSP	[15]-[18]
Transcriptional activation of dbpb from mRNA	82.35	NFKB1 NFKBIA RELA	[19]-[21]

TABLE VI
PATHWAYS FROM SVM-LASSO RANKED BY 10-FOLD CV CLASSIFICATION ACCURACY IN BREAST CANCER DATA.

Pathway	10-fold CV accuracy (%)	Selected genes	Ref.
Actions of Nitric Oxide in the Heart	60	CYP7B1 GOT2 GOT1	[24]- [25]
Acute Myocardial Infarction	60	ODC1 OAC3	[26]-[27]

Table 6 shows the top pathways with the highest accuracy of 10-fold CV in breast cancer data. Next, the accuracy for both pathway is equal which is 60%. For the first pathway in the table is Actions of Nitric Oxide in the Heart or known as Nitric oxide (NO) [28]. Nitric oxide (NO) is a pleiotropic regulator which critical to the numerous biological processes, include vasodilatation, neurotransmission, and macrophage-mediated immunity. Also a multifunctional gaseous molecule and a highly reactive free radical [29]. Furthermore, various studies have shown the combination of three isoforms which are nitric oxide synthases (NOS), comprises inducible (iNOS), and endothelial NOS (eNOS) can promote or inhibit the etiology of cancer. Interestingly, NOS activity has been detected in tumor cells of various histogenetic origins. From this pathway, there are three informative genes are selected. These genes are validated with biology database that shows the relation with cancer diseases.

Next pathway is Acute Myocardial Infarction, known as a myocardial infarction (MI). Which is proven its potential implications for cancer disease in women such as breast cancer. Previous research by Scottish Breast Cancer Committee found a significant reduction in the incidence of fatal myocardial infarction in women receiving adjuvant tamoxifen that can cause breast cancer [30]. Furthermore, not only in cancer, but it is also has been examined in chronic diseases like Heart disease and cancer in a population-based sample of middle-aged men [31]. This is supported with the informative selected genes by the proposed method that shown the relation with targeted phenotype.

As a conclusion, both pathways have different information, different selected genes and different relation that gives a lot of clues towards cancer diseases. With this knowledge, it can help the expert to identify and detect the cancer diseases easily.

IV. CONCLUSIONS

Introduction of microarray technology has enabled the capability to measure the expression of several thousand of genes with multiple samples simultaneously. This has spurred the development of various sophisticated computational methods that used to analyse the microarray data and extract useful information from the data that can help us to understand the biology of the phenotype of study. However, the early methods of microarray analysis often based on the single-gene analysis (SGA), which analyze the data in terms of individual genes. Basically, the result of the analysis often in the form of a list of differentially expressed genes (DEGs), which consists of individual genes that found differentially expressed in the dataset. This list of DEGs often failed to provide mechanistic insights regarding the biology of the phenotype of study [32].

This research focuses on the pathway-based analysis approach with support vector machine to improve gene selection and classification in cancer diseases. Pathway-based analysis has been applied to the gene expression data. Support vector machine is a machine learning that applied the computational approach in classifying the genes based on their features. In support vector machine, penalizedSVM has been implemented for gene selection process.

In this research, pathway-based analysis with support vector machine was introduced to predict the optimal gene classification result. In order to reduce the size of data, a LASSO algorithm has been used in support vector machine. There are two types dataset that used in this research which is lung cancer dataset and breast cancer dataset. Every gene expression data will be merged with pathway data. Thus, the process of selection and classification were done in sets of data. It can reduce the complexity of classification process and less time-consuming.

As a result, the accuracy of gene selection is produced, and comparison between previous methods are being made. By using lung cancer data, the accuracy of classification was higher, but for breast cancer data the accuracy was low. This shows that the types of data affect the performance of classification. Next, the selected informative gene was analyzed using the biological database in order to study the relation or contribution to cancer disease.

From the comparison, the accuracy performance of the proposed method was better in lung cancer data compared the previous methods. But for breast cancer data the

performance is not good as lung cancer data. As a conclusion, pathway-based analysis with support vector machine (LASSO) successfully classified the genes to improve the accuracy performance of gene selection and classification in cancer disease.

After that, the performance measurement is done by calculating the accuracy performance and validate the gene selection with the biological database. The implement did improve the accuracy performance of gene selection and classification process. The difference between previous works that used same data is more than 30%. This shows a good improvement to the current methods with an addition of pathway-based analysis in gene classification. As stated, the pathway-based analysis is the best analysis method in classification compared to previous methods.

ACKNOWLEDGMENT

This research is supported by Universiti Teknologi Malaysia through the Tier 1 research grants (Grant numbers: Q.J130000.2528.11H11).

REFERENCES

[1] M. Yamamoto, T. Wakatsuki, A. Hada, and A. Ryo, "Use of serial analysis of gene expression (SAGE) technology", Immunological Methods, 250(1-2), 45-66, 2001.

[2] R. K. Curtis, M. Oresic, and A. Vidal-Puig, "Pathways to the analysis of microarray data", Trends of Medicine, 23(8), 429-435, 2005.

[3] D. Nam and S. Y. Kim, "Gene-set approcah for expression pattern analysis", Briefings in Bioinformatics, 9(3), 189-197, 2008.

[4] T. C. Siang, T. W. Soon, S. Kasim, M. S. Mohamad, C. W. Howe, S. Deris, and Z. Ibrahim, " A Review of Cancer Classification Software for Gene Expression Data", International Journal of Bio-Science and Bio-Technology, 7(4), 89-108, 2015.

[5] L. Jin, Y. X. Zuo, Y. W. Su, L. X. Zhao, Q. M. Yuan, Z. L. Han, X. Zhao, D. Y. Chen, and Q. S. Rao. Pathway-based analysis tools for complex diseases: a review. Genomics Proteomics Bioinformatics. 12(5): 210-220, 2014.

[6] D. Gosh and A. M. Chinnaiyan, "Classification and selection of biomarkers in genomic data using LASSO", BioMed Research International, 2005(2), 147-154, 2005.

[7] D. G. Beer, S. L. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek and M. L. Lizyness, "Gene-expression profiles predict survival of patients with lung adenocarcinoma", Nature medicine, 8(8), 816-824, 2002.

[8] H. Farmer, N. McCabe, C. J. Lord, A. N. Tutt, D. A. Johnson, T. B. Richardson and N. M. Martin, "Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy", Nature, 434(7035), 917-921, 2005.

[9] C. F. Hardy, O. Dryga, S. Seematter, P. M. Pahl and R. A. Sclafani, "mcm5/cdc46-bobl bypasses the requirement for the S phase activator Cdc7p", Proceedings of the National Academy of Science, 94(7), 3151-3155, 1997.

[10] M. A. Madine, C. Y. Khoo, A. D. Mills and R. A. Laskey, "MCM3 complex required for cell cycle regulation of DNA replication in vertebrate cells", Nature, 375(6530), 421-424, 1995.

[11] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, and J. A. Todd, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls", Nature, 447(7145), 661-678, 2007.

[12] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies and M. S. Kuehn, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project", Nature, 447(7146), 799-816, 2007.

[13] C. Looft, S. Paul, P. D. Thomsen, M. Yerle, B. Brenig, and E. Kalm, "Isolation and assignment of the UDP-glucose pyrophosphorylase gene (UGP2) to porcine chromosome 3q21-q22 by FISH and by analysis of somatic cell and radiation hybrid panels", Cytogenic and Genome Research, 89(3-4), 154-155, 2000.

[14] C. D. Mitnick, S. S. Shin, K. J. Seung, M. L. Rich, S. S. Atwood, J. J. Furin and C. A. Bonilla, "Comprehensive treatment of the extensively drug-resistant tuberculosis, New England Journal of Medicine, 359(6), 563-574, 2008.

[15] C. R. Greenberg, C. L. Taylor, J. C. Haworth, L. E. Seargeant, S. Philips, B. Triggs-Raine and B. N. Chodirker, "A homoallelic prinatal (lethal) form of hypo-phosphatasia in Canadian Mennonites", Genomics, 17(1),215-127, 1993.

[16] D. B. Gould, F. C. Phalan, G. J. Breedveld, S. E. van Mil, R. S. Smith, J. C. Schimenti and S. W. John, "Mutation in Co14a1 cause prinatal cerebral haemorrhage and porencephaly". Science, 308(5725), 1167-1171, 2005.

[17] J. Favor, C. J. Gloeckner, D. Janik. M. Klempt, A. Neuhau-ser-Klaus, W. Pretsch and L. Quinranilla-Fend, "Type IV procollagen missense mutations associated with defects of the eye, vascular stability, the brain, kidney function and embryonic or postnatal viability in the mouse, Mus musculus: an extension of the Col4a1 allelic series and the identification of the first two Col4a2 mutant alleles", Genetics, 175(2), 725-736.

[18] J. M. Kerr, L. W. Fisher, J. D. Termine, M. G. Wang, O. W. McBride and M. F. Young, " The human bone sialoprotein gene (IBSP): genomic localization and characterization, Genomics, 17(2), 408-415, 1993.

[19] S. Landi, V. Moreno, L. Gioi-Patricola, E. Guino, M. Navarro, J. de Oca and Bellvitage Colorectal Cancer Study Group, "Association of common polymorphism in inflammatory genes interleukin (IL) 6, IL 8, tumor necrosis factor a, NFKB1, and peroxisome proliferator-activated receptor with colorectal cancer". Cancer research, 63(13), 3560-3566, 2003.

[20] M. Bredel, D. M. Scholtens, A. K. Yadav, A. A. Alvarez, J. J. Renfrow, J. P. Chandler and R. Ferrarese, "NFKBIA deletioan in glioblastomas", New England Journal of Medicine, 364(7), 627-637, 2011.

[21] N. Kerkar, N. Hadzic, E. T. Davies, B. Portmann, P. T. Donaldson, M. Rela and G. Mieli-Vergani, "De-novo autoimmune hepatitis after liver transplantation. The Lancet, 351(9100), 409-413, 1998.

[22] G. D'Urso, L. R. Marraccino, R. D. Marshak, & M. J. Roberts. Cell cycle control of DNA replication by a homologue from human cells of the p34cdc2 protein kinase. Science, 250(4982), 786-792, 1990.

[23] M. A. Bosch. Classical galactosaemia revisited. Journal of inherited metabolic disease, 29(4), 516-525, 2006.

[24] K. M. Tsaousidou, K. Ouahchi, T. T. Warner, Y. Yang, A. M. Simpson, G. N. Laing, & A. M. Patton. "Sequence alterations within CYP7B1 implicate defective cholesterol homeostasis in motor-neuron degeneration", The American Journal of Human Genetics, 82(2), 510-515, 2008.

[25] J. R. DeLorenzo, & H. F. Ruddle. "Glutamate oxalate transaminase (GOT) genetics in Mus musculus: linkage, polymorphism, and phenotypes of the Got-2 and Got-1 loci", Biochemical genetics, 4(2), 259-273, 1970.

[26] D. M. Hogarty, D. M. Norris, K. Davis, X. Liu, F. N. Evageliou, S. C. Hayes, & J. Keating. "ODC1 is a critical determinant of MYCN oncogenesis and a therapeutic target in neuroblastoma", Cancer research, 68(23), 9735-9745, 2008.

[27] A. K. Kumar, M. Kasthuraiah, S. C. Reddy, & D. C. Reddy. "Mn (OAc) 3· 2H 2 O-mediated three-component, one-pot, condensation reaction: an efficient synthesis of 4-aryl-substituted 3, 4-dihydropyrimidin-2-ones", Tetrahedron Letters, 42(44), 7873-7875, 2001.

[28] W. Xu, Z. L. Liu, M. Loizidou, M. Ahmed, & G. I. Charles. The role of nitric oxide in cancer. Cell research, 12(5-6), 311, 2002.

[29] D. Fukumura, S. Kashiwagi, & K. R. Jain. The role of nitric oxide in tumour progression. Nature reviews. Cancer, 6(7), 521, 2006.

[30] C. C. McDonald, & J. H. Stewart. Fatal myocardial infarction in the Scottish adjuvant tamoxifen trial. The Scottish Breast Cancer Committee. Bmj, 303(6800), 435-437, 1991.

[31] A. S. Everson, E. D. Goldberg, A. G. Kaplan, D. R. Cohen, E. Pukkala, J. Tuomilehto, & T. J. Salonen. Hopelessness and risk of mortality and incidence of myocardial infarction and cancer. Psychosomatic Medicine, 58(2), 113-121, 1996.

[32] P. Khatri, M. Sirota, and J. A. Butte. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 8(2): 1-10, 2012.