Solving Problems of Clustering and Classification of Cancer Diseases Based on DNA Methylation Data

A. Polovinkin[1], I. Krylov[1], P. Druzhkov[1], M. Ivanchenko[1], I. Meyerov[1], A. Zaikin[1, 2], N. Zolotykh[1]

[1] ***State University of Nizhni Novgorod, Institute of Information Technologies, Mathematics and Mechanics, Nizhni Novgorod, Russia***

[2] ***University College London Department of Mathematics, London, Great Britain***

Abstract

The article deals with the problem of diagnosis of oncological diseases based on the analysis of DNA methylation data using algorithms of cluster analysis and supervised learning. The groups of genes are identified, methylation patterns of which significantly change when cancer appears. High accuracy is achieved in classification of patients impacted by different cancer types and in identification if the cell taken from a certain tissue is aberrant or normal. With method of cluster analysis two cancer types are highlighted for which the hypothesis was confirmed stating that among the people affected by certain cancer types there are groups with principally different methylation pattern.

Introduction

In spite of all advances of modern medicine introduction of new diagnosis and treatment methods, cancer disease and mortality rates constantly keep steadily growing all over the world. Unfortunately show signs of clinical symptoms indicate the extensive-stage disease that is why pre-existing cancer detection seems to be the most promising approach. According to the international practices the selection of risk groups and screening study are the most prospective early detection of malignant neoplasms. This article discusses a new approach to the problem of early cancer diagnosis based on searching and analysis of low-level factors which can witness the development and existence of oncological disease in gene domain.

At the moment the genetic risk factors of cancer are practically elusive, however, their identification is supposed to be breakthrough advance which will result in much more effective screening methods and early diagnostics.

Epigenetic changes are DNA modifications resulted not from variations of nucleotide sequence, i. e. the variations occur not in genes but in external factors directly related to gene activities. One type of epigenetic changes is DNA methylation when methyl group ($-CH_3$) joins certain molecule regions.

Aberrant structure of DNA methylation is one of the essential cancer signs, which enables its early diagnosis [7], however the exact role of this data in cancer genesis and clinical prediction remains elusive. Cancer is characterized by both hypermethylation (increase of methylation) and hypomethylation (decrease) of DNA. However, cancer can be witnessed not only by variation of mean level of gene methylation. The hypothesis has been proposed according to which dysregulation of stem cell genes results from aberrant variability (dispersion) of intragenic DNA methylation. This correlates with the fact that not only methylation level but also variability in certain genomic locations may be highly relevant to cancer development [6]. In particular, it has been shown that the increased stochasticity and variability in regions where the methylation level changes with cancer, results in aberrant and modified gene expression, thus explaining tumor heterogeneity [3]. Also some authors have shown that the markers reflecting differential variability of DNA methylation features may provide for better diagnosis and risk assessment of precancer genesis [8, 9].

Though the importance of study of intragenic and intergenic DNA methylation structure is clearly understood at the moment only modifications between different genomes have been studied but the problem how the remodeling of intragenic and intergenic DNA methylation is related to the origin of carcinomas. This article deals with the problem how to identify the gene group, methylation patterns of which significantly change with emerging of cancer disease, and analyses the application efficiency of certain DNA methylation methods to solve problems of classification and identification of essential features. Using methods of cluster analysis the hypothesis is studied which states that among the people affected by the same cancer type there are different groups which might be treated with potentially different methods. Authors analyze accuracy of solving problems of binary and multiclass classification between different cancer types under application of ensembles of decision trees.

Methods and data

As initial data for supposed study we propose to use inspection results of examinees from the international data base The Cancer Genome Atlas [10], which contain information about methylation level received with TheIlluminaInfinium HumanMethylation450 BeadChip [5]. Data contain circa 485000 loci per genome the intragenic location for 330000 of which is known as well as the name of related gene. Thus 15-17 loci per gene are available. These data are available for 13 different cancer types (Bladder Urothelial Carcenoma, BLCA; Breast Invasive Carcenomia, BRCA; etc). The number of objects related to each cancer type varies from tens to hundreds.

The following methods and markers are proposed to study intragenic DNA methylation structure. The first marker group does not depend on probe sequence inside the gene or related gene region. This marker group includes the mean value for gene methylation and dispersion. The second group includes markers considering the intragenic probe location. These markers are the mean value of outlier derivative, degree of spatial outlier asymmetry, degree of deviation from line linking methylation levels at the gene ends. Except the computation of their value, for the markers of the first and second groups Z-score is applied that is computation of deviation from the mean value of proper degree for all samples from “normal” selection, measured in mean-square deviation. The value obtained in such manner will be an instability degree for methylation outlier for the corresponding gene.

As classifier we suggest to apply the decision trees [4] and their ensembles (in particular Random Forest [1]). Among the advantages of Random Forest the following features shall be mentioned: high quality of obtained models, similar to SVM and boosting, and better compared with neural nets [2], ability to effectively process data with large number of features and classes, invariance for monotonic transformations of features values, possibility to process both continuous and discrete features, presence of methods to evaluate the importance of specific features in the model. Moreover Random Forest model enables estimation of generalization error on-the-fly during its training (out-of-bag error [1]).

Results of Numerical Experiments

Classification results

From the practical point of view it is desirable to consider two types of problems: to define if a person is affected or normal using tissue samples of certain organ; to distinguish between different cancer types and normal cells. The developed measures of intragenic methylation are used as a sample description. Classification accuracy is expressed in terms of misclassified samples fraction and estimated using the out-of-bag error of the Random Forest model. As Table 1 shows, the achieved accuracy for problems of binary classification makes up from 93% to 100% and for problem of multiclass classification (Table 2) is 96.5% which enables practical application of these results.

Table 1. Results of Binary Classification for 13 Cancer Types

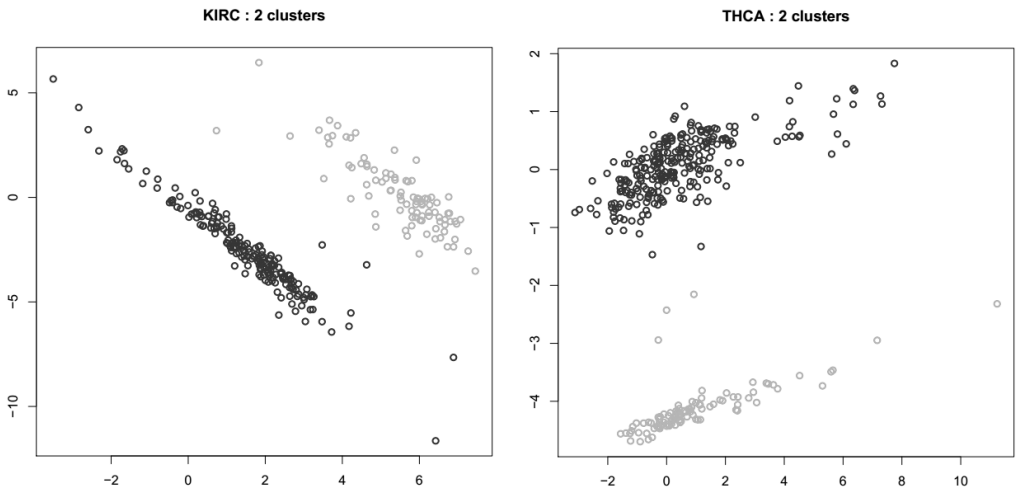
Cancer type	Accuracy	Type I error	Type II error
BLCA	0.965	0.166	0.008
BRCA	0.978	0.133	0.009
COAD	0.989	0.105	0
NHSC	0.975	0.12	0.006
KIRC	1.000	0	0
KIRP	0.984	0	0.023
LIHC	0.966	0.02	0.051
LUAD	1.000	0	0
LUSC	1.000	0	0
PRAD	0.933	0.183	0.04
READ	0.98	0.25	0
THCA	0.968	0.26	0.003
UCEC	0.983	0.139	0

Table 2. Misclassification Table to Classify Individuals Affected by Different Cancer Types.

	HEALTHY	BLCA	BRCA	COAD	HNSC	KIRC	KIRP	LIHC	LUAD	LUSC	PRAD	READ	THCA	UCEC
HEALTHY	0.96	0	0.01	0	0	0	0	0.02	0	0	0	0	0.01	0
BLCA	0.01	0.85	0.01	0	0.09	0	0	0	0	0	0.04	0	0	0
BRCA	0.01	0	0.99	0	0	0	0	0	0	0	0	0	0	0
COAD	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0
HNSC	0.01	0	0	0	0.97	0	0	0	0	0	0.02	0	0	0
KIRC	0.03	0	0	0	0	0.95	0.01	0	0	0	0.01	0	0	0
KIRP	0.02	0.03	0	0	0	0.13	0.82	0	0	0	0	0	0	0
LIHC	0.01	0	0	0	0	0	0	0.99	0	0	0	0	0	0
LUAD	0.07	0.01	0	0	0	0	0	0	0.92	0	0	0	0	0
LUSC	0.01	0	0	0	0	0	0	0	0	0.97	0.02	0	0	0
PRAD	0	0	0	0	0.08	0	0	0	0	0.08	0.84	0	0	0
READ	0.23	0	0	0.03	0	0	0	0	0	0.04	0	0.70	0	0
THCA	0.07	0	0	0	0	0	0	0	0	0	0	0	0.93	0
UCEC	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00

Clustering Results

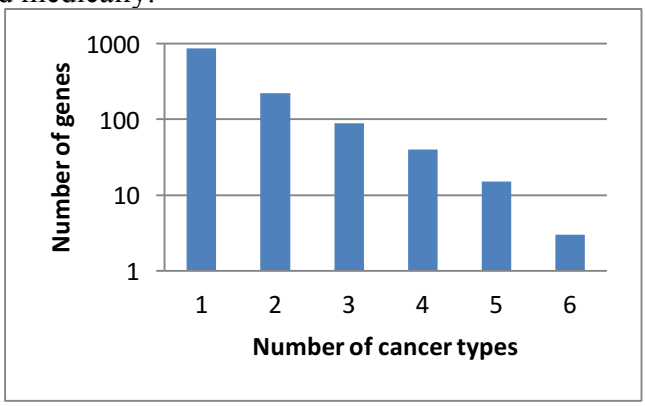
The hypothesis has been proposed that within the same type of oncological diseases there are a number of different groups which theoretically speaking might be treated differently. To test this hypothesis the following has been proposed – for each type of cancer to divide classifying description of cells, related to specified type, in groups applying the methods of cluster analysis. The degree of mean level of methylation for the gene has been used as a classifying description, the k-means algorithm has been used for clustering [4] (for practical reasons 2-3 clusters has been supposed to be available). The experiment has shown that there is a distinct separation of individuals affected by KIRC and THCA cancer types in clusters. So the further analysis of these results is required from practical point of view.



Pic.1 Clustering Results with k-means Algorithm for KIRC and THCA Cancer Types

Selection of important features

One of the problems of practical importance is the existence of specific genes responsible for the development of one or another oncological disease. In this article the importance of each gene was analyzed regarding its usefulness to solve problems of binary classification (if a cell of certain tissue is malignant or normal). A set of features has been selected for each type of cancer thereby each feature ensures the quality of binary classification (cross-validation error in decision trees of depth 1) exceeding a certain threshold (within this article this value equals 0.9). The lists of significant genes for all cancer types obtained in this manner have been combined. The picture 2 shows the overall statistics (the number of genes simultaneously important for classification of a certain number of cancer types). As the diagram shows there are relatively small sets of genes important for classification of several cancer types. In the future the obtained results shall be analyzed medically.



Pic.2 Number of Genes Important for Classification of Several Cancer Types

Conclusion

Within this article the gene group methylation patterns of which significantly change with emerging of cancer disease has been identified. Using methods of cluster analysis the hypothesis has been studied which states that among the people affected by the same cancer type there are different groups. Two cancer types have been outlined for which the hypothesis has been confirmed. The obtained solution accuracy for the problems of binary and multiclass classification enables the practical application of the results.

References

1. L. Breiman. Random Forests. Machine Learning 45 (1), 2011, p 5-32.

2. R. Caruana, A. Niculescu-Mizil. An Empirical Comparison of Supervised Learning Algorithms Using Different Performance Metrics. ICML '06 Proceedings of the 23rd international conference on Machine learning. p. 161-168.
3. K.D. Hansen, et al. Increased methylation variation in epigenetic domains across cancer types. Nat Genet, 2011. 43(8): p. 768-75.
4. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. — Springer, 2001.
5. Infinium HumanMethylation450 BeadChip | Illumina [http://products.illumina.com/products/methylation_450_beadchip_kits.ilmn]. 21.09.2014
6. A.E. Jaffe, et al. Significance analysis and statistical dissection of variably methylated regions. Biostatistics, 2012. 13(1): p. 166-78.
7. P.A. Jones, S.B. Baylin. The epigenomics of cancer. Cell, 2007. 128(4): p. 683-92.
8. A.E. Teschendorff, M. Widschwendter. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. Bioinformatics, 2012. 28(11): p. 1487-94.
9. A.E. Teschendorff, et al. Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. Genome Med, 2012. 4(3): p. 24.
10. The Cancer Genome Atlas – Cancer Genome – TCGA [http://cancergenome.nih.gov/]. 21.09.2014
11. M. Widschwendter, et al. Epigenetic stem cell signature in cancer. Nat Genet, 2007. 39(2): p. 157-8.