ORIGINAL ARTICLE

THE JOURNAL OF NEUROBEHAVIORAL SCIENCES
NÖRODAVRANIŞ BİLİMLERİ DERGİSİ

Year : 2015
Volume : 2
Issue Number : 3
Doi Number : 10.5455/JNBS.1446109872

Article history:
Received 29 October 2015
Received in revised form 4 November 2015
Accepted 12 November 2015

CLASSIFICATION OF SCHIZOPHRENIA PATIENTS BY USING GENOMIC DATA: A DATA MINING APPROACH

ŞİZOFRENİ HASTALARININ GENOMİK VERİ KULLANARAK KLASİFİYE EDİLMESİ; VERİ MADENCİĞİ YAKLAŞIMI

Kaan Yilancioglu[*1] , Muhsin Konuk[2]

JNBS 2015 Published by Üsküdar University www.jnbs.org

Abstract

Genomic information obtained from robust analysis methods such as microarray and next generation sequencing reveals underlying disease mediating factors and potential diagnostic biomarkers. Data mining methods have been widely chosen for classification and regression studies of health sciences as well as other disciplines since the beginning. In the present study, public Gene Expression Omnibus (GEO) genome wide expression dataset (ID: GSE12679) consisting of mRNA transcripts of post-mortem brain tissues in schizophrenic and normal patients were analyzed by using Multilayer Perceptron Neural Network (MLP NN) algorithm. A set of most differentially expressed genetic features (p<0.001) were used for creating the classifier which can predict disease states in test set with ~82% accuracy. Differentially expressed genes used as classifying biomarkers gain utmost importance for revealing hidden underlying genetic factors associated with important psychiatric diseases. We could also suggest that such data mining tools might be applicable for developing genome-based diagnostic tools.

Keywords: Data Mining, Schizophrenia, Neural Network

Özet

Yeni nesil sekanslama ve mikrodizilim/çip analizlerinden elde edilen genomik veriler, çeşitli hastalıkların altında yatan moleküler sebepleri açığa çıkarmakta ve potansiyel tanı biyomarkörlerinin tanımlanmasını olanaklı kılmaktadır. Ortaya çıktığından buyana klasifikasyon ve regresyon analizlerini temel alan veri madenciliği metotları çeşitli çalışmalarda sıklıkla kullanılmaktadır. Bu çalışmada, NCBI GEO veri bankasından elde edilen, post-mortem beyin dokularından elde edilmiş beyin dokularının mRNA transcript analiz verileri MLP nöral ağ algoritması kullanılarak incelenmiştir. Çalışmada, dokular arasında ki transkripsiyon düzeyleri farkları analiz edilmiştir. Transkripsiyon düzeyleri farkı kullanılarak oluşturulan klasifikatör, şizofrenik ve normal hasta gruplarını %82 kesinlikle tahmin etmiştir. Psikiyatrik hastalıkların altında yatan etmenlerin aydınlatılmasında genetik biyomarkörlerin rolü günden güne önem kazanmaktadır. Ayrıca bu çalışmada kullanılan yöntemlere benzer yöntemlerin, genom temelli tanı yöntemlerinin bulunmasında katkı sağlayacağı düşünülmektedir.

Anahtar Kelimeler: *Veri Madenciliği, Şizofreni, Nöral Ağ*

1. Introduction

Schizophrenia is a chronic brain disease that impairs normal behavior, speech and thinking processes. Diagnosis of the disease mostly relies on clinical examinations. The disease has many subcategories with various complex symptoms reflecting its biologically heterogeneous characteristics as a mental disease.

There has been an increasing effort for utilization of brain visualization and genetic variation analysis to find potential biomarkers for better understanding of underlying pathologies of brain diseases such as schizophrenia disorder. These kind of classifying perspectives in using disease related biomarkers including genomic information has gained utmost importance. Since the strong genetic association was demonstrated, more research activities have been held by using genomics combined with upgraded statistical methods (Orrù, Pettersson-Yeo, Marquand, Sartori & Mechelli, 2012). In this decade, use of machine learning algorithms, analyzing genomic data obtained from different platforms such as Microarray and Next Generation Sequencing (NGS) has gained importance. Machine learning algorithms have been suggested to be successfully utilized in training classifiers to decode genetic profiles of interest from genomic data (Lu & Han, 2003). Presently, limited work has been carried out using genotypic information to classify patients with brain disorders from normal subjects. One example demonstrated that SVMs can classify both bipolar and schizophrenia from normal subjects with high accuracy by

[*1] Address for Correspondence: Department of Bioengineering, Faculty of Engineering and Natural Sciences, Üsküdar University, Istanbul, Turkey. E-mail: kaan.yilancioglu@uskudar.edu.tr
[2]Department of Molecular Biology and Genetics, Faculty of Engineering and Natural Sciences, Üsküdar University, Istanbul, Turkey.

using gene expression data (Struyf, Dobrin & Page, 2008). It is believed in that the schizophrenia may develop as a result of reciprocal action of genetic predisposition and environmental factors. Patients who were diagnosed with schizophrenia have immediate relatives with a history which clearly reflect the importance of genetic factors in development of the disease. However, even monozygotic twins have only about 42% concordance for the disease (Johnson, 2000). High-throughput methods such as microarray and recently next generation sequencing have generated vast information on numerous disease states. Advanced computational power in parallel with the advances on data mining tools enabled us to analyze these huge datasets. Recent genomic data obtained from microarray analysis on schizophrenia might help finding new genomic biomarkers and more importantly classify disease states for better diagnosis by the use of data mining approaches. In this study we present a supervised machine learning method to classify schizophrenic individuals that incorporates publicly available microarray gene expression data.

2. Data and Analysis Methods

2.1. Data

Publicly available microarray expression data set (ID: GSE12679) was obtained from the GEO database (Harris et al., 2008). The data set was divided into two groups as endothelial and neuronal cell section groups. The samples belong to 7 schizophrenia and 9 non-schizophrenic normal patients for training and 5 schizophrenia and 6 control patients for testing respectively. For each subject, demographic and clinical information were described in the original paper (Harris et al., 2008). The expression data was obtained using Affymetrix Human Genome U133A GeneChip plus 2.0 oligonucleotide arrays containing 54,676 probe sets (Affymetrix, Santa Clara, CA). Probe level data was summarized using the robust multi-array average (RMA) method (Wu & Irizarry, 2005). The data set includes the RMA value of each probe set as a numerical feature. All computational analyses were done by using R (v3.1.2) (R-project.org, 2015). For microarray data preparation "LIMMA" package (Ritchie et al., 2015) was used.

2.2. Multilayer Perceptron Neural Network (MLP NN)

MLP neural networks are typically trained with back propagation (BP) algorithm. BP is an application of the gradient method or other numerical optimization methods to feed-forward ANN (Artificial Neural Network) so as to minimize the network errors. It is the most popular method for performing supervised learning in ANN research community. This dataset was given as an input to the most popular data mining tool WEKA 3.6 for analyzing the correct accuracy prediction of various MLP ANN algorithm (Cs.waikato.ac.nz, 2015).

2.3. Evaluation of Classifier

Performance of the classification technique was evaluated by means of Receiver Operating Characteristic (ROC) curves. ROC analysis allows to simultaneously compare classifier for different misclassification costs and class distributions. It is based on the notions of "true positive rate" (TP, also known as sensitivity or recall) and "false positive rate" (FP, also known as 1.0 – specificity). Area under the curve (AUC) value was calculated for the MLP NN classifier in order to evaluate the power of discriminative power of the method. For evaluation a discrete endothelial cell derived microarray gene expression data (top 500 DE genes with p value < 0.05) was used for training and neuronal cell derived microarray gene expression data was used for the test set.

3. Results and Discussion

3.1. Expression Analysis

2403 differentially expressed probes demonstrating p-value<0.05 (without multiple testing correction) were extracted among 54,676. First 50 most differentially expressed probes (Figure 1) were selected and pathway analysis was applied by using a free, open-source, curated and peer reviewed pathway database, Reactome (Reactome.org, 2015). Accordingly, genes namely PDK1 (226452_at) and GRIK1 (214611_at) were found to be directly related to neuronal system and they were further investigated. It was previously demonstrated that PDK1 CC genotype results to enhanced prevalence of schizophrenia. Moreover, decreased parietal P300 amplitude, which is a well-studied schizophrenic endophenotype and glutamate and glutamine concentrations are increased in the frontal lobe of PDK1 dysmorphic mice and human PDK1 CC individuals (Lang et al., 2015).

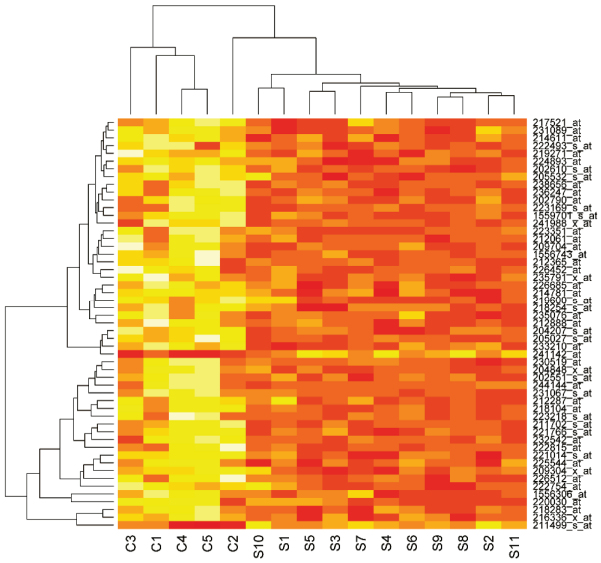

Figure 1: Heatmap of top 50 differentially expressed probes used for pathway analysis.

Recent studies have showed that a reduced abundance of Akt1 which is a crucial component of PDK1-Akt signaling in the brain was found to be significantly associated with

JNBS 2015 Published by Üsküdar University www.jnbs.org

schizophrenia, and Akt1 deficiency results to greater sensitivity of the disruption of sensorimotor gating mediated by amphetamine (Emamian, Hall, Birnbaum, Karayiorgou & Gogos, 2004). Another important finding is that the deficiency of GABAergic neurons in the prefrontal cortex is shown to be associated with schizophrenia. Loss of neocortical GABAergic neurons related to the absence of PDK1-Akt signaling was suggested to be linked with the pathogenesis of schizophrenia (Gonzalez-Burgos & Lewis, 2012). Glutamatergic function is one of the major hypotheses for schizophrenia. Within the glutamate system, the glutamate receptor ionotropic kainate-1 (GRIK1) gene is suggested to be particularly involved in schizophrenia. In this manner, the reduction of GRIK1 in the dorsolateral prefrontal cortex of schizophrenia patients was previously reported (Hirata et al., 2012).

3.2. MLP NN Classification and Classifier Performance

Microarray gene expression data (NCBI-GEO ID: GSE12679) consisting of both human endothelial and neuronal cells isolated from postmortem dorsolateral prefrontal cortex were used to generate the MLP NN classifier. A set of 500 top DE ($p < 0.05$ without multiple testing correction) genes between control (n=5) and schizophrenia (n=11) endothelial cells were used to train the classifier. Testing data used the same 500 DE gene set derived from neuronal cells as classifying attributes among 6 control and 5 schizophrenia samples. MLP NN model classifies schizophrenic and control patients at very high accuracy on testing data as ~82%. AUC was found to be 0.7. Other accuracy indicators were shown in Table 1.

The biggest limitation of the study was the shortage of sample size. Hence, the study should be considered as pilot study and sample size should be increased in further studies.

Table 1: Detailed accuracy measures of test set using MLP NN classifier. C represents control whereas S represents schizophrenia groups. On the left confusion matrix of the classifier on testing data is shown.

S	C		TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
5	0	S	0.667	0	1	0.667	0.8	0.7	C
2	4	C	1	0.333	0.714	1	0.833	0.7	S
Confusion Matrix			0.818	0.152	0.87	0.818	0.815	0.7	Weighted Avg.

4. Conclusions

Differentially expressed genetic markers used as classifying features in this MLP prediction analysis might be used for revealing important genes and gene families associated with schizophrenia disease and more importantly the classifier method might be applicable to developing effective Microarray-based diagnostic tests for this important psychiatric disease.

References

Cs.waikato.ac.nz,. (2015). Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Retrieved 7 December 2015, from http://www.cs.waikato.ac.nz/ml/weka/

Emamian, E., Hall, D., Birnbaum, M., Karayiorgou, M., & Gogos, J. (2004). Convergent evidence for impaired AKT1-GSK3β signaling in schizophrenia. Nature Genetics, 36(2), 131-137. http://dx.doi.org/10.1038/ng1296

Gonzalez-Burgos, G., & Lewis, D. (2012). NMDA Receptor Hypofunction, Parvalbumin-Positive Neurons, and Cortical Gamma Oscillations in Schizophrenia. Schizophrenia Bulletin, 38(5), 950-957. http://dx.doi.org/10.1093/schbul/sbs010

Harris, L., Wayland, M., Lan, M., Ryan, M., Giger, T., & Lockstone, H. et al. (2008). The Cerebral Microvasculature in Schizophrenia: A Laser Capture Microdissection Study. Plos ONE, 3(12), e3964. http://dx.doi.org/10.1371/journal.pone.0003964

Hirata, Y., Zai, C., Souza, R., Lieberman, J., Meltzer, H., & Kennedy, J. (2012). Association study of GRIK1 gene polymorphisms in schizophrenia: case-control and family-based studies. Human Psychopharmacology: Clinical And Experimental, 27(4), 345-351. http://dx.doi.org/10.1002/hup.2233

Johnson, R. (2000). Prevention of post-herpetic neuralgia: Can it be achieved?. Acta Anaesthesiol Scand, 44(8), 903-905. http://dx.doi.org/10.1034/j.1399-6576.2000.440801.x

Lang, U., Ackermann, T., Wolfer, D., Schubert, F., Sohr, R., & Hörtnagl, H. et al. (2015). Phosphoinositide-Dependent Protein Kinase 1 (PDK1). Zeitschrift Für Psychologie, 223(3), 165-172. http://dx.doi.org/10.1027/2151-2604/a000217

Lu, Y., & Han, J. (2003). Cancer classification using gene expression data. Information Systems, 28(4), 243-268. http://dx.doi.org/10.1016/s0306-4379(02)00072-8

Orrù, G., Pettersson-Yeo, W., Marquand, A., Sartori, G., & Mechelli, A. (2012). Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. Neuroscience & Biobehavioral Reviews, 36(4), 1140-1152. http://dx.doi.org/10.1016/j.neubiorev.2012.01.004

Reactome.org,. (2015). Reactome | Pathway Browser. Retrieved 7 December 2015, from http://www.reactome.org/PathwayBrowser/#/

Ritchie, M., Phipson, B., Wu, D., Hu, Y., Law, C., Shi, W., & Smyth, G. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47-e47. http://dx.doi.org/10.1093/nar/gkv007

R-project.org,. (2015). R: The R Project for Statistical Computing. Retrieved 7 December 2015, from http://www.R-project.org/

Struyf, J., Dobrin, S., & Page, D. (2008). Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia. BMC Genomics, 9(1), 531. http://dx.doi.org/10.1186/1471-2164-9-531

Wu, Z., & Irizarry, R. (2005). Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays. Journal Of Computational Biology, 12(6), 882-893. http://dx.doi.org/10.1089/cmb.2005.12.882