Science Signpost Publishing
Biological and Chemical Research, 28-34 | Science Signpost Publishing

Microarray Gene Expression Data-Based Bioinformatics Method for the Diagnosis of Male Hypertension

Nazli Ucunoglu[1], Arinze Akutekwe[2], Turgay Isbir[3]

1 Department of Molecular Medicine, Institute of Health Sciences, Yeditepe University, Kayısdag, 34755, Istanbul, Turkey.

2 Bio-Health Informatics Research Group, Faculty of Technology, De Montfort University, LE1 9BH, Leicester,UK.

3 Department of Medical Biology, Faculty of Medicine, Yeditepe University, Kayısdag, 34755, Istanbul, Turkey.

Received: October 14, 2016 / Accepted: November 15, 2016 / Published: January 25, 2017

Abstract: Hypertension is a chronic medical condition that the blood pressure in the arteries is elevated. Hypertension can lead to damaged organs, as well as several illnesses, such as renal failure (kidney failure), aneurysm, heart failure, stroke, or heart attack. In our investigation, ten subsets were designed for male hypertension patient and control group. In this paper we apply t-test and entropy feature selection methods using 2fold and 5fold cross validation as our model selection methods with K-Nearest neighbour classifier. Among these groups, 3 number of biomarkers set were chosen (1,3,9) for 4 tables (t-test; 2-fold and 5-fold; entropy; 2-fold and 5-fold). From these biomarker sets which has the highest accuracy which is the measurement used for the classifier assessment was analysed and taken to the best models for each sub-set table. Each sub-set tables were analysed with each other and we tried to find the most appropriate biomarker. The defined biomarker was searched within database in order to find relationship with the illness. Consequently, highly recurrent and highly accurate candidate genes can be further analysed for becoming a biomarker. Further analysis (both database and wet study) can be suggested for the highly recurrent genes like Hs. 683236 (null), Hs. 475902, 420541, 656129, 647705 and 657792.

Keywords: Hypertension, t-test, entropy, K-Nearest neighbour classifier, biomarker.

1. Introduction

Hypertension has been one of the most critical cardiovascular diseases among both men and women and also has an effect of an increased risk of stroke, myocardial infarction, and mortality [1, 2] which are the crucial risk factors that associated with cardiovascular diseases [2, 3]. This disease has a major role in blood

Corresponding author: Nazli Ucunoglu, Department of Molecular Medicine, Institute of Health Sciences, Yeditepe University,Kayisdag, 34755, Istanbul, Turkey. E-mail: naz_zli@hotmail.com.

pressure (BP) variation as well as it is a multifactorial disease in which, genetic factors and environmental effects play role to enhance the disease [4, 5]. Recently, a group of genes have been discovered to influence the mechanism of blood pressure. The result showed that there is a strong relation between these genes and the disease, but the rest of groupsdiscovered have not been viewed in an association with the evidence [6, 7, 8].

Up to now, there have been several investigations that indicated biomarkers which have own pathways and these biomarkers have not been in any relationship between them [9]. It means, it is still unknown whether multiple biomarkers are independently associated with hypertension risk. The interaction between these biomarkers and hypertension could become definite by the measurement of other biomarkers [10-12].

2. Materials and Methods

Feature Selection and Classification

In high-dimensional settings, feature selection helps to select the most important features in order to reduce feature size, creating best class discrimantory information and avoid over-fitting. In this study, we employ two different kind of featureselection methods.

i) Student's T-test

The test statistic is the number of standard error by which two sample means are separated. It is a hypothesis test which follows a normal distribution if the null hypothesis is supported and can be used to determine if two sets of data are significantly different from each other (two-sample independent or Welch's test) [13]. For feature selection, it is used to determine if the distribution of values of a feature for two different classes are distinct. If distinct, they are included in the feature vector for the classifier training. Mathematically,

$$t = \frac{difference\ between\ the\ two\ means}{std\ error\ of\ the\ difference} \tag{1}$$

$$= \frac{\overline{Y}_A - \overline{Y}_B}{Se_{diff}} \tag{2}$$

For two non-correlated (independent) variables, the variance of the difference is the sum of the separate variances. Therefore the standard error of the difference between two sample means can be written as:

$$Se_{diff} = \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}} \tag{3}$$

where S^2 is the variance and n is the sample size. A and B are two independent samples.

ii) Relative Entropy

The relative entropy or the Kullback-Leibler divergence between two probability distributions on a random variable is a measure of the distance between them [14]. For two probability distributions $p(x)$ and q(x) over a discrete random variable X, the relative entropy given by $D(p||q)$ is defined as:

$$D(p \,||\, q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{4}$$

Given that $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$. The relative entropy is always non-negative and is zero if and only if $p = q$.

iii) K-Nearest Neighbour Classifier

These are memory-based non-parametric classifiers that require no model to be fit. Given a query point x_0, we find the k training points $x_{(r)}$, $r = 1,\ldots, k$ closest to x_0, and then classify using majority vote among the k neighbours. The distance between neighbours is usually taken to be the Euclidean distance and features are standardized to have mean zero and variance 1. KNN has been successful in many classification problems such as handwritten digits, satellite image scenes and EKG patterns [15].

3. Results and Discussion

Data was collected from the paper called "A Neural Network Model for Constructing Endophenotypes of Common Complex Diseases - An Application to Male Young-onset Hypertension Microarray Data" to determine best predictors for hypertension, model evaluation using 2-fold and 5-fold CV was performed on both ttest and entropy. The tables show the selected genes and percentage accuracy.

Table I. Table Showing Two Fold Features With Highest Accuracy

Genes			
Ttest		**Entropy**	
Genes (UniGene ID)	**Acc. (%)**	**Genes (UniGene ID)**	**Acc. (%)**
Hs. 18912	60.26	Hs. 253726	66.67
Hs. 662191	60.26		
Hs. 591908	60.26		
Hs. 594099	60.26		

Only one feature selected by entropy using 2 fold had better accuracy that those selected by the ttest. 5 features selected by ttest using 5 fold had better accuracy that those selected by the entropy. According to the provided data for 2-fold, 5 different genes listed below in the Table I and 11 different genes shown in the Table II can be used for further analysis.

Table II. Tables Showing Five Fold Features With Highest Accuracy

Genes			
Ttest		**Entropy**	
Genes (UniGene ID)	**Acc. (%)**	**Genes (UniGene ID)**	**Acc. (%)**
Hs. 475902	77.42	Hs. 683236	70.00
Hs. 420541	77.42	Hs. 516484	70.00
Hs. 656129	77.42	Hs. 534373	70.00
Hs. 647705	77.42	Hs. 289123	70.00
Hs. 657792	77.42	Hs. 146809	70.00
		Hs. 411480	70.00

Table III. Wilcoxon Test ResultFor Two-Fold Cross Validation For Both Ttest And Entropy

2-fold		
Ttest		**Entropy**
P	0.2095	0.2095
W	33	33
Mu	0.5740	0.5911
Mean	0.5696	0.5961
sd	0.0360	0.0342

Wilcoxon test result showed that, when compared 2-fold and 5- fold for test and entropy, entropy 5-fold had the highest accuracy according to the mean and the median indicated in the Table III and Table IV.

Table IV. Wilcoxon Test Result For Five-Fold Cross-Validatıon For Both Ttest And Entropy

5-fold		
Ttest		**Entropy**
P	0.3226	0.3226
W	36.5	36.5
Mu	0.5645	0.6129
Mean	0.5864	0.6079
sd	0.0817	0.0550

According to these models, thioredoxin reductase gene (%66.67) had the highest accuracy in the entropy two-fold test. As Table II showed that Hs. 475902, Hs. 420541, Hs. 656129, Hs. 647705 and Hs. 657792 (gene names respectively cAMP regulated phosphoprotein, KIAA1202 protein, homo sapiens cDNA FLJ36210fis, coiled-coil domain containing 19 and homo sapiens transcribed sequence) features had the highest accuracy as a group of genes with the accuracy of %77.42 in the ttest. All these genes are presented as biomarkes candidates. They can be used for further database validation and further wet validation.

On the other hand, as the 5-fold CV of ttest implied, first 9 genes analyzed in one subset, has succeeded 77% for testing and provided the highest result. These 9 different genes in this column can be used for further analysis. In addition to the cluster of first nine genes, the gene APOBEC3F, presenting the highest accuracy rates at other subsets can be used for further database and wet study validations. It is interesting that there is no gene common in both 2 fold and 5 fold analysis. This investigation needs more finding in order to find out common genes and to be more clear to associate with stress.

As we observe in Table I, feature selection using the entropy for 2-fold as a cross validation tool supplied approximately 67% accuracy result. The thioredoxin reductase gene (Hs.235726) had the highest result for accuracy rate in 66,67%. This gene is a very important candidate as a biomarker. Because the gene itself provides 66.67 % predictive results, whereas 9 genes given in different subset, which includes the same gene and 8 others, can only provide around 62% predictive results. So probably it would be much appropriate to conduct both database and wet study validation basing on this gene.

A further analysis on the relation between cAMP regulated phosphoprotein, KIAA1202 protein, homo sapiens cDNA FLJ36210fis, coiled-coil domain containing 19 and homo sapiens transcribed sequence and hypertension could give us oppurtunity to use these gene as biomarkers. Their wet study and pathway

analysis through databases should surely be conducted. Further analysis both database and wet study can be suggested for the highly recurrent gene Hs. 683236 which is currently null.

In Table II, we acquired that accuracy results were provided around 70% - 77.42%. Hs. 475902, 420541, 656129, 647705 and 657792 could be used as biomarker. Because the group of gene provides 77.42% predictive results , however the other genes studied together for 5 fold entropy, can only provide 70% predictive results. 5 genes in the subset of 5fold CV ttest can be ranked according to their recurrency in the general testing results.

Consequently, highly recurrent and highly accurate candidate genes can be further analysed for becoming a biomarker. Further analysis (both database and wet study) can be suggested for the highly recurrent genes like Hs. 683236 (null), Hs. 475902, 420541, 656129, 647705 and 657792.

4. Conclusion

In this paper, high-quality biomarker selection for hypertension diagnosis have been carried out using ttest and Entropy feature selection techniques and K-Nearest Neighbour classifier. The result showed that this genes and selected by that method was found to be highly associated with hypertension disease. Also this genes might play significant role in diagnosing the disease.

Further investigations will be carried out in the databases to find out the relationships between these selected biomarkers and the disease. Other methods such as support vector machine and neural network will be used to conduct further analysis.

References

[1] CY. Hung, KY. Wang, TJ. Wu, YC. Hsieh, JL. Huang, El-W. Loh, CH. Lin, Resistant hypertension, patient characteristics, and risk of stroke, *PLoS One*. Aug 4; 9 (8) 2014.

[2] CM. Lawes, S. Vander Hoorn, A. Rodgers, International Society of Hypertension, Global burden of blood-pressure-related disease, *Lancet* 371 (2008): 1513-1518.

[3] S. Yusuf, S. Hawken, S. Ounpuu, T. Dans, A. Avezum, F. Lanas, M . McQueen, A . Budaj, P . Pais, J. Varigos, L . Lisheng; INTERHEART Study Investigators, Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study, *Lancet* 364 (2004): 937-952.

[4] Kh. Dhanachandra Singh, A. Jajodia, H. Kaur, R. Kukreti, M. Karthikeyan, Gender specific association of RAS gene polymorphism with essential hypertension: a case-control study, *Biomed Res Int*. 2014: 538053 2014.

[5] J. Baima, M. Nicolaou, F. Schwartz, AL. DeStefano, A. Manolis, I. Gavras, C. Laffer, F. Elijovich, L. Farrer, CT. Baldwin, H. Gavras, Evidence for linkage between essential hypertension and a putative locus on human chromosome 17, *Hypertension*. July; 34 (1) (1999): 4-7.

[6] X. Jeunemaitre, F. Soubrier, YV. Kotelevtsev, RP. Lifton, CS. Williams, A. Charru, SC. Hunt, PN. Hopkins, RR. Williams, JM. Lalouel, et al., Molecular basis of human hypertension: role of angiotensinogen, *Cell*. Oct 2; 71 (1) (1992): 169-180.

[7] JP. Forman, ND. Fisher, MR. Pollak, DG. Cox, S. Tonna, GC. Curhan, Renin-angiotensin system polymorphisms and risk of hypertension: influence of environmental factors, *Journal of Clinical Hypertension*. Jun; 10 (6) (2008): 459-466.

[8] A. Tsezou, G. Karayannis, E. Giannatou, V. Papanikolaou, F. Triposkiadis, Association of renin-angiotensin system and natriuretic peptide receptor A gene polymorphisms with hypertension in a Hellenic population, *Journal of the Renin-Angiotensin-Aldosterone System.* Dec; 9 (4) (2008): 202-207.

[9] TJ. Wang, P. Gona, MG. Larson, D. Levy, EJ. Benjamin, GH. Tofler, PF. Jacques, JB. Meigs, N. Rifai, J. Selhub, SJ. Robins, C. Newton-Cheh, RS. Vasan, Multiple biomarkers and the risk of incident hypertension. Hypertension, Mar;49 (3) (2007): 432-8.

[10] N. Srikumar, NJ. Brown, PN. Hopkins, X. Jeunemaitre, SC. Hunt, DE. Vaughan, GH. Williams, PAI-1 in human hypertension: relation to hypertensive groups, *Am J Hypertens*. Aug; 15 (2002): 683-690.

[11] R. Baldoncini, G. Desideri, C. Bellini, M. Valenti, G. De Mattia, A. Santucci, C. Ferri, High plasma renin activity is combined with elevated urinary albumin excretion in essential hypertensive patients, *Kidney Int*. Oct; 56 (4) (1999): 1499-1504.

[12] GH. Tofler, RB. D'Agostino, PF. Jacques, AG. Bostom, PW. Wilson, I. Lipinska, MA. Mittleman, J. Selhub., Association between increased homocysteine levels and impaired fibrinolytic potential: potential mechanism for cardiovascular risk, *Thromb Haemost*. Nov;88 (2002):799-804.

[13] MJ. Crawley. The R book. John Wiley & Sons, 2012.

[14] Bishop, M. Christopher. Pattern recognition and machine learning. Vol. 1. New York: springer, 2006.

[15] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. The elements of statistical learning. Vol. 2, no.1. New York: Springer, 2009.