35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

A Framework Towards Computational Discovery of Disease Sub-types and Associated (Sub-)Biomarkers

Mehmet Nadir Kurnaz*, *Member, IEEE EMBS,* Huseyin Seker, *Member, IEEE*

***Abstract*— Biomarker related patient data is generally assessed in order to determine relevant but generalized subset of the biomarkers. However, it fails to identify specific sub-groups of the patients or their corresponding (subset of) the biomarkers. This paper therefore proposes a novel framework that is capable of discovering disease sub-groups (types) and associated subset of biomarkers, which is expected to lead to enable the discovery of personalized bio-marker set. The framework is based on the utilization of a histogram obtained by using the Euclidean distances between the samples in a given data set. The t-test method is used for the selection of sub-set(s) of the biomarkers whereas the classification is performed by means of k-nearest neighbor, support vector machines and naive Bayes (NBayes) classifiers. For the assessment of the methods, leave-out-out cross validation is employed. As a case study, the method is applied in the analysis of male hypertension microarray data that consists of 159 patients and 22184 gene expressions. The method has helped identify specific sub-groups of the patients and their corresponding bio-marker sub-sets. The results therefore suggest that the generalized bio-marker sub-sets are not representative of the disease and therefore more focus should be on the sub-groups of the patients and their biomarker subsets identified through the proposed approach. It is particularly observed that the threshold values over the histogram are crucial to discover both sub-sets of the samples and biomarkers, and therefore can be used to determine complexity level of the study.**

I. Introduction

Microarray biomarker (or gene) selection that can help diagnose disease and identify new treatments is a popular problem in molecular biology and pharmaceutical manufacturing [1, 2]. However, the discovery of distinguishing gene or a set of genes is quite difficult because of complex and high-dimensional microarray data [3]. Microarray data sets have usually a small number of patient samples but a huge number of genes. Even though all the genes in microarray interact with each other in many ways, only a small number of the genes may have valuable meanings for a given problem. Due to the fact that there may be a set of genes that are not associated with the disease or conditions in biological systems, it may be useful for selecting sub-types of disease and associated biomarkers [4, 5]. In this manner, discovering discriminative and meaningful sub-sets of the genes leads to novel drug targets and candidates as well as treatment methods and strategies [6].

In the literature, various computational techniques including statistical methods, artificial neural networks and fuzzy clustering have been proposed to select biomarkers (genes) for different diseases. Feature (gene) selection techniques such as chi-square, Euclidean distance, t-test, correlation-based feature selection, beam search, genetic algorithms, weighted naive Bayes and support vector machines have also been explored for similar purposes [7]. However, they have mainly dealt with the analysis of entire data set and failed to identify sub-types of the subjects and their biomarker subsets.

In this study, a novel strategy based on measuring distances between samples is proposed for discovering biologically and clinically meaningful sub-types of the subjects and discriminative sub-sets of the biomarkers. In order to achieve it, biomarkers are selected by using the t-test ranking method. Three well-known supervised classification methods, namely k-nearest neighbours (k-NN), support vector machines (SVM) and naive Bayes (NBayes) algorithms are utilised as a classifier. To assess the statistical validity of the classifiers, leave-one-out cross-validation (LOOCV) method is performed.

II. Materials and Methods

The proposed method in the study is examined on microarray data that consists of 159 samples, each of which has 22184 gene expressions, obtained from 77 male hypertension and 82 male normotensive samples.

The scheme for discovering the disease sub-types and associated sub-biomarkers is depicted in Fig. 1. At the first stage, the widely-used biomarker selection method, t-test, is utilised to reduce the number of biomarkers and to select sub-set of the biomarkers. By performing this stage, a relevant subset of features (biomarkers) is obtained.

In the second stage, the novel strategy proposed in this study is carried out. As it is well-known from classification problems, the samples that are very different from each other in different classes can be easily distinguished leading to a more robust classification model. However, similar samples that have very similar characteristics cannot be easily separated from each other, which makes development of a classification model harder [8]. Since all biomarkers related to a sequence (or sample) cannot be important for discovering discriminative and meaningful biomarkers, irrelevant or noisy samples should be eliminated from the data set. In order to eliminate the samples, a histogram according to distance values in a matrix is plotted. The values are defined in such a way that samples are either similar to each other or not. Distances between all the samples in dataset are calculated by using the Euclidean distance to determine most appropriate sub-sets of the samples and corresponding biomarkers.

M. N. Kurnaz is with the Department of Electrical and Electronics Engineering, Faculty of Engineering, Nigde University, Nigde, 51245 Turkey (e-mail: mnkurnaz@gmail.com)

H. Seker is with the Bio-Health Informatics Research Group, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK. (e-mail: hseker@dmu.ac.uk)

*Corresponding author: M.N.Kurnaz (mnkurnaz@gmail.com)

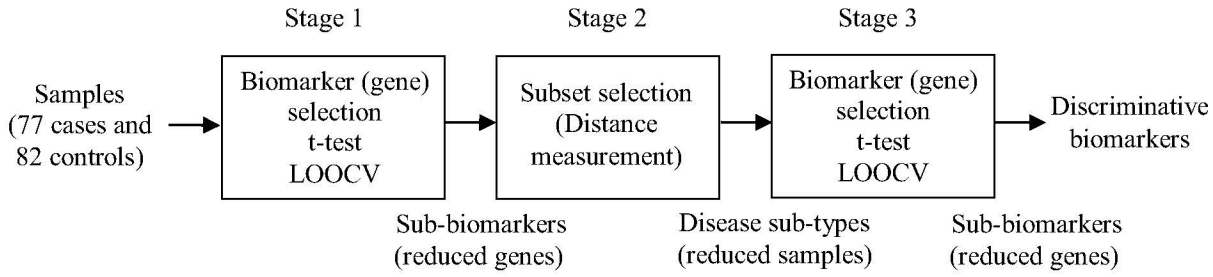

Figure 1. A framework that depicts computational discovery of disease sub-types and associated sub-biomarkers

Given an *mx*-by-*n* data matrix X and *my*-by-*n* data matrix `Y`, Euclidean distances between the vector x_s and y_c are calculated as follows:

$$d_{sc}^2 = (x_s - y_c)(x_s - y_c)'$$

where x_s and y_c represent case and control samples, respectively. In this study, since the data set consists of 77 case samples and 82 control samples with 22184 gene expressions, *mx, my* and *n* are 77, 82 and 22184, respectively.

The distance matrix that is used to measure how close a set of similar samples is obtained. The histogram to be obtained through the distance matrix helps determine the samples that are either near (similar) or far (dissimilar) from each other. By means of determining a threshold value for the plotted histogram, the samples that are far from each other, namely are easily classified, removed from the dataset. Finally, a number of disease sub-types is then obtained.

At the final stage, the same processes at the first stage is repeated to obtain the second group of sub-biomarkers for the obtained sub-types. As used in the first stage, biomarkers are selected by using the t-test ranking methods. Eventually, this process helps a set of the discriminative biomarkers for the microarray dataset.

Classification performance of the proposed method is assessed in a comparative manner by using the leave-one-out cross-validation (LOOCV) method for the k-NN, SVM and naive Bayes classifiers.

III. Results and Discussion

In order to discover disease sub-types and associated sub-biomarkers, the processes within all the stages of the proposed method as shown in Fig. 1 are carried and applied in the hypertension dataset that consists of 159 male subjects and 22184 gene expressions. In the study, k is set at 3 for k-NN classifier because of small sample size, SVM classifier is trained by using the sequential minimal optimization algorithm and the linear kernel function, and naive Bayes classifier is trained by using normal (Gaussian) distribution.

In the study, while discovering the disease sub-types and associated sub-biomarkers, initially, subsets of the biomarkers were selected by using the t-test ranking method. The subset of the biomarkers selected consists of 36 genes, and occurrence frequencies of these genes' index numbers are presented in Table I. It is observed that the genes 2782, 9984, 10621, 11249, 12376 and 14199 are selected yielding the occurrence frequency of 159 suggesting the most representative genes.

At the second stage, the samples were eliminated using the histogram based on the Euclidean distances between all the samples in the dataset in order to select the most important biomarkers that are associated with the sub-groups of the patients. The histogram shown in Fig. 2 depicts how the samples are distributed and threshold values can be set to select sub-groups of the samples. As an example, since the threshold value is selected as 3.2, the samples that have the distances greater than 3.2 can be eliminated from the dataset. As the threshold value is 3.2, 68 of 77 disease subjects and 74 of 82 control subjects are found to be similar samples whereas 9 of 77 disease subjects and 8 of 82 control subjects are regarded as dissimilar samples. As a consequence, the number of both disease and control subjects can be decreased or increased by changing the threshold values. The numbers of similar and dissimilar subjects obtained for different threshold values are presented in Table II. It is observed that higher value of the threshold decreases the similar subjects.

At the later stage, having obtained sub-sets of the subjects via the histogram, biomarker selection process was carried out again by using the t-test ranking method. At this stage, sub-sets of the biomarkers are selected separately for both similar and dissimilar biomarkers. The selected biomarkers among 36 sub-biomarkers and occurrence frequencies of them for different threshold values, and similar and dissimilar subjects are represented in Table III. It is observed from the histogram and results that the number of selected biomarkers changed for similar and dissimilar subjects according to the threshold values specified. The number of selected sub-biomarkers for similar subjects belonging to various threshold values, which are 3.2, 3.0, 2.8 and 2.6, were obtained to be 15, 15, 14 and 16, respectively. Similarly, for the dissimilar subjects, they were found to be 15, 17, 16 and 14, respectively. For instance, for the similar subjects, while the most dominant biomarkers were found to be 6428, 9735, 10118, 10621, 12376 and 16524 when the threshold value was set at 3.4 whereas they were obtained to be 5179, 5838, 9735, 10621, 12376 and 16524 for the threshold value of 2.8, as shown in Table III.

TABLE I. SELECTED BIOMARKER INDEX VALUES WITH THEIR OCCURRENCE FREQUENCIES

SB	OFoSB
2782	159
9984	159
10621	159
11249	159
12376	159
14199	159
386	154
16926	132
1711	130
2097	92
9856	29
16524	29
3734	15
8136	7
7382	6
81	5
6428	5
9437	5
8515	4
12642	4
3983	2
14828	2
17392	2
3010	1
4232	1
4376	1
5127	1
5179	1
5838	1
6208	1
7732	1
8451	1
9735	1
10118	1
13903	1
18591	1

SB: Selected biomarkers (First group of selected sub-biomarkers)
OFoSB: Occurrence frequencies of the selected biomarkers

In order to evaluate the classification accuracy of the proposed method, k-NN, SVM and naive Bayes classifiers with the leave-one-out cross-validation (LOOCV) method were examined for each selected sub-sets (disease sub-types and sub-biomarkers). Table IV presents classification accuracies of the proposed method for each classifier and various threshold values. It is observed that the accuracies obtained for each model with these different threshold values are obtained to be different, therefore the results further suggest the importance of the concept and framework proposed in this paper.

The SVM and NBayes classifiers yielded slightly better classification accuracies than those of k-NN classifier for the similar and dissimilar sub-biomarkers. However, when all the biomarkers were taken into consideration for the analyses, NBayes classifier gave the

TABLE II. THE NUMBERS OF SIMILAR AND DISSIMILAR SUBJECTS OBTAINED FOR DIFFERENT THRESHOLD VALUES

	# of hypertension (case) subjects		# of nermotensive (control) subjects	
	similar	dissimilar	similar	dissimilar
thr=3.2	68	9	74	8
thr=3.0	66	11	64	18
thr=2.8	55	22	39	43
thr=2.6	36	41	20	62

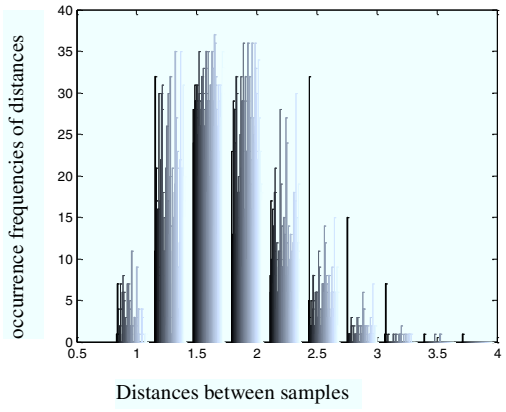

Figure 2. The histogram of the distances between samples.

best classification accuracy, and k-NN classifier resulted in a better classification accuracy than that of SVM classifier. Both similar and dissimilar subjects that consist of disease sub-types and sub-biomarkers gave better classification accuracies when all the subjects and biomarkers were analysed. However, classification accuracies of the similar subjects are less than those of the dissimilar subjects since the similar subjects have similar biomarkers and it is therefore quite difficult to separate them from each other. Sub-biomarkers in the similar data group can be regarded as more decisive. Therefore, these biomarkers seem able to discover more meaningful and discriminative biomarkers.

IV. CONCLUSIONS

In this study, a histogram-based method is presented in order to discover sub-types of patients, which is expected to reveal sub-types of disease, and their associated sub-biomarker sets. It is shown that the results obtained through the generalised analysis of the given post-genome data may yield misleading information, therefore the approach proposed should be carried out for more reliable and natural outcome. While the sub-biomarker sets are selected by using the t-test feature selection method, the sub-types of patients (disease) are obtained by using the histogram that is based on the Euclidean distance between the samples. Various threshold values over the histogram are examined in order to successfully determine similar/dissimilar subjects. It is observed that the selected sub-types of the samples and their corresponding sub-biomarker sets are sensitive to the

threshold values, therefore the future research is being carried out towards the development of a model that will be more capable of determining optimum threshold value.

Given the promising results, the proposed method will be applied to other post-genome data sets for its fine-tuning and to further establish its robustness.

REFERENCES

[1] M. Baker, "In biomarkers we trust?" *Nat. Biotechnol.,* vol. 23, pp. 297–304, 2005.

[2] R. Frank and R. Hargreaves "Clinical biomarkers in drug discovery and development", *Nat. Rev. Drug Discovery,* vol.2, pp. 566–80, 2003.

[3] K. S. Lynn, et al., "A neural network model for constructing endophenotypes of common complex diseases - an application to male young-onset hypertension microarray data", *Bioinformatics,* vol. 25(8), pp. 981–8, 2009.

[4] I. Guyon, "An introduction to variable and feature selection", *J. Mach. Learn. Res.,* vol. 3, pp. 1157–82, 2003.

[5] Kim H. et al. "Biomarker discovery using statistically significant gene sets", *J. Computational Biology,* vol. 18(10), pp. 1329-38, 2011.

[6] H. Seker, "Computational discovery of personalised biomarkers", *Proc. of the International Congress on Bioinformatics and Biomics*. Kusadasi, Turkey, 2011.

[7] Y. Saeys et al., "A review of feature selection techniques in bioinformatics", *Bioinformatics,* vol. 23(19), pp. 2507–17, 2007.

[8] L. Yu et al., "Stable gene selection from microarray data via sample weighting", *IEEE/ACM Transactions on Computational Biology And Bioinformatics*, vol. 9(1), pp. 262-72, 2012.

TABLE III. THE SELECTED SUB-BIOMARKERS FOR DIFFERENT THRESHOLD VALUES

thr=3.2				thr=3.0				thr=2.8				thr=2.6			
similar subjects		dissimilar subjects		similar subjects		dissimilar subjects		similar subjects		dissimilar subjects		similar subjects		dissimilar subjects	
NoSB	OFoSB	NoSB	OFoSB	NoSB	OFoSB	NoSB	OFoSB	NoSB	OFoSB	NoSB	OFoSB	NoSB	OFoSB	NoSB	OFoSB
6428	142	2782	17	5179	130	7382	29	3010	94	2097	65	6208	56	1711	103
9735	142	7382	17	5838	130	8515	29	5179	94	7382	65	7382	56	7382	103
10118	142	8515	17	9735	130	9984	29	9437	94	9984	65	9735	56	9984	103
10621	142	11249	17	10621	130	12642	29	9735	94	11249	65	14199	56	10621	103
12376	142	12642	17	12376	130	13903	29	14828	94	12642	65	14828	56	12642	103
16524	142	14199	17	16524	130	2782	28	17392	94	14199	65	8136	55	14199	103
14828	141	2097	16	4232	124	4376	28	10621	90	18591	65	9437	55	18591	103
17392	141	18591	14	10118	115	11249	28	10118	89	1711	61	7732	50	11249	101
5838	128	3983	10	14828	113	14199	28	9856	88	81	54	10118	50	386	89
386	119	8136	6	17392	75	2097	23	3734	78	16926	53	5127	26	12376	65
3010	31	9984	6	6428	64	8136	4	8451	14	8515	16	17392	22	81	41
5179	5	13903	6	9856	15	81	1	6428	11	8136	7	5838	8	8136	7
4232	1	1711	5	3010	8	1711	1	4232	5	2782	1	4232	6	4376	5
8451	1	6208	3	8451	5	3734	1	12376	1	3983	1	4376	3	16524	1
11249	1	16926	2	386	1	6208	1			4376	1	5179	3		
						16926	1			13903	1	6428	2		
						18591	1								

NoSB: Number of Selected Biomarkers, OFoSB: Occurence Frequencies of Selected Biomarkers

TABLE IV. CLASSIFICATION PERFORMANCES OF THE PROPOSED METHOD FOR THREE CLASSIFIER MODELS AND FOUR THRESHOLD VALUES

	Accuracy (%)		Accuracy (%)	
Classifier	All subjects	Threshold	Similar subjects	Dissimilar subjects
k-NN	49.69		64.08	76.47
SVM	45.91	3.2	71.83	82.35
NBayes	51.57		71.13	88.24
k-NN	49.69		60.77	75.86
SVM	45.91	3	62.31	82.76
NBayes	51.57		62.31	89.66
k-NN	49.69		61.70	80.00
SVM	45.91	2.8	62.77	80.00
NBayes	51.57		64.89	84.62
k-NN	49.69		50.00	75.73
SVM	45.91	2.6	53.57	79.61
NBayes	51.57		55.36	75.73