MATHEMATICAL BIOSCIENCES
AND ENGINEERING
Volume **10**, Number **3**, June **2013**

doi:10.3934/mbe.2013.10.667

pp. **667–690**

ON THE SENSITIVITY OF FEATURE RANKED LISTS FOR LARGE-SCALE BIOLOGICAL DATA

Danuta Gaweł and Krzysztof Fujarewicz

Silesian University of Technology, Institute of Automatic Control
Akademicka 16, 44-100 Gliwice, Poland

Abstract. The problem of feature selection for large-scale genomic data, for example from DNA microarray experiments, is one of the fundamental and well-investigated problems in modern computational biology. From the computational point of view, a selected gene list should be characterized by good predictive power and should be understood and well explained from the biological point of view. Recently, another feature of selected gene lists is increasingly investigated, namely their stability which measures how the content and/or the gene order change when the data are perturbed. In this paper we propose a new approach to analysis of gene list stability, termed the sensitivity index, that does not require any data perturbation and allows the gene list that is most reliable in a biological sense to be chosen.

1. **Introduction.**

Present techniques in molecular biology such as DNA microarrays, mass spectrometry, and deep sequencing deliver vast data sets. A common property of these data sets is that the number of features (genes, peptides, etc.) is much greater than the number of samples (observations). This requires careful feature selection as the very first step of supervised data analysis.

Many methods of gene selection have been proposed in the literature, which may be divided into two groups: univariate methods and multivariate methods. Univariate methods grade all genes separately by using statistical methods, and create a gene ranking in which the top genes are assumed to be the most informative (discriminative). Multivariate methods, roughly speaking, try to find the best gene set instead of a set of the best genes [12, 13, 17], and theoretically should give better gene signatures due to the fact that they may take into account interactions between genes. However, practice shows that for real data, especially with a small number of observations, this is not always true. So far there is no agreement which approach to feature selection is better, univariate or multivariate, mainly because of the problem of multiple hypothesis testing. For univariate methods we test G hypotheses where G is the number of genes (features), which is huge (thousands). For multivariate methods, this problem is even more evident and severe because the number of tested hypotheses is equal to the number of all possible sets of genes. For this reason, the risk of false-positive results increases and sometimes multivariate methods appear less suitable than univariate methods, see for example [20]. This explains why univariate methods are still popular. In this paper we focus on these methods.

2010 *Mathematics Subject Classification.* 92-08.

Key words and phrases. Feature selection, feature ranking, gene list stability, sensitivity, large-scale data, DNA microarrays.

Recently, the so-called stability of gene lists [3, 18] has drawn the attention of scientists dealing with large-scale biological data. Gene list stability is very important from a biological point of view. It is known that many different gene subsets may give comparable predictive power of a classifier trained on these sets, which sometimes leads to confusion and problems with biological interpretation of the genes identified. Although classification quality is associated with stability, the connection is still not well understood. Moreover, a higher stability may not indicate higher quality of classification. For example, we can imagine a ranking method that always chooses the same genes as top-genes; such a gene list would be perfectly stable, but the quality of classification based on those selected genes would be poor. Taking this into account, it is better to use the ranking method that gives more stable lists, while at the same time preserving good predictive power.

The stabilities of gene lists are measured by introducing a perturbation to the original data set and examining how the content of the top-gene list and its order were changed. When the list created on the basis of the perturbed data set is exactly the same as that formed on the basis of the original data, we say that this gene list is perfectly stable. The data may be perturbed in different ways; typically the original data set may be re-sampled, using for example a bootstrap technique, or the data may be changed by adding noise to the data matrix. Both these methods are computationally intensive because they require generation of many altered lists (as a result of data perturbation) and comparing them to the original list.

Perturbation methods may also affect the results, and among other things this was the reason for the creation of the new approach described in this paper. Unlike all existing methods of gene list stability analysis, this method does not require any data perturbation, resulting in a significant reduction of computational time. Moreover, to use standard methods a data set must contain enough observations to enable perturbation, introduced for example by re-sampling. In the case of the sensitivity index the minimum number of observations is determined by a ranking method (eg. for Student's t-test the number is 4, and for Fold Change 2). The basic idea of the approach proposed here is to calculate the sensitivity of the statistics used for creation of a gene ranking with respect to changes of the feature values (for example, gene expression). Then an aggregate sensitivity index for the whole data set is defined, whose value specifies the average percentage of data variation that may change the list order. The proposed method usually gives different results of stability than existing stability indexes, i.e. indicates different ranking methods as the most stable. This is because of different numerical approach to data analysis (lack of data perturbation). To examine this differences a biomedical analysis of genes from obtained lists was performed. The approach was tested on two DNA microarray data sets, a colon data set and an ovarian data set. For both data sets the sensitivity index indicated as the most stable lists of genes with stronger association with the particular disease than other methods. The sensitivity index was also tested on artificially generated data.

This article is organized as follows: Section 2 describes the gene ranking methods analyzed in this paper, and Section 3 contains a description of known gene list stability assessment. Our new approach (sensitivity index) is presented in Section 4, and finally the results of our study are presented in Section 5.

2. Gene ranking methods.

We concentrate on comparisons between two groups of samples (patients), T for treated and C for control. In this section the most common statistics used for generation of ranked lists are shortly described.

2.1. Fold Change (FC).

Fold change is one of the most popular gene ranking methods. The formula used in this work for FC is [35]:

$$fc_j = |\overline{x_{tj}} - \overline{x_{cj}}| \tag{1}$$

where $\overline{x_{tj}}$ and $\overline{x_{cj}}$ are means of $\log_2$ values of expression for the j-th gene in groups T and C respectively.

2.2. Probability Fold Change (PFC).

The Probability Fold Change is the modification of FC [7]:

$$S_{\theta,j} = \max(fc_j - t_{1-\theta,k}se_j, 0) \tag{2}$$

where t is the critical value in Student's t-test for k degrees of freedom ($k = n_t + n_c - 2$), n_t and n_c are the numbers of observations in groups T and C respectively, and θ is the confidence degree (in this work we have used $\theta = 0.95$). se in equation (2) is the standard error:

$$se_j = sp_j\sqrt{\frac{1}{n_t} + \frac{1}{n_c}} \tag{3}$$

where sp^2 is the pooled sample variance:

$$sp_j^2 = \frac{(n_t - 1)s_{tj}^2 + (n_c - 1)s_{cj}^2}{n_t + n_c - 2} \tag{4}$$

s_t and s_c denotes the standard deviations for the j-th gene in groups T and C respectively.

2.3. Student's t-test.

Student's t-test, like Fold Change, is one of the most popular gene ranking methods and can be calculated as:

$$t_j = \frac{fc_j}{se_j} \tag{5}$$

2.4. Welch t-test.

The Welch t-test is a modification of the Student's t-test, and in contrast to the Student's t-test allows unequal variances in groups. This statistic can be calculated by the formula:

$$tg_j = \frac{fc_j}{\sqrt{\frac{var(x_{tj})}{n_t} + \frac{var(x_{cj})}{n_c}}} \tag{6}$$

where $var(x_{tj})$ and $var(x_{cj})$ are the variances of the j-th gene in groups T and C respectively.

2.5. Bayesian t-test (BAYT).

The Bayesian t-test, also known as Cyber-T or BL, can be calculated by the formula:

$$bayt_j = \frac{fc}{se'_j} \tag{7}$$

where

$$se'_j = \sqrt{v_0 se_0^2 + \frac{(n_t + n_c - 1)se_j^2}{v_0 + n_t + n_c - 2}} \tag{8}$$

v_0 and se_0 are the global variables [7].

$$se_0 = \sqrt{\frac{\sum_{m=1}^{g} se_m^2}{g}} \tag{9}$$

g in the above formula is the total number of all genes.
The default value of v_0 is $10 - n_t - n_c$. Since v_0 cannot take a negative value and the used data sets used consists of more than 10 observations, in this work we have used a value $v_0 = |10 - n_t - n_c|$.
In the case when $v_0 = 0$ the BAYT statistic gives the same results as Student's t-test.

2.6. **Significance analysis of microarrays (SAM).** Significance analysis of microarrays is another modification of Student's t-test and is given by equation [7, 32]:

$$sam_j = \frac{fc_j}{se_j + c} \tag{10}$$

where c is a constant estimated as 90% percentile of the values of standard error estimated for all genes.

2.7. **Signal to Noise ratio (SN).** SN is another method where each gene is examined separately and genes are ranked by the value of sn [29]:

$$sn_j = \frac{fc}{s_{tj} + s_{cj}} \tag{11}$$

3. **Methods for gene list stability assessment.** In this section several existing methods for gene list stability assessment will be defined. In all these methods the original gene list is compared to lists generated by the ranking method based on perturbed data sets, for example by using the bootstrap technique.

3.1. **Stability index s.** The value of the index s is calculated according to the formula [30]:

$$s = 1 - \sum_{b=1}^{B} \sum_{j=1}^{G} \frac{2|r_j - r'_{bj}|}{BG(G+1)}. \tag{12}$$

where B is the number of bootstrap probes, G is the chosen number of genes falling within the list, r_j is the rank of gene j in the original gene list[1] and r'_{bj} is the rank of the j-th gene on the b-th bootstrap gene list[2].

If the gene does not exist in the b-th bootstrap G-top genes list, it receives the rank $r'_{bj} = G + 1$.

The higher the value of s obtained, the more stable is the gene list, and for a perfectly stable gene list s equals 1.

[1] Gene list generated on the base of the original data set
[2] Gene list generated on the base of the bootstrap probe

3.2. **Stability index s_1.** Another stability index is s_1 [3]:

$$s_1 = \sum_{j=1}^{G} I(r_j \leq G \wedge r'_{bj} \leq G). \tag{13}$$

Function I equals 1 when the j-th gene form the original gene list falls within the bootstrap gene list, regardless its position on both of those lists.

Formally, for an unstable gene list s_1 equals 0, and for a stable gene list G. However, in this paper the s_1 value is scaled so that its maximum value was 1 for a perfectly stable gene list.

3.3. **Stability index s_2 and CAT plot.** In order to calculate values of the s_2 index [3], we use the s_1 index (13) according to the formula:

$$s_2 = \frac{s_1}{2G - s_1}. \tag{14}$$

The maximum value of s_2 for a perfectly stable gene list is 1.

To obtain the CAT (Correspondence At the Top) plot we need to calculate values of s_2 for a gene list containing $g = 1, 2, \ldots, G$ top-genes. This plot shows the change of gene list stability depending on the selected number of genes that fall within the list.

3.4. **Union number.** Union Number (UN) [18] is the number of unique genes that falls within the B bootstrap gene lists. In this paper we have scaled this index so that for a perfectly stable gene list it equals 0 and for an unstable gene list 1 [18].

3.5. **Bootstrap based feature ranking plot.** In this paper we use the Boostrap Based Feature Ranking (BBFR) score Q_j [30, 11]:

$$Q_j = \frac{1}{B} \sum_{b=1}^{B} q_{bj}. \tag{15}$$

q_{bj} equals 1 if the j-th gene from the original gene list is on the b-th bootstrap list, and 0 otherwise.

Values of Q_j are sorted in descending order and scaled by $1/B$. The maximum value of Q_j is 1, which informs that the j-th gene falls within all of the B bootstrap lists.

When on the abscissa we mark the number of gene (j), and on the ordinate Q_j we obtain BBFR plot.

4. **Sensitivity index.** All of those described indexes are based on comparison between the original gene list and bootstrap gene lists. The bootstrap method implements into data changes so large that they cannot be assumed to be only simple noise. If we would add gaussian noise to the data set so it would be a little different from the original one, there would be a problem of choosing the level of the noise and the percentage of the data set to be changed. Therefore it is worth considering a method of assessing the stability of a gene list without changing the original data set.

We assume that all genes are sorted according to the statistic values in descending order i.e. $st_1 \geq st_2 \geq \ldots \geq st_G$. The distances between the j-th statistic and its two neighboring statistics are $st_{j-1} - st_j$ and $st_j - st_{j+1}$ respectively. The mean of these two distances is

$$D_j = \frac{st_{j-1} - st_{j+1}}{2} \tag{16}$$

and its value denotes how the statistic value should change (on average) so that it would change the position of the j-th gene on the list.

After calculation of the derivative of the statistics used for ranking genes with respect to the expression value x_{ji} (for the j-th gene and the i-th observation) we obtained the sensitivity of ranking methods to a small change in the observed expression values:

$$S_{ji} = \frac{\partial st}{\partial x_{ji}}. \tag{17}$$

For all popular statistics used for gene rankings it is possible to derive sensitivities (17) analytically, which are presented in the Appendix. Alternatively, the finite difference approximation may be used.

The change of the statistic st_j depends on the change of x_{ji} and may be expressed approximately using the sensitivity (17):

$$\Delta st_j \approx S_{ji} \Delta x_{ji} \tag{18}$$

Substituting Δst_j by D_j and solving (18) with respect to Δx_{ji} one can estimate Δx_{ji} sufficient to change the position of the j-th gene in the ranking as $\Delta x_{ji} \approx |D_j / S_{ji}|$. This quantity expressed as the relative percentage change of x_{ji} gives

$$\% x_{ji} \approx \left| \frac{D_j}{S_{ji} x_{ji}} \right| \cdot 100\%. \tag{19}$$

Now, let us define the sensitivity index for the j-th gene as the right side of (19) averaged over all N observations:

$$W_j = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{D_j}{S_{ji} x_{ji}} \right| \cdot 100\%. \tag{20}$$

The sensitivity index W_j measures the stability of the position of the j-th gene in the ranking. Therefore the average value of W_j over all G genes:

$$W = \frac{1}{G} \sum_{j=1}^{G} W_j \tag{21}$$

will be a measure of the stability of the whole gene list and consequently the stability of the ranking method which was used to establish the list. The gene list for which we obtain the highest value of W is the most stable. It is worth noting that this method is very useful when our data set consist of a small number of observations (too small to use a bootstrap method for generating probes). The minimal number of observations in this case determines ranking methods (for those methods used in this paper the minimum number of observations is 2 in each group).

5. **Results.** In this section we present the results of gene list stability analysis and assessment of classifier performance on the base of microarray data sets from experiments on samples of colon and ovarian cancer, and an artificial data set. The same methods of gene selection, classifiers and classification quality indexes were used for all data sets. Then the biomedical analysis were done for two real data sets. The next two subsections explain in details how above analysis were performed.

5.1. **Methods of feature selection, classification and classification quality assessment.** Gene lists are generated with ranking methods. In this paper we use the most common ranking methods: Fold Change (FC), Probability Fold Change (PFC), Student's *t*-test, Welch *t*-test, Bayesian *t*-test, Significance Analysis of Microarrays (SAM) and Signal to Noise ratio (SN).

Although stability measure is very important, it cannot be properly interpreted without evaluation of classifier performance and for this reason we have implemented two classifiers: Support Vector Machine (SVM) and Diagonal Linear Discriminant Analysis (DLDA) [15, 8, 9].

For classifier performance assessment we use specificity, sensitivity, accuracy and area under ROC curve (AUC). We used values of $B = 1000$ for the number of bootstrap probes generated and $G = 50$ for the number of top genes within the gene list.

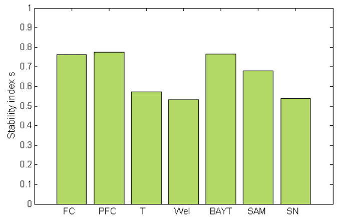
5.2. **Biomedical analysis.** In order to assess the reliability of a gene list (in the biological sense), we performed comparisons with the biomedical data. Primarily we selected the most stable gene lists on the basis of the sensitivity index W or on the basis of indexes s and s_1, and then we selected those genes that were on only one of these two lists, omitting genes common to both lists. We therefore obtained as the result two gene lists with less than 50 genes that were unique for the particular original gene list (one created with the method selected with indexes s and s_1, and one created with the method selected with sensitivity index W). Subsequently, because of the large number of genes in the lists for colon cancer we selected 10 with the highest ranks from those gene lists and then we searched in databases and the literature whether there exist any relationships between the selected genes and a particular disease. For the ovarian cancer data set the number of examined genes was shorter and was equal to 5.

5.3. **Colon cancer data set.** The colon cancer data set was collected in the Cancer Center and Institute of Oncology in Warsaw to identify genes differentially expressed in normal and cancer cells. The data set includes expression levels for 19058 genes from 82 samples, 34 from colon cancer tissue and 48 from normal tissue. In Table 1 are presented the results of stability analysis. Similarly, in Figs. 1 - 3 and in Figs. 5 - 6 are shown the values of stability indexes based on comparisons between the original and bootstrap gene lists with confidence intervals estimated by the percentile method marked for indexes s_1,s_2. The values of index W are presented in Fig. 4.

Table 1. Stability indexes values.

	FC	PFC	Student's *t*-test	Bayesian *t*-test	SAM	SN	Welch *t*-test
s	0.7626	**0.7747**	0.5734	0.7665	0.6788	0.5378	0.5315
s_1	0.8776	**0.9045**	0.7158	0.8705	0.8018	0.644	0.6315
s_2	0.7842	**0.8273**	0.5607	0.7733	0.6718	0.4791	0.4659
UN	0.0017	**0.0015**	0.0064	0.0018	0.0028	0.0094	0.0096
W	0.3898	0.4131	0.6037	1.4900	**3.2160**	1.7480	1.1520

In Table 2 are presented values of those scores for the SVM classifier and in Table 3 for the DLDA classifier. Bar charts with the AUC values are shown in Figs. 7 and 8.



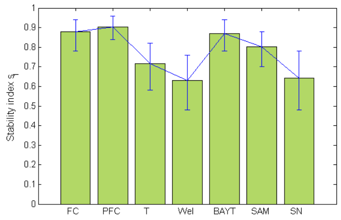
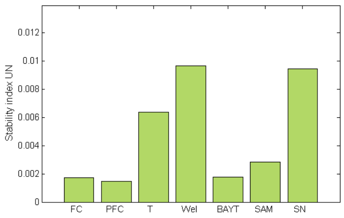


Figure 1.
Index s

Figure 2.
Index s_1



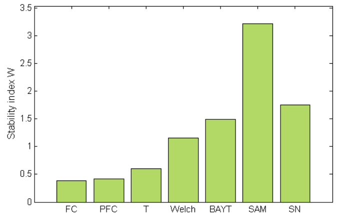


Figure 3.
Index UN

Figure 4.
Index W

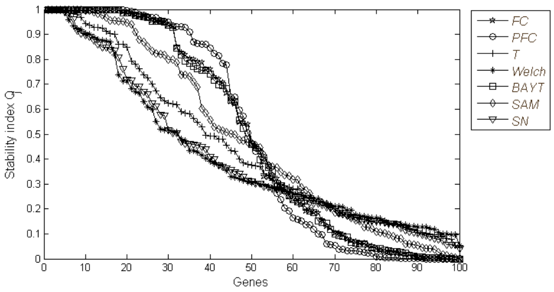


Figure 5. BBFR chart

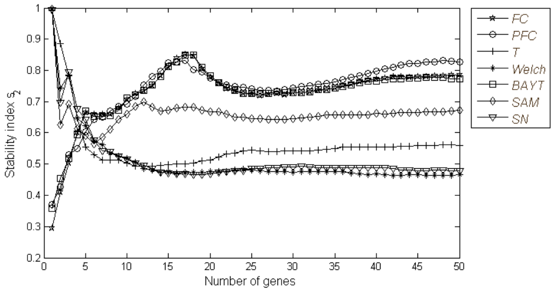


Figure 6. CAT chart

Table 2. Evaluation of SVM classifier performance.

	FC	PFC	Student's	Bayesian	SAM	SN	Welch
			t-test	*t*-test			*t*-test
sensitivity	0.9989	0.9988	0.9989	0.9992	**0.9999**	0.9994	0.9993
specificity	0.9883	0.9882	**0.9933**	0.9883	0.9884	0.9883	0.9879
accuracy	0.994	0.9939	**0.9964**	0.9942	0.9947	0.9944	0.9941
AUC	0.9998	0.9998	**0.9999**	0.9998	0.9998	0.9996	0.9996

Table 3. Evaluation of DLDA classifier performance.

	FC	PFC	Student's	Bayesian	SAM	SN	Welch
			t-test	*t*-test			*t*-test
sensitivity	0.9845	0.9890	0.9912	0.9844	**0.9913**	0.9875	0.9844
specificity	0.9623	0.9695	**0.9965**	0.9613	0.9671	0.9787	0.9799
accuracy	0.9745	0.9802	**0.9933**	0.974	0.9804	0.9834	0.9823
AUC	0.9939	0.9962	**0.9989**	0.9937	0.9897	0.9913	0.9918

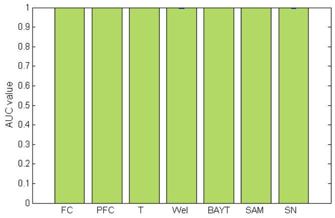


Figure 7.
Values of AUC.
SVM classifier.

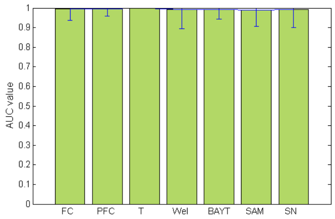


Figure 8.
Values of AUC.
DLDA classifier.

As the result, we obtained a set of values of stability indexes and classifier performance scores for gene lists generated with the ranking methods fold change (FC), Probability Fold Change (PFC), Student's t-test, Welch t-test, Bayesian t-test, Significance Analysis of Microarrays (SAM) and signal to noise ratio (SN).

In the case of the microarray data set obtained for samples of colon cancer and comparisons between normal and cancer cells, the indexes s,s_1,s_2,UN and plots: CAT and BBFR plots show that the most stable gene list was generated with the PFC method. However, the W index indicates SAM as the best method to rank genes. Because evaluation of classifier performance shows no statistically significant difference in those two ranking methods, both of them can be used to generate stable gene lists.

However, before making a final decision one should consider whether in the biological sense the first items on the lists of genes are verified.

The comparison with biomedical data is presented in Tables 4 and 5.

On the basis of the comparisons with biomedical data we can say that the better method to rank genes is that indicated by the sensitivity index W which is SAM.

Table 4. Comparison with biomedical data for colon cancer. Gene list created with the PFC method.

Colon cancer PFC list (indexes s and s_1)			
Gene name	Alias	Gene rank	Association with colorectal cancer
CA2	CAC, CAII, Car2, CA-II	5	Association with colon cancer was not found.
CA7	CAVII	7	According to [39] there is association of this gene with colon cancer.
CLDN23	hCG1646163, CLDNL	11	The protein encoded by CLDN23 is a member of integral membrane proteins and tight junction strands. According to [40] CLDN23 is expressed in colon tumors, germinal center B-cells, placenta, and stomach. Moreover this gene is down-regulated in intestinal-type gastric cancer [40]
HEPACAM2	MIKI, UNQ305/PRO346	13	HEPACAM2 takes part in mitotic division. Lack of HEPACAM2 may lead to mitotic arrest, disorganized spindles, and scattered chromosomes. It is thought that this gene may be a myeloid tumor suppressor [41]. However, direct association with colon cancer was not found.
CLCA1	CACC, CACC1, CLCRG1, CaCC-1, GOB5, hCLCA1, hCaCC-1	19	CLCA1 is a member of the protein family calcium-sensitive chloride conductance channel [40]. Moreover, according to [38] this gene is significantly down regulated in colon cancer.
CHI3L1	ASRT7, CGP-39, GP-39, GP39, HC-gp39, HCGP-3P, YKL-40, YKL40, YYL-40, hCGP-39	20	CHI3L1 encodes a glycoprotein which is thought to be involved in the processes of tissue remodeling and inflammation. CHI3L1 is expressed in cancer cells according to immunohistochemical analysis [19]. Moreover, the authors of [19] claim that it promotes cancer cell proliferation, angiogenesis, and macrophage recruitment in colon cancer cells. The association of CHI3L1 with colon cancer is also mentioned in [6] and [5]
CEACAM7	CEA, CGM2	21	CEACAM7 (carcinoembryonic antigen-related cell adhesion molecule 7) plays a role in tumor differentiation [36]. Furthermore, according to [27] CEACAM7 is down-regulated in colon cancer cells.Additionally, the association between CEACAM7 gene and colon cancer is mentioned in [31]
HHLA2	B7H7	22	Association with colon cancer was not found.
C1orf115	RP11-322F10.4	23	Association with colon cancer was not found.
ITLN1	UNQ640/PRO1270, HL-1, HL1, INTL, ITLN, LFR, hIntL, omentin	24	According to [37] ITLN1 is expressed on colorectal cancer cells. Moreover, the association between ITLN1 and colon cancer is mentioned in [33]
Number of genes for which association with colon cancer was found	**6**		

TABLE 5. Comparison with biomedical data for colon cancer. Gene list created with the SAM method.

Colon cancer SAM list (index W)			
Gene name	Alias	Gene rank	Association with colorectal cancer
NFE2L3	NRF3	5	NFE2L3 is expressed in colorectal tumors [37] Furthermore, it is reported that NFE2L3 mRNA is over-expressed in colorectal tumors when compared to normal cells [25]
OSTbeta	OSTB, OST-BETA	6	The organic solute transporter beta is expressed in colorectal tumors [37]. Moreover, according to [2] it plays a role in transportation of inter alia bile acids. Up-regulation of OSTbeta leads to toxic accumulation of bile acids, and colon cancer is related to their concentration [2]
TGFBI	BIGH3, CDB1, CDG2, CDGG1, CSD, CSD1, CSD2, CSD3, EBMD, LCD1	12	According to [37] is expressed in colorectal tumors. Furthermore, TGFBI might be a promoter or suppressor of cancer growth depending on the tissue [16]. In colon cancers the protein encoded by TGFBI is up-regulated which increases cell migration and metastatic potential [16]. Additionally, the association between TGFBI is mentioned in [21].
SPIB	SPI-B	14	Association with colon cancer was not found.
MT1M	MT-1M, MT-IM, MT1, MT1K	17	Association with colon cancer was not found.
UGT2B17	UDPGT2B17	20	UGT2B17 belongs to the UDP glucuronosyltransferase 2 polypeptide family [41] which catalyze the transfer of glucuronic acid. Acording to [44] it is highly expressed in colon, and according to [37] it is expressed in colorectal tumors.
TSPAN7	A15, CCG-B7, CD231, DXS1692E, MRX58, MXS1, TALLA-1, TM4SF2, TM4SF2b,	22	The protein encoded by TSPAN7 belongs to the tetraspanin family. According to [43] inhibition of TSPAN7 correlates with decreased proliferation of colon cancer cells. Moreover according to [37] is expressed in colorectal tumors.
SCNN1B	hCG_23853, BESC1, ENaCb, ENaCbeta, SCNEB	23	According to [38] one of the functions of SCNN1B is to control the reabsorption of sodium in kidney, colon, lung, and sweat glands. Although SCNN1B was not considered as a carcinogenic protein it is expressed in colorectal tumors [37] but the authors of [14] noticed significant under-expression in colorectal tumors in comparison with normal tissue. The association of SCNN1B with colon cancer is also mentioned in [10]
TESC	CHP3, TSC	25	According to [37] TESC is expressed in colorectal tumors.Although according to [28] the relationship of TESC with colorectal cancer has not yet been examined, these authors claim that it is one of the highly putative colon cancer markers.
SCIN		27	According to [37] SCIN is expressed in colorectal tumors. Furthermore, it is up-regulated in colon cancer cells in comparison with normal cells [42]. Moreover the same authors claim that high expression of this gene is an independent risk factor for prognosis of colorectal cancer liver metastasis.
Number of genes for which association with colon cancer was found	**8**		

5.4. **Ovarian cancer data set.** In this subsection we present the results of stability analysis and assessment of classifier performance based on an ovarian cancer microarray data set that was collected in the Cancer Center in Warsaw. Genes expressed differentially between cells with a mutated $P53$ gene and cells with the proper sequence of gene $P53$ were sought. This data set consists of 54613 genes for 90 patients, 75 with a mutated $P53$ gene and 15 with the proper sequence. In Table 6 are presented the results of stability analysis. As in Figs. 9 - 11 and Figs. 13 - 14, the values of stability indexes that are based on comparisons between the original gene list and bootstrap gene lists with confidence intervals estimated with the percentile method indicated are shown. The values of index W are presented in Fig. 12.

Table 6. Stability indexes values.

	FC	PFC	Student's	Welch	Bayesian	SAM	SN
			t-test	t-test	t-test		
s	0.3090	**0.3276**	0.1734	0.1621	0.3087	0.2541	0.1691
s_1	0.4631	**0.5337**	0.3134	0.2452	0.4598	0.3695	0.2545
s_2	0.3088	**0.3701**	0.1891	0.1425	0.3059	0.2327	0.1482
UN	0.0098	**0.0071**	0.0131	0.0204	0.0099	0.01292	0.0237
W	0.2055	0.3802	0.2062	0.4512	**0.6886**	0.2733	0.1372

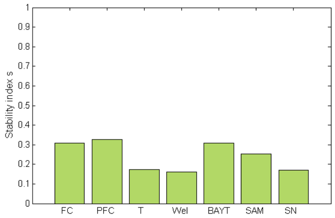


Figure 9.
Index s

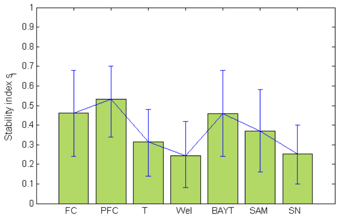


Figure 10.
Index s_1

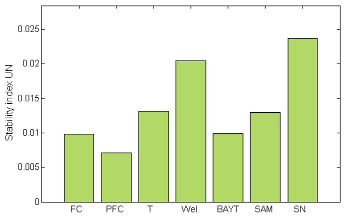


Figure 11.
Index UN

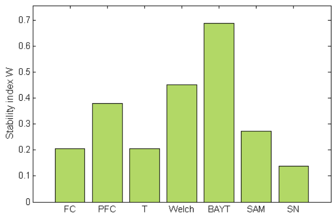


Figure 12.
Index W

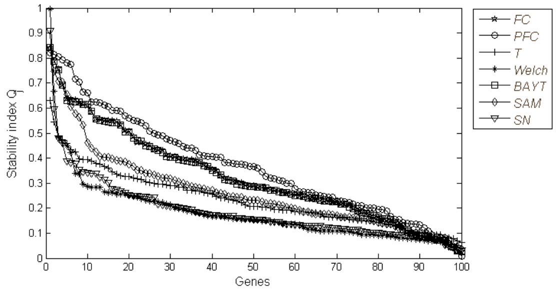


Figure 13. BBFR chart

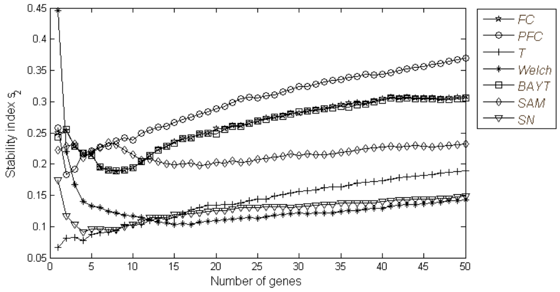


Figure 14. CAT chart

In Table 7 are presented the values of those scores for the SVM classifier and in Table 8 for the DLDA classifier. Bar charts with the AUC value are shown in Figs. 15 and 16.

Table 7. Evaluation of SVM classifier performance.

	FC	PFC	Student's *t*-test	Welch *t*-test	Bayesian *t*-test	SAM	SN
Sensitivity	0.9125	0.9044	0.9066	**0.9484**	0.9128	0.9128	0.9443
Specificity	0.3525	0.3299	0.4431	**0.5167**	0.3536	0.3705	0.4902
Accuracy	0.8179	0.8077	0.8291	**0.8765**	0.8183	0.8216	0.8688
AUC	0.7011	0.6821	0.72	**0.801**	0.7007	0.7018	0.784

Table 8. Evaluation of DLDA classifier performance.

	FC	PFC	Student's *t*-test	Welch *t*-test	Bayesian *t*-test	SAM	SN
Sensitivity	0.9351	0.9217	0.9099	0.9502	0.935	0.947	**0.9631**
Specificity	0.3791	0.3724	0.4997	**0.5647**	0.3819	0.3976	0.5268
Accuracy	0.8402	0.8282	0.8399	0.8852	0.8405	0.853	**0.8888**
AUC	0.7051	0.7005	0.7394	**0.7815**	0.7047	0.7186	0.7783

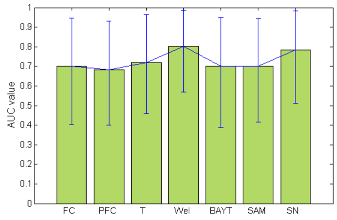


FIGURE 15. Values of AUC. SVM classifier.

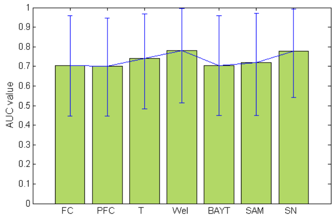


FIGURE 16. Values of AUC. DLDA classifier.

In the case of the microarray data set from samples of ovarian cancer and comparison between cells with a mutated $P53$ gene and cells with the proper sequence of gene $P53$ the indexes s,s_1,s_2,UN and CAT and $BBFR$ plots show that the most stable gene list was generated with the PFC method. However, the W index indicates the Bayesian t-test as the best method to rank genes. Because the evaluation of classifier performance shows no statistically significant difference in those two ranking methods (which was tested using the Wilcoxon test), both of them can be used to generate stable gene lists. The comparison with biomedical data is presented in Tables 9 and 10.

TABLE 9. Comparison with biomedical data for ovarian cancer. Gene list created with the PFC method.

Ovarian cancer PFC list (indexes s and s_1)			
Gene name	Alias	Gene rank	Association with ovarian cancer
EFCAB10		35	Association with ovarian cancer was not found.
DKFZp761K1021	P50, P85, PAK3, PIXB, COOL1, P50BP, COOL-1, P85SPR, BETA-PIX, P85COOL1, Nbla10314, ARHGEF7	38	According to [37] ARHGEF7 is expressed in ovary and ovarian tumor. Moreover the authors of [22] noticed overexpression of ARHGEF7 in breast cancer and although they did not find this expression as significant trend of ARHGEF7, they claim that there might be a correlation between ARHGEF7 and breast cancer.However direct association with ovarian cancer was not found.
PHA1	BESC3, ENaCg, ENaCgamma, SCNN1G, SCNEG	43	Association with ovarian cancer was not found.
MPD1	CMH1, MYH7, SPMD, SPMM, CMD1S, MYHCB	46	According to [37] MPD1 is expressed in ovary and ovarian tumors. However direct association with ovarian cancer was not found.
ERP	ELK3, NET, SAP2	48	Association with ovarian cancer was not found.
Number of genes for which an association with colon cancer was found	**0**		

TABLE 10. Comparison with biomedical data for ovarian cancer. Gene list created with the Bayesian t-test method.

Ovarian cancer Bayesian t-test list (index W)			
Gene name	Alias	Gene rank	Association with ovarian cancer
SMT3H4	IDDM5; SUMO4; SUMO-4; dJ281H8.4	27	Association with ovarian cancer was not found.
XPG	RP11-484I6.5, COFS3, ERCM2, UVDR, ERCC5, XPGC	33	The official symbol for XPG [40] is ERCC5 which is "a novel biomarker of ovarian cancer prognosis" according to [34], additionally according to [37] ERCC5 is expressed in ovary and ovarian tumor. Moreover in [26] the authors claim that significant proportion of ovarian tumors have the XPG promoter methylated.
CAFS	DGS; TGA; TBX1; CTHM; DGCR; DORV; VCFS; TBX1C	38	According to [37] TBX1 is expressed in the ovary and ovarian cancer tissue.Moreover the association of TBX1 with ovarian cancer is also mentioned in [23].
MGC149559	FAM27E3	42	Association with ovarian cancer was not found.
NCRNA00158	C21orf42, LINC00158	43	Association with ovarian cancer was not found.
Number of genes for which an association with colon cancer was found	**2**		

On the basis of the comparisons with biomedical data, we can say that the better method to rank genes is that indicated by the sensitivity index W which is the Bayesian t-test.

5.5. **Artificial data set.** The artificial data set was created with the assumption that for 900 genes there are no differences between two types of cells. For those features expression values were drawn from a normal distribution with a standard deviation equal to 0.1 and an average equal to 4. For 100 features we assumed a constant standard deviation of 0.2 and different averages changed with the schema as in Fig. 17.

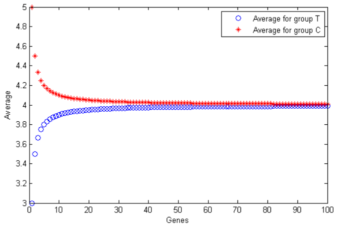


FIGURE 17. Average for 100 genes of artificial data set.

We assumed 10 observations in the group T and 15 observations in the group C.

The indexes based on comparisons between the original gene list and the bootstrap lists suggest that the most stable gene lists for the artificial data set were

Table 11. Stability indexes values.

	FC	PFC	Student's t-test	Welch t-test	Bayesian t-test	SAM	SN
s	0.5426	**0.5769**	0.4402	0.3745	0.5423	0.4596	0.3784
s_1	0.5440	**0.6669**	0.5194	0.4416	0.5354	0.501	0.4427
s_2	0.3759	**0.5029**	0.3545	0.287	0.3681	0.337	0.2877
UN	0.0126	**0.0065**	0.0123	0.0171	0.0129	0.0154	0.0171
W	0.6296	0.8889	0.02652	0.681	**1.223**	1.108	0.6394

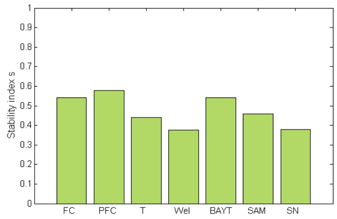


Figure 18.
Index s

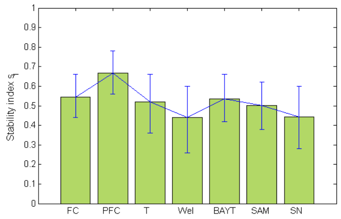


Figure 19.
Index s_1

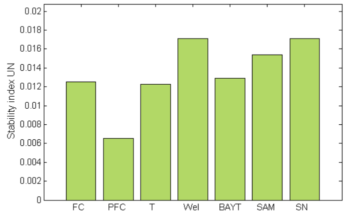


Figure 20.
Index UN

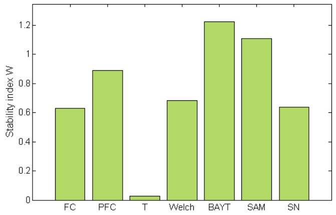


Figure 21.
Index W

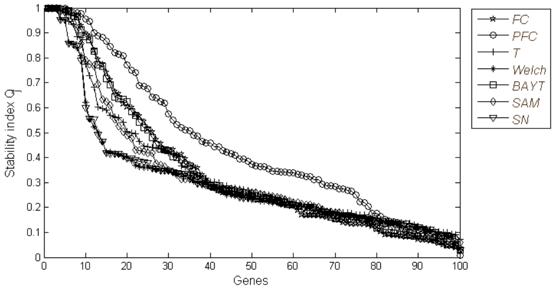


Figure 22. BBFR chart

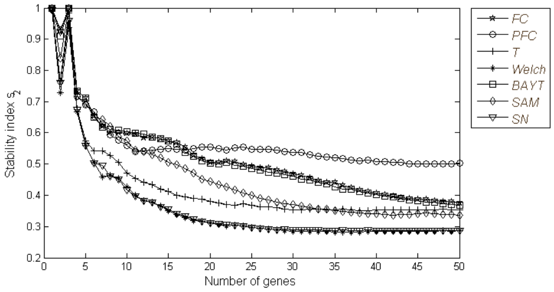


Figure 23. CAT chart

Table 12. Evaluation of SVM classifier performance.

	FC	PFC	Student's t-test	Welch t-test	Bayesian t-test	SAM	SN
Sensitivity	0.9691	**0.9787**	0.9589	0.93	0.9675	0.9571	0.9318
Specificity	0.8822	**0.9288**	0.7751	0.7062	0.8764	0.8023	0.7005
Accuracy	0.9222	**0.9531**	0.8683	0.8187	0.9183	0.8787	0.8168
AUC	0.9914	**0.9951**	0.9729	0.954	0.9907	0.9823	0.9525

Table 13. Evaluation of DLDA classifier performance.

	FC	PFC	Student's t-test	Welch t-test	Bayesian t-test	SAM	SN
Sensitivity	1	1	0.9999	0.9999	1	1	0.9999
Specificity	0.9991	1	0.998	0.9968	0.9991	0.9965	0.9952
Accuracy	0.9994	1	0.9986	0.9978	0.9994	0.9975	0.9967
AUC	0.9999	1	0.9997	0.9995	0.9999	0.9997	0.9988

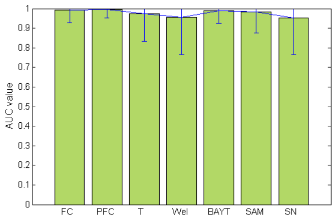


Figure 24. Values of AUC. SVM classifier.

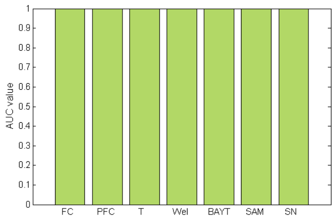


Figure 25. Values of AUC. DLDA classifier.

those created with the PFC method. However, the index W suggests that the best method is the Bayesian t-test. Based on classifier assessment we cannot decide which of those gene lists is more reliable.

6. Conclusions.

Continually developing techniques allow thousands of genes to be examined in one experiment. High dimensional data and frequently small numbers of observations make the problem of choosing the best ranking method more difficult. It is obvious that the method of introducing changes (bootstrap re-sampling, adding noise, etc.) into a data set has an influence on the results of the stability analysis. Usually the order of ranking methods, starting from the most stable, is different for bootstrap re-sampling and for altering the data by adding noise. Moreover, in the case of adding noise there is a major problem of deciding what kind of noise it should be, how large a power would be the best, and what percentage of the data set to change. For this reason, we created the sensitivity index W whose value informs about the average percentage of change that has to be introduced into the original data set to completely transform the gene list.

A major advantage of this index is that there is no need to change the original data set, and consequently the computational time is much shorter. Moreover, the minimum number of observations is limited only by the statistical method used and not by the method of introducing changes into the data set (for example, by the bootstrap method). Additionally, the comparisons with biomedical data show that for a larger number of genes on the list created with a ranking method pointed by the sensitivity index W, there is a stronger association with the particular disease than for genes on the list created with the method pointed by indexes s and s_1, confirming that the sensitivity index W helps to identify a gene list which is more reliable in a biological sense.

Acknowledgments. This work was supported by the Polish National Science Center under grants: N N519 647840 (D.G.) and UMO-2012/05/B/ST6/03472 (K.F.). The authors would also like to thank R. Hancock, J. Rzeszowska and R. Jaksik for valuable discussion and suggestions leading to substantial improvement of the manuscript as well as J. Ostrowski and K. Lisowska for the opportunity of DNA microarray data analysis.

Appendix A. Appendix.

A.1. Auxiliary symbols and their derivatives.

Before the derivatives (17) of the most common statistics will be presented, some auxiliary formulas are introduced:

$$sumT = \sum_{i=1}^{n_t} (x_{ti} - \overline{x_t})^2 \tag{22}$$

$$sumC = \sum_{j=1}^{n_c} (x_{cj} - \overline{x_c})^2 \tag{23}$$

Derivatives:

$$\frac{\partial sumT}{\partial x_{ti}} = 2(x_{ti} - \overline{x_t}) \tag{24}$$

$$\frac{\partial sumC}{\partial x_{cj}} = 2(x_{cj} - \overline{x_c}) \tag{25}$$

Standard deviation in group T and C respectively:

$$s_t = \sqrt{\frac{\sum_{i=1}^{n_t} (x_{ti} - \overline{x_t})^2}{n_t - 1}} \tag{26}$$

$$s_c = \sqrt{\frac{\sum_{j=1}^{n_c}(x_{cj} - \overline{x_c})^2}{n_c - 1}} \tag{27}$$

Derivatives:

$$\frac{\partial s_t}{\partial x_{ti}} = \frac{1}{2s_t(n_t - 1)}\frac{\partial sumT}{\partial x_{ti}} \tag{28}$$

$$\frac{\partial s_c}{\partial x_{cj}} = \frac{1}{2s_c(n_c - 1)}\frac{\partial sumC}{\partial x_{cj}} \tag{29}$$

Square pooled sample variance:

$$sp = \sqrt{\frac{(n_t - 1)s_t^2 + (n_c - 1)s_c^2}{n_t + n_c - 2}} \tag{30}$$

Derivatives:

$$\frac{\partial sp}{\partial x_{ti}} = \frac{1}{(n_t + n_c - 2)sp}(n_t - 1)s_t\frac{\partial s_t}{\partial x_{ti}} \tag{31}$$

$$\frac{\partial sp}{\partial x_{cj}} = \frac{1}{(n_t + n_c - 2)sp}(n_c - 1)s_c\frac{\partial s_c}{\partial x_{cj}} \tag{32}$$

Standard error:

$$se = sp\sqrt{\frac{1}{n_t} + \frac{1}{n_c}} \tag{33}$$

Derivatives:

$$\frac{\partial se}{\partial x_{ti}} = \sqrt{\frac{1}{n_c} + \frac{1}{n_t}}\frac{\partial sp}{\partial x_{ti}} \tag{34}$$

$$\frac{\partial se}{\partial x_{cj}} = \sqrt{\frac{1}{n_c} + \frac{1}{n_t}}\frac{\partial sp}{\partial x_{cj}} \tag{35}$$

A.2. **Derivatives of chosen statistics. Fold Change:**

$$fc = |\overline{x_t} - \overline{x_c}| \tag{36}$$

Derivatives for $\overline{x_t} < \overline{x_c}$:

$$\frac{\partial fc}{\partial x_{ti}} = \frac{-1}{n_t} \tag{37}$$

$$\frac{\partial fc}{\partial x_{cj}} = \frac{1}{n_c} \tag{38}$$

Derivatives dla $\overline{x_t} > \overline{x_c}$:

$$\frac{\partial fc}{\partial x_{ti}} = \frac{1}{n_t} \tag{39}$$

$$\frac{\partial fc}{\partial x_{cj}} = \frac{-1}{n_c} \tag{40}$$

Probability fold change:

$$pfc = fc - tse \tag{41}$$

Derivatives:

$$\frac{\partial pfc}{\partial x_{ti}} = \frac{\partial fc}{\partial x_{ti}} - t\frac{\partial se}{\partial x_{ti}} \tag{42}$$

$$\frac{\partial pfc}{\partial x_{cj}} = \frac{\partial fc}{\partial x_{cj}} - t\frac{\partial se}{\partial x_{cj}} \tag{43}$$

The conditional mean:

$$srw = \frac{\sum_{z=1}^{n_2}(fc|inten)}{n_2} \tag{44}$$

Derivatives for $\overline{x_t} < \overline{x_c}$:

$$\frac{\partial srw}{\partial x_{ti}} = \frac{-1}{n_t n_2} \tag{45}$$

$$\frac{\partial srw}{\partial x_{cj}} = \frac{1}{n_c n_2} \tag{46}$$

Derivatives for $\overline{x_t} > \overline{x_c}$:

$$\frac{\partial srw}{\partial x_{ti}} = \frac{1}{n_t n_2} \tag{47}$$

$$\frac{\partial srw}{\partial x_{cj}} = \frac{-1}{n_c n_2} \tag{48}$$

Student's t-test:

$$t = \frac{fc}{se} \tag{49}$$

Derivatives:

$$\frac{\partial t}{\partial x_{ti}} = \frac{\frac{\partial fc}{\partial x_{ti}} se - fc \frac{\partial se}{\partial x_{ti}}}{se^2} \tag{50}$$

$$\frac{\partial t}{\partial x_{cj}} = \frac{\frac{\partial fc}{\partial x_{cj}} se - fc \frac{\partial se}{\partial x_{cj}}}{se^2} \tag{51}$$

Supporting calculations to calculate the derivative of Welch t-test:

$$varT = var(x_t) \tag{52}$$

$$varC = var(x_c) \tag{53}$$

Derivatives:

$$\frac{\partial varT}{\partial x_{ti}} = \frac{1}{n_t - 1} \frac{\partial sumaT}{\partial x_{ti}} \tag{54}$$

$$\frac{\partial varC}{\partial x_{cj}} = \frac{1}{n_c - 1} \frac{\partial sumaC}{\partial x_{cj}} \tag{55}$$

Square of the sum of the variance:

$$psw = \sqrt{\frac{var(x_t)}{n_t} + \frac{var(x_c)}{n_c}} \tag{56}$$

Derivatives:

$$\frac{\partial psw}{\partial x_{ti}} = \frac{1}{2psw \cdot n_t} \frac{\partial varT}{\partial x_{ti}} \tag{57}$$

$$\frac{\partial psw}{\partial x_{cj}} = \frac{1}{2psw \cdot n_c} \frac{\partial var}{\partial x_{cj}} \tag{58}$$

Welch t-test:

$$tg = \frac{fc}{\sqrt{\frac{var(x_t)}{n_t} + \frac{var(x_c)}{n_c}}} \tag{59}$$

Derivatives:

$$\frac{\partial tg}{\partial x_{ti}} = \frac{\frac{\partial fc}{\partial x_{ti}} psw - fc \frac{\partial psw}{\partial x_{ti}}}{(psw)^2} \tag{60}$$

$$\frac{\partial tg}{\partial x_{cj}} = \frac{\frac{\partial fc}{\partial x_{cj}} psw - fc \frac{\partial psw}{\partial x_{cj}}}{(psw)^2} \tag{61}$$

Supporting calculations to calculate the derivative of Bayesian t-test:

Original formula:

$$se_0 = \sqrt{\frac{\sum_{m=1}^{g} se_m^2}{g}} \tag{62}$$

Derivatives:

$$\frac{\partial se_0}{\partial x_{ti}} = \frac{1}{2se_0 g} 2se_m \frac{\partial se}{\partial x_{ti}} \tag{63}$$

$$\frac{\partial se_0}{\partial x_{cj}} = \frac{1}{2se_0 g} 2se_m \frac{\partial se}{\partial x_{cj}} \tag{64}$$

Original formula:

$$se' = \sqrt{v_0 se_0^2 + \frac{(n_t + n_c - 1)se^2}{v_0 + n_t + n_c - 2}} \tag{65}$$

Derivatives:

$$\frac{\partial se'}{\partial x_{ti}} = \frac{1}{2se'}(v_0 2se_0 \frac{\partial se_0}{\partial x_{ti}} + \frac{n_c + n_t - 1}{v_0 + n_t + n_c - 2} 2se \frac{\partial se}{\partial x_{ti}}) \tag{66}$$

$$\frac{\partial se'}{\partial x_{cj}} = \frac{1}{2se'}(v_0 2se_0 \frac{\partial se_0}{\partial x_{cj}} + \frac{n_c + n_t - 1}{v_0 + n_t + n_c - 2} 2se \frac{\partial se}{\partial x_{cj}}) \tag{67}$$

Bayesian t-test:

$$bayt = \frac{fc}{se'_j} \tag{68}$$

Derivatives:

$$\frac{\partial bayt}{\partial x_{ti}} = \frac{\frac{\partial fc}{\partial x_{ti}} se' - fc \frac{\partial se'}{\partial x_{ti}}}{(se')^2} \tag{69}$$

$$\frac{\partial bayt}{\partial x_{cj}} = \frac{\frac{\partial fc}{\partial x_{cj}} se' - fc \frac{\partial se'}{\partial x_{cj}}}{(se')^2} \tag{70}$$

Significance analysis of microarrays:

$$sam_j = \frac{fc_j}{se_j + c} \tag{71}$$

Derivatives:

$$\frac{\partial sam}{\partial x_{ti}} = \frac{\frac{\partial fc}{\partial x_{ti}}(se + c) - fc \frac{\partial se}{\partial x_{ti}}}{(se + c)^2} \tag{72}$$

$$\frac{\partial sam}{\partial x_{cj}} = \frac{\frac{\partial fc}{\partial x_{cj}}(se + c) - fc \frac{\partial se}{\partial x_{cj}}}{(se + c)^2} \tag{73}$$

Signal to Noise ratio:

$$sn = \frac{fc}{s_t + s_c} \tag{74}$$

Derivatives:

$$\frac{\partial sn}{\partial x_{ti}} = \frac{\frac{\partial fc}{\partial x_{ti}}(s_t + s_c) - fc \frac{\partial s_t}{\partial x_{ti}}}{(s_t + s_c)^2} \tag{75}$$

$$\frac{\partial sn}{\partial x_{cj}} = \frac{\frac{\partial fc}{\partial x_{cj}}(s_t + s_c) - fc \frac{\partial s_t}{\partial x_{cj}}}{(s_t + s_c)^2} \tag{76}$$

REFERENCES

[1] C. Alvarez-Baron, P. Jonsson, C. Thomas, S. Dryer and C. Williams, *The two-pore domain potassium channel KCNK5: Induction by estrogen receptor alpha and role in proliferation of breast cancer cells*, Molecular Endocrinology, **25** (2011), 1326–1336.

[2] N. Ballatori, N. Li, F. Fang, J. Boyer, W. Christian and C. Hammond, *OST alpha-OST beta: A key membrane transporter of bile acids and conjugated steroids*, Frontiers in Bioscience, **14** (2009), 2829–2844.

[3] A. Boulesteix and M. Slawski, *Stability and aggregation of ranked gene lists*, Briefings in Bioinformatics, **10** (2009), 556–568.

[4] R. Buckanovich, D. Sasaroli, A. O'Brien-Jenkins, J. Botbyl, R. Hammond, D. Katsaros, R. Sandaltzopoulos, L. Liotta, P. Gimotty and G. Coukos, *Tumor vascular proteins as biomarkers in ovarian cancer*, Journal of Clinical Oncology, **25** (2007), 852–861.

[5] V. Catalán, J. Gómez-Ambrosi, A. Rodríguez, B. Ramírez, F. Rotellar, V. Valentí, C. Silva, M. Gil, J. Salvador and G. Frühbeck, *Up-regulation of the novel proinflammatory adipokines lipocalin-2, chitinase-3 like-1 and osteopontin as well as angiogenic-related factors in visceral adipose tissue of patients with colon cancer*, The Journal of Nutritional Biochemistry, **22** (2011), 634–641.

[6] F. Coffman, *Chitinase 3-Like-1 (CHI3L1): A putative disease marker at the interface of proteomics and glycomics*, Critical Reviews in Clinical Laboratory Sciences, **45** (2008), 531–562.

[7] X. Deng, J. Xu, J. Hui and C. Wang, *Probability fold change: A robust computational approach for identifying differentially expressed gene lists*, Computer Methods and Programs in Biomedicine, **93** (2009), 124–139.

[8] S. Dudoit, J. Fridlyand and T. Speed, *Comparison of discrimination methods for the classification of tumors using gene expression data*, Journal of the American Statistical Association, **97** (2002), 77–87.

[9] S. Dudoit and R. Gentleman, "Bioconductor Short Course," 2003. Available from: `http://www.bioconductor.org/help/course-materials/2003/Milan/Lectures/classif.pdf`.

[10] T. J. Farr, S. J. Coddington-Lawson, P. M. Snyder and F. J. McDonald, *Human Nedd4 interacts with the human epithelial Na+ channel: WW3 but not WW1 binds to Na+-channel subunits*, The Biochemical Journal, **345** (2000), 503–509.

[11] K. Fujarewicz, et al, *A multigene approach to differentiate papillary thyroid carcinoma from benign lesions: Gene selection using bootstrap-based support vector machines*, Endocrine - Related Cancer, **14** (2007), 809–826.

[12] K. Fujarewicz, M. Kimmel and J. Rzeszowska-Wolny, *Improved classification of microarray gene expression data using support vector machines*, Journal of Medical Informatics and Technologies, **2** (2001), MI9–MI17.

[13] K. Fujarewicz and M. Wiench, *Selecting differencially expressed genes for colon tumor classification*, International Journal of Applied Mathematics and Computer Science, **13** (2003), 327–335.

[14] J. Harvey, A. Gannon, Z. Li, C. Beard and C. Burgess, *Identification of a novel methylation marker, SCNN1B*, AACR Meeting Abstracts, (2005), 217-c-218 .

[15] T. Hastie, R. Tibshirani and J. Friedman, "The Elements of Statistical Learning. Data Mining, Inference, and Prediction," 2nd edition, Springer-Verlag, 2009.

[16] M. Irigoyen, M. Pajares, J. Agorreta, M. Ponz-Sarvisé E. Salvo, M. Lozano, R. Pío, I. Gil-Bazo and A. Rouzaut, *TGFBI expression is associated with a better response to chemotherapy in NSCLC*, Molecular Cancer, **9** (2010).

[17] B. Jarząb, M. Wiench, K. Fujarewicz, K. Simek, M. Jarząb, M. Oczko-Wojciechowska, J. Włoch, A. Czarniecki, E. Chmielik, D. Lange, A. Pawlaczek, S. Szpak, E. Gubała and A. Świerniak, *Gene expression profile of papillary thyroid Ccncer: sources of variability and diagnostic implications*, Cancer Research, **65** (2005), 1587–1597.

[18] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea and C. Furlanello, *Algebraic stability indicators for ranked lists in molecular profiling*, Bioinformatics, **24** (2008), 258–264.

[19] M. Kawada, H. Seno, K. Kanda, Y. Nakanishi, R. Akitake, H. Komekado, K. Kawada, Y. Sakai, E. Mizoguchi and T. Chiba, *Chitinase 3-like 1 promotes macrophage recruitment and angiogenesis in colorectal cancer*, Oncogene, **31** (2012), 3111–3123.

[20] C. Lai, M. Reinders, L. Veer and L. Wessels, *A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets*, BMC Bioinformatics, **7** (2006), 235–244.

[21] C. Ma, Y. Rong, D. Radiloff, M. Datto, B. Centeno, S. Bao, A. Cheng, F. Lin, S. Jiang, T. Yeatman and X Wang, *Extracellular matrix protein betaig-h3/TGFBI promotes metastasis of colon cancer by enhancing cell extravasation*, Genes Dev., **22** (2008), 308–321.

[22] L. Melchor, L. Saucedo-Cuevas, I. Munoz-Repeto, S. Rodrǵuez-Pinilla, E. Honrado, A. Campoverde, J. Palacios, K. Nathanson, M. García and J. Benítez, *Comprehensive characterization of the DNA amplification at 13q34 in human breast cancer reveals TFDP1 and CUL4A as likely candidate target genes*, Breast Cancer Research, **11** (2009).

[23] C. Palena, D. Polev, K. Tsang, et al., *The human T-box mesodermal transcription factor Brachyury is a candidate target for T-cell-mediated cancer immunotherapy*, Clin. Cancer Res., **13** (2007), 2471–2478.

[24] T. Palma, A. Conti, T. Cristofaro, S. Scala, L. Nitsch and M. Zannini, *Identification of novel Pax8 targets in FRTL-5 thyroid cells by gene silencing and expression microarray analysis*, PLoS ONE, **6** (2011), 1–10.

[25] M. Palma, L. Lopez, M. García, N. de Roja, T. Ruiz, J. García, E. Rosell, C. Vela, P. Rueda and M. Rodriguez, *Detection of collagen triple helix repeat containing-1 and nuclear factor (erythroid-derived 2)-like 3 in colorectal cancer*, BMC Clinical Pathology, **12** (2012), 2–14.

[26] M. Sabatino, M. Marabese, M. Ganzinelli, E. Caiola, C. Geroni and M. Broggini, *Down-regulation of the nucleotide excision repair gene XPG as a new mechanism of drug resistance in human and murine cancer cells*, Molecular Cancer, **9** (2010).

[27] S. Scholzel, W. Zimmermann, G. Schwarzkopf, F. Grunert, B. Rogaczewski and J. Thompson, *Carcinoembryonic antigen family members CEACAM6 and CEACAM7 are differentially expressed in normal tissues and oppositely deregulated in hyperplastic colorectal polyps and early adenomas*, Am. J. Pathol., **156** (2000), 595–605.

[28] E. Y. Song, H. G. Lee, Y. II Yeom, N. Y. Ji, J. W. Kim, S. Y. Kim, M. S. Won, K. S. Chung, Y. H. Kim, H. K. Chun and J. H. Kim, *Diagnostic kit of colon cancer using colon cancer related marker, and diagnostic method therof*, 2010, Patent WO/2010/061996.

[29] G. Stiglic and P. Kokol, *Stability of ranked gene lists in large microarray analysis studies*, Journal of Biomedicine and Biotechnology, **2010** (2010), 556–568.

[30] S. Student and K. Fujarewicz, *Stable feature selection and classification algorithms for multiclass microarray data*, Biology Direct, **7** (2012).

[31] J. Thompson, M. Seitz, E. Chastre, M. Ditter, C. Aldrian, C. Gespach and W. Zimmermann, *Down-regulation of carcinoembryonic antigen family member 2 expression is an early event in colorectal tumorigenesis*, Cancer Research, **57** (1997), 1776–1784.

[32] V. G. Tusher, R. Tibshirani and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response*, Proceedings of the National Academy of Sciences of the United States of America, **98** (2001), 5116–5121.

[33] A. Wali, P. Morin, C. Hough, F. Lonardo, T. Seya, M. Carbone and H. Pass, *Identification of intelectin overexpression in malignant pleural mesothelioma by serial analysis of gene expression (SAGE)*, Lung Cancer, **48** (2005), 19–29.

[34] C. Walsh, S. Ogawa, H. Karahashi, D. Scoles, J. Pavelka, H. Tran, C. Miller, N. Kawamata, C. Ginther, J. Dering, M. Sanada, Y. Nannya, D. Slamon, P. Koeffler and B. Karlan, *ERCC5 is a novel biomarker of ovarian cancer prognosis*, Journal of Clinical Oncology, **26** (2008), 2952–2958.

[35] D. Witten and R. Tibshirani, *A comparison of fold-change and the t-statistic for microarray data analysis*, Stanford University, (2007), 1–13.

[36] J. Zhou, L. Zhang, Y. Gu, K. Li, Y. Nie, D. Fan and Y. Feng, *Dynamic expression of CEACAM7 in precursor lesions of gastric carcinoma and its prognostic value in combination with CEA*, World Journal of Surgical Oncology, **9** (2011).

[37] "GEDI (Genetic Diseases/Gene Discovery)," Available from: `http://gedi.ci.uchicago.edu/`.

[38] "GeneCards (Human Gene Compendium)," Available from: `http://www.genecards.org/`.

[39] "MalaCards," Available from: `http://www.malacards.org/`.

[40] " NCBI (National Center for Biotechnology Information) Gene Database," Available from: `http://www.ncbi.nlm.nih.gov/`.

[41] "OMIM (Online Mendelian Inheritance in Man)," Available from: `http://www.ncbi.nlm.nih.gov/omim`.

[42] "The Clinical Correlation Between Scin, Cdkl1, Cugbp1, Slc16a7 With Colorectal Cancer Liver Metastasis," 2012. Available from: http://www.globethesis.com/?t=2154330335497716 and http://www.res-medical.com/oncology/93581.

[43] "USGENE BLAST Search Portal," Available from: https://usgene.sequencebase.com/.

[44] "WikiGenes," Available from: www.wikigenes.org/.

Received June 30, 2012; Accepted January 24, 2013.

E-mail address: danuta.gawel@polsl.pl
E-mail address: krzysztof.fujarewicz@polsl.pl