Exploiting scale-free information from expression data for cancer classification

Alexey V. Antonov,[1] Igor V. Tetko,[1] Denis Kosykh,[1]
Dmitrij Surmeli,[1] Hans-Werner Mewes[1,2]

[1]GSF National Research Center for Environment and Health,
Institute for Bioinformatics, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany

[2]Department of Genome-Oriented Bioinformatics, Wissenschaftszentrum
Weihenstephan, Technische Universität München, 85350 Freising, Germany

antonov@gsf.de

Abstract: In most studies concerning expression data analyses information on the variability of gene intensity across samples is usually exploited. This information is sensitive to initial data processing which affects the final conclusions. However expression data contains scale free information which is directly comparable between different samples. We propose to use the pairwise ratio of gene expression values rather than their absolute intensities for classification of expression data. This information is stable to data processing and thus more attractive for classification analyses. In proposed schema of data analyses only information on relative gene expression levels in each sample is exploited. Testing on publicly available datasets leads to superior classification results.

1 INTRODUCTION

An important goal of DNA microarray research is to develop tools to diagnose cancer more accurately based on the genetic profile. This information can be used subsequently for therapeutic decisions and for prognostic judgments. The method became a modern tool well recognized in cancer research and diagnosis [1-3].

Two technologies, namely Affymetrix (oligonucleotide) and spotted arrays (cDNA) received a wide usage in the recent years. Affymetrix technology is based on absolute measurements of gene expression values. The cDNA technology is based on relative measurements of mRNA abundance. Usually Affymetrix technology requires only one step normalization procedure. This step makes the gene expression values comparable between different samples. The genes intensities within a chip are measured directly. On the contrary normalization for cDNA technology require two steps procedure: correction of gene intensities within a slide and correction of gene intensities between different slides.

In this study we propose to use not gene expression profiles for classification analyses but the profiles, which represent ratios of expression value of two different genes. Gene pairwise ratio (GPR) profile is scale free, i.e. the ratio of expression values of two genes automatically comparable between different samples. The information, which is contained in GPR profile is different from both gene expression profiles although dependent. At the same time it is scale free and thus more robust for classification purposes.

The Gene pairwise ratios for cancer diagnostic have been used by Gordon et al[4,5]. The ratios of the expression levels of selected genes (<10) were employed for cancer classification. Nevertheless demonstrating promising classification results the idea is not widely known within the bioinformatics community and is not systematically examined.

We demonstrate the use of (GPR) profiles as a powerful and selective tool for cancer classification of expression data. We perform a systematic comparison between the proposed and standard classification schemas by applying different classification algorithms on several publicly available datasets. Our results suggest that using GPR rather then gene expression profiles are preferable for classification purposes.

2 METHODS

Let g_i^m be the expression matrix of *N* genes *i = 1,...,N* and *n* samples *m = 1,...,n*. Each matrix row $\boldsymbol{g_i}$ represents the expression profile of gene *i*, and each matrix column $\boldsymbol{g^m}$ represents the expression values of the array *m.* For each pair of genes *i* and *j (i<j)* we compute pairwise ratios $r_l^m = g_m^i / g_m^j$ where *l = 1,...,N(N-1)/2* is the index for a new ratio feature. Thus r_l^m represents the matrix of gene pairwise ratios (GPR). Each matrix row $\boldsymbol{r_l}$ represents the GPR profile of a pair of genes *(i,j)* and each matrix column $\boldsymbol{r^m}$ represents the GPR values for the array *m.* In this study we propose that for classification and feature selection purposes, analyzing r_l^m is preferable to g_i^m**.** The arguments are straightforward. The dataset of GPR profiles contains information on the variation of relative expression values of different pairs of genes across samples. These values are automatically comparable between different samples. Thus the classification of GPR data uses only information on relative expression values of different genes in each sample and discards information about gene absolute (normalized) expression values between different samples which is the case when one is analyzing the matrix g_i^m directly. The price for this advantage is a huge dimensionality of matrix r_l^m. The number of all possible ratios for a microarray with *N* genes is *N(N-1)/2*. For example, a chip with ten thousand genes is transformed into an object with approximately fifty million GPR features. In this case, an application of most classification procedures directly is virtually infeasible. However, we do not need to use all the features and can select a limited number of related ones only. Thus in reality the dimension $\boldsymbol{r^m}$ is much smaller**.** Further we will refer to classification methods utilizing information from the matrix g_i^m directly as *traditional* or *gene schema* and classification analysis of the GPR matrix r_l^m as *ratio schema*.

We start by developing a feature selection procedure to reduce the dimension of feature space. Let us consider a general case of multi classification task and denote $k = 1,..,K$ class labels. K is total number of different classes. For each class k and feature l (whether gene or GPR profile) we may define a correlation measure. For this purpose each class k is associated with a binary vector $\mathbf{z_k} = (z_1,...,z_n)$. The components of vector $\mathbf{z_k}$ correspond to sample labels and are 0 or 1 depending on whether or not the corresponding sample belongs to the class k. The correlation between feature l and class k is defined as *Pearson* correlation coefficient between corresponding feature vector $\mathbf{f_l}$ ($\mathbf{f_l} = \mathbf{r_l}$ for GPR data (r_l^m) or $\mathbf{f_l} = \mathbf{g_i}$ for gene data (g_i^m)) and corresponding binary vector $\mathbf{z_k}$: $corr = \dfrac{(f_l - \bar{f}_l, z_k - \bar{z}_k)}{\|f_l - \bar{f}_l\| \|z_k - \bar{z}_k\|}$. Thus each feature l is ranked in relation to class k. We can select the top p_{corr} features for each class and use them for classification purposes. The number p_{corr} of top correlated genes was used as a feature selection parameter which was varied to select different feature sets.

It is well documented that low expressed genes represent more noisy measurements than those which are highly expressed. Therefore, most studies on classification of expression data apply filtering procedures as an initial step. Genes with invariable profiles across samples are also of less interest for classification purposes since they have small discriminating power. For this reason we applied a standard deviation (SD) filter to reduce the dimensionality of matrix g_i^m. Genes with a standard deviation of expression profile across the training samples less than a given SD threshold were removed from the set g_i^m. The number of genes p_{sd} selected with SD threshold was used as a parameter of the feature selection procedure.

It might happen that some genes are overrepresented among the top correlated ratios, i.e. most ratio features represent the ratio of one particular gene with other genes. For example, the ratio of a differentially expressed gene to genes with constant profiles will also be differential. In this case only one such feature (ratio) bears the desired information and all others are redundant. To get rid of this redundancy we filter the selected ratios by restricting the maximal number of features which may be formed by any one gene. This number p_{max_gene} was used as a parameter of the feature selection procedure for the *ratio* schema.

Thus feature selection for the *traditional gene* schema was controlled by two parameters (p_{sd}, p_{corr}) and for the *ratio* schema by three parameters (p_{sd}, p_{corr}, p_{max_gene}). To select an optimal set of classification features the parameters values were optimized during a 5-fold cross validation procedure using the training data only. We optimized the number of selected features in terms of the training set and thus no bias was introduced for test set classification performance.

We have investigated the performance of the proposed *ratio* schema in comparison to the traditional approach. For this reason we analyzed several machine learning procedures employing both "*ratio*" and standard *"gene"* classification schemas for several publicly available expression datasets.

We compare the classification performance of three classification procedures preferably used in the field: Support Vector Machines (SVM) [6], Naive Bayes (NB) and k-nearest neighbor (KNN). In our study we use the WEKA data mining tool [7]. To evaluate the performance of both schemas and not to bias the results due to parameter optimization we use each classification procedure with default internal parameters (we also screened several parameters of the used algorithms and did not observe significant differences in the performances of the methods). For example, the polynomial kernel of degree 1 was used for SVM. The number of nearest neighbors for KNN was 4.

In our study we investigated 3 publicly available expression datasets: *Multiple tumor type Data (MTT)* [8], *Malignant Gliomas (MG)* [2], *Breast Cancer (BC)* [3]. In all cases to obtain reliable results we randomly reshuffle test and training sets. 50 percent of the samples were randomly assigned to the training set and the remaining samples were used as a test set. All numeric experiments were performed with $j = 1,\ldots,100$ randomizations.

For each classification method, randomization j and a set of parameters value ($p_{sd}, p_{corr}, p_{max\_gene}$) the 5-fold cross validation accuracy for *ratio* $CV_r(p_{sd}, p_{corr}, p_{max\_gene}, j)$ and *gene* $CV_g(p_{sd}, p_{corr}, j)$ schema was computed along with the test set accuracy $T_r(p_{sd}, p_{corr}, p_{max\_gene}, j)$ and $T_g(p_{sd}, p_{corr}, j)$. The parameters $p^*_{sd}, p^*_{corr}, p^*_{\max\_gene}$ for *ratio schema* and p^*_{sd}, p^*_{corr} for *gene schema* corresponding to the best cross validation prediction rate was used to define the test set performance rate $T_r(p^*_{sd}, p^*_{corr}, p^*_{\max\_gene}, j)$ *and* $T_g(p^*_{sd}, p^*_{corr}, j)$. The general accuracy (cross validation and test) was computed by averaging $CV_r(p^*_{sd}, p^*_{corr}, p^*_{\max\_gene}, j)$, $T_r(p^*_{sd}, p^*_{corr}, p^*_{\max\_gene}, j)$ *and* $CV_g(p^*_{sd}, p^*_{corr}, j)$, $T_g(p^*_{sd}, p^*_{corr}, j)$ across the randomizations j.

We evaluated the statistical significance of the difference in prediction accuracy between the two tested schemas using a Wilcoxon signed rank test. Since the observations on the randomizations are not independent, we remark that this standard Wilcoxon signed rank test is used only as a common heuristic [9] to indicate statistical difference between average accuracies on 100 randomizations.

2 DATA

The data were downloaded from the public sources. The downloaded data were already normalized as specified in the original publications. The data may contain negative or zero expression values. In this case expression values were restricted to be greater than a given threshold, i.e. expression levels which are smaller than 50 were set to 50. The affimetrix data were additionally log transformed. No further data preprocessing was done. Next we present a brief description of each dataset used. The first two sets were generated with Affymetrix technology and the last one with cDNA spotted arrayext.

2.1 Multiple tumor type Data (MTT)

The multiple tumor type classification data [8] provides measurements for 16,063 probes in 198 tumor samples representing 14 abundant human cancer classes. The dataset is split into training and test sets. The training set contains 144 samples and the test set comprises another 54 samples.

2.2 Malignant Gliomas (MG)

The Malignant Gliomas dataset [2] contains expression profiles of approximately 12,000 genes in a set of 50 gliomas, 28 glioblastomas and 22 anaplastic oligodendrogliomas. Supervised learning approaches were used in the original study to build a two-class prediction model based on a subset of 14 glioblastomas and 7 anaplastic oligodendrogliomas based on classic histology methods. This model was then used to predict the classification of 29 clinically common, histologically nonclassic samples, 14 glioblastomas and 15 anaplastic oligodendrogliomas.

2.3 Breast Cancer (BC)

The Breast Cancer dataset [3] contains measurements for approximately 25,000 human genes using cDNA microarray technology. The supervised classification was applied to identify a gene expression signature strongly predictive of a short interval to distant metastases. The publicly available dataset consists of 97 primary breast cancers: 34 from patients who developed distant metastases within 5 years and 44 from patients who continued to be disease-free after a period of at least 5 years. To validate the prognosis, an additional independent set of primary tumors from 19 young, lymph-node-negative breast cancer patients was selected. This group consisted of 7 patients who remained metastasis free for at least five years, and 12 patients who developed distant metastases within five years.

3. RESULTS

We pursue a random reshuffling strategy in all cases. The feature selection procedure was controlled by three parameters in the case of *ratio* schema and by two in the case of *gene* schema. The parameters values tested during 5-fold cross validation procedure are presented in table 1.

Parameter	*Value*
p_{sd}	*500,1000,1500,2000,2500,3000*
p_{corr}	*5,10,20,35,55,75,100,200,300,400,...,1900,2000*
p_{max_gene}	*1, 3, 6, 9, 12, 15*

Table 1. Parameters tested during feature optimization procedure. The p_{max_gene} was used only for the ratio schema.

The classification results for *MTT,MG, BC* data sets can be found in table 2. For each classification problem, the results represent the statistical summary (mean and standard deviation) of the numerical experiments on 100 randomizations of the original dataset. The classification performance of three algorithms for both schemas is presented. Last column indicate the statistical significance of the difference between the performance of the two schemas.

Method	*ratio schema*		*gene schema*		
	training set	***test set***	*training set*	***test set***	*p –value*
		MTT dataset			
SVM	*87.0±3.3*	***85.2±3.9***	*80.4±3.9*	***77.2±3.6***	$<10^{-2}$
NB	*77.1±4.2*	***74.7±5.3***	*72.7±4.1*	***68.1±5.4***	$<10^{-2}$
KNN	*78.0±3.5*	***77.1±4.3.***	*72.4±4.4*	***69.8±4.1***	$<10^{-2}$
		MG dataset			
SVM	*77.8±3.9*	***75.2±6.7***	*71.4±4.1*	***70.1±8.6***	$<10^{-2}$
NB	*76.9±4.1*	***76.0±6.1***	*68.3±4.3*	***67.3±8.1***	$<10^{-2}$

KNN	*73.5±3.7*	***72.2±8.5***	*69.0±3.7*	***64.7±8.3***	$<10^{-2}$
BC dataset					
SVM	*74.3±3.7*	***73.5±5.7***	*68.5±5.1*	***67.3±5.7***	$<10^{-2}$
NB	*74.5±4.4*	***73.4±5.8***	*63.5±4.7*	***62.1±7.9***	$<10^{-2}$
KNN	*70.9±3.9*	***69.2±8.1***	*60.3±5.0*	***59.2±7.1***	$<10^{-2}$

Table 2. Summary of the results of the numerical experiments on data sets *MTT, MG, BC*. For both schemas the training (5-fold cross validation) and test set accuracy are reported. The last column presents the statistical significance of the difference between the classification performance of ratio and gene schemas for the test set prediction accuracy (Wilcoxon signed rank test).

As one can see from table 2 the *ratio* schema outperformed the traditional approach for all datasets. We would like to point out that in the case of each dataset using *traditional* approach we got classification results close or approximately equal to the ones reported previously. For example [10] employing LS-SVM applying the same protocol (random splitting to test and training sets) and different feature selection procedures get the following results (only best are reported): *BC – 68.4 ± 7.6, MG -70±9.* This result is very similar to the best rates reported in Table 2 for *traditional schema*.

The *MTT* dataset was originally split into training and test sets. The training set contains 144 samples and the test set comprises another 54 samples. In this form the dataset was analyzed in a number of studies and represents an established benchmark. A number of different classification procedures were applied. The best result, a 78% prediction rate on the test set (12 misclassifications of 54 test samples), was obtained using support vector machines. The classification rates for KNN varied from 60% to 68% depending on the number of preprocessed genes. For Naive Bayes the reported classification performance was 63% [11]. In table 3 we present classification results related to the original split into test and training sets of *MTT* data. As one can see the *ratio* schema yielded the best ever reported results for each particular machine learning algorithm.

MTT dataset (Original split)	*Ratio schema*		*Gene schema*	
	Training set	***Test set***	*Training set*	***Test set***
SVM	*80*	***85***	*74*	***78***
NB	*76*	***70***	*70*	***63***
KNN	*80*	***78***	*70*	***63***

Table 3. Summary of classification results for MTT dataset (original split [8])

4 DISCUSSION

This work examines a new schema for the classification of biological samples using microarray data. It exploits a principally different type of information. The standard schemas consider microarray samples as objects embedded into gene space and exploit information on the variability of mRNA concentrations across samples. In this case each classification procedure relies on prior data normalization to make gene expression values comparable between different samples. The "*ratio*" schema of data analysis embeds samples into the gene pairwise ratio data space. In contrast to the former method, we exploit only information on the variability of the ratio of expression levels of two genes among different samples. This value is automatically comparable between different samples. Therefore the "*ratio*" schema eliminates the problem of cross sample comparison without prior data normalization. The price for this is the huge dimensionality of "*ratio*" objects. To overcome this problem we apply the standard correlation based feature selection procedure.

The study also provides a comparison between standard and newly proposed classification schemas. It suggests that the ratio schema provides higher performance for the analyzed machine learning algorithms. To compare both schemas fairly we employ the same feature selection procedure and select the same number of features. The "*ratio*" schema in most cases achieved better performance and, for a number of cases, gets the best classification results ever reported. It should be mentioned that for a fair comparison of machine learning methods we did not optimize their parameters for each particular scheme, but just used the standard values recommended by WEKA.

We analyzed three data sets from both available microarray technologies, namely Affymetrix (2 sets, *MTT, MG*) and cDNA chips (*BC*). The ratio schema provided a significant improvement in the prediction ability of the investigated methods. This result indicates that the proposed methodology can be applied for datasets generated by both technologies.

Using "differentially expressed" gene pairwise ratios as class markers is also promising. Such markers are more stable to data normalization. Moreover, gene pairwise ratios are more sensitive in detecting changes in expression states of the cell. From a biological point of view "differentially expressed" gene pairwise ratios represent pairs of genes whose relative concentrations were changed. This may be a direct or indirect consequence of the different cell states. For example, let us assume that we have a pair of genes regulated by the same transcription factor (TF), one positively and the other negatively. Suppose the change in the cell state causes a small reduction in the TF concentration. This will result in an increase in concentration of the negatively regulated gene and decrease in concentration of the positively regulated one. If this effect is not strong enough those two genes could not be found as differentially expressed. However, their pairwise ratio will double the "differential" effect. Thus using ratios not only leads to noise reduction but also has the potential to increase sensitivity in detection of gene expression changes within different cell states. This may explain the higher prediction ability of features selected with the "*ratio*" schema compared to the standard

REFERENCES

[1] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537

[2] Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, et al. (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res 63: 1602-1607.

[3] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530-536.

[4] Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, et al. (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res 62: 4963-4967.

[5] Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, et al. (2003) Using gene expression ratios to predict outcome among patients with mesothelioma. J Natl Cancer Inst 95: 598-605.

[6] Vapnik VN (1998) Statistical Leaning Theory. New York: Wiley.

[7] Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20: 2479-2481.

[8] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, et al. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 98: 15149-15154.

[9] van Gestel T, Suykens JAK, Baesens B, Viaene S, Vanthienen J, et al. (2004) Benchmarking least squares support vector machine classifiers. Machine Learning 54: 5-32.

[10] Pochet N, De Smet F, Suykens JA, De Moor BL (2004) Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. Bioinformatics 20: 3185-3195.

[11] Li T, Zhang C, Ogihara M (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20: 2429-2437.