Artificial Intelligence in Medicine (2007) **40**, 29–44

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Selection of relevant genes in cancer diagnosis based on their prediction accuracy

Rosalia Maglietta [a], Annarita D'Addabbo [a], Ada Piepoli [b], Francesco Perri [b], Sabino Liuni [c], Graziano Pesole [c,d], Nicola Ancona [a,*]

[a] *Istituto di Studi sui Sistemi Intelligenti per l'Automazione, CNR Via Amendola 122/D-I, 70126 Bari, Italy*
[b] *Unità Operativa di Gastroenterologia, IRCCS, "Casa Sollievo della Sofferenza"-Ospedale, Viale Cappuccini, 71013 San Giovanni Rotondo (FG), Italy*
[c] *Istituto di Tecnologie Biomediche, Sede di Bari, CNR Via Amendola 122/D, 70126 Bari, Italy*
[d] *Dipartimento di Biochimica e Biologia Molecolare, Universitá di Bari, Via E. Orabona 4, 70126 Bari, Italy*

Received 27 January 2006; received in revised form 1 June 2006; accepted 6 June 2006

KEYWORDS
Cancer diagnosis;
Gene selection;
DNA microarray;
Supervised learning;
Classification

Summary

Motivations: One of the main problems in cancer diagnosis by using DNA microarray data is selecting genes relevant for the pathology by analyzing their expression profiles in tissues in two different phenotypical conditions. The question we pose is the following: how do we measure the relevance of a single gene in a given pathology?
Methods: A gene is relevant for a particular disease if we are able to correctly predict the occurrence of the pathology in new patients on the basis of its expression level only. In other words, a gene is informative for the disease if its expression levels are useful for training a classifier able to generalize, that is, able to correctly predict the status of new patients. In this paper we present a selection bias free, statistically well founded method for finding relevant genes on the basis of their classification ability.
Results: We applied the method on a colon cancer data set and produced a list of relevant genes, ranked on the basis of their prediction accuracy. We found, out of more than 6500 available genes, 54 overexpressed in normal tissues and 77 over-expressed in tumor tissues having prediction accuracy greater than 70% with p-value ≤ 0.05.
Conclusions: The relevance of the selected genes was assessed (a) statistically, evaluating the p-value of the estimate prediction accuracy of each gene; (b)

* Corresponding author. Tel.: +39 080 5929428; fax: +39 080 5929460.
E-mail address: ancona@ba.issia.cnr.it (N. Ancona).

0933-3657/$ – see front matter
doi:10.1016/j.artmed.2006.06.002

biologically, confirming the involvement of many genes in generic carcinogenic processes and in particular for the colon; (c) comparatively, verifying the presence of these genes in other studies on the same data-set.

1. Introduction

The analysis of gene expression profiles with DNA microarrays has become a mainstay of genomics research [1]. In fact, this technology allows to measure the expression levels of thousands of genes simultaneously, providing a molecular snapshot of the status of a sample of cells in a given tissue. One of the main problems in gene expression analysis is to determine genes which are differentially expressed either in different tissues [2] or in the same tissue in two phenotypically different conditions [3]. The latter problem is particularly relevant in oncology [4] where one attempts to correlate gene expression profiles to different types of tumors [5,6] or to different stages of the same pathology [7]. Selecting the most informative genes in cancer diagnosis and prognosis is relevant for several reasons both biological and computational. Finding genes whose expression levels correlate with a particular disease is important for choosing the most appropriate treatment and for predicting recurrence of the disease [8]. Moreover, it would allow the design of ad hoc and more economic DNA microarrays tailored for the particular pathology, recording the expression levels of some tens of genes only. Furthermore, the selection of a subset of genes to work with reduces in principle the risk of "overfitting" [9], which arises when the number d of features (gene expression levels) is extremely larger than the number ℓ of specimens.

In this paper we focus on the problem of finding differentially expressed genes, relevant for a particular pathology, by analyzing their expression profiles in tissues belonging to two different groups, for example disease versus normal tissues, or one subtype versus another subtype [3,5,7]. In this problem, we typically have a sample of ℓ labelled data $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_\ell, y_\ell)\}$, called examples, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$ for $i = 1, 2, \ldots, \ell$. Here $\mathbf{x}_i$ represents the gene expression profile of the i th tissue and the label y_i indicates its class, for example $y_i = 1$ for normal and $y_i = -1$ for disease tissue. In general, we use a suitable statistics[1] for producing a ranked gene list, having the most differentially expressed genes at the top positions of the list. Successively, the expression levels of the most significant genes are used for training and testing a classifier by using the examples in S. The prediction error of the classifier is used as a measure of the relevance of the selected genes with respect to the pathology at hand. Signal-to-noise ratio [3,10], entropy [11], probability of selection [12], recursive feature elimination in its various forms [9,13,12] are examples of statistics commonly used for gene ranking. All these studies suggest that a few number of examples [14,15] and, more important, a few number of genes [16] are sufficient for obtaining high classification accuracy. Such experimental evidence leads to studying procedures that examine one gene at a time with respect to a given pathology (see [16] and references therein). Although some information could be lost by not considering genes jointly, focusing on single genes often simplifies the biological interpretation of the results. The question we pose is the following: how do we measure the relevance of a single gene in a given pathology? A gene is relevant for a given disease if we are able to correctly predict the occurrence of the pathology in new patients on the basis of that gene only. In other words, a gene is informative for a given disease if its expression levels are useful for training a classifier able to generalize, that is, able to correctly predict the status of new patients [17]. So, generalization ability of a predictor trained by using the expression levels of a single gene is a measure of the relevance of the gene in the pathology at hand. Moreover it provides an estimate of the differentiation degree of the expression levels of the gene in the examined tissues. In fact, differentially expressed genes should exhibit higher prediction accuracy than uniformly distributed genes.

The idea of selecting and ranking genes according to their classification ability is not new and it is present in literature under different guises. In [16] a logistic regression model is used for fitting the data and the accuracy of the fitted model is measured by the maximum likelihood. In [18] a simple threshold classifier is proposed and the accuracy of the model is measured by the number of misclassified training examples. The main problem of these methods is that they use the training error as a measure of prediction accuracy that, in general, does not coincide with the generalization error [17]. In fact, it is

[1] A statistics is any function of the sample S.

well known that if the hypothesis space[2] is too large, functions can be found which fit exactly the data, but they will have poor generalization capabilities on new data [19]. Our objective, on the contrary, is to train classifiers which are able to generalize, that is to predict the status of new patients not belonging to the training set, by using the expression levels of a single gene. So, it makes sense to use prediction accuracy as a measure of the relevance of a gene in the pathology at hand and it can be measured by using a finite number of data.

In this paper we present a selection bias free [20], statistically well founded method for finding relevant genes on the basis of their classification ability. We use regularized least squares (RLS) classifiers [21–23], a valuable alternative to support vector machine (SVM) classifiers [17] for tumor classification by DNA microarray data. Moreover, we use the leave-k-out cross validation (LKOCV) procedure which provides a statistically significant estimate of the generalization error of a learning machine [14,24]. Finally, we assess the statistical significance of the measured classification accuracies with non parametric permutation tests [25,26].

We applied the proposed method on the well known colon cancer data set [5] and produced a list of relevant genes, ranked on the basis of their prediction accuracy. We found, out of more than 6500 available genes, 54 overexpressed in normal tissues and 77 overexpressed in tumor tissues having prediction accuracy greater than 70% with p-value ≤ 0.05. The relevance of many selected genes was assessed both on biological basis and by verifying their presence in lists appeared on three papers recently published on the same data-set [5,18,11].

2. Materials and methods

2.1. Data set description

The data set we have analyzed is composed of gene expression profiles relative to 40 tumor and 22 normal colon tissue samples obtained with the Affimetrix Hum6000 oligonucleotide array [5]. Each sample consists of more than 6500 human gene expression levels. The data set and more detailed information are available on the web at site http://www.molbio.princeton.edu/colondata. In literature one can commonly find papers focused on the analysis of a restricted version of this data set composed of only 2000 gene expression levels. Differently, we choose to analyze the complete data set with the objective of selecting the most significant genes involved in this pathology starting from a more general and extensive point of view.

[2] This term indicates the class of functions we use for classifying the data.

2.2. RLS classifiers

RLS models were proposed mainly for facing regression problems with the objective of recovering a real valued function $y = f(\mathbf{x})$, starting from the knowledge of a finite number of observations of the function $(\mathbf{x}_i, y_i)$ at sparse locations of its domain and in the presence of noise [27]. The main difference between a regression and classification problem is that in the former the output variable y can assume any real value; in the latter, it can assume a finite number of possible values. In our case, y assumes only two values $\{-1, 1\}$. This means that every classification problem can be considered as a regression problem [19]. We are given a set of ℓ independent, identically distributed data $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_\ell, y_\ell)\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$ for $i = 1, 2, \ldots, \ell$. Data are drawn from a fixed but unknown probability density function $p(\mathbf{x}, y)$. Let us first consider the class of the linear functions $y = \langle \mathbf{w}, \mathbf{x} \rangle$, where $\mathbf{w} \in \mathbb{R}^d$ is a vector of parameters and the symbol $\langle \mathbf{w}, \mathbf{x} \rangle$ denotes the scalar product: $\langle \mathbf{w}, \mathbf{x} \rangle = \mathbf{w}^T \mathbf{x}$. If the random variables $\mathbf{x}$ or y have a non-vanishing mean, a bias term has to be included in the model. This is done by including a supplementary variable (constant and equal to 1) to the input vector. In the regularized least-squares approach, $\mathbf{w}$ is chosen so as to minimize the following functional:

$$L(\mathbf{w}) = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2 + \lambda \|\mathbf{w}\|^2 \tag{1}$$

where $\|\mathbf{w}\| = \sqrt{\langle \mathbf{w}, \mathbf{w} \rangle}$ is the Euclidean norm induced by the scalar product. The first term in functional L is called the empirical risk, the mean square error of the predictor $y = \langle \mathbf{w}, \mathbf{x} \rangle$ evaluated on the training data. The presence of the second term, called regularization term, can be motivated geometrically by the following considerations. An example $(\mathbf{x}_i, y_i)$ can be thought of as a point in a $\mathbb{R}^{d+1}$. Each function $y = \langle \mathbf{w}, \mathbf{x} \rangle$ determines an hyperplane in this space, approximating the examples in S. The prediction square error on point i is $\varepsilon_i = (y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2$. Let d_i the square distance between the point $\mathbf{x}_i$ and the approximating hyperplane. It is easy to see that (see Fig. 1):

$$d_i = \frac{\varepsilon_i}{1 + \|\mathbf{w}\|^2} \tag{2}$$

This equation shows that the smaller $\|\mathbf{w}\|^2$, the better the deviation ε_i approximates the true distance d_i. Hence the role of the regularization term, whose relevance in (1) depends on the value of parameter

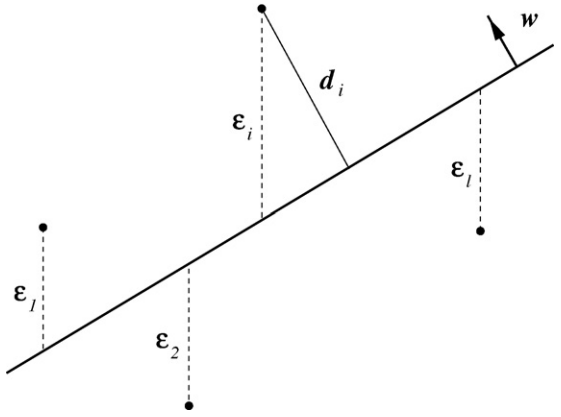

Figure 1 Geometrical interpretation of regularization.

λ, is to let the linear estimator be chosen as the hyperplane minimizing the mean square distance with the data points. The vector $\mathbf{w}$ minimizing (1) is solution of the following linear system of order d:

$$(\mathbf{X}\mathbf{X}^{\mathsf{T}} + \lambda\ell\mathbf{I}_d)\mathbf{w} = \mathbf{X}\mathbf{y}, \tag{3}$$

where $\mathbf{X}$ is a $d \times \ell$ matrix having the examples $\mathbf{x}_i$ as its columns, $\mathbf{y} = (y_1, y_2, \ldots, y_\ell)^{\mathsf{T}}$ and $\mathbf{I}_d$ is the $d \times d$ identity matrix. Note that, since the matrix $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ is positive semidefinite, then for $\lambda > 0$ the matrix $\mathbf{X}\mathbf{X}^{\mathsf{T}} + \lambda\ell\mathbf{I}_d$ is definite positive and therefore invertible (see Appendix A). Then the vector $\mathbf{w}^*$ minimizing (1) exists and it is given by:

$$\mathbf{w}^* = (\mathbf{X}\mathbf{X}^{\mathsf{T}} + \lambda\ell\mathbf{I}_d)^{-1}\mathbf{X}\mathbf{y} \tag{4}$$

It is possible to show that the value of λ controls the influence of the noise present in the data on the estimation of the solution $\mathbf{w}^*$. The parameter λ is the only free parameter and its value can be chosen by using cross validation. The classification of a new data $\mathbf{x}$ involves the evaluation of the decision function:

$$y = \text{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle) \tag{5}$$

As Eq. (4) shows, determining $\mathbf{w}^*$ requires the solution of a linear system of d order, where d is the number of components of each $\mathbf{x}_i$. In some cases d could be extremely large and so any direct method can be adopted for estimating $\mathbf{w}^*$. This occurs in the problem at hand where the number of genes d of each specimen is order of tens of thousand and the number ℓ of specimens is order of ten or hundred. We will show that the models we are describing allow to rewrite a linear system of d order as a linear system of ℓ order, overcoming the difficulties connected to problems with a huge number of features. At this aim, let us suppose $\mathbf{w}$ to be expressed as linear combination of the vectors $\mathbf{x}_i$ for $i = 1, 2, \ldots, \ell$. This means that there exist ℓ coefficients $\mathbf{c} = (c_1, c_2, \ldots, c_\ell)^{\mathsf{T}}$ such that:

$$\mathbf{w} = \mathbf{X}\mathbf{c} \tag{6}$$

Substituting (6) in (3) we have:

$$(\mathbf{K} + \lambda\ell\mathbf{I}_\ell)\mathbf{c} = \mathbf{y} \tag{7}$$

where $\mathbf{K} = \mathbf{X}^{\mathsf{T}}\mathbf{X}$ is a $\ell \times \ell$ matrix with generic element $K_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ and $\mathbf{I}_\ell$ is the identity matrix of ℓ order. Also in this case, since $\mathbf{K}$ is a positive semidefinite matrix, then for $\lambda > 0$ the matrix $\mathbf{K} + \lambda\ell\mathbf{I}_\ell$ is positive definite and so invertible. Then the vector $\mathbf{c}^* \in \mathbb{R}^\ell$ solution of (7) is given by:

$$\mathbf{c}^* = (\mathbf{K} + \lambda\ell\mathbf{I}_\ell)^{-1}\mathbf{y} \tag{8}$$

obtained by solving a linear system of ℓ order. Note that the normal $\mathbf{w}^*$ to the optimal approximating hyperplane can be recovered by using (6). In this case the classification of a new data $\mathbf{x}$ involves the evaluation of the decision function:

$$y = \text{sign}\left(\sum_{i=1}^{\ell} c_i^* \langle \mathbf{x}_i, \mathbf{x} \rangle\right) \tag{9}$$

Then the class $y \in \{-1, 1\}$ of $\mathbf{x}$ is expressed evaluating the scalar product between the data and all the elements of the training set S.

The extension of the model to the general case of non linear predictors is done by mapping the input vectors $\mathbf{x}$ in a higher dimensional feature space and looking for a linear predictor in this new space. The mapping is implicitly done by suitable kernel functions which compute scalar products in the feature space [21].

2.3. Estimating the prediction accuracy of classifiers

In this section we focus on the problem of estimating the generalization error of a classifier and of assessing the statistical significance of this estimate. The method we use consists of a cross validation procedure for estimating the error rate of a classifier with a given number of training examples and of a permutation test for assessing the statistical significance of the obtained classification performance. In particular, let n be the training set size, with $n = 1, 2, \ldots, \ell - 1$, and let $\ell - n$ be the resulting test

Table 1 Error rate e and its p-value for different training set sizes

Training set size	e	p
20	0.23	0.017
25	0.24	0.015
30	0.22	0.019
35	0.22	0.011
40	0.20	0.017
45	0.20	0.024
50	0.20	0.041
55	0.20	0.086
61	0.18	0.506

Table 2 List of overexpressed genes in normal colon tissue, their error rate *e* and the corresponding *p*-value

GAN	*e*	*p*	B	A	F	Description
M97496	0.16	0.020				Homo sapiens guanylate cyclase activator 2A (guanylin) mRNA, complete cds
M63391	0.17	0.013	*			Human desmin gene, complete cds
M76378	0.17	0.013	*		*	Human cysteine-rich protein (CRP) gene, exons 5 and 6
J02854	0.18	0.015	*		*	MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN); contains element TAR1 repetitive element
U17077	0.18	0.011				Human BENE mRNA, partial cds
M83670	0.19	0.035				Human carbonic anhydrase IV mRNA, complete cds
H54425	0.19	0.040				METALLOTHIONEIN-II (Homo sapiens)
X93349	0.20	0.039				Homo sapiens mRNA for PEP-19
H84249	0.20	0.019				ADENYLATE CYCLASE, TYPE VI (*Canis familiaris*)
T64297	0.20	0.027				FATTY ACID-BINDING PROTEIN, LIVER (HUMAN)
L02785	0.20	0.033				Homo sapiens colon mucosa-associated (DRA) mRNA, complete cds
U25138	0.20	0.038	*			Human MaxiK potassium channel beta subunit mRNA, complete cds
J03037	0.20	0.028				Human carbonic anhydrase II mRNA, complete cds
J04040	0.21	0.028				Human glucagon mRNA, complete cds
M63603	0.22	0.027				Human phospholamban mRNA, complete cds
U03749	0.22	0.039				Human chromogranin A (CHGA) gene, exon 8
R93176	0.22	0.022				CARBONIC ANHYDRASE I (HUMAN)
T55741	0.23	0.039				MYOSIN LIGHT CHAIN KINASE, SMOOTH MUSCLE (*Gallus gallus*)
U37019	0.23	0.046				Human smooth muscle cell calponin mRNA, complete cds
X52001	0.23	0.034				Homo sapiens endothelin 3 mRNA
M21221	0.23	0.031				Human follicle-stimulating hormone beta-subunit gene, exon 3
M16801	0.24	0.046				MINERALOCORTICOID RECEPTOR (HUMAN)
T60778	0.24	0.041	*			MATRIX GLA-PROTEIN PRECURSOR (*Rattus norvegicus*)
H52207	0.24	0.043				MATRIX GLA-PROTEIN PRECURSOR (HUMAN)
T71025	0.25	0.029	*			METALLOTHIONEIN-1G (Homo sapiens)
H43887	0.25	0.046	*			COMPLEMENT FACTOR D PRECURSOR (Homo sapiens)
H88665	0.25	0.041				PLEIOTROPHIN PRECURSOR (Homo sapiens)
T60155	0.25	0.038	*			ACTIN, AORTIC SMOOTH MUSCLE (HUMAN)
X54162	0.25	0.049				Human mRNA for a 64 Kd autoantigen expressed in thyroid and extra-ocular muscle
R51912	0.25	0.034				SOMATOSTATIN I PRECURSOR (HUMAN)
Z50753	0.25	0.049	*		*	Homo sapiens mRNA for GCAP-II/uroguanylin precursor
T96548	0.25	0.045				ACTIN, GAMMA-ENTERIC SMOOTH MUSCLE (HUMAN)
M74509	0.25	0.043				Human endogenous retrovirus type C oncovirus sequence
H57136	0.25	0.036				SODIUM/POTASSIUM-TRANSPORTING ATPASE GAMMA CHAIN (*Bos taurus*)
R48602	0.26	0.037				Human Smoothelin mRNA
M36634	0.26	0.027	*			Human vasoactive intestinal peptide (VIP) mRNA, complete cds
X86693	0.26	0.045	*			Homo sapiens mRNA for hevin like protein
X53416	0.26	0.030				Human mRNA for actin-binding protein (filamin) (ABP-280)
T72257	0.26	0.044				LIVER 60 KD CARBOXYLESTERASE 1 PRECURSOR (*Rattus norvegicus*)
H26655	0.26	0.026				VON WILLEBRAND FACTOR PRECURSOR (Homo sapiens)
M76424	0.26	0.019				Homo sapiens Alu repeatitive element
M55618	0.26	0.048				Homo sapiens hexabrachion (HXB) mRNA, completecds
U20325	0.26	0.039				Human cocaine and amphetamine regulated transcript CART (hCART) gene, complete cds
M87770	0.27	0.040				Human fibroblast growth factor receptor (K-sam) mRNA, complete cds
J04621	0.27	0.044				SYNDECAN-2 PRECURSOR (HUMAN); contains Alu repetitive element
U16811	0.27	0.038				Human Bak mRNA, complete cds

Table 2 (*Continued*)

GAN	*e*	*p*	B	A	F	Description
T54547	0.27	0.037				COMPLEMENT FACTOR D PRECURSOR (Homo sapiens)
L05144	0.27	0.034	*			PHOSPHOENOLPYRUVATE CARBOXYKINASE, CYTOSOLIC (HUMAN); contains Alu repetitive element; contains element PTR5 repetitive element
D42047	0.27	0.039	*			Human KIAA0089 mRNA, partial cds
M18533	0.27	0.044				Homo sapiens dystrophin (DMD) mRNA, complete cds
L11708	0.28	0.043				Human 17 beta hydroxysteroid dehydrogenase type 2 mRNA, complete cds
R52030	0.29	0.029				VON WILLEBRAND FACTOR PRECURSOR (Homo sapiens)
X80754	0.29	0.041				Homo sapiens mRNA for GTP-binding protein
H01420	0.30	0.028				AMINE OXIDASE (HUMAN)

* Indicates the presence of the gene in B, A, F (see text).

set size. We build T_1 pairs of training and test sets with n and $\ell - n$ examples, respectively, by random sampling without replacement the data set S. In the training/test split of the data, we preserve the same proportion of positive and negative examples as S. For each of these T_1 random splits, we evaluate the error rate e_{n_i} of the classifier trained on n examples, testing it on $\ell - n$ examples. So, the LKOCV error e_n is given by $e_n = \frac{1}{T_1}\sum_{i=1}^{T_1} e_{n_i}$. The second step consists in evaluating the statistical significance of the error rate e_n. In a nutshell, we are interested to measure how the observed accuracy is due to the existing correlation between gene expression levels $\mathbf{x}_i$ and class labels y_i, and how it is observed by chance because of the high dimensionality of the space where the examples live. At the aim of assessing the statistical significance of the error rate we apply the classical method of hypothesis testing. Let H_0 be the null hypothesis in which we assume that the random variables $\mathbf{x}$ and y are independent. For evaluating the p-value corresponding to e_n, we need to know the probability density function of e_n under the null hypothesis. Since it is unknown, we invoke non-parametric permutation tests [25] which allow to estimate the empirical probability density function of any statistic under H_0 from the available data. In the context of classification, the methods consists of (a) permuting randomly the labels of the training set, (b) training a random classifier on this randomly labelled training set, and (c) testing the obtained classifier on a test set having correctly labelled examples. The reason which justifies this procedure is that under the null hypothesis all the training sets generated through label permutations are equally likely to be observed, because the random variables $\mathbf{x}$ and y are independent. So the permutation test technique allows to determine how a classifier with error rate e_n works better than a classifier trained on randomly labelled data in classifying correctly labelled data. In particular we carry out the following steps. For every random split of S in training and test sets, we perform T_2 random permutations of the labels in the training set. For each permutation, we build a random classifier and test the classifier on the test set having correctly labelled examples. Let us indicate with $e_{n_{i,j}}$ the error rate of the random classifier trained on n examples in the i th cross validation and in the j th random permutation. Then the empirical probability density function of the error rate under the null hypothesis is:

$$p_n(e) = \frac{1}{T_1 T_2}\sum_{i=1}^{T_1}\sum_{j=1}^{T_2}\delta(e - e_{n_{i,j}}) \tag{10}$$

composed of a sum of delta functions[3] centered on the measured errors. The statistical significance (p-value) of the error rate e_n is given by the percentage of error rates $e_{n_{i,j}}$ smaller than e_n.

3. Results

In this section we illustrate the performances of our method for selecting relevant genes based on prediction accuracy, applied on colon cancer data set [5]. For reducing the risk of overfitting, we used linear RLS classifiers in all the experiments. At the aim of estimating the prediction accuracy of a single gene, we needed to fix the number n of examples to use in training phase. Instead of fixing this value *a priori*, we established it by using data. In particular, we measured the LKOCV error of RLS classifiers trained with a different number n of training examples and tested on the remaining $\ell - n$ examples. In this phase, all the available genes were used. Their expression levels were appropriately normalized to have zero mean and unit var-

[3] The delta function also called Dirac's delta function or impulse function [28] is usually defined by the equation $\int_{-\infty}^{\infty}\delta(t)\,dt = 1$ and $\delta(t) = 0$ for $t \neq 0$.

Table 3 List of overexpressed genes in tumor colon tissue, their error rate *e* and the corresponding *p*-value

GAN	*e*	*p*	B	A	F	Description
R94588	0.16	0.005				CELL DIVISION PROTEIN KINASE 2 (HUMAN)
T50501	0.23	0.005				EUKARYOTIC INITIATION FACTOR 1A (*Saccharomyces cerevisiae*)
J05032	0.23	0.015	*			Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds
D13642	0.23	0.040				Human splicing factor 3b, subunit 3 (SF3B3) mRNA, complete cds
L18960	0.23	0.005				Human protein synthesis factor (eIF-4C) mRNA, complete cds
H08393	0.23	0.003	*		*	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)
M31523	0.24	0.022				Human transcription factor (E2A) mRNA, complete cds
T51961	0.24	0.001				PROLIFERATING CELL NUCLEAR ANTIGEN (HUMAN)
R71676	0.24	0.003				CURVED DNA-BINDING PROTEIN (*Schizosaccharomyces pombe*)
T51621	0.24	0.015				Human ionizing radiation resistance conferring protein mRNA, complete cds
M22382	0.25	0.002	*			MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (HUMAN)
D14657	0.25	0.002				Human KIAA0101 mRNA, partial cds
H09351	0.25	0.008				Human MCM7 minichromosome maintenance deficient 7 (*S. cerevisiae*)
X12671	0.25	0.011	*			Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1
X63629	0.25	0.003	*			Homo sapiens mRNA for p cadherin
H40095	0.25	0.005	*			MACROPHAGE MIGRATION INHIBITORY FACTOR (HUMAN)
T87871	0.25	0.017				MYOBLAST CELL SURFACE ANTIGEN 24.1D5 (Homo sapiens)
R37741	0.26	0.003				Human deoxyhypusine synthase mRNA, complete cds
M34458	0.26	0.003				LAMIN B1 (HUMAN); gb:X14170 murine mRNA for lamin C (MOUSE)
Y00285	0.26	0.014				Human mRNA for insuline-like growth factor II receptor
H65842	0.26	0.004				RED CELL ACID PHOSPHATASE 1, ISOZYME F (Homo sapiens)
L19183	0.26	0.025				Human MAC30 mRNA, 3'-end
M63904	0.26	0.013				Human G-alpha 16 protein mRNA, complete cds
X66171	0.26	0.024				Homo sapiens CMRF35 mRNA, complete CDS
T64148	0.27	0.002				POLYADENYLATE-BINDING PROTEIN (*Xenopus laevis*)
H48051	0.27	0.013				SEVENLESS PROTEIN (*Drosophila virilis*)
T61949	0.27	0.010				INORGANIC PYROPHOSPHATASE (*Bos taurus*)
M80244	0.27	0.034				INTEGRAL MEMBRANE PROTEIN E16 (HUMAN)
T94764	0.27	0.006				SET PROTEIN (Homo sapiens)
M26697	0.27	0.013	*			Human nucleolar protein (B23) mRNA, complete cds
R50976	0.27	0.033				CARBOXYPEPTIDASE H (*Bos taurus*)
T51023	0.27	0.016	*		*	HEAT SHOCK PROTEIN HSP 90-BETA (HUMAN)
T52185	0.28	0.037	*	*		P17074 40S RIBOSOMAL PROTEIN
T95018	0.28	0.024	*			40S RIBOSOMAL PROTEIN S18 (Homo sapiens)
U09564	0.28	0.007	*			Human serine kinase mRNA, complete cds
X14958	0.28	0.010	*			Human hmgI mRNA for high mobility group protein Y
R37660	0.28	0.041				STATHMIN (Homo sapiens)
H55916	0.28	0.044	*		*	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR (HUMAN)
D43948	0.28	0.021				Human KIAA0097 mRNA, complete cds
H09599	0.28	0.031				MITOCHONDRIAL IMPORT RECEPTOR MOM38 (*Neurospora crassa*)
M81934	0.28	0.016				Human cdc25B mRNA, complete cds
T91563	0.28	0.009				CD44 ANTIGEN, EPITHELIAL FORM PRECURSOR (Homo sapiens)
T70920	0.28	0.015				P59 PROTEIN (Homo sapiens)
X12466	0.28	0.034	*			Human mRNA for snRNP E protein
L26953	0.29	0.034				Homo sapiens chromosomal protein mRNA, complete cds
X74987	0.29	0.040				Homo sapiens mRNA for 2'—5' oligoadenylate binding protein
T89692	0.29	0.015				COMPLEMENT FACTOR I PRECURSOR (Homo sapiens)
X55715	0.29	0.026	*	*		Human Hums3 mRNA for 40S ribosomal protein s3

Table 3 (*Continued*)

GAN	*e*	*p*	B	A	F	Description
R60357	0.29	0.007				Human mRNA for alanyl tRNA synthetase
D31885	0.29	0.009	*			Human KIAA0069 mRNA, partial cds
M77836	0.29	0.005				PYRROLINE-5-CARBOXYLATE REDUCTASE (HUMAN)
H86045	0.29	0.030				STATIN S1 (*Rattus norvegicus*)
H84154	0.29	0.045				G1/S-SPECIFIC CYCLIN D2 (Homo sapiens)
T47377	0.29	0.004	*		*	S100 calcium binding protein P (Human)
R08183	0.29	0.012	*			Q04984 10 KD HEAT SHOCK PROTEIN, MITOCHONDRIAL
R26668	0.29	0.034				PENICILLIN-BINDING PROTEIN 1A (*Haemophilus influenzae*)
R98945	0.29	0.050				SODIUM CHANNEL PROTEIN PARA (*Drosophila melanogaster*)
T67257	0.29	0.048				KILLER TOXIN-RESISTANCE PROTEIN 5 PRECURSOR (*Saccharomyces cerevisiae*)
M36981	0.29	0.009	*			Human putative NDP kinase (nm23-H2S) mRNA, complete cds
H64427	0.29	0.008				60S RIBOSOMAL PROTEIN L7 (Homo sapiens)
X07994	0.29	0.018				Human mRNA for lactase-phlorizin hydrolase LPH (EC 3.2.1.23-62)
U18934	0.29	0.044				Human receptor tyrosine kinase (DTK) mRNA, complete cds
M15476	0.29	0.028				UROKINASE-TYPE PLASMINOGEN ACTIVATOR PRECURSOR (HUMAN)
H50129	0.30	0.016				PROTEIN PHOSPHATASE PP2A, 55 KD REGULATORY SUBUNIT, NEURONAL ISOFORM (*Oryctolagus cuniculus*)
U11700	0.30	0.017				Human copper transporting ATPase mRNA, complete cds
H61535	0.30	0.012				TRANSCRIPTION FACTOR E2-ALPHA (Homo sapiens)
T61609	0.30	0.010	*	*		LAMININ RECEPTOR (HUMAN)
T86473	0.30	0.004	*			NUCLEOSIDE DIPHOSPHATE KINASE A (HUMAN)
X78627	0.30	0.008				Homo sapiens mRNA for translin
R43728	0.30	0.009				G2/MITOTIC-SPECIFIC CYCLIN A (Homo sapiens)
R54097	0.30	0.008			*	TRANSLATIONAL INITIATION FACTOR 2 BETA SUBUNIT (HUMAN)
H29320	0.30	0.025				HYPOTHETICAL GTP-BINDING PROTEIN IN PMI40-PAC2 INTERGENIC REGION (*Saccharomyces cerevisiae*)
U22055	0.30	0.021				Human 100 kDa coactivator mRNA, complete cds
R84411	0.30	0.025	*			SMALL NUCLEAR RIBONUCLEOPROTEIN ASSOCIATED PROTEINS B AND B' (HUMAN)
X13293	0.30	0.016				MYB-RELATED PROTEIN B (HUMAN)
T51571	0.30	0.036	*			P24480 CALGIZZARIN
M61763	0.30	0.048				Human alanine:glyoxylate aminotransferase (AGT1) gene, exon 11 and mRNA

* Indicates the presence of the gene in B, A, F (see text).

iance. In particular, for each pair of training and test sets with n and $\ell - n$ examples, respectively, we used the n training examples for computing the mean and variance of each gene and used these parameters for normalizing the genes in both training and test sets. Table 1 shows the error rate e and the p-value of RLS classifiers, obtained by varying the number of training examples. The error values were estimated performing $T_1 = 500$ cross validations and $T_2 = 500$ random permutations of the labels. Note that, by virtue of Eq. (8), for a given cross validation, we do not need to ''retrain'' the classifier in each random permutation. In fact, the new vector $\mathbf{c}^*$ associated to a permutation of the labels in $\mathbf{y}$ is obtained by multiplying $\mathbf{y}$ with the matrix $(\mathbf{K} + \lambda\ell\mathbf{I}_\ell)^{-1}$ which is independent of $\mathbf{y}$. This simple property of RLS classifiers considerably reduces the computational complexity of our method. The best performances were obtained with $n = 40$ training examples, with an error rate $e = 20\%$ and p-value = 0.017. In fact, although the error rate is constant in the range [40, 55], it reached the smallest p-value for $n = 40$. Note that the leave-one-out (LOO) error (last row in Table 1), although it exhibits poor statistical significance, has a value comparable to the ones of the LKOCV error when n is in the range [40, 55]. This means that the LOO error provides a good estimate of the generalization error of a learning machine [29] and it can be used as a valid alternative to LKOCV error for comparing the performances of different classification rules. This aspect is relevant for RLS classifiers which require just one training for evaluating the LOO error [23]. Moreover, our results are in agreement with the ones described in [14] where it is shown that 10—20 examples suffice for training

Table 4 Error rate e and p-value of the relevant genes reported in [5]

GAN	e	p	Description
T52185	0.29	0.035	P17074 40S RIBOSOMAL PROTEIN
X55715	0.29	0.029	Human Hums3 mRNA for 40S ribosomal protein s3
T61609	0.29	0.010	LAMININ RECEPTOR (HUMAN)
T57619	0.31	0.011	40S RIBOSOMAL PROTEIN S6 (*Nicotiana tabacum*)
T58861	0.32	0.017	60S RIBOSOMAL PROTEIN L30E (*Kluyveromyces lactis*)
T57633	0.34	0.046	40S RIBOSOMAL PROTEIN S8 (HUMAN)
R50158	0.34	0.024	MITOCHONDRIAL LON PROTEASE HOMOLOG PRECURSOR (Homo sapiens)
T52015	0.35	0.028	ELONGATION FACTOR 1-GAMMA (HUMAN)
T72879	0.35	0.034	60S RIBOSOMAL PROTEIN L7A (HUMAN)
T48804	0.36	0.024	40S RIBOSOMAL PROTEIN S24 (HUMAN)
T49423	0.36	0.045	BREAST BASIC CONSERVED PROTEIN 1 (HUMAN)
U14971	0.36	0.026	Human ribosomal protein S9 mRNA, complete cds
T72938	0.34	0.051	QM PROTEIN (HUMAN)
T51560	0.36	0.054	40S RIBOSOMAL PROTEIN S16 (HUMAN)
R22197	0.34	0.060	60S RIBOSOMAL PROTEIN L32 (HUMAN)
T63591	0.35	0.061	60S ACIDIC RIBOSOMAL PROTEIN P0 (HUMAN)
T63484	0.38	0.757	Human ornithine decarboxylase antizyme (Oaz). mRNA, complete cds
H09263	0.37	0.763	ELONGATION FACTOR 1-ALPHA 1 (Homo sapiens)
T51496	0.37	0.772	60S RIBOSOMAL PROTEIN L37A (HUMAN)
T56934	0.37	0.774	Homo sapiens alpha NAC mRNA
T47144	0.37	0.774	JN0549 RIBOSOMAL PROTEIN YL30
R86975	0.37	0.789	40S RIBOSOMAL PROTEIN S28 (HUMAN)
R01182	0.38	0.791	60S RIBOSOMAL PROTEIN L38 (HUMAN)
R85464	0.37	0.796	ATP SYNTHASE LIPID-BINDING PROTEIN P2 PRECURSOR (HUMAN)
H77302	0.36	0.803	60S RIBOSOMAL PROTEIN (HUMAN)
R02593	0.37	0.807	60S ACIDIC RIBOSOMAL PROTEIN P1 (*Polyorchis penicillatus*)
H54676	0.37	0.807	60S RIBOSOMAL PROTEIN L18A (HUMAN)
T52642	0.36	0.816	GUANYLATE KINASE HOMOLOG (*Vaccinia virus*)

classification rules with high statistical significance.

For measuring the prediction accuracy of each gene singularly, we measured the LKOCV error of one-dimensional RLS classifiers by using $n = 40$ training and $\ell - n = 22$ test examples. In the case of one-dimensional, linear RLS classifiers, the training and test phases are extremely cheap from a computational point of view because they require very simple computations (see Appendix B for details). The LKOCV error was measured by performing $T_1 = 500$ cross validations of the examples in the data set. Also in this case, each gene was normalized to have zero mean and unit variance, and mean and variance were computed by using the training examples only. The statistical significance (p-value) was computed only for those genes associated to an error rate $e \leq 40\%$. In fact, classifiers with $e > 40\%$ have performances close to random classifiers and so very poor statistical significance. This property significantly reduces the number of genes for which it makes sense to evaluate the p-value. For these genes we applied the permutation test performing $T_2 = 500$ random permutations of the labels. Tables 2 and 3 show the error rate e and the corresponding p-value for genes most expressed in normal and tumor tissues, respectively. A gene g is most expressed in normal tissues if $\mu_+(g) \geq \mu_-(g)$, where $\mu_+(g)$ and $\mu_-(g)$ are the averages of the expression levels of g in normal and tumor tissues, respectively. Only the genes having $e \leq 30\%$ and $p \leq 0.05$ are shown, sorted according to their error rate. Besides reporting the accession number (GAN) and a brief description of the gene, we point out with an asterisk those genes appeared in papers recently published on the same data set [5,18,11], indicated with the letters A, B, F, respectively. Note that some genes, for example guanylate cyclase activator 2A (guanylin) (GUCA2A; gene ID: M97496) and cell division protein kinase 2 (CDK2; gene ID: R94588) with the lowest error rates in the Tables 2 and 3, respectively, have a statistically significative prediction error of 16% ($p = 0.02$) and 16% ($p = 0.005$), lower than 20%, that is the error rate obtained using all the genes (see Table 1). Such genes are effectively related to colon cancer.

Table 5 Error rate *e* and *p*-value of the relevant genes reported in [11]

GAN	*e*	*p*	Description
M76378	0.17	0.029	Human cysteine-rich protein (CRP) gene, exons 5 and 6
J02854	0.18	0.031	MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN)
H08393	0.23	0.004	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)
Z50753	0.25	0.050	Homo sapiens mRNA for GCAP-II/uroguanylin precursor
T51023	0.28	0.017	HEAT SHOCK PROTEIN HSP 90-BETA (HUMAN)
H55916	0.28	0.046	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR (HUMAN)
T47377	0.30	0.004	S100 calcium binding protein P (Human)
R54097	0.30	0.008	TRANSLATIONAL INITIATION FACTOR 2 BETA SUBUNIT (HUMAN)
T79152	0.30	0.019	60S RIBOSOMAL PROTEIN L19 (HUMAN)
T62947	0.31	0.021	60S RIBOSOMAL PROTEIN L24 (*Arabidopsis thaliana*)
M80815	0.31	0.040	Homo sapiens a-L-fucosidase gene, exon 7 and 8, and complete cds
D13315	0.32	0.022	LACTOYLGLUTATHIONE LYASE (HUMAN); contains Alu repetitive element
T58861	0.32	0.015	60S RIBOSOMAL PROTEIN L30E (*Kluyveromyces lactis*)
T57619	0.32	0.031	40S RIBOSOMAL PROTEIN S6 (*Nicotiana tabacum*)
R36977	0.32	0.005	P03001 TRANSCRIPTION FACTOR IIIA
R15447	0.33	0.024	CALNEXIN PRECURSOR (Homo sapiens)
T51261	0.34	0.009	GLIA DERIVED NEXIN PRECURSOR (*Mus musculus*)
D14812	0.35	0.032	Human mRNA for ORF, complete cds
J04102	0.35	0.012	Human erythroblastosis virus oncogene homolog 2 (ets-2) mRNA, complete cds
U37012	0.36	0.015	Human cleavage and polyadenylation specificity factor mRNA, complete cds
T41204	0.36	0.027	P14780 92 KD TYPE V COLLAGENASE PRECURSOR
M35878	0.36	0.035	Human insulin-like growth factor-binding protein-3 gene, complete cds, clone HL1006d
M26383	0.36	0.000	Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds
R62549	0.36	0.002	PUTATIVE SERINE/THREONINE-PROTEIN KINASE B0464.5 IN CHROMOSOME III (*Caenorhabditis elegans*)
R01755	0.36	0.037	TRANSTHYRETIN PRECURSOR (Homo sapiens)
M82919	0.35	0.055	Human gamma amino butyric acid (GABAA) receptor beta-3 subunit mRNA, complete cds
T94579	0.36	0.059	Human chitotriosidase precursor mRNA, complete cds
T61661	0.30	0.064	PROFILIN I (HUMAN)
T51849	0.34	0.067	TYROSINE-PROTEIN KINASE RECEPTOR ELK PRECURSOR (*Rattus norvegicus*)
R87126	0.25	0.067	MYOSIN HEAVY CHAIN, NONMUSCLE (*Gallus gallus*)
Z49269	0.32	0.077	Homo sapiens gene for chemokine HCC-1
H20709	0.35	0.089	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (HUMAN)
R44418	0.36	0.096	EBNA-2 NUCLEAR PROTEIN (Epstein-barr virus)
T47383	0.36	0.098	ALKALINE PHOSPHATASE, PLACENTAL TYPE 1 PRECURSOR (Homo sapiens)
M20543	0.34	0.110	Human skeletal alpha-actin gene, complete cds
H81558	0.35	0.117	PROCYCLIC FORM SPECIFIC POLYPEPTIDE B1-ALPHA PRECURSOR (*Trypanosoma brucei brucei*)
R88740	0.34	0.130	ATP SYNTHASE COUPLING FACTOR 6, MITOCHONDRIAL PRECURSOR (HUMAN)
T51539	0.34	0.131	HEPATOCYTE GROWTH FACTOR-LIKE PROTEIN PRECURSOR (Homo sapiens)
M92287	0.32	0.137	Homo sapiens cyclin D3 (CCND3) mRNA, complete cds
T94993	0.33	0.139	FIBROBLAST GROWTH FACTOR RECEPTOR 2 PRECURSOR (Homo sapiens)
H64489	0.33	0.148	LEUKOCYTE ANTIGEN CD37 (Homo sapiens)
X02875	0.36	0.150	Human mRNA (3′-fragment) for (2'—5') oligo A synthetase E (1.8 kb RNA)
T67406	0.34	0.173	COMPLEMENT C4 PRECURSOR (Homo sapiens)
R80427	0.34	0.201	C4-DICARBOXYLATE TRANSPORT SENSOR PROTEIN DCTB (*Rhizobium leguminosarum*)
L07648	0.32	0.221	Human MXI1 mRNA, complete cds
X68314	0.39	0.742	Homo sapiens mRNA for glutathione peroxidase-GI
R81170	0.39	0.758	TRANSLATIONALLY CONTROLLED TUMOR PROTEIN (Homo sapiens)
M81651	0.39	0.772	Human semenogelin II (SEMGII) gene, complete cds
H06061	0.37	0.775	VOLTAGE-DEPENDENT ANION-SELECTIVE CHANNEL PROTEIN 1 (Homo sapiens)
R59583	0.37	0.780	PRE-MRNA SPLICING FACTOR SRP75 (Homo sapiens)
M31303	0.37	0.782	Human oncoprotein 18 (Op18) gene, complete cds

Table 5 (*Continued*)

GAN	*e*	*p*	Description
T72863	0.37	0.783	FERRITIN LIGHT CHAIN (HUMAN)
J03210	0.37	0.785	Human collagenase type IV mRNA, 3'-end
T88902	0.38	0.786	COT PROTO-ONCOGENE SERINE/THREONINE-PROTEIN KINASE (Homo sapiens)
H20289	0.37	0.788	G25K GTP-BINDING PROTEIN, BRAIN ISOFORM (Homo sapiens)
H81864	0.37	0.790	CELL DIVISION CONTROL PROTEIN 2 HOMOLOG (*Plasmodium falciparum*)
H16096	0.37	0.792	MITOCHONDRIAL PROCESSING PROTEASE BETA SUBUNIT PRECURSOR (*Rattus norvegicus*)
T57882	0.37	0.793	MYOSIN HEAVY CHAIN, NONMUSCLE TYPE A (Homo sapiens)
K02268	0.36	0.805	Human enkephalin B (enkB) gene, exon 4 and 3'-flank and complete cds
X17025	0.37	0.808	Human homolog of yeast IPP isomerase
R39531	0.37	0.814	PROBABLE 26S PROTEASE SUBUNIT SUG1 (*Xenopus laevis*)
H01418	0.37	0.836	SON OF SEVENLESS PROTEIN (*Drosophila melanogaster*)
H64807	0.37	0.836	PLACENTAL FOLATE TRANSPORTER (Homo sapiens)
T47424	0.40	0.836	INSULIN RECEPTOR SUBSTRATE-1 (Homo sapiens)
M28219	0.37	0.838	Homo sapiens low density lipoprotein receptor (FH 10 mutant causing familial hypercholesterolemia) mRNA, 3'-end
R33481	0.39	0.838	TRANSCRIPTION FACTOR ATF-A AND ATF-A-DELTA (Homo sapiens)
H24401	0.36	0.843	MAP KINASE PHOSPHATASE-1 (Homo sapiens)
U00968	0.37	0.845	STEROL REGULATORY ELEMENT BINDING PROTEIN 1 (HUMAN)
D17532	0.37	0.846	Human mRNA for RCK, complete cds
T40507	0.37	0.848	CCAAT-BINDING TRANSCRIPTION FACTOR I SUBUNIT A (Homo sapiens)
R67275	0.38	0.857	COLLAGEN ALPHA 1(XI) CHAIN PRECURSOR (Homo sapiens)
T74556	0.37	0.877	ATP SYNTHASE ALPHA CHAIN, MITOCHONDRIAL PRECURSOR (HUMAN)
T84051	0.37	0.887	G25K GTP-BINDING PROTEIN, PLACENTAL ISOFORM (Homo sapiens)
H49870	0.39	0.891	MAD PROTEIN (Homo sapiens)
M64231	0.37	0.896	Human spermidine synthase gene, complete cds
T98835	0.38	0.896	80.7 KD ALPHA TRANS-INDUCING PROTEIN (Bovine herpesvirus type 1)
K03474	0.37	0.908	Human Mullerian inhibiting substance gene, complete cds
T79831	0.37	0.934	MAP KINASE PHOSPHATASE-1 (Homo sapiens)
T64012	0.37	0.935	ACETYLCHOLINE RECEPTOR PROTEIN, DELTA CHAIN PRECURSOR (*Xenopus laevis*)
M23115	0.37	0.939	Homo sapiens calcium-ATPase (HK2) mRNA, complete cds

We found that guanylate cyclase activator 2A and uroguanylin (GUCA2B; gene ID: Z50753), included in Table 2 with error rate $e = 25\%$ ($p = 0.049$), are up-regulated in normal tissues. In fact, guanylin and uroguanylin are markedly reduced in early colon tumor with very low expression in adenocarcinoma of the colon and also in its begin precursor, adenoma [30,31]. Treatment with uroguanylin has recently been found to have possible therapeutic significance with a significant reduction in the number of pre-cancerous colon polyps (adenomas), shrinkage in the remainder and observed apoptosis of adenocarcinoma cells [31].

The expression of CDK2 gene (see Table 3) increases progressively during the carcinogenic process and its overexpression takes part in colorectal carcinogenesis [32]. Mechanisms have been proposed for reducing the expression of CDK2. In fact, in [33], the authors evaluated the effect of aspirin (ASA) and three other structurally unrelated NSAIDs (indomethacin, naproxen, and piroxicam) on cell proliferation, cell cycle phase distribution, and the development of apoptosis in HT-29 colon adenocarcinoma cells in vitro. They found that, parallel to their effect on cell cycle, ASA and indomethacin reduced the levels of cyclin-dependent kinases 2 (CDK2) that is important for cell cycle progression. The findings presented in this paper suggest possible mechanisms for the cancer preventive effects of these compounds in human.

Other genes determined by our method are relevant for colon cancer. Among highly expressed genes in tumor tissues (see Table 3), Collagen alpha 2(XI) chain (COL9A2; gene ID: H08393) and CD44 antigen epithelial form precursor (CD44; gene ID: T91563) genes are particularly significant having error rates $e = 23\%$ ($p = 0.003$) and $e = 28\%$ ($p = 0.009$), respectively. COL9A2 is involved in cell adhesion. Colon carcinoma cells have collagen degrading activity as part of the metastatic process [9,31]. Dysregulated expression of CD44 isoforms is involved in promoting cell transformation into colon carcinogenesis, and in most other types of cancer, and this implicates an acquisition of resistance to apoptosis [34]. An other important gene over-expressed in tumor tissue and found by our method

Table 6 Error rate e and p-value of the relevant genes reported in [18]

GAN	e	p	Description
M63391	0.17	0.013	Human desmin gene, complete cds
M76378	0.17	0.025	Human cysteine-rich protein (CRP) gene, exons 5 and 6
J02854	0.18	0.013	MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN)
U25138	0.21	0.036	Human MaxiK potassium channel beta subunit mRNA, complete cds
H08393	0.23	0.004	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)
J05032	0.24	0.014	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds
T60778	0.24	0.044	MATRIX GLA-PROTEIN PRECURSOR (*Rattus norvegicus*)
T60155	0.24	0.035	ACTIN, AORTIC SMOOTH MUSCLE (HUMAN)
Z50753	0.24	0.050	Homo sapiens mRNA for GCAP-II/uroguanylin precursor
T71025	0.24	0.029	METALLOTHIONEIN-1G (Homo sapiens)
M22382	0.25	0.002	MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (HUMAN)
H40095	0.25	0.005	MACROPHAGE MIGRATION INHIBITORY FACTOR (HUMAN)
H43887	0.25	0.043	COMPLEMENT FACTOR D PRECURSOR (Homo sapiens)
X86693	0.26	0.047	Homo sapiens mRNA for hevin like protein
X63629	0.26	0.003	Homo sapiens mRNA for p cadherin
X12671	0.26	0.011	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1
M36634	0.26	0.024	Human vasoactive intestinal peptide (VIP) mRNA, complete cds
M26697	0.27	0.006	Human nucleolar protein (B23) mRNA, complete cds
D42047	0.27	0.042	HumanKIAA0089 mRNA, partial cds
L05144	0.27	0.037	PHOSPHOENOLPYRUVATE CARBOXYKINASE, CYTOSOLIC (HUMAN)
M64110	0.27	0.042	Human caldesmon mRNA, complete cds
T51023	0.27	0.006	HEAT SHOCK PROTEIN HSP 90-BETA (HUMAN)
U09564	0.27	0.008	Human serine kinase mRNA, complete cds
T95018	0.27	0.024	40S RIBOSOMAL PROTEIN S18 (Homo sapiens)
X14958	0.28	0.011	Human hmgI mRNA for high mobility group protein Y
T52185	0.28	0.035	P17074 40S RIBOSOMAL PROTEIN
H55916	0.29	0.039	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR (HUMAN)
R08183	0.29	0.014	Q04984 10 KD HEAT SHOCK PROTEIN, MITOCHONDRIAL
X55715	0.29	0.028	Human Hums3 mRNA for 40S ribosomal protein s3
T61609	0.29	0.010	LAMININ RECEPTOR (HUMAN)
D31885	0.29	0.009	HumanKIAA0069 mRNA,partial cds
T47377	0.29	0.005	S100 calcium binding protein P (Human)
X12466	0.29	0.031	Human mRNA for snRNP E protein
M36981	0.30	0.010	Human putative NDP kinase (nm23-H2S) mRNA, complete cds
T86473	0.30	0.004	NUCLEOSIDE DIPHOSPHATE KINASE A (HUMAN)
T79152	0.30	0.018	60S RIBOSOMAL PROTEIN L19 (HUMAN)
T51571	0.30	0.035	P24480 CALGIZZARIN
R84411	0.30	0.027	SMALL NUCLEAR RIBONUCLEOPROTEIN ASSOCIATED PROTEINS B AND B′ (HUMAN)
H89087	0.30	0.013	SPLICING FACTOR SC35 (Homo sapiens)
X70326	0.31	0.016	Homo sapiens MacMarcks mRNA
M80815	0.31	0.041	Homo sapiens a-L-fucosidase gene, exon 7 and 8, and complete cds
T83368	0.31	0.015	MEMBRANE COFACTOR PROTEIN PRECURSOR (Homo sapiens)
D63874	0.31	0.024	Human mRNA for HMG-1
R36977	0.31	0.003	P03001 TRANSCRIPTION FACTOR IIIA
T51529	0.31	0.008	ELONGATION FACTOR 1-DELTA (*Artemia salina*)
R75843	0.32	0.011	TRANSLATIONAL INITIATION FACTOR 2 GAMMA SUBUNIT (Homo sapiens)
T62947	0.32	0.040	60S RIBOSOMAL PROTEIN L24 (*Arabidopsis thaliana*)
U30825	0.32	0.009	Human splicing factor SRp30c mRNA, complete cds
U17899	0.32	0.017	Human chloride channel regulatory protein mRNA, complete cds
X56597	0.32	0.013	Human humFib mRNA for fibrillarin
U32519	0.32	0.045	Human GAP SH3 binding protein mRNA, complete cds
T96873	0.32	0.009	HYPOTHETICAL PROTEIN IN TRPE 3′REGION (*Spirochaeta aurantia*)
U26312	0.32	0.008	Human heterochromatin protein HP1Hs-gamma mRNA, partial cds
X54942	0.33	0.008	Homo sapiens ckshs2 mRNA for Cks1 protein homologue
X15183	0.33	0.017	Human mRNA for 90-kDa heat-shock protein
H20819	0.33	0.046	26S PROTEASE REGULATORY SUBUNIT 6 (Homo sapiens)
L41559	0.33	0.039	Homo sapiens pterin-4a-carbinolamine dehydratase (PCBD) mRNA, complete cds

Table 6 (*Continued*)

GAN	*e*	*p*	Description
T57633	0.33	0.046	40S RIBOSOMAL PROTEIN S8 (HUMAN)
T86749	0.33	0.006	Human (clone PSK-J3) cyclin-dependent protein kinase mRNA, complete cds
R42501	0.34	0.019	INOSINE-5′-MONOPHOSPHATE DEHYDROGENASE 2 (HUMAN)
R64115	0.34	0.005	ADENOSYLHOMOCYSTEINASE (Homo sapiens)
X70944	0.34	0.039	Homo sapiens mRNA for PTB-associated splicing factor
L08069	0.35	0.024	Human heat shock protein, E. coli DnaJ homologue mRNA, complete cds
H40560	0.35	0.019	THIOREDOXIN (HUMAN)
X13482	0.35	0.011	U2 SMALL NUCLEAR RIBONUCLEOPROTEIN A' (HUMAN); contains MER22 repetitive element
U29092	0.35	0.023	Human ubiquitin conjugating enzyme mRNA, complete cds
X53586	0.36	0.008	Human mRNA for integrin alpha 6
H87135	0.36	0.026	IMMEDIATE-EARLY PROTEIN IE180 (*Pseudorabies virus*)
T40454	0.36	0.004	ANTIGENIC SURFACE DETERMINANT PROTEIN OA3 PRECURSOR (Homo sapiens)
D00596	0.36	0.004	Human thymidylate syntase (EC 2.1.1.45) gene, complete cds
M26383	0.36	0.000	Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds
X62048	0.36	0.038	Homo sapiens Wee1 hu gene
R52081	0.36	0.009	TRANSCRIPTIONAL ACTIVATOR GCN5 (*Saccharomyces cerevisiae*)
R87126	0.25	0.070	MYOSIN HEAVY CHAIN, NONMUSCLE (*Gallus gallus*)
L25941	0.36	0.070	Homo sapiens integral nuclear envelope inner membrane protein (LBR) gene, complete cds
T92451	0.25	0.071	TROPOMYOSIN, FIBROBLAST AND EPITHELIAL MUSCLE-TYPE (HUMAN)
U19969	0.30	0.078	Human two-handed zinc finger protein ZEB mRNA, partial cds
Z49269	0.31	0.080	Homo sapiens gene for chemokine HCC-1
X12496	0.35	0.086	Human mRNA for erythrocyte membrane sialoglycoprotein beta (glycophorin C)
R78934	0.28	0.088	ENDOTHELIAL ACTIN-BINDING PROTEIN (Homo sapiens)
H06524	0.28	0.098	GELSOLIN PRECURSOR, PLASMA (HUMAN)
M91463	0.28	0.102	Human glucose transporter (GLUT4) gene, complete cds
T67077	0.27	0.102	SODIUM/POTASSIUM-TRANSPORTING ATPASE GAMMA CHAIN (Ovis aries)
D29808	0.33	0.125	Human mRNA for T-cell acute lymphoblastic leukemia associated antigen 1 (TALLA-1), complete cds
H77597	0.33	0.135	Homo sapiens mRNA for metallothionein (HUMAN)
H64489	0.34	0.147	LEUKOCYTE ANTIGEN CD37 (Homo sapiens)
X74295	0.33	0.151	Homo sapiens mRNA for alpha 7B integrin
T57630	0.37	0.791	S34195 RIBOSOMAL PROTEIN L3
X74262	0.37	0.801	Homo sapiens RbAp48 mRNA encoding retinoblastoma binding protein
H11719	0.38	0.849	MONOCYTE DIFFERENTIATION ANTIGEN CD14 PRECURSOR (HUMAN)

is S100 calcium binding protein P (S100P; gene ID: T47377) having $e = 29\%$ ($p = 0.004$) (see Table 3). This gene can stimulate cellular proliferation and may function as a tumor growth factor [12].

Moreover seven muscle-related genes, highly expressed in normal tissues, were selected (see Table 2): J02854, T55741, U37019, T60155, X54162, T96548, R48602. This result confirms the evidence that the normal colon tissue have higher muscle content than cancer colon tissue [5,12]. Furthermore, the presence in Table 3 of T52185, T95018, X55715 and H64427 corresponding to ribosomal protein, confirms the observation that these genes have lower expression in normal than in cancer colon tissue [5,12].

Finally we point out down-regulated in adenoma (DRA; gene ID: L02785), carbonic anhydrase I (CA1; gene ID: R93176) and carbonic anhydrase II (CA2; gene ID: J03037), metallothionein (MT; genes ID: H54425; T71025) genes present in Table 2. DRA down-regulation was positively associated with colonic tumor progression and was particularly significant in the early transition from normal mucosa to polyp to adenocarcinoma. DRA expression does not appear to be strictly associated with colonic cell differentiation; rather, its absence and down-regulation were associated with the proliferating component of the crypt epithelium and with neoplastic transformation, respectively [35]. The expressions of carbonic anhydrase I (R93176) and II (J03037) are correlated with biological aggressiveness of colorectal cancer and synchronous distant metastasis, especially carbonic anhydrase I for colon cancer and carbonic anhydrase II for rectal cancer [36].

A physiological role for metallothionein is observed in cellular proliferation and in regulation of protein during the mitotic cell cycle point suggesting that it may also serve as a proliferation marker [37].

3.1. Related works

Our approach to gene selection provides a valuable method for assessing the statistical significance of genes found to be relevant by other studies, which do not explicitly face the problem of the statistical significance of their results. This is the case, for example, of the genes reported in [5,18,11]. We found that half of the relevant genes selected by [5](see Table 4) and only one-third of the ones selected by [11](see Table 5) had statistically significant prediction accuracies. On the contrary, this property holds true for almost all the relevant genes selected in [18](see Table 6).

4. Conclusions

In this paper we propose prediction accuracy as a measure of the relevance of a single gene in the pathology at hand. The rationale is that a gene can be thought of as relevant if it is differentially expressed in normal/disease tissues and its expression levels con be used for training classifiers able to correctly predict the status of new specimens. We have presented an unbiased, statistically well founded method based on RLS classifiers, a valuable alternative to SVM for tumor classification by DNA microarray data. We have used LKOCV error as estimate of the generalization error of a classifier and non parametric permutation tests for assessing the statistical significance of the obtained estimates. The relevance of the selected genes, obtained on a well known colon cancer data set, has been assessed (a) statistically, evaluating the p-value of the estimate prediction accuracy of each gene; (b) biologically, confirming the involvement of many genes in generic carcinogenic processes and in particular for the colon; (c) comparatively, verifying the presence of these genes in other studies on the same data-set. We plan to test the method on other case/control studies for cancer diagnosis based on DNA microarray data and to apply the method to functionally correlated classes of genes.

Acknowledgements

We would like to thank Sebastiano Stramaglia for valuable and illuminating discussions on numerous theoretical and experimental aspects of the paper. Laura Castellana made numerous and useful comments on the earlier version of the paper. We want to thank Cosimo Marzo for his contribution. This work was supported by Cluster C03 Studio di geni di interesse biomedico e agroalimentare (Ministero dell'Istruzione, Universitá e Ricerca Scientifica, Italy).

Appendix A. Property of the smallest eigenvalue of $\mathbf{A}+\mathbf{B}$

In this section we show that if $\mathbf{B}$ is positive definite, then the smallest eigenvalue of $\mathbf{A}+\mathbf{B}$ is larger than the smallest eigenvalue of $\mathbf{A}$ [38]. In fact, by definition of positive definite matrix, for every non zero vector $\mathbf{x}$, we have: $\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}>0$. Then we have:

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}<\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}+\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}$$

Dividing both members by $\mathbf{x}^{\mathrm{T}}\mathbf{x}>0$ we have:

$$\frac{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}}<\frac{\mathbf{x}^{\mathrm{T}}(\mathbf{A}+\mathbf{B})\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}}$$

By definition of Rayleigh's quotient, this is equivalent to say that for every $\mathbf{x}\neq 0$:

$$R_{\mathrm{A}}(\mathbf{x})<R_{\mathrm{A}+\mathrm{B}}(\mathbf{x})$$

where $R_{\mathrm{A}}(\mathbf{x})$ and $R_{\mathrm{A}+\mathrm{B}}(\mathbf{x})$ are the Rayleigh's quotients of the matrixes $\mathbf{A}$ and $\mathbf{A}+\mathbf{B}$, respectively. Knowing that the minimum value of the Rayleigh's quotient coincides with the smallest eigenvalue, then we can write that:

$$\lambda_{\mathrm{A}}<\lambda_{\mathrm{A}+\mathrm{B}}$$

where λ_{A} and λ_{A} are the smallest eigenvalues of the matrixes $\mathbf{A}$ and $\mathbf{A}+\mathbf{B}$, respectively.

Note that if the matrix $\mathbf{A}$ is positive semidefinite then $\lambda_{\mathrm{A}}=0$. So from the previous property follows that $\lambda_{\mathrm{A}}>0$ and then the matrix $\mathbf{A}+\mathbf{B}$ is positive definite.

Appendix B. Analysis of computational complexity

Let us consider the case in which the data set S contains the expression levels of a single gene measured in two phenotypically different conditions. In this case $S=\{(x_1,y_1),(x_2,y_2),\ldots,(x_\ell,y_\ell)\}$, where $x_i\in\mathbb{R}$ and $y_i\in\{-1,1\}$ for $i=1,2,\ldots,\ell$. At the aim of including a bias term implicitly in the linear model, we add a supplementary variable (constant and equal to 1) to each input x_i. In particular, defining the vectors $\mathbf{x}_i=(x_i,1)^{\mathrm{T}}$ and $\mathbf{w}=(w_1,w_2)^{\mathrm{T}}$, the function (1) to minimize becomes:

$$L'(\mathbf{w})=\frac{1}{\ell}\sum_{i=1}^{\ell}(y_i-w_1x_i-w_2)^2+\lambda(w_1^2+w_2^2) \quad \text{(B.1)}$$

It easy to see that the vector $\mathbf{w}^*$ minimizing (B.1) is given by:

$$w_1^* = \frac{\left(\sum_i x_i\right)\left(\sum_i y_i\right) - \ell(\lambda+1)\left(\sum_i x_i y_i\right)}{\left(\sum_i x_i\right)^2 - \ell(\lambda+1)\left(\sum_i x_i^2 + \lambda\ell\right)},$$

$$w_2^* = \frac{\left(\sum_i x_i\right)\left(\sum_i x_i y_i\right) - \left(\sum_i x_i^2 + \lambda\ell\right)\left(\sum_i y_i\right)}{\left(\sum_i x_i\right)^2 - \ell(\lambda+1)\left(\sum_i x_i^2 + \lambda\ell\right)}$$

This shows that the learning phase involves a number of multiplications proportional to the number ℓ of specimens and only two divisions. For determining the class y of an expression level x we only need to evaluate $y = \mathrm{sign}(w_1^* x + w_2^*)$, indicating that the test phase is inexpensive. During the permutation test in which we randomly permute the labels y_i of the training set, the quantities in both the denominators do not change. Only some factors in the numerators need to be computed at each random permutation.

Finally, since the evaluation of the prediction accuracy involves the expression levels of each gene singularly, the scheme can be easily executed in parallel on different computers, drastically reducing the computational load.

References

[1] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. Science 1995;270:467—70.

[2] Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 2005;21(5):650—9.

[3] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531—7.

[4] Ramaswamy S, Golub TR. DNA microarrays in clinical oncology. J Clin Oncol 2001;20:1932—41.

[5] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci 1999;96:6745—50.

[6] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci 2001;98:15149—54.

[7] Birkenkamp-Demtroder K, Christensen LL, Olesen SH, Frederiksen CM, Laiho P, Aaltonen LA, et al. Gene expression in colorectal cancer. Cancer Res 2002;62:4352—63.

[8] Wang Y, Jatkoe T, Zhang Y, Mutch MG, Talantov D, Jiang J, et al. Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. J Clin Oncol 2004;22(9):1564—71.

[9] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389—422.

[10] Slonim DK, Tamayo P, Mesirov JP, Golub TR, Lander ES. Class prediction and discovery using gene expression data. In: Shamir R, Miyano S, Istrail S, Pevzner P, Waterman M, editors. RECOMB '00: Proceedings of the fourth annual international conference on computational molecular biology. New York, NY, USA: ACM Press; 2000. p. 263—72.

[11] Furlanello C, Serafini M, Merler S, Jurman G. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. BMC Bioinformatics 2003;4(1):54.

[12] Fu LM, Fu-Liu CS. Evaluation of gene importance in microarray data based upon probability of selection. BMC Bioinformatics 2005;6:67.

[13] Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. Feature selection for svms. In: Leen TK, Dietterich TG, Volker T, editors. Advances in neural information processing systems, vol. 13. Cambridge, MA: MIT Press; 2001 p. 668—74.

[14] Mukherjee S, Tamayo P, Rogers S, Rifkin R, Engle A, Campbell C, et al. Estimating dataset size requirements for classifying dna microarray data. J Comp Biol 2003;10:119—42.

[15] Fu WJ, Dougherty ER, Mallick B, Carroll RJ. How many samples are needed to build a classifier: a general sequential approach. Bioinformatics 2005;21(1):63—70.

[16] Li W, Sun F, Grosse I. Extreme value distribution based gene selection criteria for discriminant microarray data analysis using logistic regression. J Comp Biol 2004;11(2-3):215—26.

[17] Vapnik V. The nature of statistical learning theory New York: Springer Verlag Inc.; 1995.

[18] Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. Tissue classification with gene expression profiles. J Comp Biol 2000;7:559—83.

[19] Evgeniou T, Pontil M, Poggio T. Regularization networks and support vector machines. Adv Comp Math 2000;13(1):1—50.

[20] Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci 2002;99:6562—6.

[21] Rifkin R, Yeo G, Poggio T. Regularized least squares classification. In: Suykens, Horvath, Basu, Micchelli, Vandewalle, editors. Advances in learning theory: methods model and applications NATO Science Series III: computer and systems sciences, vol. 90. Amsterdam: IOS Press; 2003. p. 131—53.

[22] Zhang P, Peng J. Svm vs regularized least squares classification. In: Kittler J, Petrou M, Nixon M, editors. Proceedings of the 17th international conference on pattern recognition (ICPR '04). Los Alamitos, CA, USA: IEEE Computer Society; 2004. p. 176—9.

[23] Ancona N, Maglietta R, D'Addabbo A, Liuni S, Pesole G. Regularized least squares cancer classifiers from DNA microarray data. BMC Bioinformatics 2005;6(Suppl 4):S2.

[24] Golland P, Liang F, Mukherjee S, Panchenko D. Permutation tests for classification. In: Auer P, Meir P, editors. Lecture notes in computer science, vol. 3559. Heidelberg: Springer Berlin; 2005. p. 501—15.

[25] Good P. Permutation tests: a practical guide to resampling methods for testing hypotheses New York: Springer Verlag Inc.; 1994.

[26] Nichols TE, Holmes AP. Non-parametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp 2001;15:1—25.

[27] Tikhonov AN, Arsenin VY. Solutions of ill-posed problems Washington, DC: W.H. Winston; 1977.

[28] Papoulis A. The Fourier integral and its applications New York: McGraw-Hill Book Company; 1962.

[29] Luntz A, Brailovsky V. On estimation of characters obtained in statistical procedure of recognition. Technicheskaya Kibernetica 1969;3.

[30] Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res 2001;61:3124—30.

[31] Li Y, Campbell C, Tipping M. Bayesian automatic relevance determination algorithms for classifying gene expression data. Bioinformatics 2002;18:1332—9.
[32] Li JQ, Miki H, Ohmori M, Wu F, Funamoto Y. Expression of cyclin e and cyclin-dependent kinase 2 correlates with metastasis and prognosis in colorectal carcinoma. Hum Pathol 2001;32:945—53.
[33] Shiff SJ, Koutsos MI, Qiao L, Rigas B. Non-steroidal antiinflammatory drugs inhibit the proliferation of colon adenocarcinoma cells: effects on cell cycle and apoptosis. Exp Cell Res 1996;222:179—88.
[34] Lakshman M, Subramaniam V, Wong S, Jothy S. Cd44 promotes resistance to apoptosis in murine colonic epithelium. J Cell Physiol 2005;203(3):583—8.
[35] Antalis TM, Reeder JA, Gotley DC, Byeon MK, Walsh MD, Henderson KW, et al. Down-regulation of the down-regulated in adenoma (dra) gene correlates with colon tumor progression. Clin Cancer Res 1998;4:1857—63.
[36] Bekku S, Mochizuki H, Yamamoto T, Ueno H, Takayama E, Tadakuma T. Expression of carbonic anhydrase i or ii and correlation to clinical aspects of colorectal cancer. Hepatogastroenterology 2000;47:998—1001.
[37] Nagel WW, Vallee BL. Cell cycle regulation of metallothionein in human colonic cancer cells. Proc Natl Acad Sci 1995;17/92(2):579—83.
[38] Strang G. Linear algebra and its applications Wellesley, MA: Wellesley-Cambridge Press; 1988.