Journal of Biomedical Informatics 49 (2014) 32–44

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Robust gene signatures from microarray data using genetic algorithms enriched with biological pathway keywords

R.M. Luque-Baena [a,*], D. Urda [a,b], M. Gonzalo Claros [c], L. Franco [a,b], J.M. Jerez [a,b]

[a] *Departmento de Lenguajes y Ciencias de la Computación, University of Málaga, Bulevar Louis Pasteur, 35, 29071 Málaga, Spain*
[b] *Instituto de Investigación Biomédica de Málaga (IBIMA), Málaga, Spain*
[c] *Supercomputing and Bioinformatics Centre, University of Málaga, C/ Severo Ochoa, 34, 29590 Málaga, Spain*

ARTICLE INFO

Article history:
Received 24 July 2013
Accepted 16 January 2014
Available online 27 January 2014

Keywords:
DNA analysis
Evolutionary algorithms
Biological enrichment
Feature selection

ABSTRACT

Genetic algorithms are widely used in the estimation of expression profiles from microarrays data. However, these techniques are unable to produce stable and robust solutions suitable to use in clinical and biomedical studies. This paper presents a novel two-stage evolutionary strategy for gene feature selection combining the genetic algorithm with biological information extracted from the KEGG database. A comparative study is carried out over public data from three different types of cancer (leukemia, lung cancer and prostate cancer). Even though the analyses only use features having KEGG information, the results demonstrate that this two-stage evolutionary strategy increased the consistency, robustness and accuracy of a blind discrimination among relapsed and healthy individuals. Therefore, this approach could facilitate the definition of gene signatures for the clinical prognosis and diagnostic of cancer diseases in a near future. Additionally, it could also be used for biological knowledge discovery about the studied disease.

1. Introduction

The term cancer encompasses more than 100 potentially life-threatening diseases affecting nearly every part of the body. Cancer is a complex, multifactorial, genetic disease involving structural and expression abnormalities of both coding and non-coding genes. In this sense, gene expression profiling plays an important role in a wide range of areas in biological science for handling cancer diseases [1–4]. The analysis of DNA microarray data requires a selection of features (genes) due to the small number of samples available (mostly less than a hundred) and the large number of features (in the order of thousands). This problem is well-known in the literature as the “large-p-small-n” paradigm or the curse of dimensionality [5].

Evolutionary models have been proposed in several works [6–12] and constitute one of the most widely used techniques for feature selection and prognosis analysis in microarray datasets. Despite all the variety of feature selection techniques proposed in the literature, it still remains a problematic intrinsic to the domain of DNA microarrays. Genetic algorithms (GAs) [13–18], as a particular case of evolutionary models, use classification techniques within the algorithm to evaluate and evolve the population. Producing stable or robust solutions is a desired property of feature selection algorithms, in particular for clinical and biomedical studies. Nevertheless, robustness is a property difficult to be analyzed and is often overlooked. In [19–21] different approaches are proposed, addressing the main drawbacks related to overfitting and robustness, through a modified GA that includes an early-stopping criteria and establishing a feature ranking method that leads to more robust solutions. Although some proposals use biological information to analyze DNA microarray data [22], none of them includes it into the mechanisms that guide the searching procedure in the GA. In our opinion, this strategy would, on one hand, produce more robust feature subset selections and, on the other hand, permit to obtain signatures more relevant for clinicians and biomedical researchers.

In this approach, a two-stage procedure is proposed in order to obtain robust feature subset selections with good performance rates in test future data. Bootstrap Cross-Validation (BCV) is used since its good behavior related to misclassification error with small samples has been previously demonstrated [23,24], including DNA microarray datasets. A novel feature scoring method within the GA is also proposed, taking into account biological information related to the studied disorders. One widely used source of biological information is the Gene Ontology (GO) database [25] since it

* Corresponding author. Address: Department of Computer Languages and Computer Science, University of Málaga, Bulevar Louis Pasteur, 35, 29071 Málaga, Spain. Fax: +34 952131397.

E-mail addresses: rmluque@lcc.uma.es (R.M. Luque-Baena), durda@lcc.uma.es (D. Urda), claros@uma.es (M. Gonzalo Claros), lfranco@lcc.uma.es (L. Franco), jja@lcc.uma.es (J.M. Jerez).

http://dx.doi.org/10.1016/j.jbi.2014.01.006
1532-0464/© 2014 Elsevier Inc. All rights reserved.

provides a controlled vocabulary for the description of cellular components, molecular functions, and biological processes. However, GO is sub-classified using a hierarchy of unclear reasoning with no validation analysis, contains insufficient number of rules for determining whether a given concept is present or not in GO, and most importantly, most GO terms have been assigned by sequence similarity through an automatic analysis, without laboratory validation [26]. Therefore, we have discarded the use of GO and moved to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [27]. Since many years, this database has been one of the most important sources for building initial pathway models because it can be used as a reference knowledge base for deciphering the genome and linking genes/proteins to biological systems and also to the environment. Its main strength is that it is manually drawn and the assignment of a KEGG code to a sequence implies experimental evidence support. On the contrary, if a protein coded by a sequence does not produce enzymatic activity or is not part of a signaling pathway, it will never have a KEGG code. Fortunately, most genes involved in cancer have enzyme activity and belong to signaling pathways. This makes KEGG a valuable and highly reliable source of pathways and lead us to obtain robust and biologically important feature subset selections. In fact, KEGG codes 05200 to 05223 are specifically dedicated to cancer.[1] As an example, KEGG pathways have allowed the generation of systems biology models [28], the identification of disease virulence factors [29], the emergence of molecular pathway perturbations in sporadic amyotrophic lateral sclerosis [30], or the analysis of the lipidomic and transcriptomic changes showing the distinct roles of STAT1 and STAT3 on apoptosis, immunity and lipid metabolism [31].

The rest of the paper is structured as follows. Section 2 presents the methodology of our approach and Section 3 shows the experimental results over different databases. Section 4 provides the final conclusions of the work drawn from the analysis of the selected genes and from the study of the influence of the biological information in the performance of the strategy.

2. Materials and methods

2.1. Materials

Three free-public high-dimensional biomedical datasets have been used within this work. Each of them is related to an specific cancer study disorder: leukemia,[2] prostate[3] and lung[4] cancer diseases.

2.1.1. Leukemia dataset

This dataset was taken from a collection of leukemia patient samples reported in [32] and it often serves as benchmark for microarray analysis methods. It contains measurements corresponding to acute lymphoblast leukemia (ALL) and acute myeloid leukemia (AML) samples from bone marrow and peripheral blood. The dataset consists of 72 samples (25 of them of AML and 47 samples of ALL) and each one is measured over 7129 genes. The ID for the leukemia Affymetrix GeneChip HuGeneFL array is hu6800. In particular, the R package "hu6800.db" [33] has been used to manage and preprocess the biological information related to this microarray.

2.1.2. Prostate dataset

This dataset was reported in [34]. Prostate tumors are among the most heterogeneous of cancers, both histologically and with respect to highly divergent clinical outcomes. The dataset consists of 102 samples (52 of them are tumor samples and 50 samples are non-tumor ones) and each one is represented by 12,600 genes. The Affymetrix ID for the prostate cancer microarray is HGU95av2 and the R package "hgu95av2.db" [35] was used to manage and preprocess biological information related to this microarray.

2.1.3. Lung dataset

This dataset presents a classification between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung, being reported in [36]. It consists of 181 tissue samples (31 corresponds to MPM samples and 150 to ADCA) and each one is described by 12,533 genes. The Affymetrix ID for the prostate cancer microarray is hgu95a and the R package "hgu95a.db" [37] was used to manage and preprocess biological information related to this microarray.

2.2. Methodology

In this paper, a novel two-step methodology is applied in a strategy based on the use of GA with the addition of biological information, with the aim of obtaining a robust subset of features with high prediction capabilities. The first stage uses a filtering approach based on the KEGG database to retain those features representing enzymes and to establish a ranking for the different pathways available. The second stage implements the feature selection procedure, executing a GA for each of the best ranked pathways.

A high level description of our methodology approach is shown in Fig. 1 as well as a brief pseudocode of the algorithm is described in Algorithm 1. It is important to highlight the choice of a BCV strategy to obtain an accuracy measure on both stages of our approach because it has been previously demonstrated in [23,24] its good behavior under estimating misclassification error with small samples, as is the particular case of DNA microarray datasets. In concrete, the developed procedure executes a 50 bootstrap resampling and 5-k-fold validation techniques. Therefore, this approach tries to, on one hand, discover a robust subset of features with biological relevance on the studied disorder; on the other hand, good generalization rates in the prediction stage are essential to determine the probability of suffering from a specific condition.

Algorithm 1. Pseudocode of the two-step methodology used for gene feature selection

```
{initialization}
$[Train, Test]\{1..50\} \Leftarrow BCV(dataset, 50)$
$[Pathways]\{1..N\} \Leftarrow KEGG(chip(dataset))$
$[Keywords]\{1..M\} \Leftarrow SetKeywords(dataset)$

{first-step: for each pathway, set a prediction ability and
  find occurrences of keywords}
for $i = 1 \rightarrow N$ do
    $P_i \Leftarrow Pathways[i]$
    for $j = 1 \rightarrow 50$ do
        $TR_j \Leftarrow Train[j]$
        $P_i\_TR_j\_PredictionAbility \Leftarrow CrossValidation(TR_j, Genes(P_i))$
    end for
    $P_i\_PredictionAbility \Leftarrow mean(P_i\_TR_j\_PredictionAbility)$
    $P_i\_DetectedKeywords \Leftarrow TextMining(P_i, Keywords)$
end for

{second-step: for the most promising pathways, make a
  feature selection using a GA}
```

(continued on next page)

[1] http://www.genome.jp/kegg-bin/get_htext?htext=br08901&query="Human%20Diseases"&option=-s.

[2] http://datam.i2r.a-star.edu.sg/datasets/krbd/Leukemia/ALLAML.html.

[3] http://datam.i2r.a-star.edu.sg/datasets/krbd/ProstateCancer/ProstateCancer.html.

[4] http://cilab.ujn.edu.cn/datasets.htm.

```
$[SelectedPathways]\{1,\ldots,K|K<N\} \Leftarrow ChoosePathways$
    $(P_i\_PredictionAbility, P_i\_DetectedKeywords)$
for $i = 1 \rightarrow K$ do
    $SP_i \Leftarrow SelectedPathways[i]$
    for $j = 1 \rightarrow 50$ do
        $TR_j \Leftarrow Train[j]$
        $T_j \Leftarrow Test[j]$
        $SP_i\_TR_j\_SelectedFeatures \Leftarrow GeneticAlgorithm(TR_j, SP_i))$
        $SP_i\_T_j\_Prediction \Leftarrow Accuracy(T_j, SP_i\_TR_j\_SelectedFeatures)$
    end for
    $SP_i\_Prediction \Leftarrow mean(SP_i\_T_j\_Prediction)$
end for
```

2.2.1. Pathway prediction ability

Several pathways involved in the studied disorder are represented in a DNA Affymetrix chip. On this stage, our approach sets a prediction ability for each pathway on two ways: first, by obtaining an accuracy measure representing the capability of the genes of every pathway to generalize the problem; and second, by doing a text mining procedure searching for some keywords that may appear on the description of a pathway.

Statistical analysis were performed using R,[5] in particular those R packages mentioned in [38,35,37] have been used to obtain the pathways related to the studied diseases (Leukemia, Prostate and Lung). Each pathway is scored by the generalization rate after filtering features of the dataset keeping only those genes that are contained in the pathway and giving them as input to a classifying model. Furthermore, a text mining procedure is executed for each pathway in order to localize those pathways that may be more correlated to the studied disorder. Table 1 shows the keywords used within this procedure that have been obtain through biological support tools using Ingenuity Pathways Analysis (IPA®.[6]) Then, the main purpose of the text mining process is to analyze the content of the webpages of each pathway and to search on it for some keywords. As a result, those pathways containing a higher number of keywords would lead us to think that are more correlated to the studied disease.

2.2.2. Evolutionary strategy

GAs are a class of optimization procedure inspired by the biological mechanisms of reproduction. In this kind of optimization problems, a fitness function $f(\mathbf{x})$ should be maximized or minimized over a given space X of arbitrary dimension. On this stage of our approach, the most promising pathways are selected according to the prediction ability and the number of keywords found on the text mining procedure. A GA is executed for each of these pathways in order to find a robust feature subset selection taking into account biological information, preponderating the activation of genes included in the studied pathway without discarding the selection of the rest of genes.

2.2.2.1. Encoding and initial population. A simple encoding scheme to represent as much as possible of the available information was employed, in which the chromosome is a string of bits whose length is determined by the total number of genes. Each variable is associated with one bit in the string. If the ith bit is active (value 1), then the ith gene is selected in the chromosome. Otherwise, a value of 0 indicates that the corresponding feature is ignored. In this way, each chromosome represents a different feature subset. Both, the active features and the number of them are generated randomly. In all the experiments, the population size of 100 individuals was used and the number of active features for a certain chromosome limited to 100, thus generating chromosomes representing signatures of few genes.

2.2.2.2. Selection, crossover and mutation. A selection strategy based on roulette wheel and uniform sampling is applied. Additionally, the E best chromosomes should be retained for the next generation. The E parameter is called elite count or sometimes referred as reproduction operator p_e (probability of the retained chromosomes in the population, between 0 and 1), since involves the insertion of a copy of a chromosome in the next generation. Scattered crossover, in which each bit of the offspring is chosen randomly, was the choice for combining parents of the previous generation. The crossover rate p_c can be found in the interval $(0,1)$, with values close to 1. In addition to that, a traditional mutation operator which flips a specific bit with a probability rate of p_m is considered. Usually, the mutation rate is rather lower than the crossover rate [39]. A modification which involves mutating a random number of bits between 1 and the number of active features of the individual is introduced. Since it was empirically verified that the best subsets include few features, this change avoids the increment on the number of active features in the last generations of the GA. Furthermore, the activation of those genes included in the studied pathway is prioritized without discarding the activation of the rest of genes. On the same way, the deactivation of genes not included in the pathway is prioritized without discarding the deactivation of genes that are present on the studied pathway. The following rule needs to be satisfied: $p_e + p_c + p_m = 1$. A comparative study for the selection of these rates is conducted in Section 3.

2.2.2.3. Fitness function. The fitness function assesses each chromosome in the population so that it may be ranked against all the other chromosomes. The main goal of feature subset selection is to use less features to achieve the same or better performance that provides more biological relevance for the studied disease. Therefore, the fitness function should contain three terms, so for a certain chromosome x to be analyzed, the function to be minimized is represented as follows:

$$fitness(\mathbf{x}) = (1 - ACC(\mathbf{x})) + \lambda\frac{k}{100} + \beta score(\mathbf{x}), \tag{1}$$

where k is the number of selected features, 100 is a normalization factor due to the limited number of active features in a chromosome, and function "score" that estimates the biological relevance of the selected features according to the number of selected genes that are included on the studied pathway and how many of them are not included in it. The "score" function is compute as shown in next equation:

$$score(\mathbf{x}) = \left(1 - \frac{i}{M}\right) + \frac{j}{N}, \tag{2}$$

where M and N are normalization factors representing respectively the number of genes on the studied pathway and the total number of genes on the dataset ($M \ll N, i$ is the number of selected genes that are included in the pathway, and j the number of selected genes that are not contained in the studied pathway ($i + j = k$).

The accuracy rate (ACC) in Eq. (1) is obtained after the application of a classification algorithm to the datasets. We have considered in this work two standard and well-known classifiers: a low complexity method named Linear Discriminant Analysis (LDA) [40], whose aim is to find a linear combination of features which separates two or more classes of patterns; and Support Vector Machines (SVM), a more sophisticated method that find the

[5] http://www.r-project.org/.

[6] http://www.ingenuity.com/products/ipa.

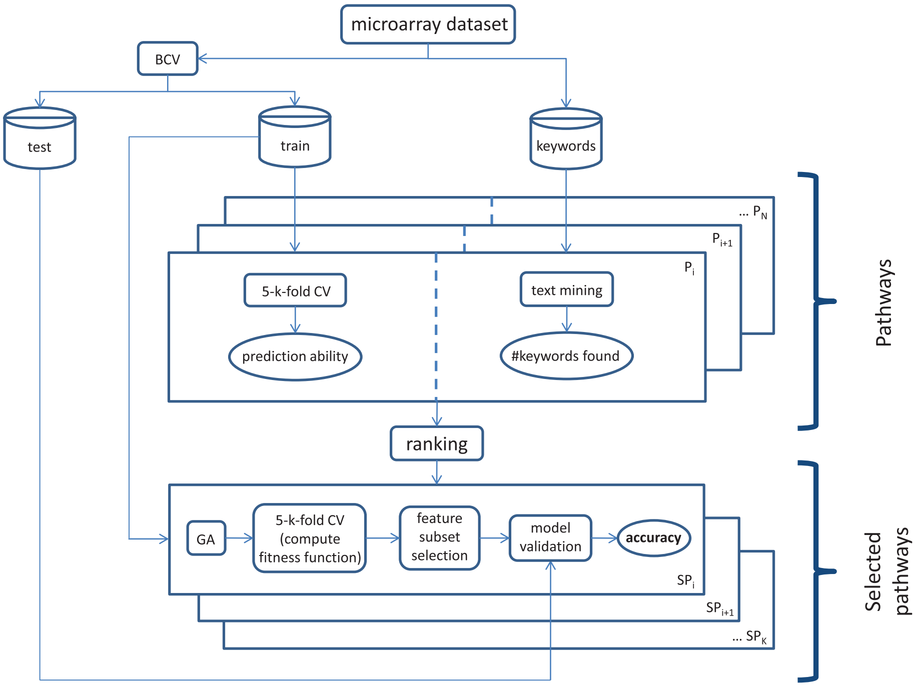


Fig. 1. Framework of the proposed approach that includes the two stages related to the ranking of the pre-selected pathways and to the final model selection.

Table 1
Information about the keywords used in the text mining procedure.

	Keywords
Common	Apoptosis, cancer, tumor, tumorigenesis, carcinoma, malignant, metastasis, infection, hypoplasia, neoplasia
Leukemia	Leukemia, lymphocytic, myeloid, lymphoblastic, T-cell, B-cell, myelogenous, leuke, immun, lymph, nodule
Prostate	Prostate, prosta, epithelial, psa, kallikrein, urin, erect, hypertrophy
Lung	Lung, AT2, interalveolar, pleura, pulmo, alveo, pneumo, epithelial, small-cell, nodule, squamous

optimal separation margin between two classes, and which has been widely used in microarray analysis [41,42].

Furthermore, since determining a robust gene signature to predict outcome disease can be considered as a feature selection problem, the authors also include a performance comparative analysis between the proposed strategy and three filter methods commonly used to do variable selection: ReliefF [43], extension of the original Relief algorithm [44] which works by randomly sampling an instance from the data and then locating its nearest neighbor from the same and opposite class; Consistency-based Filter (Cons) [45], which evaluates the worth of a subset of features by the level of consistency in the class values when the training instances are projected onto the subset of attributes; and Information Gain (IG) [46], which provides an ordered ranking of all the features and then a threshold is required.

3. Results

This section shows the results of our approach on three selected data sets, *Leukemia*, *Lung* and *Prostate*. Table 2 outlines the result of evaluating and sorting each pathway by its predictive ability, following the scheme shown in Section 2.2.1. The ranking of the pathways in terms of this predictive ability (fifth column) is indicated in the first column. Columns 2, 3 and 4 contain information on the analyzed pathway, such as its code, description and number of genes. The last two columns show the number of keywords and keywords found after the text mining procedure for each database. This information provides an idea of the relationship between the pathway and the disease, based on the number of keywords found. The first ten pathways are listed, although, in some cases, pathways in lower positions in the ranking are added due to its influence with the disease, which is measured by the number of keywords (pathway 04062 in *Leukemia* and pathway 05215 in *Prostate*). The list of keywords used can be found in Table 1. The selected pathways, in bold in Table 2, are analyzed in the next phase of our methodology described in Section 2.2.2. This selection is carried out taking into account the predictive capacity and number of keywords of each pathway.

Before applying our methodology based on genetic algorithms, it is necessary to estimate the parameters related to the selection, mutation and crossover operators referred in Section 2.2.2. For this, the standard genetic algorithm (GA), which is quite common in the literature, is considered as the reference strategy and use as comparative framework for the parameter estimation procedure. This estimation is carry out by analyzing the *Leukemia* dataset, and different combinations of the p_e, p_c and p_m parameters together with the accuracy results and number of selected genes are shown in Table 3.

It is possible to observe that the differences in the accuracy rates for each parameter combination are not statistically significant, which implies that, for these cancer datasets, any combination of parameters can be chosen. Specifically, the authors have selected the parameters $p_c = 0.72,\ p_e = 0.18$ and $p_m = 0.1$ (Table 3, in italic), since they lead to the obtention of the largest success rate (Table 3, in bold).

Table 2
Pathways ranked by their predictive ability for each data set. The selected pathways to be analyzed and integrated in the genetic algorithm are shown in bold.

Rank	Code	Pathway Name	#Genes	Predictive Ability	#Keys	Keywords
Leukemia dataset						
1	**04640**	**Hematopoietic cell lineage**	**105**	**0.945 ± 0.024**	**3**	Myeloid, Immune, lymphoid
2	04614	Renin-angiotensin system	16	0.911 ± 0.028	0	
3	00480	Glutathione metabolism	31	0.899 ± 0.033	0	
4	**05340**	**Primary immunodeficiency**	**31**	**0.899 ± 0.027**	**4**	B-cell, immunodeficiency, lymphocyte, infection
5	**04662**	**B cell receptor signaling pathway**	**68**	**0.894 ± 0.030**	**5**	B-cell, immunity, lymphedema, tumorigenesis
6	00590	Arachidonic acid metabolism	32	0.891 ± 0.026	0	
7	01100	Metabolic pathways	658	0.881 ± 0.025	0	
8	**04670**	**Leukocyte transendothelial migration**	**90**	**0.878 ± 0.029**	**2**	Immune, lymphocyte
9	00030	Pentose phosphate pathway	21	0.876 ± 0.034	0	
10	04145	Phagosome	124	0.874 ± 0.031	2	Immune, lymphocyte
11	04666	Fc gamma R-mediated phagocytosis	67	0.869 ± 0.027	1	Immune
12	**05200**	**Pathways in cancer**	**319**	**0.867 ± 0.024**	**13**	Leukemia, myeloid, myelogenous, immunohistochemical, apoptosis, cancer, tumor, tumorigenesis, carcinoma, malignant, metastasis, neoplasia
13	05146	Amoebiasis	116	0.864 ± 0.027	3	Immune, apoptosis, infection
14	04141	Protein processing in endoplasmic reticulum	88	0.863 ± 0.031	1	apoptosis
15	04970	Salivary secretion	77	0.863 ± 0.027	0	
26	**04062**	**Chemokine signaling pathway**	**161**	**0.845 ± 0.029**	**6**	Leukemia, lymphocytic, immune, lymphedema, tumor
Lung dataset						
1	**04144**	**Endocytosis**	**244**	**0.988 ± 0.009**	**0**	
2	01100	Metabolic pathways	970	0.977 ± 0.012	0	
3	**04530**	**Tight junction**	**158**	**0.977 ± 0.009**	**1**	Epithelial
4	**04514**	**Cell adhesion molecules (CAMs)**	**154**	**0.975 ± 0.010**	**0**	
5	04360	Axon guidance	166	0.974 ± 0.008	0	
6	**04610**	**Complement and coagulation cascades**	**73**	**0.971 ± 0.009**	**3**	Cancer, tumor, metastasis
7	**04010**	**MAPK signaling pathway**	**423**	**0.971 ± 0.010**	**2**	AT2, tumor
8	00240	Pyrimidine metabolism	83	0.970 ± 0.010	0	
9	04062	Chemokine signaling pathway	254	0.969 ± 0.013	1	tumor
10	**05200**	**Pathways in cancer**	**557**	**0.969 ± 0.012**	**11**	Lung, small-cell, squamous, apoptosis, cancer, tumor, tumorigenesis, carcinoma, malignant,metastasis, neoplasia
Prostate dataset						
1	**00480**	**Glutathione metabolism**	**41**	**0.754 ± 0.029**	**0**	
2	00750	Vitamin B6 metabolism	2	0.741 ± 0.032	0	
3	**00040**	**Pentose and glucuronate interconversions**	**18**	**0.740 ± 0.054**	**1**	Tumoral
4	04974	Protein digestion and absorption	80	0.739 ± 0.038	0	
5	00330	Arginine and proline metabolism	62	0.726 ± 0.032	0	
6	**04610**	**Complement and coagulation cascades**	73	0.724 ± 0.036	**3**	Cancer, tumor, metastasis
7	00340	Histidine metabolism	22	0.722 ± 0.044	0	
8	04964	Proximal tubule bicarbonate reclamation	24	0.721 ± 0.029	0	
9	00270	Cysteine and methionine metabolism	33	0.721 ± 0.028	0	
10	00071	Fatty acid metabolism	45	0.720 ± 0.032	0	
11	00380	Tryptophan metabolism	45	0.719 ± 0.046	0	
12	00350	Tyrosine metabolism	44	0.716 ± 0.032	0	
13	00640	Propanoate metabolism	29	0.716 ± 0.044	0	
14	00010	Glycolysis/gluconeogenesis	65	0.713 ± 0.036	0	
15	**00980**	**Metabolism of xenobiotics by cytochrome P450**	**65**	**0.713 ± 0.025**	**1**	Cancer
16	00982	Drug metabolism – cytochrome P450	75	0.712 ± 0.028	1	Cancer
17	**04512**	**ECM-receptor interaction**	**123**	**0.712 ± 0.043**	**3**	Apoptosis, cancer, tumor
53	**05215**	**Prostate cancer**	**166**	0.654 ± 0.027	**7**	Prostate, apoptosis, cancer, tumor,metastasis, neoplasia

Table 3
Parameter estimation for the crossover, reproduction and mutation operators of the GA for the *Leukemia* dataset.

Crossover rate (p_c)	Reproduction rate (p_e)	Mutation rate (p_m)	Accuracy	#Genes
0.375	0.375	0.25	0.9523 ± 0.0155	4.49 ± 0.61
0.45	0.45	0.1	0.9489 ± 0.0156	4.37 ± 0.66
0.49	0.49	0.02	0.9529 ± 0.0169	4.18 ± 0.62
0.6	0.15	0.25	0.9500 ± 0.0144	4.52 ± 0.64
0.7125	0.0375	0.25	0.9510 ± 0.0137	4.58 ± 0.73
0.72	*0.18*	*0.1*	**0.9539 ± 0.0162**	*4.54 ± 0.76*
0.784	0.196	0.02	0.9495 ± 0.0134	4.41 ± 0.63
0.855	0.045	0.1	0.9487 ± 0.0157	4.74 ± 0.64
0.931	0.049	0.02	0.9485 ± 0.0138	4.48 ± 0.52

Different evolutionary strategies are considered for comparing the results. The first one, the standard GA, whose objective is to minimize the number of genes and training error of the classification model for each combination of genes. As a second strategy a two-stage approach named (Filter + GA) is included. Initially, a filter based on biological information which selects those genes

Table 4
Performance comparison among different feature selection strategies for each cancer dataset for LDA and SVM classifiers. On average, columns three and five present the number of genes and the accuracy for each framework in the format of *mean ± standard deviation*. Additionally, the four column shows a robustness measure in terms of low variability of the selected genes when the strategy is executed repeatedly. The starred values indicate that the results are statistically significant.

Classifier	Database	Strategy	#Genes	Robustness	Accuracy
LDA	Leukemia	GA	4.54 ± 0.76	0.0773	0.9539 ± 0.0162
		Filter + GA	4.48 ± 0.56	0.0954	0.9531 ± 0.014
		Filter + GA + Pathway 04640	4.47 ± 0.71	0.1753	***0.9638 ± 0.0126**
		Filter + GA + Pathway 05340	31.83 ± 1.36	0.7022	***0.9713 ± 0.0116**
		Filter + GA + Pathway 04662	5.40 ± 0.95	0.1900	***0.9606 ± 0.0145**
		Filter + GA + Pathway 04670	4.97 ± 0.80	0.1171	0.9521 ± 0.0142
		Filter + GA + Pathway 05200	4.80 ± 0.54	0.0877	0.9491 ± 0.0156
		Filter + GA + Pathway 04062	4.70 ± 0.70	0.0926	0.9463 ± 0.0159
	Lung	GA	3.53 ± 0.35	0.0826	0.9753 ± 0.0048
		Filter + GA	3.88 ± 0.55	0.0706	0.9772 ± 0.0058
		Filter + GA + Pathway 04144	4.29 ± 0.53	0.1397	***0.9809 ± 0.0068**
		Filter + GA + Pathway 04530	3.84 ± 0.46	0.1797	***0.9826 ± 0.0046**
		Filter + GA + Pathway 04514	4.41 ± 0.56	0.1274	0.9767 ± 0.0069
		Filter + GA + Pathway 04610	5.69 ± 0.93	0.1453	0.9759 ± 0.0088
		Filter + GA + Pathway 04010	4.04 ± 0.55	0.0981	**0.9785 ± 0.0055**
		Filter + GA + Pathway 05200	4.03 ± 0.62	0.0790	0.9754 ± 0.0062
	Prostate	GA	6.10 ± 0.68	0.0836	0.9120 ± 0.0139
		Filter + GA	5.91 ± 0.86	0.0916	0.9060 ± 0.0130
		Filter + GA + Pathway 00480	14.30 ± 2.63	0.3022	0.9080 ± 0.0136
		Filter + GA + Pathway 00040	23.24 ± 1.52	0.4851	0.9107 ± 0.0153
		Filter + GA + Pathway 04610	6.97 ± 1.15	0.1701	0.9103 ± 0.0128
		Filter + GA + Pathway 00980	8.27 ± 0.83	0.1636	**0.9137 ± 0.0115**
		Filter + GA + Pathway 04512	7.62 ± 0.96	0.1228	0.9001 ± 0.0181
		Filter + GA + Pathway 05215	6.96 ± 0.95	0.1474	**0.9122 ± 0.0136**
SVM	Leukemia	GA	4.88 ± 0.80	0.0858	0.9164 ± 0.0178
		Filter + GA	4.94 ± 1.01	0.0931	0.9214 ± 0.0199
		Filter + GA + Pathway 04640	4.05 ± 0.80	0.1364	0.9486 ± 0.0113
		Filter + GA + Pathway 05340	30.82 ± 1.62	0.9033	0.9387 ± 0.0202
		Filter + GA + Pathway 04662	5.41 ± 1.20	0.1141	0.9277 ± 0.0212
		Filter + GA + Pathway 04670	5.32 ± 1.03	0.0917	0.9136 ± 0.0281
		Filter + GA + Pathway 05200	4.86 ± 0.76	0.0750	0.9153 ± 0.0217
		Filter + GA + Pathway 04062	4.98 ± 0.99	0.0847	0.9088 ± 0.0242
	Lung	GA	3.77 ± 0.87	0.0750	0.9678 ± 0.0069
		Filter + GA	3.91 ± 0.65	0.0740	0.9696 ± 0.0068
		Filter + GA + Pathway 04144	4.15 ± 0.57	0.1033	0.9625 ± 0.0097
		Filter + GA + Pathway 04530	3.55 ± 0.64	0.2085	0.9705 ± 0.0090
		Filter + GA + Pathway 04514	3.84 ± 0.78	0.1300	0.9680 ± 0.0073
		Filter + GA + Pathway 04610	5.29 ± 1.06	0.1334	0.9625 ± 0.0108
		Filter + GA + Pathway 04010	4.00 ± 0.73	0.1070	0.9656 ± 0.0086
		Filter + GA + Pathway 05200	4.12 ± 0.63	0.0711	0.9621 ± 0.0091
	Prostate	GA	7.63 ± 1.24	0.1126	0.8705 ± 0.0310
		Filter + GA	8.17 ± 1.46	0.1060	0.8645 ± 0.0250
		Filter + GA + Pathway 00480	26.24 ± 4.02	0.4722	0.8890 ± 0.0229
		Filter + GA + Pathway 00040	24.54 ± 1.18	0.7890	0.8820 ± 0.0241
		Filter + GA + Pathway 04610	9.14 ± 1.32	0.1012	0.8713 ± 0.0210
		Filter + GA + Pathway 00980	11.15 ± 2.10	0.1289	0.8796 ± 0.0239
		Filter + GA + Pathway 04512	9.02 ± 1.58	0.0903	0.8613 ± 0.0227
		Filter + GA + Pathway 05215	8.34 ± 1.30	0.1046	0.8659 ± 0.0232

considered as enzymes (KEGG database) is applied. It is important to note that no information about the class (relapse or not) is used to carry out the filtering process, unlike other statistical techniques such as CFS (Correlation-based Feature Selection) [47], mRMR (minimum Redundancy Maximum Relevance) [48] or Relief [44]. Thus, for *Leukemia* database, we move from 7129 to 3413 genes, for *Lung* from 12,533 to 5470 variables, and *Prostate* from 12,600 to 5489, obtaining a reduction of the 50% of the total. Subsequently, over the reduced set of features a standard genetic algorithm is applied. The aim is to check if the selection of relevant information from a biological point of view can guide the search for solutions with greater predictive capacity. Finally, we have many strategies as pathways selected in the first phase of the methodology, naming the strategies (Filter + GA + Pathway *code*). In this case, the standard genetic algorithm is modified to give advantage to the genes of the pathway analyzed, using the techniques discussed in Section 2.2.2.

According to the classification models to be used in the fitness function of the proposed strategies, the authors have considered to carry out the simulations by applying both LDA and SVM classifiers. Since LDA has no parameters, no adjustment has been required. On the other hand, for the SVM method, a grid search strategy is applied for finding optimal parameter values for each of the fifty resampling for each cancer dataset, and is performed before the genetic algorithm. The tentative parameters to be selected are, namely: the kernel type, t = {linear, polynomial, radial base function, sigmoid}, cost, Co = {1, 3, 5, 7, 9, 10, 12, 15}, degree, d = {1, 2, 3, 4, 5}, gamma, g = {0.001, 0.005, 0.1, 0.15, 0.2, 0.4, 0.6,

Table 5
Performance comparison among the "Filter + GA + Pathway" combined strategy and three well-known filtering methods (Cons, IG and ReliefF). ACC and number of genes (*mean* ± *std*) are reported for LDA and SVM classifiers on the three analyzed datasets.

Strategy	Leukemia			
	LDA		SVM	
	ACC	#Genes	ACC	#Genes
Filter + GA + Pathway 05340	97.13 ± 1.16	31.83 ± 1.86	93.87 ± 2.02	30.82 ± 1.62
Filter + GA + Pathway 04640	96.38 ± 1.26	4.47 ± 0.71	94.86 ± 1.13	4.05 ± 0.80
Cons	85.85 ± 8.55	1.84 ± 0.51	88.24 ± 5.95	1.84 ± 0.51
IG	93.13 ± 4.40	9 ± 0	93.36 ± 4.33	9 ± 0
ReliefF	93.31 ± 4.37	9 ± 0	90.48 ± 5.15	9 ± 0
Lung				
Filter + GA + Pathway 04144	98.09 ± 0.68	4.29 ± 0.53	96.25 ± 0.97	4.15 ± 0.57
Filter + GA + Pathway 04530	98.26 ± 0.46	3.84 ± 0.46	97.05 ± 0.90	3.55 ± 0.64
Cons	94.08 ± 3.36	1.84 ± 0.42	94.57 ± 2.55	1.84 ± 0.42
IG	98.68 ± 1.51	22 ± 0	98.88 ± 1.39	22 ± 0
ReliefF	97.89 ± 1.81	22 ± 0	98.47 ± 1.43	22 ± 0
Prostate				
Filter + GA + Pathway 00980	91.37 ± 1.15	8.27 ± 0.83	87.96 ± 2.39	11.15 ± 2.10
Filter + GA + Pathway 00480	90.80 ± 1.36	14.30 ± 2.63	88.90 ± 2.29	26.24 ± 4.02
Cons	81.51 ± 7.57	3.20 ± 0.67	82.49 ± 6.72	3.20 ± 0.67
IG	91.66 ± 4.07	12 ± 0	85.86 ± 4.86	12 ± 0
ReliefF	90.22 ± 4.53	12 ± 0	88.50 ± 5.17	12 ± 0

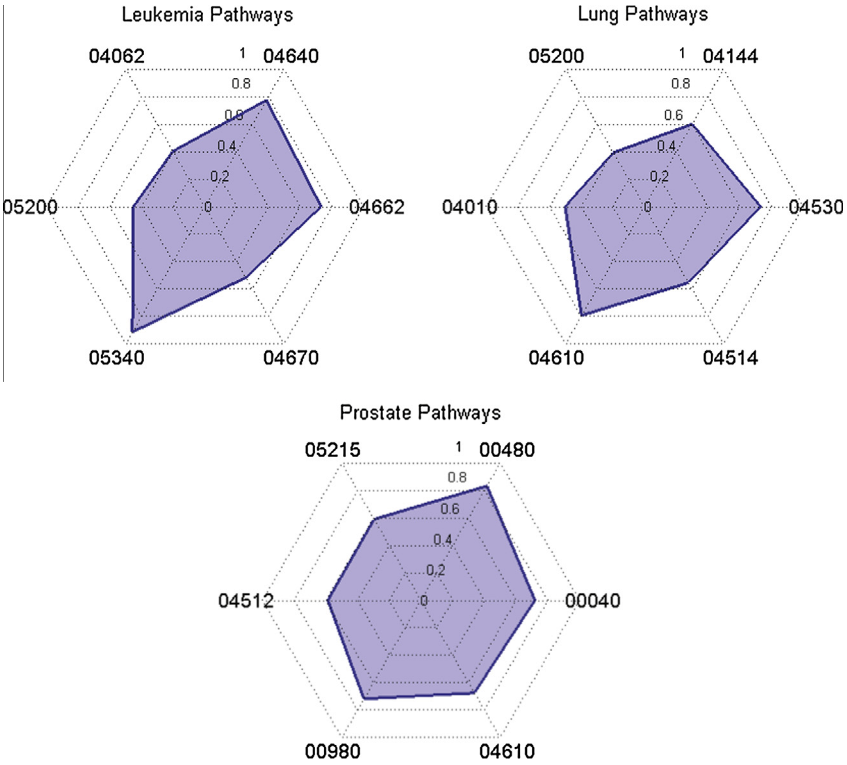


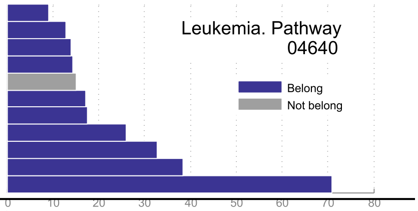
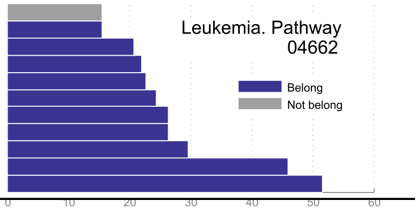
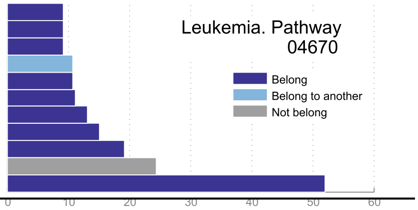
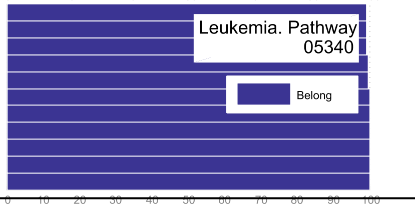
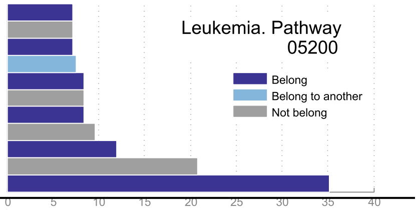
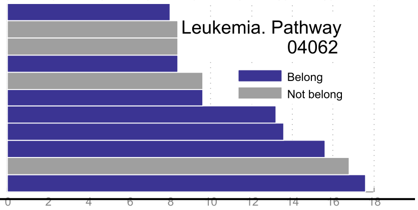
Fig. 2. Proportion of the final selected genes which belong to the analyzed pathway for the databases Leukemia, Lung and Prostate.

0.8, 1, 2, 3, 5} and coef0, $r = \{0, 1, 2\}$. It should be noted that not all the parameters are required for each kernel type. For further information please visit [49].

Table 4 shows a comparison of the results after applying different strategies. The first three columns show the classification method, the dataset and the strategy used. The fourth column represents the number of genes, on average, after executing the method fifty resamplings and five repetitions for each resampling. Robustness column in Table 2 indicates the average frequency of the most selected genes, which are those that appear more than 5% of the time in any of the solutions. The last column shows the result of prediction of the disease over a test set not used during all the process.

The accuracy results for the LDA method are, in general, slightly better than those obtained by applying SVM, although LDA has lower complexity. This is not surprising since it has been shown before that simpler classification techniques can lead to competitive or even better results [50]. Therefore, the following analysis done to

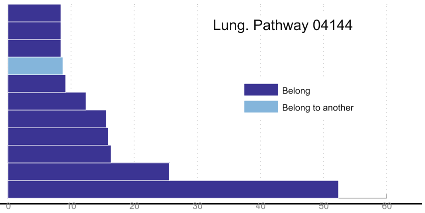
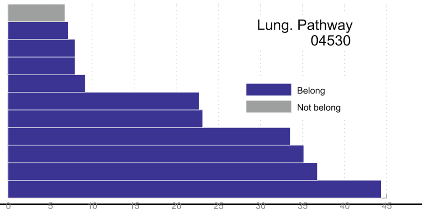
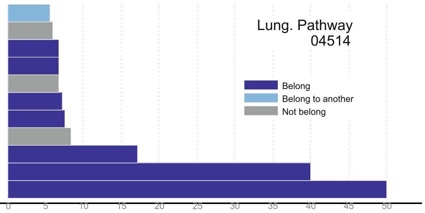
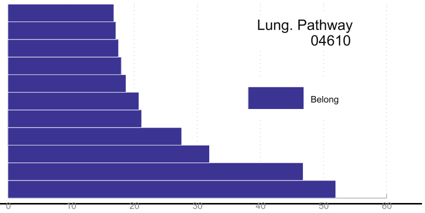
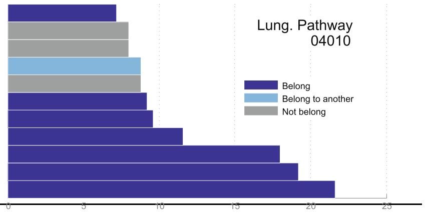
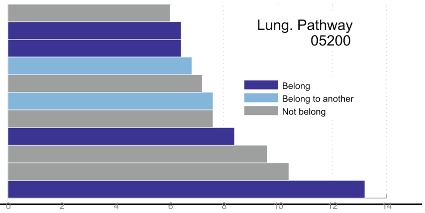
Table 6
The ten most selected features for the Leukemia database. Frequency selection is represented by an horizontal bar, where blue color indicates that the gene belong to the pathway, cyan color which belong to other of the selected pathways and gray color means that this gene do not belong to any of them. The index, gene symbol and probe set ID of each gene are shown in columns one to three. (Note that the axes of the bar graphs are different for different pathways).

ID	*Symbol*	*Probe Set ID*	*Freq.(%)*	*Bar Representation*
7128	GYPA	M71243_f_at	9.20	
50	TFRC	M11507_3_at	12.80	
49	TFRC	M11507_M_at	14.00	
5765	KIT	X06182_s_at	14.40	
4847	ZYX	X95735_at	15.20	
6974	CD19	M28170_at	17.20	
2063	CSF3R	M59820_at	17.60	
6225	CD19	M84371_rna1_s_at	26.00	
1975	ITGA2B	M34344_at	32.80	
1685	DNTT	M11722_at	38.40	
1834	CD33	M23197_at	70.80	
4847	ZYX	X95735_at	15.60	
6999	PRKCB	X06318_at	15.60	
6974	CD19	M28170_at	20.80	
2439	PPP3CC	S46622_at	22.00	
4324	CD22	X59350_at	22.80	
2642	CD79A	U05259_rna1_at	24.40	
1109	IFITM1	J04164_at	26.40	
6207	PRKCB	M18255_cds2_s_at	26.40	
6225	CD19	M84371_rna1_s_at	29.60	
1745	LYN	M16038_at	46.00	
1962	CD81	M33680_at	51.60	
1775	ACTG1	M19283_at	9.20	
3784	PTPN11	U79291_at	9.20	
6967	ITGB2	M15395_at	9.20	
1834	CD33	M23197_at	10.80	
4653	PIK3R2	X80907_at	10.80	
6207	PRKCB	M18255_cds2_s_at	11.20	
2793	PXN	U14588_at	13.20	
5130	RHOH	Z35227_at	15.20	
5552	CXCR4	L06797_s_at	19.20	
4847	ZYX	X95735_at	24.40	
4211	EZR	X51521_at	52.00	
4352	CD40	X60592_at	99.20	
4928	PTPRC	Y00062_at	99.20	
5767	CD79A	X13451_s_at	99.20	
2642	CD79A	U05259_rna1_at	99.60	
5041	RFXAP	Y12812_at	99.60	
181	IL2RG	D11086_at	100.00	
2717	JAK3	U09607_at	100.00	
4050	CD3D	X03934_at	100.00	
6228	LCK	M26692_s_at	100.00	
6236	CD3E	M23323_s_at	100.00	
6510	LCK	U23852_s_at	100.00	
2063	CSF3R	M59820_at	7.20	
2288	CFD	M84526_at	7.20	
4714	FADD	X84709_at	7.20	
1834	CD33	M23197_at	7.60	
4447	MAX	X66867_cds1_at	8.40	
4951	NME4	Y07604_at	8.40	
6801	RB1	L49229_f_at	8.40	
1779	MPO	M19507_at	9.60	
5765	KIT	X06182_s_at	12.00	
4847	ZYX	X95735_at	20.80	
1975	ITGA2B	M34344_at	35.20	
6200	IL8	M28130_rna1_s_at	8.00	
3252	MGST1	U46499_at	8.40	
4951	NME4	Y07604_at	8.40	
6201	IL8	Y00787_s_at	8.40	
1779	MPO	M19507_at	9.60	
5445	GNB1	X04526_at	9.60	
1800	CCL5	M21121_at	13.20	
1745	LYN	M16038_at	13.60	
5552	CXCR4	L06797_s_at	15.60	
4847	ZYX	X95735_at	16.80	
3938	TIAM1	U90902_at	17.60	

extract the most significant genes for each cancer dataset is conducted only with the LDA classifier, since it provides a better accuracy rate, does not require any parameter setting, and is a simple and fast classification method. Additionally, note that the objective of the present work is not the comparison of different classification algorithms, but the extraction of robust feature subsets with potential biological relevance. It is remarkable that the use of biological knowledge by means of the pathways (information obtained from the KEGG database) more related to the analyzed diseases improves the GA strategy in all three data sets, being this improvement statistically significant in two of them (Leukemia and Lung).

The statistical test used to determine this significant difference involves a balanced two-way ANOVA followed by a multiple comparison procedure with a Bonferroni correction (p-value = 0.05). Thus, for *Leukemia* dataset, selected genes with the strategies based on the pathways 04640, 05340 and 04662 provide a better prediction than the reference strategy (GA). Furthermore, the incorporation of biological information improves the robustness, in the sense that there is less variability in the final subset of selected genes after executing several times the algorithm. In the case of the *Lung* database the strategies with pathways 04144 and 04530 also improve the forecast of the standard genetic algorithm.

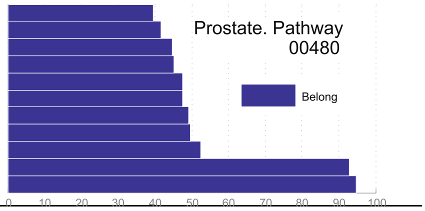
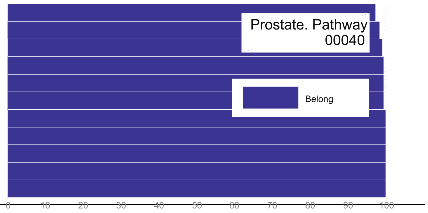
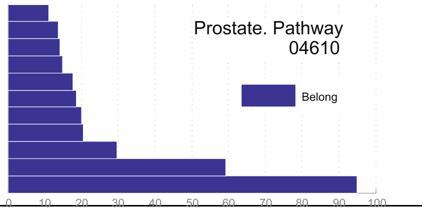
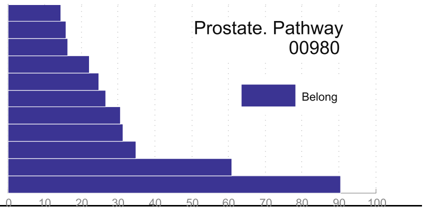
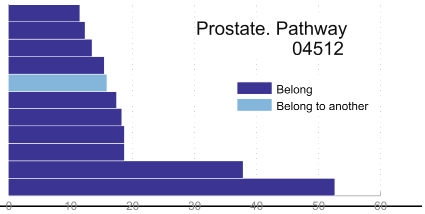
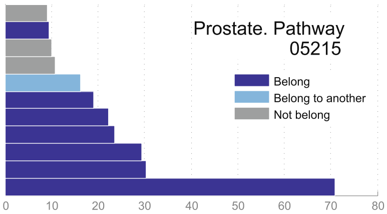
Table 7
The ten most selected features for the Lung database.

ID	Symbol	Probe Set ID	Freq.(%)	Bar Representation
633	ERBB3	1585_at	8.40	
833	TGFB3	1767_s_at	8.40	
9758	SH3GLB1	39691_at	8.40	
3844	SPTAN1	33833_at	8.80	
10168	EHD1	40098_at	9.20	
2521	CLTB	32523_at	12.40	
1182	ERBB3	2089_s_at	15.60	
425	RHOA	1394_at	16.00	
2520	CLTB	32522_f_at	16.40	
9371	CLTB	39307_s_at	25.60	
9863	AP2M1	39795_at	52.40	
7748	SEMA3C	376_at	6.80	
4174	ACTG1	34160_at	7.20	
881	PRKCD	1810_s_at	8.00	
7354	RHOA	37309_at	8.00	
11052	PARD3	40973_at	9.20	
425	RHOA	1394_at	22.80	
8393	RRAS	38338_at	23.20	
2039	PRKCD	32046_at	33.60	
8537	CLDN7	38482_at	35.20	
5301	CLDN4	35276_at	36.80	
3844	SPTAN1	33833_at	44.40	
8393	RRAS	38338_at	5.60	
8370	ALDH1A2	38315_at	6.00	
1245	SELE	265_s_at	6.80	
8508	ICAM2	38453_at	6.80	
9707	GFPT2	39640_at	6.80	
3173	NEO1	33169_at	7.20	
3583	CD226	33575_at	7.60	
7748	SEMA3C	376_at	8.40	
1143	CDH2	2053_at	17.20	
5301	CLDN4	35276_at	40.00	
8537	CLDN7	38482_at	50.00	
9843	SERPING1	39775_at	16.80	
5727	CFI	35698_at	17.20	
3459	PLAT	33452_at	17.60	
9373	BDKRB2	39309_at	18.00	
5925	CR1	35894_at	18.80	
6581	F3	36543_at	20.80	
6821	SERPINA1	36781_at	21.20	
8496	CD46	38441_s_at	27.60	
5853	CFB	35822_at	32.00	
8178	SERPINE1	38125_at	46.80	
9474	C1R	39409_at	52.00	
12532	IL1R2	998_s_at	7.20	
6571	PTGIS	36533_at	8.00	
9707	GFPT2	39640_at	8.00	
3844	SPTAN1	33833_at	8.80	
8370	ALDH1A2	38315_at	8.80	
1146	FGFR1	2056_at	9.20	
5356	FLNC	35330_at	9.60	
8130	FLNB	38078_at	11.60	
667	FGF9	1616_at	18.00	
3250	MAPK13	33245_at	19.20	
8393	RRAS	38338_at	21.60	
8370	ALDH1A2	38315_at	6.00	
1136	JUP	2047_s_at	6.40	
5104	FGF9	35081_at	6.40	
9863	AP2M1	39795_at	6.80	
12298	PTGIS	759_at	7.20	
3844	SPTAN1	33833_at	7.60	
9707	GFPT2	39640_at	7.60	
12047	STAT5A	506_s_at	8.40	
6571	PTGIS	36533_at	9.60	
7748	SEMA3C	376_at	10.40	
667	FGF9	1616_at	13.20	

The authors also compared the performance of some well-known methods to do feature selection (Cons, IG and ReliefF) regards to the ‘‘Filter + GA + Pathway’’ combined strategy. Table 5 shows the ACC for the best solution obtained by the combined strategy and the methods Cons, IG and ReliefF, all of them by using LDA and SVM classifiers on the three analyzed datasets. The ‘‘Filter + GA + Pathway’’ strategy equalizes or outperforms the feature selection methods in terms of prediction accuracy, but with the advantage of incorporating some biological knowledge about the dynamic of the disease. Regarding the number of selected genes, the Cons method behaves very aggressive and extracts a very small set as significant genes, whereas IG and ReliefF, as ranked methods, provide sorted solutions with a higher number of genes that makes necessary to establish a cut-off criteria ($N/8$ with N as the sample size to retain a similar number of genes regarding to the other strategies).

Not only is it important to analyze the robustness and prediction of the solutions for each strategy. Another aspect to consider is the choice of genes in the selected feature subsets. So, those strategies whose solutions include more genes of the pathway analyzed indicate that this pathway may have a greater influence on the disease. This information is shown in Fig. 2 for the selected pathways for each dataset as a ratio, where the closer to one the greater number of genes in the pathway are in the solutions obtained. Thus, in the *Leukemia* dataset the strategies of pathways 05340 (Primary immunodeficiency) and 04640 (Hematopoietic cell lineage) include many of those genes from these pathways, which may imply that its relationship with *Leukemia* disease could be significant. In fact, the biological meaning of both pathways seems to be highly related to leukemia. Other important pathways in relation to the disease are found in the same Fig. 2, such as the 04610 and 04530 for *Lung* database and the 00480 pathway for *Prostate*.

Table 8
The ten most selected features for the Prostate database.

ID	*Symbol*	*Probe Set ID*	*Freq.(%)*	*Bar Representation*
7956	RRM1	34314_at	39.60	
6734	ANPEP	39385_at	41.60	
11629	ODC1	1081_at	44.80	
4584	GCLC	31850_at	45.20	
5418	GPX2	35194_at	47.60	
6008	GSTT1	37222_at	47.60	
4473	GGCT	41696_at	49.20	
11438	GSTM5	1290_g_at	49.60	
8527	ODC1	36203_at	52.40	
7139	GSTA4	40508_at	92.80	
11871	GSTP1	829_s_at	94.80	
8857	UGP2	37373_at	97.39	
5438	UGDH	35214_at	98.51	
4274	UGT2B7	41377_f_at	99.25	
143	UGT2B11	31382_f_at	99.63	
3090	RPE	37797_at	99.63	
7584	ALDH2	32747_at	99.63	
528	XYLB	31767_at	100.00	
1001	UGT2B17	33673_r_at	100.00	
1449	UGP2	35558_at	100.00	
4969	GUSB	33308_at	100.00	
7040	ALDH3A2	40409_at	100.00	
6700	CD59	39351_at	11.07	
4682	C3AR1	32068_at	13.83	
8316	PROS1	35752_s_at	14.23	
9058	F13A1	38052_at	15.02	
5988	F2	37202_at	17.79	
2675	C8B	36304_at	18.58	
5469	F5	35245_at	20.16	
6805	CD55	39695_at	20.55	
8611	PLG	36646_at	29.64	
8878	C7	37394_at	59.29	
9850	CFD	40282_s_at	94.86	
11854	CYP1B1	859_at	14.53	
11497	GSTZ1	1212_at	15.88	
8651	ALDH1A3	36686_at	16.22	
4274	UGT2B7	41377_f_at	22.30	
7139	GSTA4	40508_at	24.66	
6941	CYP1B1	40071_at	26.69	
10968	CYP3A4	1756_f_at	30.74	
11245	CYP2C18	1477_s_at	31.42	
2895	CYP3A5	37124_i_at	34.80	
8952	ADH5	37707_i_at	60.81	
11871	GSTP1	829_s_at	90.54	
12054	THBS2	659_g_at	11.60	
3015	ITGA1	37484_at	12.40	
9133	SDC1	38127_at	13.60	
4805	COL4A5	32667_at	15.60	
9850	CFD	40282_s_at	16.00	
8085	COL6A2	34802_at	17.60	
10689	ITGB5	2058_s_at	18.40	
1244	GP5	34633_s_at	18.80	
7032	COMP	40162_s_at	18.80	
12594	THBS4	103_at	38.00	
3794	COL4A6	39939_at	52.80	
12495	PTGDS	216_at	9.16	
9264	IGF1	38737_at	9.56	
5890	PEX3	36864_at	9.96	
9172	PTGDS	38406_f_at	10.76	
9850	CFD	40282_s_at	16.33	
11215	KLK3	1513_at	19.12	
12378	PIK3R3	322_at	22.31	
11204	IGF1	1501_at	23.51	
9900	FOXO1	40570_at	29.48	
8610	RELA	36645_at	30.28	
11871	GSTP1	829_s_at	70.92	

Tables 6–8 present the ten most selected genes for the six pathways considered for each database, where each pathway is represented in a row of the table. The selection in our two-step approach does not take into account up- or down-regulation. This implies that, after the analysis, the user must look at the original data to know the sense of regulation. The first three columns show information about the gene, such as the internal index (ID), the gene symbol (name of the gene although it is not unique) and the probe set ID, which is related to the chip where the database has been extracted (e.g., Affymetrix). The bar graph of the last column represents the frequency of selection (fourth column) of each feature in the generated solutions. Blue color indicates that the gene belongs to the pathway analyzed; cyan color which the gene belongs to another of the selected pathway; and gray color assumes the gene does not belong to any of the analyzed pathways. A higher frequency of selection might imply a higher relevance of the gene in the prognosis of the disease.

It should be highlighted that the genes of the analyzed pathway are stimulated to be selected. However, they are discarded if its prediction ability of the disease (together with the remaining selected genes) is poor. Therefore, those genes which are selected out of the pathway and are rather frequent, could be considered as relevant genes associated with the prognosis of the disease.

4. Discussion and conclusions

The authors have analyzed in this work three cancer data sets using a combined approach of genetic algorithms and biological

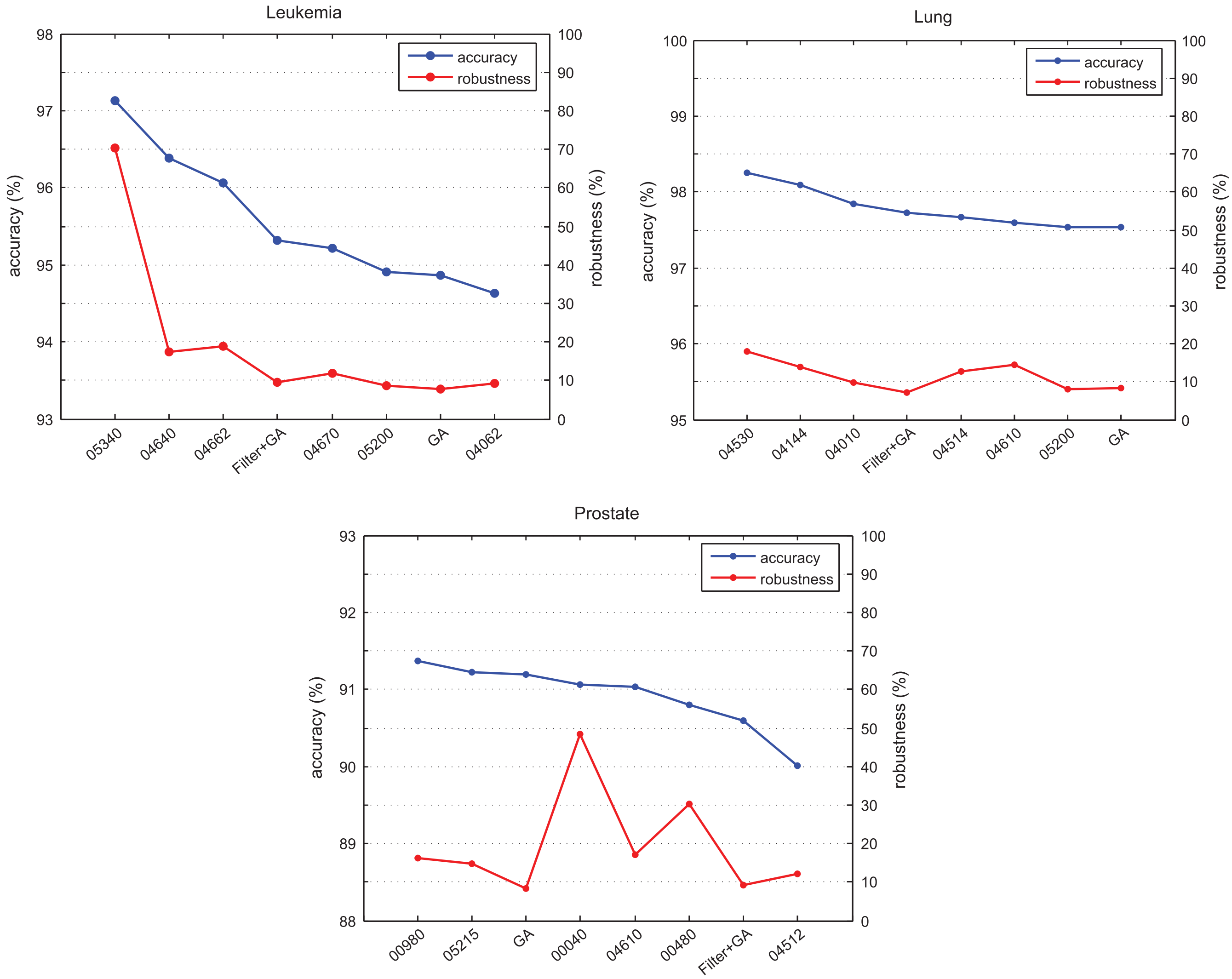


Fig. 3. Accuracy and robustness obtained for the selected pathways for each considered database (Leukemia, Lung and Prostate). The graphs include the results obtained when using a strategy based only on genetic algorithms (GA) and on genetic algorithms plus the filtering approach (Filter + GA) (see text for more details).

relevant information, in order to obtain a robust feature subset selection with good performance rates. The approach incorporates a novel feature scoring method within the GA, taking into account biological information about proteins (mostly enzymes) involved in the pathways of the studied disorders. The most remarkable finding is that our proposal improves the standard GA strategy regardless of the classification model used (LDA or SVM) in the three analyzed data sets (Table 4, Accuracy column), leading to statistically significant results in two of them (Leukemia and Lung). Even more important from the biological and clinic point of view, the robustness, in terms of the most selected genes that can be used to define gene signatures, is also improved in all three analyzed databases (Table 4, Robustness column). The main consequence of both facts is that the results of a KEGG-improved GA can provide more repetitive and consistent results that will facilitate the definition of gene signatures for further clinical diagnostic and prognostic. Moreover, the comparative analysis done among the KEGG-improved GA (Table 5) and three alternative filter methods (Cons, IG and ReliefF) demonstrated a similar or higher performance of the KEGG-improved GA, with the additional benefit of the biological information about the disease dynamics provided by this new GA-based strategy.

Regarding the summarizing results of Fig. 3 it can be seen that the best placed pathways in Table 4 provide more accurate and robust results. This opens the possibility of a deeper study of which KEGG-pathway(s) provide(s) the better results for any disease dataset. It should be noted that those feature subsets that include more genes of the analyzed pathways analyzed might indicate that this particular pathway has a greater biological impact on the disease.

But the proposed KEGG-improved GA not only can be used for diagnostic and prognostic, but also for biological knowledge discovery about the disease. Regarding the most remarkable genes of Tables 6–8 that even not originally present in the selected pathways, form part of the final selection thus playing an important role for obtaining robust and accurate prediction results. For example, in Table 6 (Leukemia set), the gene ZYX[7] is repetitively selected in all but one pathways; it codes zyxin, a adhesion plaque protein that prompts the formation of actin-rich structures at which signal transduction assemble. In the case of the lung database (Table 7), several adhesion pathways are involved in this cancer (cf., 04530, 04514) while the ZYX gene does not seem to be significant. The gene SEMA3C[8] corresponds to a semaphorin, a protein including an inmunoglobulin domain. It

[7] http://www.genecards.org/cgi-bin/carddisp.pl?gene=ZYX.

[8] http://www.genecards.org/cgi-bin/carddisp.pl?gene=SEMA3C.

seems to play an important role in the regulation of developmental processes and axon growing. Its presence suggests that pathways 04360 and others should be considered for future analysis. Also, gene ALDH1A2[9] is related to an aldehyde dehydrogenase enzyme that synthesises retinoic acid (RA) from retinaldehyde. RA is a hormonal signaling molecule that functions in developing and adult tissues and has been involved in spina bifida. As a result, might high levels of RA be involved in lung cancer? Gene GFPT2[10] corresponds to D-fructose-6-phosphate amidotransferase, an enzyme involved in regulating the availability of precursors for N- and O-linked glycosylation of proteins. Protein glycosilation might be affected in lung cancer, and thus it deserves further analysis. PTGIS,[11] although selected only in two pathways, is a prostaglandin I2 (prostacyclin) synthase, a protein of cytochrome P450 superfamily of enzymes, involved in the synthesis of prostacyclin, a potent vasodilator and inhibitor of platelet aggregation that is also related to myocardial infarction, stroke, and atherosclerosis, and thus could be also involved in lung cancer.

As an overall conclusion, the results obtained suggest the important role that the incorporation of biological information might play for carrying out a robust feature selection procedure for cancer (and may be any other disease) diagnostic. Moreover, this may open the way to use GA for the prognosis of cancer diseases in a near future, a clinical aspect that is still concerning most oncologist and cancer patients.

Acknowledgments

The authors acknowledge support through Grants TIN2010-16556 from MICINN-SPAIN and P08-TIC-04026 and P10-CVI-6075 (Junta de Andalucía), all of which include FEDER funds.

References

[1] Zhang B-H, Liu J, Zhou Q-X, Zuo D, Wang Y. Analysis of differentially expressed genes in ductal carcinoma with DNA microarray. Eur Rev Med Pharmacol Sci 2013;17(6):758–66.

[2] Harper KN, Peters B, Gamble MV. Batch effects and pathway analysis: two potential perils in cancer studies involving DNA Methylation array analysis. Cancer Epidemiol Biomarkers Prevent 2013;22(6):1052–60.

[3] Cooley LD, Lebo M, Li MM, Slovak ML, Wolff DJ. American college of medical genetics and genomics technical standards and guidelines: microarray analysis for chromosome abnormalities in neoplastic disorders. Genetics Med 2013;15(6):484–94.

[4] Wong H-S, Wang H-Q. Constructing the gene regulation-level representation of microarray data for cancer classification. J Biomed Inform 2008;41(1):95–105.

[5] West M. Bayesian factor regression models in the large p, small n paradigm. Bayesian Stat 2003;7(2003):723–32.

[6] Castellanos-Garzón JA, Díaz F. An evolutionary computational model applied to cluster analysis of DNA microarray data. Expert Syst Appl 2013;40(7): 2575–91.

[7] Sungheetha A, Suganthi J. An efficient clustering-classification method in an information gain NRGA-KNN algorithm for feature election of micro array data. Life Sci J 2013;10(Suppl. 7):691–700.

[8] Keedwell E, Narayanan A. Gene expression rule discovery and multi-objective ROC analysis using a neural-genetic hybrid. Int J Data Mining Bioinform 2013;7(4):376–96.

[9] Gupta S, Garg S. Multiobjective optimization using genetic algorithm. Adv Chem Eng 2013;43:206–45.

[10] Liu H, Liu L, Zhang H. Ensemble gene selection by grouping for microarray data classification. J Biomed Inform 2010;43(1):81–7.

[11] Kim JH, Jeoung D, Lee S, Kim H. Discovering significant and interpretable patterns from multifactorial DNA microarray data with poor replication. J Biomed Inform 2004;37(4):260–8.

[12] Pan F, Wang B, Hu X, Perrizo W. Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis. J Biomed Inform 2004;37(4):240–8.

[13] Feng T, Xuezheng F, Yanqing Z, Bourgeois A. Improving feature subset selection using a genetic algorithm for microarray gene expression data. In: IEEE congress on evolutionary computation, CEC; 2006. p. 2529–34.

[14] Dolled-Filhart M, Rydén L, Cregger M, Jirström K, Harigopal M, Camp R, et al. Classification of breast cancer using genetic algorithms and tissue microarrays. Clin Cancer Res 2006;12(21):6459–68.

[15] Melita N, Popescu I, Holban S. A genetic algorithm approach to DNA microarrays analysis of pancreatic cancer. Adv Electr Comput Eng 2008;8(2):43–8.

[16] Bonilla Huerta E, Duval B, Hao J-K. A hybrid LDA and genetic algorithm for gene selection and classification of microarray data. Neurocomputing 2010;73(13–15):2375–83.

[17] Liu H, Wang W. Genetic algorithm and support vector machine-based gene microarray analysis. J Clin Rehabil Tissue Eng Res 2010; 14(17):3099–103.

[18] Lee C-P, Lin W-S, Chen Y-M, Kuo B-J. Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method. Expert Syst Appl 2011;38(5):4661–7.

[19] Loughrey J, Cunningham P. Overfitting in wrapper-based feature subset selection: the harder you try the worse it gets. In: Bramer M, Coenen F, Allen T, editors. Research and development in intelligent systems XXI. London: Springer; 2005. p. 33–43.

[20] Yang F, Mao KZ. Robust feature selection for microarray data based on multicriterion fusion. IEEE/ACM Trans Comput Biol Bioinform 2011;8(4):1080–92.

[21] Abedini M, Kirley M, Chiong R. Incorporating feature ranking and evolutionary methods for the classification of high-dimensional DNA microarray gene expression data. Austr Med J 2013;6(5):272–9.

[22] Panteris E, Swift S, Payne A, Liu X. Mining pathway signatures from microarray data and relevant biological knowledge. J Biomed Inform 2007; 40(6):698–706.

[23] Fu WJ, Carroll RJ, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics 2005;21(9):1979–86.

[24] Pan W. Bootstrapping likelihood for model selection with small samples; 1998.

[25] Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25(1):25–9.

[26] Smith B, Williams J, Schulze-Kremer S. The ontology of the gene ontology. In: Proceedings of AMIA symposium; 2003. p. 609–13.

[27] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucl Acids Res 2012;40(Database-Issue):109–14.

[28] Wrzodek C, Büchel F, Ruff M, Dräger A, Zell A. Precise generation of systems biology models from KEGG pathways. BMC Syst Biol 2013;7(1):1–12.

[29] Cui W, Chen L, Huang T, Gao Q, Jiang M, Zhang N, et al. Computationally identifying virulence factors based on KEGG pathways. Mol Biosyst 2013;9:1447–52.

[30] Mougeot J-L, Li Z, Price A, Wright F, Brooks B. Microarray analysis of peripheral blood lymphocytes from ALS patients and the SAFE detection of the KEGG ALS pathway. BMC Med Genom 2011;4(1):74.

[31] Dinasarapu AR, Gupta S, Maurya MR, Fahy E, Min J, Sud M, et al. A combined omics study on activated macrophages – enhanced role of stats in apoptosis, immunity and lipid metabolism. Bioinformatics 2013; 29(21):2735–43.

[32] Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531–7.

[33] Carlson M, Falcon S, Pages H, Li N. hu6800.db: Affymetrix HuGeneFL Genome Array annotation data (chip hu6800), R package version 2.6.3.

[34] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002;1(2): 203–9.

[35] Carlson M. hgu95av2.db: Affymetrix Human Genome U95 Set annotation data (chip hgu95av2), R package version 2.8.0.

[36] Gordon GJ, Jensen RV, li Hsiao L, Gullans SR, Blumenstock JE, Ramaswamy S, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res 2002;62:4963–7.

[37] Carlson M. hgu95a.db: Affymetrix Human Genome U95 Set annotation data (chip hgu95a), R package version 2.8.0.

[38] Hu Y-L, Fong S, Ferrell C, Largman C, Shen W-F. HOXA9 modulates its oncogenic partner Meis1 to influence normal hematopoiesis. Mol Cell Biol 2009;29(18):5181–92.

[39] Goldberg DE. Genetic algorithms in search. Optimization and machine learning. Addison-Wesley Longman Publishing Co., Inc.; 1989.

[40] Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. Wiley; 2001.

[41] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46(1-3):389–422.

[42] Brown M, Grundy W, Lin D, Cristianini N, Sugnet C, Furey T, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Nat Acad Sci USA 2000;97(1):262–7.

[43] Kononenko I. Estimating attributes: analysis and extensions of relief. Springer Verlag; 1994. p. 171–82.

[44] Kira K, Rendell LA. A practical approach to feature selection. In: Proceedings of the ninth international workshop on machine learning; 1992. p. 249–56.

[9] http://www.genecards.org/cgi-bin/carddisp.pl?gene=ALDH1A2.

[10] http://www.genecards.org/cgi-bin/carddisp.pl?gene=GFPT2.

[11] http://www.genecards.org/cgi-bin/carddisp.pl?gene=PTGIS.

[45] Dash M, Liu H. Consistency-based search in feature selection. Artif Intell 2003;151(1–2):155–76.

[46] Hall M, Smith L. Practical feature subset selection for machine learning. J Comput Sci 1998;98:4–6.

[47] Hall M. Correlation-based feature selection for machine learning. Ph.D. thesis, University of Waikato; 1999.

[48] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27(8):1226–38.

[49] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2011;2:1–27 [Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm].

[50] Luque-Baena RM, Elizondo D, López-Rubio E, Palomo EJ, Watson T. Assessment of geometric features for individual identification and verification in biometric hand systems. Expert Syst Appl 2013;40(9):3580–94.