



RESEARCH

Open Access

# Application of genetic algorithms and constructive neural networks for the analysis of microarray cancer data

Rafael Marcos Luque-Baena<sup>1,2\*</sup>, Daniel Urda<sup>1,2</sup>, Jose Luis Subirats<sup>1,2</sup>, Leonardo Franco<sup>1,2</sup>, Jose M Jerez<sup>1,2</sup>

From 1st International Work-Conference on Bioinformatics and Biomedical Engineering-IWBBIO 2013 Granada, Spain. 18-20 March 2013

\* Correspondence: rmluque@cc.uma.es  
<sup>1</sup>Department of Computer Science, University of Málaga, Málaga, Spain

## Abstract

**Background:** Extracting relevant information from microarray data is a very complex task due to the characteristics of the data sets, as they comprise a large number of features while few samples are generally available. In this sense, feature selection is a very important aspect of the analysis helping in the tasks of identifying relevant genes and also for maximizing predictive information.

**Methods:** Due to its simplicity and speed, Stepwise Forward Selection (SFS) is a widely used feature selection technique. In this work, we carry a comparative study of SFS and Genetic Algorithms (GA) as general frameworks for the analysis of microarray data with the aim of identifying group of genes with high predictive capability and biological relevance. Six standard and machine learning-based techniques (Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Naive Bayes (NB), C-MANTEC Constructive Neural Network, K-Nearest Neighbors (kNN) and Multilayer perceptron (MLP)) are used within both frameworks using six free-public datasets for the task of predicting cancer outcome.

**Results:** Better cancer outcome prediction results were obtained using the GA framework noting that this approach, in comparison to the SFS one, leads to a larger selection set, uses a large number of comparison between genetic profiles and thus it is computationally more intensive. Also the GA framework permitted to obtain a set of genes that can be considered to be more biologically relevant. Regarding the different classifiers used standard feedforward neural networks (MLP), LDA and SVM lead to similar and best results, while C-MANTEC and k-NN followed closely but with a lower accuracy. Further, C-MANTEC, MLP and LDA permitted to obtain a more limited set of genes in comparison to SVM, NB and kNN, and in particular C-MANTEC resulted in the most robust classifier in terms of changes in the parameter settings.

**Conclusions:** This study shows that if prediction accuracy is the objective, the GA-based approach lead to better results respect to the SFS approach, independently of the classifier used. Regarding classifiers, even if C-MANTEC did not achieve the best overall results, the performance was competitive with a very robust behaviour in terms of the parameters of the algorithm, and thus it can be considered as a candidate technique for future studies.

## Introduction

DNA microarray technology has been widely used in cancer studies for prediction of disease outcome [1]. It is a powerful platform successfully used for the analysis of gene expression in a wide variety of experimental studies [2]. However, due to the large number of features (in the order of thousands) and the small number of samples (mostly less than a hundred) in this kind of datasets, microarray data analysis faces the “large-p-small-n” paradigm [3] also known as the curse of dimensionality. In this sense, feature selection preprocessing refers to decide which genes to include in the prediction, and it is a crucial step in developing a class predictor. Including too many features could reduce the model accuracy and may lead to overfit the data [4]. Two different algorithms have been widely used in literature to carry out feature selection, the Stepwise Forward Selection algorithm (SFS) and the Genetic Algorithms (GA). In the SFS algorithm the choice of predictive features is carried out by an automatic procedure that starts from single variable models and tests the addition of each feature using a comparison criterion. This algorithm has been used to identify a predictive gene signature whose size is minimum [5,6]. GA are also well considered as suitable evolutionary strategies for feature selection in problems with a large number of features [7,8], and are applied to different areas, from object detection [9] to gene selection in microarray data [10].

On the other hand, model selection is another important step in the estimation of expression profiles to predict diseases outcome [11]. In this regards, different well-known machine learning-based techniques have been used recently in literature wrapped into features selection algorithms to develop a class predictor, e.g. Support Vector Machines (SVM) [12], Multilayer Perceptron (MLP) [13], k-Nearest Neighborhood (kNN) [14], Linear Discriminant Analysis (LDA) [15] and NaiveBayes. Nevertheless, few of these related works brings together different learning algorithms, features selection schemes and input datasets. Besides, some of them are focused mainly on optimising the prediction accuracy, and lack of any biological analysis for the resulting molecular signatures via specialised software as Ingenuity Pathway Analysis (IPA), GeneOntology (GO) or KEGG [16].

This paper presents an exhaustive analysis of performance for SFS and GA as general frameworks to estimate expression genes profiles from microarray data with high predictive capability and biological relevance. Five standard and machine learning-based techniques (MLP, SVM, kNN, LDA, NaiveBayes) are used within both frameworks using six free-public cancer datasets (breast, colon, leukemia, lung, ovarian and prostate cancer) for the task of predicting cancer outcome. Moreover, an important goal of the present study is to test the performance of a new constructive neural network classification algorithm (C-MANTEC) in the mentioned framework. C-MANTEC have been previously proved to get similar classification results than traditional multi-layer perceptrons (MLP) or support vector machines (SVM), with the advantage that the architecture is dynamically estimated [17]. This is a critical factor in the wrapper selection methods combined with neural networks, because the subsets analysed are different sizes (or even the complexity of the features selected need projections in higher spaces), which implies that the use of the same architecture is not always appropriate. On the other hand, considering that using non redundant variables is commonly preferable in feature selection, the evolutionary strategy presented in this work

incorporates a mutual information filter to minimise the correlation between the selected features while increasing the classifier performance. Furthermore, a biological analysis of the relevance of the selected genes is performed using IPA tool, which can lead us to conduct an understanding of microarray data.

### **Methodology**

Feature selection techniques can be organised into three broad categories: filter, wrapper and embedded methods [18]. Filter methods use statistical properties of the variables to discard poorly descriptive features and are independent of the classifier. Wrapper methods are more computationally demanding than filter methods, as subsets of features are evaluated with a classification algorithm in order to obtain a measure of goodness to be used as the improvement criteria. Embedded methods are also classifier dependent, but they can be viewed as a search in the combined space of feature subsets and classifier models, with the additional restriction that it is not possible to replace the classifier used since feature selection and classification methods work as a whole.

In this work a comparison between a SFS and GA based approach is done. As the data input space is quite large for microarray data a pre-selection approach is first applied in order to reduce the size of the input features to a 5% of the total. After this reduction, six different classifiers are applied within both frameworks.

### **Pre-selection step**

Since cancer datasets normally contain a large number of genes, a pre-selection step to reduce the initial number of variables is required. In terms of the quality of the features ranked, it has been found that using the Student t-test is generally more successful than other filter methods[19]. In particular, the Welch t-test [20], an adaptation of the t-test, is used for the pre-selection step assuming the two classes (patient has cancer or not) have unknown and unequal variances, as it is not advisable to use the basic t-test if both requirements are not clearly satisfied [18]. A 5% of the total number of genes are retained (between 400 and 2000 genes, approximately, in the datasets selected), which will be the input for the two approaches (SFS and GA) applied, and described below.

### **Stepwise forward selection procedure**

An exhaustive evaluation of all the possible subsets of  $n$  features involves a complexity of  $O(2^n)$  which becomes infeasible for large values for  $n$ . In this sense, several heuristic algorithms have been proposed to reduce the computational complexity of wrapper algorithms. Stepwise forward procedures for feature selection analyse the inclusion of one or several features in order to improve the performance of the classification task. Thus, sequential forward selection [21] chooses the best variable in each iteration by minimising the misclassification rate, and includes it in the final subset of features. The algorithm will continue to add variables until the performance stops to improve.

### **Evolutionary approach**

GAs are a class of optimisation procedure inspired by the biological mechanisms of reproduction. One of the key aspects of GA is the so called fitness function  $f(\mathbf{x})$ , that should be maximised or minimised over a given space  $X$  of arbitrary dimension, in an

iterative search process in which the population of selected genes evolves as described in detail below.

#### **Encoding and initial population**

A simple encoding scheme to represent as much as possible of the available information was employed. A string of bits whose length is equal to the total number of genes is used, using a binary variable associated with each bit. If the  $i^{th}$  bit is active (value 1), then the  $i^{th}$  gene is selected in the chromosome (a value of 0 indicates that the corresponding feature is ignored). Both, the active features and the number of them were generated randomly, and in all the experiments a population size of 100 individuals was used.

#### **Selection, crossover and mutation**

A selection strategy based on roulette wheel and uniform sampling was applied, while an elite count value of 10 (number of chromosomes which are retained for the next generation) was selected. Scattered crossover, in which each bit of the offspring is chosen randomly, was the choice for combining parents of the previous generation, using a crossover rate set to 0.8. In addition to that, a traditional mutation operator which flips a specific bit with a probability rate of 0.2 was considered. Since it was empirically verified that the best subsets include few features, a modification which involves mutating a random number of bits between 1 and the number of active features of the individual was also applied, as this change avoids the increment on the number of active features in the last generations of the GA.

#### **Fitness function**

The fitness function assesses each chromosome in the population so that it can be ranked against all the other chromosomes. Three aspects were considered for constructing the fitness function: i) The main objective is to obtain the highest performance ii) Among two subsets that achieve equal performance, the one that contains a lower number of features is preferred. iii) The combination of features with low redundancy among them and with a certain resemblance to the target class, are beneficial for improving performance rates [22]. Therefore, the fitness function contains three terms: the misclassification error, the number of features selected and a redundancy measure among them. Datasets are splitted into training and testing sets in order to evaluate the generalisation ability of the proposed chromosome.

Statistical techniques such as mutual information [23] can be used for measuring the correlation between a pair of features. The mutual information between two continuous random variables  $y$  and  $z$  is given by the following equation:

$$I(y, z) = \int \int p(y, z) \log \left( \frac{p(y, z)}{p(y)p(z)} \right) dy dz \quad (1)$$

where  $p(y, z)$  is the joint probability density function of  $y$  and  $z$ , and  $p(y)$  and  $p(z)$  are the marginal probability density functions of  $y$  and  $z$  respectively.

Mutual information is a non-negative quantity, with a zero value indicating that the variables are completely independent. The more correlated two variables are, the greater their mutual information. Advantages of this measure are that the dependency between variables is no longer restricted to linear correlation and that it can handle nominal or discrete features. Although it is hard to compute it for continuous data, the probability densities can be well estimated by discretising it through the use of

histograms[24]. A measure which incorporates the correlation of features with the target class and penalises the redundancy among the selected features is described as follows [22]:

$$corr(\mathbf{x}) = \frac{1}{t} \sum_{i=1}^k \sum_{j=i+1}^k I(x_j, x_i) - \frac{1}{k} \sum_{j=1}^k I(x_j, C) \quad (2)$$

where  $k$  is the number of features selected,  $C$  is the target class and  $t$  is the number of combinations between the pairs of the chromosome  $x$  analysed. Finally, the function to be minimised ( the  $fitness(\mathbf{x})$  function) is represented as follows for a given subset  $\mathbf{x}$ .

$$fitness(\mathbf{x}) = (1 - ACC(\mathbf{x})) + \lambda \frac{k}{\mathcal{N}} + \beta corr(\mathbf{x}) \quad (3)$$

where  $ACC(\mathbf{x})$  is the accuracy rate obtained by the classifier on the test set (the percentage of well-classified patterns with regards to the total patterns analysed);  $\mathcal{N}$  is the total number of extracted features; and finally,  $corr(\mathbf{x})$  defines the correlation among the features and the target class, with the aim of avoiding the redundancy in the feature vector (equation 2). The parameters  $\lambda$  and  $\beta$  can take values in the interval (0, 1) and show how influential are the terms *minimisation of the number of genes* and *mutual information* in the fitness function. Further information is provided in the results section.

### C-MANTEC algorithm

C-MANTEC (Competitive Majority Network Trained by Error Correction) [17] is a novel neural network constructive algorithm that utilises competition between neurons and a modified perceptron learning rule to build compact architectures with good prediction capabilities. The novelty of C-MANTEC is that the neurons compete for learning the new incoming data, and this process permits the creation of very compact neural architectures. At the single neuronal level, the algorithm uses the thermal perceptron rule, introduced by Marcus Frean in 1992 [25], that improves the convergence of the standard perceptron for non-linearly separable problems. C-MANTEC, as a CNN algorithm [26,27], has in addition the advantage of generating online the topology of the network by adding new neurons during the training phase, resulting in faster training times and more compact architectures. Its network topology consists of a single hidden layer of thermal perceptrons that maps the information to an output neuron that uses a majority function.

The C-MANTEC algorithm has 3 parameters to be set at the time of starting the learning procedure. Several experiments have shown that the algorithm is very robust against changes of the parameter values and thus C-MANTEC operates fairly well in a wide range of values. The three parameters of the algorithm to be set are: (i) *Imax* as maximum number of iterations allowed for each neuron present in the hidden layer per learning cycle, (ii) *gfac* a growing factor that determines when to stop a learning cycle and include a new neuron in the hidden layer, and (iii) Phi ( $\phi$ ) that determines in which case an input example is considered as noise and removed from the training dataset according to Eq. 4:

$$\forall X \in \{X_1, \dots, X_N\}, \text{ delete}(X) | NTL \geq (\mu + \phi\sigma) \quad (4)$$

where  $X$  represents a given pattern among the  $N$  patterns of the dataset,  $N_{TL}$  is the number of times that pattern  $X$  has been learnt on the current learning cycle, and the pair  $\{\mu, \sigma\}$  corresponds to the mean and variance of the normal distribution that represents the number of times that each pattern of the dataset has been learnt during the learning cycle. Thus, Eq. 4 specifies that if a given pattern ( $X$ ) has been tried to be learnt by the network a number of times larger than  $\phi$  standard deviations above the mean for the population it should be removed from the training set.

### Experimental results

In this section, six free-public cancer datasets (<http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html>) have been used to test the proposed methodology. The main characteristics (# genes, # samples, and class distribution) for each dataset is shown in Table 1. A comparison between the two analyzed frameworks is conducted, where for each methodology six classification techniques are applied, namely: LDA, SVM, NaiveBayes, C-MANTEC, kNN and MLP.

Before applying the methodology based on genetic algorithms, it is necessary to estimate the parameters  $\alpha$  and  $\beta$  associated with the fitness function and referred in a previous section. This estimation is carry out for all the cancer datasets, although only the information related to the *Lung* and *P rostate* datasets are shown by the sake of simplicity. Different combinations of the  $\lambda$  and  $\beta$  parameters together with the accuracy results on average and number of selected genes are shown in Table 2. 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 combinations  $\alpha = 0.4, \beta = 0.25$  and  $\alpha = 0.1, \beta = 0.25$  (Table 2, in italic), lead to the obtention of the largest success rate, taking into account that when  $\alpha$  is reduced ( $\alpha = 0.1$ ) the number of genes in the solution is a little higher (12.78 in *P rostate* and 4.73 in *Lung*) than when we try to minimise the solution with more emphasis ( $\alpha = 0.4, 9.32$  genes in *P rostate* and 4.25 in *Lung*, on average).

Table 3 shows the set of parameters that have to be set for each classifier, together with the different values that have been tested in this paper. For each classifier, a hold-out validation strategy is used by dividing the entire dataset on a 60 – 40% proportion; the first set to train the model and the second to obtain the accuracy in the prediction of cancer outcome. The training-testing procedure is repeated 50 times randomly varying the training and testing set to avoid a biased evaluation, permitting also to analyse the dispersion of the results.

A thorough analysis of the parameter setting is presented in Figure 1, where its influence for the different algorithms is evaluated in the variability of the classification

**Table 1 Cancer datasets**

<i>Dataset</i>	<i>#Genes</i>	<i>Samples</i>	<i>Class 0 (normal)</i>	<i>Class 1 (cancer)</i>	<i>Data Proportion</i>
<b>Leukemia</b>	7129	72	25	47	0.347
<b>Lung</b>	12533	181	150	31	0.829
<b>Colon</b>	2000	62	22	40	0.355
<b>Breast</b>	24481	78	33	44	0.423
<b>Ovarian</b>	15154	253	91	162	0.360
<b>Prostate</b>	12600	102	50	52	0.490

Main characteristics of the six cancer datasets analysed.

**Table 2 Parameters estimation for GA**

Prostate dataset				Lung dataset			
$\alpha$	$\beta$	Accuracy	#Genes	$\alpha$	$\beta$	Accuracy	#Genes
0.8	0.6	0.9838±0.0097	2.67±1.19	0.8	0.6	0.9730±0.0107	8.65±2.82
0.8	0.4	0.9899±0.0072	3.30±1.02	0.8	0.4	0.9748±0.0093	7.28±1.20
0.8	0.25	0.9914±0.0054	3.52±0.91	0.8	0.25	0.9801±0.0106	9.85±3.12
0.4	0.6	0.9827±0.0086	2.56±1.01	0.4	0.6	0.9743±0.0103	8.80±3.18
0.4	0.4	0.9912±0.0069	3.75±1.44	0.4	0.4	0.9763±0.0094	9.55±1.08
0.4	0.25	0.9938 ± 0.0061	4.25 ± 1.95	0.4	0.25	0.9849 ± 0.0089	9.32 ± 1.64
0.1	0.6	0.9837 ± 0.0104	3.04 ± 1.71	0.1	0.6	0.9770 ± 0.0095	7.83 ± 2.06
0.1	0.4	0.9895 ± 0.0065	2.88 ± 0.70	0.1	0.4	0.9763 ± 0.0118	9.63 ± 2.53
0.1	0.25	0.9966 ± 0.0041	4.73 ± 2.10	0.1	0.25	0.9854 ± 0.0101	12.78 ± 1.61

Parameter estimation for the  $\alpha$  and  $\beta$  parameters of the fitness function of the GA for the *Lung* and *Prostate* datasets.

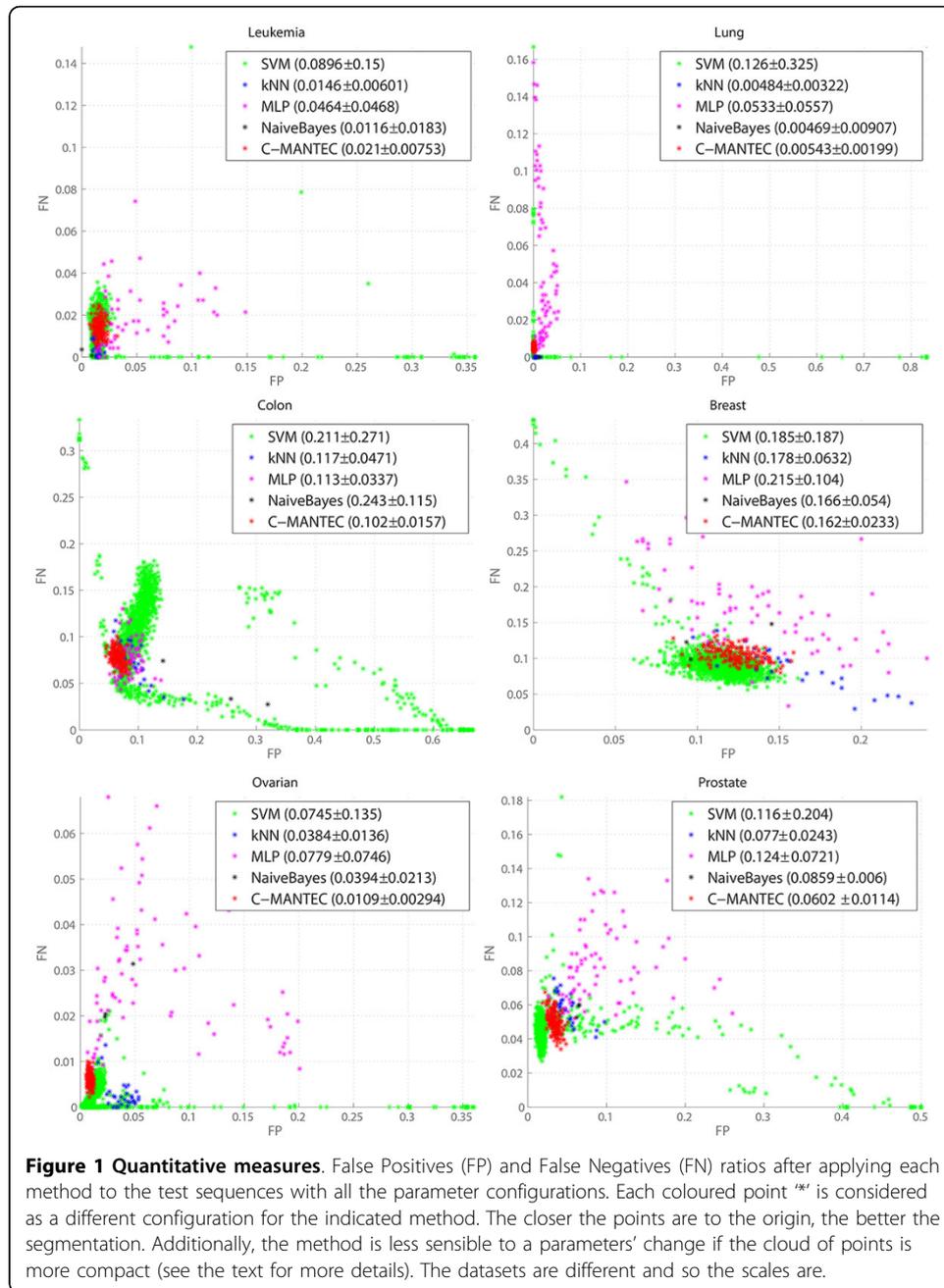
accuracy. The horizontal axis corresponds to the average percentage across the 50 samples considered of the false positives (*FP*) of the data, while the vertical axis is associated with the false negatives values (*FN*). Each point of the plot represents the *FP* and *FN* values of a generated configuration with a given parameter setting. The closer the points are to the origin, the better the classification accuracy, with optimum performance occurring for  $FN = FP = 0$  (a perfect match between the output of the algorithm and the observed outcome of the dataset). All points are located always below the contradiagonal of the plot ( $FN + FP = 1$ ) as it is verified that  $FN + FP \leq 1$ .

The variability observed for each classifier depends largely on the analysed dataset, but with the robustness of each of the method having also a strong influence, as more robust methods yield to more compact configuration clouds of points (a compact configuration cloud means that the results do not vary significantly after a change in the classifier parameters). Thus, the compactness can be defined as the standard deviation of the accuracy measures. As shown in Figure 1, the compactness for kNN, SVM and MLP methods is rather poor in general, while the C-MANTEC approach leads to configurations that are very close together, indicating clearly that the performance of this method is not very sensitive to the parameter selection. Additionally, C-MANTEC lead to the lowest values for the distance of the mean of the configuration values (*FP* and

**Table 3 Parameters settings**

Algorithm	Test Parameters
LDA	No parameters
SVM	Kernel type, $t = \{\text{linear, polynomial, radial base function, sigmoid}\}$ Cost, $C = \{1, 3, 5, 7, 9, 10, 12, 15\}$ Degree, $d = \{1, 2, 3, 4, 5\}$ Gamma, $g = \{0.001, 0.005, 0.1, 0.15, 0.2, 0.4, 0.6, 0.8, 1, 2, 3, 5\}$ Coef0, $r = \{0, 1, 2\}$
NaiveBayes	Kernel density, $K = \{0, 1\}$ Supervised discretization, $D = \{0, 1\}$
C-MANTEC	Max. Iterations, $I_{max} = \{1000, 10000, 100000\}$ GFac, $g_{fac} = \{0.01, 0.05, 0.1, 0.2, 0.25, 0.3\}$ Phi, $\phi = \{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6\}$
kNN	Neighbours, $k = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ Distance type, $d = \{\text{Euclidean, chi-squared, cosine-similarity}\}$
MLP	Hidden neurons, $N_{Hidden} = \{2, 3, 4, 5, 6\}$ Alpha, $\alpha = \{0.05, 0.1, 0.2, 0.3, 0.5\}$ Number cycles, $N_{Cycles} = \{10, 25, 50\}$

Parameter settings tested during evaluation of the classification algorithms. The combination of all the values of the parameters generate a set of configurations for each method.



FN) to the origin, confirming the robustness in the parameter setting (the LDA classifier does not have parameters to be set and thus it is not represented in the graph). In order to quantify the distribution of the prediction accuracy observed for the several configuration analysed, the legend for each classifier shows the distance to the plot origin plus/minus the standard deviation  $\left(\sqrt{(FP)^2 + (FN)^2} \pm \text{std-dev}\right)$ . For example, for the *Ovarian*, *Colon* and *Prostate* datasets, the distance to the origin for the mean value observed for the C-MANTEC algorithm is significantly lower than for the rest of alternatives (0.0109, 0.102 and 0.0602, respectively).

Comparison results between the two frameworks are shown in Table 4, where the best parameter configuration for each classification model is selected to perform the evaluation over the six datasets. In both frameworks, the accuracy rates for the *Leukemia*, *Lung* and *Ovarian* datasets are close to 100% regardless of the classifier applied, suggesting a low data complexity (in prediction terms). The complexity the *Breast*, *Colon* and *Prostate* seems higher, permitting to observe clear differences between the two approaches. The prediction accuracy obtained with the GA methodology was in almost all cases higher than the obtained within the SFS approach. Additionally, the robustness of the selected features is considerably higher in the GA (lower standard deviation

**Table 4 Performance comparison of classification techniques**

	Classifier	Parameters	GA		SFS	
			mean ± std	#genes	mean ± std	#genes
<b>Leukemia</b>	LDA	-	99.959 ± 0.07	12	97.609 ± 2.86	2
	SVM	{polynomial,15,1,0.6,0}	99.982 ± 0.06	8	<b>99.918 ± 0.52</b>	4
	NaiveBayes	{1,0}	99.974 ± 0.03	12	98.060 ± 2.19	3
	C-MANTEC	{1000,0.01,4.5}	99.038 ± 0.25	7	98.837 ± 2.46	3
	kNN	{1,Euclidean}	<b>99.994 ± 0.02</b>	10	99.844 ± 0.77	5
	MLP	{3,0.5,50}	99.944 ± 0.05	5	95.784 ± 3.38	2
<b>Lung</b>	LDA	-	99.971 ± 0.03	5	99.057 ± 1.00	3
	SVM	{linear,10,-,-,-}	<b>100 ± 0</b>	11	99.828 ± 0.70	3
	NaiveBayes	{1,0}	99.998 ± 0.01	4	<b>99.991 ± 0.07</b>	3
	C-MANTEC	{100000,0.25,2}	99.678 ± 0.08	6	99.673 ± 0.94	2
	kNN	{1,Euclidean}	99.969 ± 0.02	4	99.969 ± 0.22	4
	MLP	{4,0.1,50}	99.996 ± 0.01	4	99.778 ± 0.79	2
<b>Colon</b>	LDA	-	98.676 ± 0.35	11	87.179 ± 6.15	2
	SVM	{polynomial,1,1,0.4,2}	89.917 ± 1.26	20	91.738 ± 5.21	5
	NaiveBayes	{0,1}	90.583 ± 0.49	15	89.076 ± 7.79	4
	C-MANTEC	{10000,0.01,1}	94.315 ± 0.48	11	87.593 ± 6.69	2
	kNN	{3,cosine-similarity}	95.060 ± 0.38	19	<b>93.577 ± 4.43</b>	6
	MLP	{5,0.3,50}	<b>99.026 ± 0.30</b>	12	88.733 ± 5.51	2
<b>Breast</b>	LDA	-	99.788 ± 0.12	15	74.169 ± 6.52	1
	SVM	{polynomial,7,2,0.001,2}	99.744 ± 0.14	31	<b>81.029 ± 5.80</b>	3
	NaiveBayes	{0,0}	97.759 ± 0.23	27	73.499 ± 6.34	1
	C-MANTEC	{10000,0.01,1.5}	97.342 ± 0.39	23	76.645 ± 6.53	1
	kNN	{3,Euclidean}	97.485 ± 0.30	34	80.975 ± 6.37	2
	MLP	{4,0.3,50}	<b>99.828 ± 0.09</b>	18	79.191 ± 6.43	2
<b>Ovarian</b>	LDA	-	99.980 ± 0.01	4	<b>100 ± 0</b>	3
	SVM	{polynomial,9,1,0.2,0}	<b>100 ± 0</b>	4	99.978 ± 0.13	4
	NaiveBayes	{1,0}	99.951 ± 0.03	5	99.980 ± 0.13	4
	C-MANTEC	{1000,0.3,1.5}	99.844 ± 0.05	4	99.659 ± 0.75	3
	kNN	{1,Euclidean}	99.984 ± 0.01	4	99.982 ± 0.11	3
	MLP	{5,0.3,50}	99.999 ± 0	3	<b>100 ± 0</b>	3
<b>Prostate</b>	LDA	-	99.720 ± 0.12	9	95.677 ± 2.81	4
	SVM	{polynomial,5,1,3,1}	99.428 ± 0.31	20	<b>98.622 ± 1.79</b>	5
	NaiveBayes	{0,0}	98.817 ± 0.16	14	98.331 ± 2.13	7
	C-MANTEC	{1000,0.25,4}	98.681 ± 0.24	8	95.351 ± 3.40	4
	kNN	{3,cosine-similarity}	99.633 ± 0.11	20	97.146 ± 2.28	6
	MLP	{3,0.5,50}	<b>99.996 ± 0.02</b>	12	96.921 ± 2.37	4

Performance comparison among the two different feature selection frameworks used (GA and SFS) and the six classifiers analyzed (LDA, SVM, NaiveBayes, C-MANTEC, kNN and MLP) for each cancer microarray dataset. The results correspond to the best simulation for each dataset, showing the accuracy for method in the format of *mean ± standard deviation* and the number of selected genes.

values), fact that can be partially attributed to the larger set of genes selected. Regarding the computational complexity of both approaches, the SFS strategy involves approximately a number of comparisons of  $n_{sel} \times \overline{\#genes}$  ( $n_{sel}$ : number of pre-selected features,  $\overline{\#genes}$ : mean number of genes selected), while the GA approach utilises a maximum of 20.000 profile comparisons regardless of the dataset (length of the chromosome (100)  $\times$  number of generations (200)). For example, for the *Prostate* dataset in the SFS approach, approximately 3000 comparison are needed in the present study since  $n_{sel} \approx 600$ ,  $\overline{\#genes} = 5$ , unlike the genetic proposal which requires a greater number of combinations. However, if the number of pre-selected genes increases, the SFS method begins to loose its efficiency and may be intractable when handling thousands of genes.

Table 5 shows average results across all six datasets for the both frameworks used, noting that C-MANTEC lead to competitive classification performance with a reduced number of genes.

Further, we analyzed the differences between classifiers for the SFS and GA feature selection procedures used and for the six datasets, showing the results in Table 6. The corresponding p-value obtained after applying a Friedman's test is indicated in the third column [28]. In case this p-value is lower than 0.05, the lowest performant classifier is taken as a control group and the last column of the table lists the classifiers that lead to statistically significant results (from the lowest to the highest difference);

**Table 5 Performance comparison of feature selection frameworks**

Classifier	GA		SFS	
	mean $\pm$ std	#genes	mean $\pm$ std	#genes
LDA	99.682 $\pm$ 0.12	9.33	92.282 $\pm$ 3.22	2.5
SVM	99.082 $\pm$ 0.25	15.67	95.185 $\pm$ 2.36	4
NaiveBayes	97.847 $\pm$ 0.16	12.83	93.156 $\pm$ 3.11	3.67
C-MANTEC	98.150 $\pm$ 0.25	9.83	92.960 $\pm$ 3.46	2.5
kNN	98.688 $\pm$ 0.14	15.17	95.249 $\pm$ 2.36	4.33
MLP	99.798 $\pm$ 0.08	9	93.401 $\pm$ 3.08	2.5

Average performance comparison among two different feature selection frameworks (GA and SFS) and six classifiers (LDA, SVM, NaiveBayes, C-MANTEC, kNN and MLP) over all dataset.

**Table 6 Differences between classifiers.**

FS procedure	Dataset	p-value	Control	Statistically different classifiers
SFS	Leukemia	$< e^{-16}$	LDA	SVM
	Lung	$< e^{-16}$	LDA	kNN, NB
	Colon	$< e^{-16}$	LDA	SVM, kNN
	Breast	$< e^{-16}$	NB	kNN, SVM
	Ovarian	$< e^{-16}$	CM	LDA, NN
	Prostate	$< e^{-16}$	CM	NB, SVM
GA	Leukemia	$< e^{-16}$	CM	NB, NN, LDA, SVM, kNN
	Lung	$< e^{-16}$	CM	SVM, NB, NN
	Colon	$< e^{-16}$	SVM	LDA, NN
	Breast	$< e^{-16}$	SVM	NN, LDA
	Ovarian	$< e^{-16}$	CM	SVM, NN
	Prostate	$< e^{-16}$	CM	NN, LDA

Differences between classifiers for the two feature selection (FS) procedures used (first column). The lowest performance classifier is taken as control group and the last column of the table lists the classifiers that lead to statistically significant results (corresponding p-value indicated in the third column).

otherwise, non statistically significant results are reached (represented with a “-” on the table).

Table 7 shows a similar comparative analysis but among the SFS and GA feature selection procedures when a common classifier is used (first column of the table).

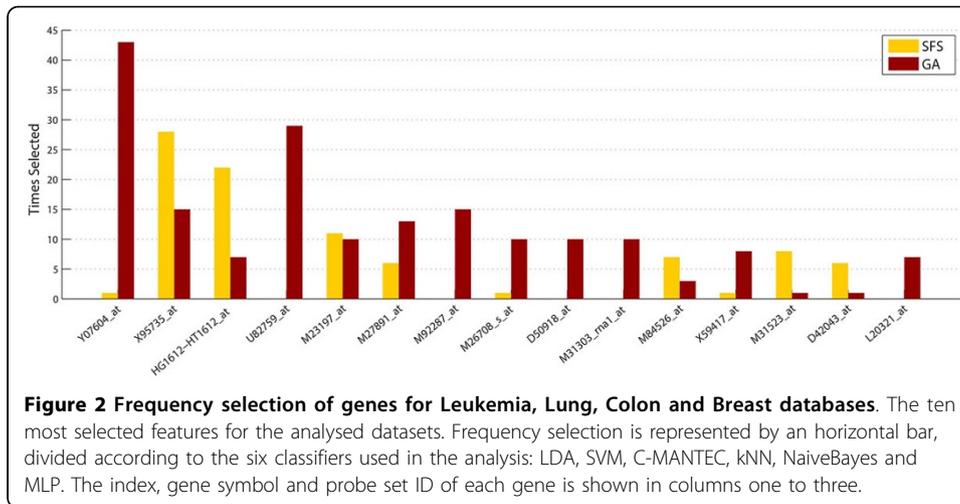
### Biological analysis

Figures 2 and 3 present the ten most selected genes for each of the six datasets considered, where each dataset is represented in a row of the table. The first three columns show information about the gene, such as the internal index (ID), the gene symbol (name of the gene) and the probe set ID, which is related to the chip where the dataset

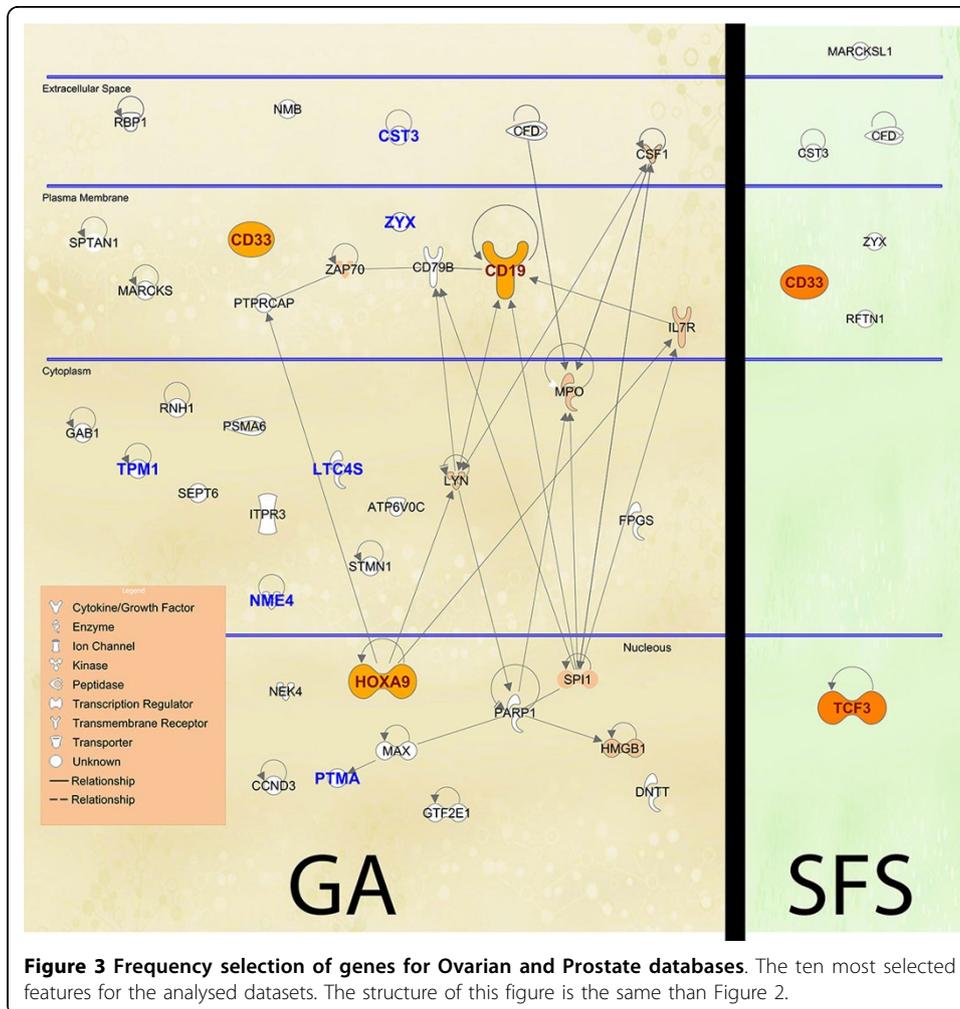
**Table 7 Differences between feature selection algorithms**

Classifier	Dataset	p-value	Control	Statistically different FS procedures
LDA	Leukemia	$1.54e^{-12}$	SFS	GA
	Lung	$1.54e^{-12}$	SFS	GA
	Colon	$1.54e^{-12}$	SFS	GA
	Breast	$1.54e^{-12}$	SFS	GA
	Ovarian	$3.28e^{-11}$	GA	SFS
	Prostate	$1.54e^{-12}$	SFS	GA
SVM	Leukemia	$3.65e^{-5}$	SFS	GA
	Lung	$1.54e^{-12}$	SFS	GA
	Colon	$2.86e^{-9}$	GA	SFS
	Breast	$1.54e^{-12}$	SFS	GA
	Ovarian	$9.13e^{-11}$	SFS	GA
	Prostate	$1.54e^{-12}$	SFS	GA
NB	Leukemia	$4.71e^{-9}$	SFS	GA
	Lung	$1.54e^{-12}$	SFS	GA
	Colon	$1.54e^{-12}$	SFS	GA
	Breast	$1.54e^{-12}$	SFS	GA
	Ovarian	0.157	-	-
	Prostate	$1.54e^{-12}$	SFS	GA
CM	Leukemia	$4.71e^{-9}$	SFS	GA
	Lung	$1.54e^{-12}$	SFS	GA
	Colon	$1.54e^{-12}$	SFS	GA
	Breast	$1.54e^{-12}$	SFS	GA
	Ovarian	0.157	-	-
	Prostate	$1.54e^{-12}$	SFS	GA
kNN	Leukemia	$1.54e^{-12}$	SFS	GA
	Lung	0.0897	-	-
	Colon	$1.54e^{-12}$	SFS	GA
	Breast	$1.54e^{-12}$	SFS	GA
	Ovarian	0.6547	-	-
	Prostate	$1.54e^{-12}$	SFS	GA
NN	Leukemia	$4.71e^{-9}$	SFS	GA
	Lung	$1.54e^{-12}$	SFS	GA
	Colon	$1.54e^{-12}$	SFS	GA
	Breast	$1.54e^{-12}$	SFS	GA
	Ovarian	0.157	-	-
	Prostate	$1.54e^{-12}$	SFS	GA

Differences between SFS and GA feature selection algorithms for the six different classification methods used (first column). The lowest performant FS procedure is taken as control group (fourth column) while the last column of the table lists the procedures that lead to statistically significant results (corresponding p-value indicated in the third column)



has been extracted (e.g., Affymetrix). The bar graph of the last column splits the frequency of selection (fourth column) of each feature according to the GA-LDA, GA-SVM, GA-CMANTEC, GA-kNN, GA-NaiveBayes and GA-MLP strategies. Most of the gene symbols have been found from their probe set ID by using tools as IPA (Ingenuity®



Systems, <http://www.ingenuity.com>) or NCBI (<http://www.ncbi.nlm.nih.gov/gene/>), although it has not been possible for the *Ovarian* dataset (first row of Figure 3) because there is no reference of the chip from which the data have been extracted.

A higher frequency of selection might imply a higher relevance of the gene in the prognosis of the disease. Those genes that are selected with similar frequency for all classifiers are considered independent with respect to the classification method. For instance, in the *Prostate* dataset (second row of Figure 3), the *MAF* gene is more significant than the *JUNB* gene, since it has been selected more times and all the classifiers selects it with the same frequency. Thus, *NaiveBayes* barely takes into account the *JUNB* gene whereas for MLP classifier it is one of the main genes.

Not only are we interested in getting good results in prognosis prediction but also in examining whether the selected genes provide biological information related to the disease studied. Therefore, if the proposed models provide this consistency between the computational and biological field, the results would be more confident and the selected genes would be more reliable from a clinical perspective, in order to their implementation in microchips and treatment in real patients. We can see that this statement is true in the proposed model using genetic algorithms.

In the case of the *Prostate* dataset is possible to find references in the literature where the genes *MAF*, which encodes a protein related to DNA-binding (most frequent gene, 99.67%) [29], *SERPINB5*, a serpin peptidase inhibitor (second most frequent, 58%) [30], *HPN*, officially named hepsin which encodes a type II transmembrane serine protease (fourth most frequent, 50%) [31] and *GSTP1*, belonging to the family of Glutathione S-transferases (GSTs) enzymes (sixth most frequent, 36.33%) [32] are biologically related to the absence or presence of prostate cancer. This supports the idea that our computational approach is robust and consistent with the results obtained in biological studies.

For the *Breast* dataset, several of the most selected genes among which are *UBC* [33,34], *ZNF222* [35] and *EWSR1* [36], are biologically associated with breast cancer. The same happen for the *Leukemia* disease, where the enforced expression of the *CD19* molecule (fifth selected, 19%) can reduce the proliferation of the malignant plasma cells [37]; the gene homeobox A9 (*HOXA9*, second selected, 33%) influences hematopoietic progenitors and acute leukemias [38]; and the *CD33* molecule (seventh selected, 17.33%) has been shown to sharply inhibit the in vitro proliferation of both normal myeloid cells and chronic myeloid leukemias [39].

From a computational point of view, Table 8 shows the best selected genes obtained by the genetic approach which also have been extracted in several related papers (last column of the table) for the particular case of the *Leukemia* dataset. It should be noted that the applied methodology is different from one paper to another. For instance, five of the ten genes are also reported in the list of the 50 most important genes (selected from 7129) suggested in [40].

Focusing on the *Leukemia* dataset (one of the most studied dataset in the literature), and as a biological analysis of the features selected, Figure 4 displays a comparison between the most selected genes, after 50 independent executions and with independence of the classifier used, for both GA and SFS selection procedures. Moreover, the IPA tool is used in order to explore the functional involvement of each gene set obtained by GA and SFS in the studied disease. In concrete, three of the fifteen most frequently genes are highlighted in bold on the *x*-axis in Figure 4 as

**Table 8 Selected genes for the Leukemia dataset**

ID	Probe Set ID	Gene Description	References
4951	<i>Y07604_at</i>	NME/NM23 nucleoside diphosphate kinase 4	[41-43]
3847	<i>U82759_at</i>	Homeo box A9	[40,44,43]
6169	<i>M13690_s_at</i>	C1NH Complement component 1 inhibitor	[43,45]
6184	<i>M26708_s_at</i>	PTMA Prothymosin alpha	[45]
6225	<i>M84371_rna1_s_at</i>	CD19 Molecule	[46]
1882	<i>M27891_at</i>	CST3 Cystatin C	[40,41,43,44]
1834	<i>M23197_at</i>	CD33 antigen	[40,44,46]
4847	<i>X95735_at</i>	Zyxin	[40,41,44,46]
3320	<i>U50136_rna1_at</i>	LTC4 synthase	[40,43,44,46]
5094	<i>Z24727_at</i>	TPM1 Tropomyosin alpha chain	[47]

The best selected genes ranked with the GA approach for the *Leukemia* dataset which also appear in other studies in the literature.

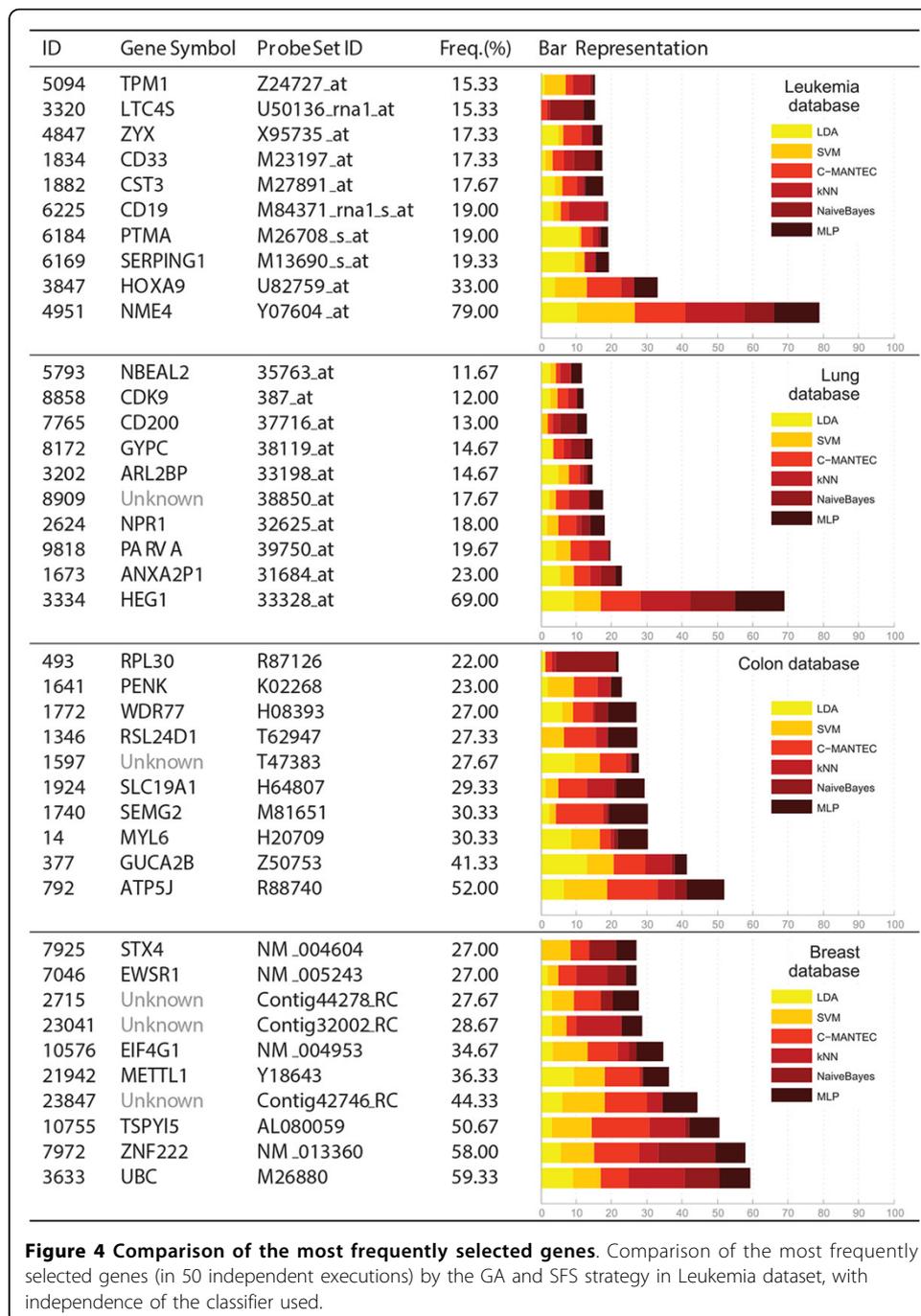
founded genes in the IPA database with biological relevance on the Leukemia cancer disease.

A deeper biological analysis is performed using the IPA tool for the GA-CMANTEC strategy considering the *Leukemia* dataset. Figure 5 shows those genes that are selected at least a 5% of the times both with GA-CMANTEC or SFS-CMANTEC strategy after 50 independent executions. The names shown on this figure correspond to the symbol of each gene according to Figure 2. It is important to highlight the difference on the number of genes selected through the GA and SFS strategy due to the casuistic of each algorithm. Additionally, on the left side are represented in bold nine of the ten most frequently selected genes with independence of the classifier used. Moreover, using C-MANTEC as classifier allow to obtain these nine most selected genes. Finally, filled in genes represent those genes that have demonstrate biological relevance on the Leukemia disease. In this sense, the GA-CMANTEC strategy presents 10 out of 37 genes as a result while the SFS-CMANTEC strategy presents 2 out of 7. Although these results are similar in proportion, the GA-CMANTEC strategy could be considered more explicative from a biological point view with no detriment on the classification performance. Furthermore, the connections among the selected genes (represented by links in Figure 5), which are more numerous in the GA approach, suggest as well a significant relationship with the occurrence of the disease.

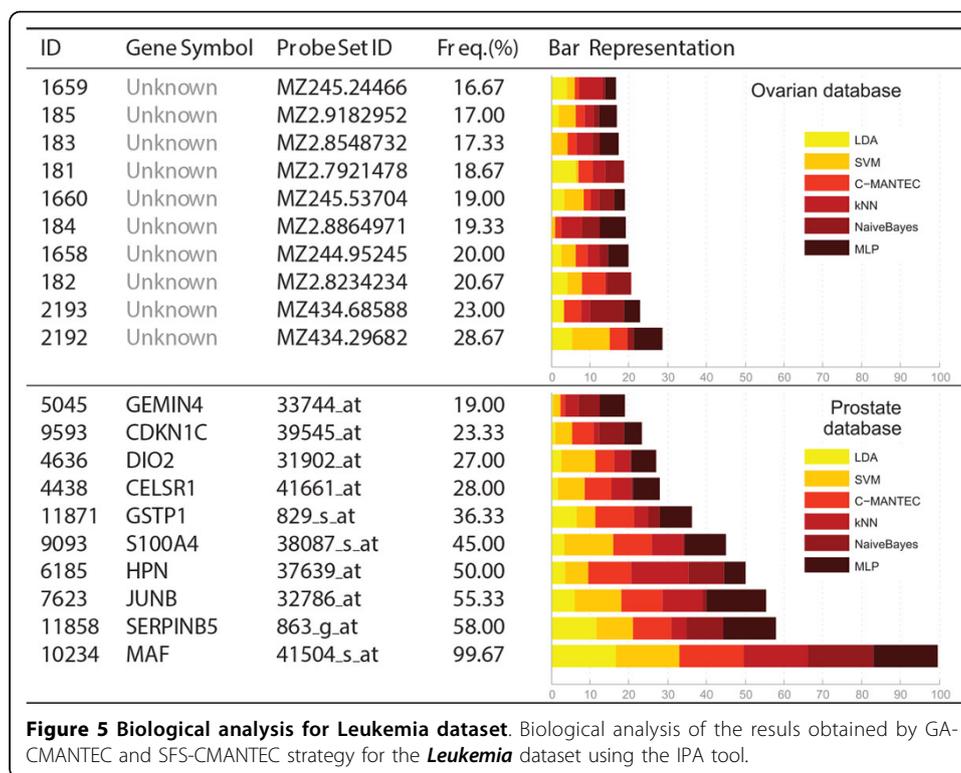
## Conclusions

In this work, a new methodology approach combining genetic algorithm with constructive neural networks has been proposed in order to predict cancer outcome. For six free-public cancer datasets, we compared under GA and SFS frameworks the prediction accuracy of the C-MANTEC algorithm against the following five standard classifiers: LDA, SVM, NaiveBayes, kNN or MLP.

On average, the strategy based on the GA approach leads to better prediction rates, observing that these results are independent of the classifier used, noting also that prediction results under the GA framework show lower variability, and thus can be considered as more robust. On the other hand, it should be noted that the SFS approach is less computationally intensive, involving in the present study approximately seven times less gene comparisons, and also leading to a group of selected genes much smaller than those selected under the GA approach. Nevertheless, when complex datasets



are studied like *Breast* or *Colon*, cancer prognosis results are quite poor when using the SFS approach, presumably since the search in the state space is much more restrictive. Additionally, an analysis done using the IPA methodology suggests that the biological relevance of the genes selected under the GA framework is higher than the observed using the SFS approach, as indicated by the reference frequency in the literature and also regarding the relationship between them (even if this effect might be due to the size of both selected sets).



Regarding the comparison between the different classifiers implemented, standard feed-forward neural networks (MLP), LDA and SVM lead to similar and best results while C-MANTEC and kNN followed closely but with a bit lower accuracy. C-MANTEC, MLP and LDA permitted to obtain a more reduced set of genes in comparison to SVM, NB and kNN. Further, C-MANTEC resulted in the most robust classifier in terms of changes in the parameter settings, a relevant feature for its use in wrapper feature selection methods (as it will reduce execution times related to parameter tuning). Additionally, we are considering the use of an ensemble of all these classifiers as a further work, in order to obtain a greater consensus on the classification result, which could lead to greater robustness and accuracy of the resulting model.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

RML, DU and JMJ contributed to the conception and design of the study. RML, DU and JLS contributed to write the programming code. RML, DU, JMJ and LF contributed to the analysis and interpretation of the data, and RML, DU, LF and JMJ contributed to the drafting of the manuscript. All authors approved the manuscript.

#### Declarations

Publication funding for this article has come from grants TIN2010-16556 (MICINN-Spain) and TIC-4026/2008 (Consejería de Innovación, Ciencia y Empresa. Junta de Andalucía), both including FEDER funds. Additionally, the authors acknowledge support through these grants.

This article has been published as part of *Theoretical Biology and Medical Modelling* Volume 11 Supplement 1, 2014: Selected articles from the 1st International Work-Conference on Bioinformatics and Biomedical Engineering-IWBBIO 2013. The full contents of the supplement are available online at <http://www.tbiomed.com/supplements/11/S1>.

#### Authors' details

<sup>1</sup>Department of Computer Science, University of Málaga, Málaga, Spain. <sup>2</sup>Biomedical Research Institute of Málaga (IBIMA), Spain.

Published: 7 May 2014

## References

1. Wei JS, Greer BT, Westermann F, Steinberg SM, Son CG: **Prediction of Clinical Outcome Using Gene Expression Profiling and Artificial Neural Networks for Patients with Neuroblastom.** *Cancer Research* 2004, **64**:6883-6891.
2. Pellagatti A, Vetrie D, Langford CF, Gama S, Eagleton H, Wainscoat JS, Boulwood J: **Gene Expression Profiling in Polycythemia Vera Using cDNA Microarray Technology.** *Cancer Research* 2003, **63**:3940-3944.
3. West M: **Bayesian factor regression models in the "large p, small n" paradigm.** *Bayesian statistics* 2003, **7**(2003):723-732.
4. Ransohoff D: **Rules of evidence for cancer molecular-marker discovery and validation.** *Nature Reviews Cancer* 2004, **4**(4):309-314.
5. Lancashire LJ, Rees RC, Ball GR: **Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach.** *Artificial Intelligence In Medicine* 2008, **43**(2):99-111.
6. Peng H, Fu Y, Liu J, Fang X, Jiang C: **Optimal gene subset selection using the modified SFFS algorithm for tumor classification.** *Neural Computing and Applications* 2012, 1-8.
7. Raymer M, Punch W, Goodman E, Kuhn L, Jain A: **Dimensionality reduction using genetic algorithms.** *IEEE Transactions on Evolutionary Computation* 2000, **4**(2):164-171.
8. Chiang Y, Chiang H, Lin S: **The application of ant colony optimization for gene selection in microarray-based cancer classification.** *Proceedings of the 7th International Conference on Machine Learning and Cybernetics, ICMLC* 2008, 7:4001-4006.
9. Sun Z, Bebis G, Miller R: **Object detection using feature subset selection.** *Pattern Recognition* 2004, **37**(11):2165-2176.
10. McLachlan G, Bean R, Peel D: **A mixture model-based approach to the clustering of microarray expression data.** *Bioinformatics* 2002, **18**(3):413-422.
11. Molinaro A, Simon R, Pfeiffer R: **Prediction error estimation: a comparison of resampling methods.** *Bioinformatics* 2005, **21**(15):3301-3307.
12. Zhang F, Kaufman HL, Deng Y, Drabier R: **Recursive SVM biomarker selection for early detection of breast cancer in peripheral blood.** *BMC Medical Genomics* 2013, **6**(1).
13. Tong DL, Schierz AC: **Hybrid genetic algorithm-neural network: Feature extraction for unprocessed microarray data.** *Artificial Intelligence in Medicine* 2011, 53:47-56.
14. Su Y, Wang R, Li C, Chen P: **A dynamic subspace learning method for tumor classification using microarray gene expression data.** *Proceedings - 2011 7th International Conference on Natural Computation, ICNC 2011* 2011, 1:396-400.
15. Student S, Fajarewicz K: **Stable feature selection and classification algorithms for multiclass microarray data.** *Biology Direct* 2012, **7**.
16. Werner T: **Bioinformatics applications for pathway analysis of microarray data.** *Current Opinion in Biotechnology* 2008, **19**:50-54.
17. Subirats JL, Franco L, Jerez JM: **C-Mantec: A novel constructive neural network algorithm incorporating competition between neurons.** *Neural Networks* 2012, **26**:130-140.
18. Saeyns Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, **23**(19):2507-2517.
19. Huerta EB, Duval B, Hao JK: **A hybrid LDA and genetic algorithm for gene selection and classification of microarray data.** *Neurocomputing* 2010, **73**:2375-2383.
20. Welch BL: **The generalization of Student's problem when several different population variances are involved.** *Biometrika* 1947, **34**(1-2):28-35.
21. Webb AR: *Statistical Pattern Recognition*. 2 edition. John Wiley & Sons; 2011, third edition.
22. Peng H, Long F, Ding C: **Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2005, **27**(8):1226-1238.
23. Guo B, Nixon M: **Gait Feature Subset Selection by Mutual Information.** *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 2009, **39**:36-46.
24. Moddemeijer R: **On Estimation of Entropy and Mutual Information of Continuous Distributions.** *Signal Processing* 1989, **16**(3):233-246.
25. Frean M: **A "thermal" perceptron learning rule.** *Neural Comput.* 1992, **4**(6):946-957.
26. García-Pedrajas N, Ortiz-Boyer D: **A cooperative constructive method for neural networks for pattern recognition.** *Pattern Recogn.* 2007, **40**:80-98.
27. Subirats JL, Jerez JM, Franco L: **A New Decomposition Algorithm for Threshold Synthesis and Generalization of Boolean Functions.** *IEEE Transactions on Circuits and Systems* 2008, **1**(55):3188-3196.
28. García S, Fernández A, Luengo J, Herrera F: **Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power.** *Information Sciences* 2010, **180**(10):2044-2064.
29. Steele VE, Arnold JT, Le H, Izmirlian G, Blackman MR: **Comparative Effects of DHEA and DHT on Gene Expression in Human LNCaP Prostate Cancer Cells.** *Anticancer Research* 2006, **26**(5A):3205-3215.
30. Sheng S, Carey J, Sefter EA, Dias L, Hendrix MJ, Sager R: **Maspain acts at the cell membrane to inhibit invasion and motility of mammary and prostatic cancer cells.** *Proc Natl Acad Sci USA* 1996, **93**(21):11669-74.
31. Srikantan V, Valladares M, Rhim JS, Srivastava S: **HEPSIN Inhibits Cell Growth/Invasion in Prostate Cancer Cells.** *Cancer Research* 2002, **62**(23):6812-6816.
32. Lin X, Tascilar M, Lee WH, Vles WJ, Lee BH, Veeraswamy R, Asgari K, Freije D, van Rees B, Gage WR, Bova GS, Isaacs WB, Brooks JD, DeWeese TL, Marzo AMD, Nelson WG: **{GSTP1} CpG Island Hypermethylation Is Responsible for the Absence of {GSTP1} Expression in Human Prostate Cancer Cells.** *The American Journal of Pathology* 2001, **159**(5):1815-1826.

33. Ramachandran C, Rodriguez S, Ramachandran R, Nair PR, Fonseca H, Khatib Z, Escalon E, Melnick SJ: **Expression Profiles of Apoptotic Genes Induced by Curcumin in Human Breast Cancer and Mammary Epithelial Cell Lines.** *Anticancer Research* 2005, **25**(5):3293-3302.
34. Kroll T, Odyvanova L, Clement J, Platzer C, Naumann A, Marr N, Höffken K, Wölfl S: **Molecular characterization of breast cancer cell lines by expression profiling.** *Journal of Cancer Research and Clinical Oncology* 2002, **128**(3):125-134.
35. Klein A, Olendrowitz C, Schmutzler R, Hampl J, Schlag PM, Maass N, Arnold N, Wessel R, Ramser J, Meindl A, Scherneck S, Seitz S: **Identification of brain-and bone-specific breast cancer metastasis genes.** *Cancer Letters* 2009, **276**(2):212-220.
36. Menon R, Omenn GS: **Proteomic Characterization of Novel Alternative Splice Variant Proteins in Human Epidermal Growth Factor Receptor 2 neu Induced Breast Cancers.** *Cancer Research* 2010, **70**(9):3440-3449.
37. Mahmoud MS, Fujii R, Ishikawa H, Kawano MM: **Enforced CD19 Expression Leads to Growth Inhibition and Reduced Tumorigenicity.** *Blood* 1999, **94**(10):3551-3558.
38. Hu YL, Fong S, Ferrell C, Largman C, Shen WF: **HOXA9 Modulates Its Oncogenic Partner Meis1 To Influence Normal Hematopoiesis.** *Molecular and Cellular Biology* 2009, **29**(18):5181-5192.
39. Vitale C, Romagnani C, Puccetti A, Olive D, Costello R, Chiossone L, Pitto A, Bacigalupo A, Moretta L, Mingari MC: **Surface expression and function of p75/AIRM-1 or CD33 in acute myeloid leukemias: Engagement of CD33 induces apoptosis of leukemic cells.** *Proceedings of the National Academy of Sciences* 2001, **98**(10):5764-5769.
40. Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E: **Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**(5439):531-537.
41. Yang P, Zhou BB, Zhang Z, Zomaya AY: **A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data.** *BMC Bioinformatics* 2010, **11**(1).
42. García-Nieto J, Alba E, Jourdan L, Talbi E: **Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis.** *Information Processing Letters* 2009, **109**(16):887-896.
43. Krishnapuram B, Carin L, Hartemink A: **Joint classifier and feature optimization for comprehensive cancer diagnosis using gene expression data.** *Journal of Computational Biology* 2004, **11**(2-3):227-242.
44. Chen Z, Li J, Wei L: **A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue.** *Artificial Intelligence in Medicine* 2007, **41**(2):161-175.
45. Shen Q, Shi WM, Kong W: **Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data.** *Computational Biology and Chemistry* 2008, **32**:53-60.
46. Momin B, Mitra S, Gupta R: **Reduct Generation and Classification of Gene Expression Data.** *Hybrid Information Technology, 2006 ICHIT '06 International Conference on* 2006, 1:699-708.
47. Gwinn M, Keshava C, Olivero O, Hums J, Poirier M, Weston A: **Transcriptional signatures of normal human mammary epithelial cells in response to benzo[a]pyrene exposure: a comparison of three microarray platforms.** *OMICS* 2005, **9**(4):334-50.

doi:10.1186/1742-4682-11-S1-S7

**Cite this article as:** Luque-Baena et al.: Application of genetic algorithms and constructive neural networks for the analysis of microarray cancer data. *Theoretical Biology and Medical Modelling* 2014 **11**(Suppl 1):S7.

**Submit your next manuscript to BioMed Central  
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

