Selection of biologically relevant genes with a wrapper stochastic algorithm

Kim-Anh Lê Cao, Olivier Gonçalves, Philippe Besse, Sébastien Gadat

Statistical Applications in Genetics and Molecular Biology

Volume 6, *Issue* 1 2007 *Article* 29

Selection of Biologically Relevant Genes with a Wrapper Stochastic Algorithm

Kim-Anh Lê Cao* Olivier Gonçalves†

Philippe Besse‡ Sébastien Gadat**

*Université de Toulouse, CNRS (UMR 5219) and INRA, k.lecao@imb.uq.edu.au

†LBP UMR CNRS 6023, Blaise Pascal University, olivier.goncalves@iut.u-clermont1.fr

‡Université de Toulouse, CNRS (UMR 5219), besse@math.ups-tlse.fr

**Université de Toulouse, CNRS (UMR 5219), sebastien.gadat@math.univ-toulouse.fr

Selection of Biologically Relevant Genes with a Wrapper Stochastic Algorithm*

Kim-Anh Lê Cao, Olivier Gonçalves, Philippe Besse, and Sébastien Gadat

Abstract

We investigate an important issue of a meta-algorithm for selecting variables in the framework of microarray data. This wrapper method starts from any classification algorithm and weights each variable (i.e. gene) relative to its efficiency for classification. An optimization procedure is then inferred which exhibits important genes for the studied biological process.

Theory and application with the SVM classifier were presented in Gadat and Younes, 2007 and we extend this method with CART. The classification error rates are computed on three famous public databases (Leukemia, Colon and Prostate) and compared with those from other wrapper methods (RFE, lo norm SVM, Random Forests). This allows the assessment of the statistical relevance of the proposed algorithm. Furthermore, a biological interpretation with the Ingenuity Pathway Analysis software outputs clearly shows that the gene selections from the different wrapper methods raise very relevant biological information, compared to a classical filter gene selection with T-test.

KEYWORDS: gene selection, classification, stochastic algorithm, cancer databases

*We are grateful to ACI IMPBio (ENV-STAT-EXP) who supported this research. We would also like to thank "Projet Calcul en MIdi-Pyrenées" (CALMIP) for the intensive computations and the anonymous reviewers for their helpful comments on the manuscript.

1 Introduction

Performing a feature selection algorithm has several important applications in the field of microarray data analysis. First, to determine which genes contribute the most for the biological outcome (*e.g.* cancerous *vs.* normal cells) and in which way they interact to determine this outcome. Second, to predict the outcome when a new observation is presented. It is unlikely that thousands of genes do explain the class membership of a microarray and it is hence wise to use a dimensional reduction technique. This also provides practical aspects with machine learning methods: it avoids the "curse of dimensionality" that leads to overfitting when the number of variables is too large.

Features can generally be selected with two different approaches: either explicitly (filter methods) or implicitly (wrapper methods). The aim of the filter methods is to measure the relevance of each gene. Variables are usually ordered with statistical tests and microarrays are classified with the few good-ranked selected variables. In this case, note that the selection is totally independent from the classification method (Dudoit et al., 2000 and Golub et al., 1999). The main advantages are robustness against overfitting and low cost computation, but these methods may fail to select the most "useful" features and usually disregard the interactions between the features. On the other hand, wrapper methods measure the usefulness of a set of features by exploring the subsets space. This search can be performed either with heuristic or stochastic techniques (*e.g.* simulated annealing, genetic algorithms). These methods find the "useful" variables, but are prone to overfit. Moreover, when dealing with numerous variables, an exhaustive subspace search is computationally untractable. They generally yield greedy and costly algorithms since each iteration consists in selecting smaller and smaller subsets of variables (Guyon et al., 2001, Diaz-Uriarte and Alvarez de Andrés, 2006).

These latter wrapper methods have been successfully applied on several benchmarks but suffer from lack of mathematical justification. Furthermore, they are all dedicated to one special baseline classifier that is used for constructing the decision rule. Gadat and Younes (2007) proposed a wrapper approach which does not depend on the classifier and can numerically quantify the efficiency of each gene. It uses stochastic approximations that still cover a large portion of the search space to avoid local minima. This reaches to subset selections of discriminative genes that hence hold useful information on the microarray experiment.

The two main objectives of this paper are first to numerically compare the performances of different wrapper methods by estimating the classification error rate with the e.632+ bootstrap method (Efron and Tibshirani, 1997)

and second, to provide a comparison of the different gene selections based on their biological relevance. Note that we do not intend to optimize the size of the gene subset. We rather focus on the biological interpretation of the 50 first selected genes.

The optimal feature weighting procedure (ofw) from Gadat and Younes (2007) was initially applied with the classifier Support Vector Machines (SVM: Vapnik, 2000). We investigate the application of another classifier, Classification and Regression Trees (CART) on public microarray data sets (*Leukemia*: Golub et al., 1999, *Colon*: Alon et al., 1999 and *Prostate*: Singh, D. et al., 2002). We compare the results from these two wrapper methods ofw+SVM or ofw+CART to those obtained with other well known procedures: Recursive Feature Elimination (RFE: Guyon et al., 2002), Random Forests, (RF: Breiman, 2001) and l_0 norm SVM (l_0: Weston et al., 2003), as well as the widely used T-statistics. The classification error rates are displayed for each public data set and the biological relevancies of the gene selections are discussed with the Ingenuity Pathways Analysis software.

2 Method

We introduce the optimal feature weighting meta-algorithm (ofw) from Gadat and Younes (2007) that treats several classification problems with a feature selection task. In this section we explain the main theoretical derivations that are necessary to fully understand the algorithm and its application.

2.1 Optimal Feature Weighting Model

The particularity of this algorithm is that it does not depend on the classification procedure $\mathbb{A}$ used for classification. We consider a large set of genes $\mathcal{G}$ of size N expressed on two biological conditions (or classes) $\{\mathcal{C}_1, \mathcal{C}_2\}$. $\mathcal{G}$ can be either the total number of genes spotted on the microarray or a rather large gene subset. These N genes describe a signal $\mathcal{I}$. The optimization of any given classification algorithm $\mathbb{A}$ (*e.g.* SVM, CART, Nearest Neighbors ...) is explored by passing through $\mathbb{A}$ different subspaces of genes to improve its performance with time.

System energy

Let us define a positive weight parameter $\mathbb{P}$ on each of the genes in $\mathcal{G}$. After a normalization step, we can consider $\mathbb{P}$ as a discrete probability on the N

DOI: 10.2202/1544-6115.1312

genes. The goal is to learn a probability that fits the efficiency of each gene for the classification of $\mathcal{I}$ in $\{\mathcal{C}_1, \mathcal{C}_2\}$, so that important weights are given to genes with high discriminative power and lower weights to those that have a poor influence on the classification task.

Denote p any small integer compared to N (*e.g.* $p = 2\% * N$), a gene subset of size p has to be extracted from $\mathcal{G}$. Next definition properly establishes how to measure the goodness of $\mathbb{P}$ for the set of genes $\mathcal{G}$ and the two classes $\{\mathcal{C}_1, \mathcal{C}_2\}$.

Definition of system energy:
Given a probability $\mathbb{P}$ on $\mathcal{G}$ and $\epsilon(\omega)$ the measure of classification efficiency with any p-uple $\omega \subset \mathcal{G}^p$, the energy of the system at the point $\mathbb{P}$ is defined as the mean classification performance when ω is drawn with respect to $\mathbb{P}^{\otimes p}$ (with replacement) in $\mathcal{G}^p$, that is:

$$\mathcal{E}(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[\epsilon] = \sum_{\omega \subset \mathcal{G}^p} \mathbb{P}(w)\epsilon(w) \tag{1}$$

Note here that the energy $\mathcal{E}$ depends on the way we measure the classification efficiency on ω, denoted $\epsilon(\omega)$ all along this paper. Given any standard classification algorithm $\mathbb{A}$, $\epsilon(\omega)$ will be the error of $\mathbb{A}$ computed on the training set using the set of extracted features ω. For instance, if $\mathbb{A}$ is a SVM with a linear kernel, $\epsilon(\omega)$ will be the classification error of a linear SVM using only genes in ω to describe the signal in the training set.
The computation of the sum (1) is untractable since one cannot enumerate all subsets ω of $\mathcal{G}^p$, but we will provide a stochastic algorithm to optimize $\mathcal{E}$ in next section.

Remark The more $\mathbb{P}$ enables to hold a discriminative gene g for classification (important weight on g and $\epsilon(\omega)$ small each time ω contains this gene g), the less $\mathcal{E}$. Minimizing $\mathcal{E}$ with respect to $\mathbb{P}$ will thus permit to exhibit the most weighted and thus the most discriminative genes. Hence, a natural measure of variable importance ranking will be read on the weight distribution $\mathbb{P}^\star$ minimizing $\mathcal{E}$.

2.2 Stochastic optimization method

This part provides an efficient way to minimize the energy $\mathcal{E}$ with a stochastic version of the standard gradient descent technique.

Remark first that the function $\mathcal{E}$ has to be minimized up to the constraints defined by a discrete probability measure on $\mathcal{G}$. Thus, the most natural way

to optimize (1) is to use a gradient descent of $\mathcal{E}$ projected on the set of constraints. This leads to the next definitions.

DEFINITIONS:
We define the set $\mathcal{S}$ as the simplex of probability map on $\mathcal{G}$. We also denote by $\Pi_{\mathcal{S}}$ the affine projection of any point of $\mathbb{R}^N$ on the simplex $\mathcal{S}$. This natural projection $\Pi_{\mathcal{S}}$ of any point x can be computed in a finite number of steps as mentioned in Gadat and Younes (2007).
The Euclidean gradient of $\mathcal{E}$ is:

$$\forall g \in \mathcal{G} \qquad \nabla\mathcal{E}(\mathbb{P})(g) = \sum_{\omega \subset \mathcal{G}^p} \frac{C(\omega, g)\mathbb{P}(\omega)}{\mathbb{P}(g)}\epsilon(\omega) \tag{2}$$

where $C(\omega, g)$ is the number of occurrences of g in ω. The iterative procedure to update $\mathbb{P}$ is then given by

$$\mathbb{P}_{t+dt} = \mathbb{P}_t - \nabla\mathbb{P}_t dt \tag{3}$$

Of course (2) is numerically impossible to calculate, as one cannot enumerate all possible ω in $\mathcal{G}^p$ and a stochastic approximation is needed: the Euclidian gradient expression (2) can actually be seen as an expectation. Then, to deal with such gradient, a computable Robbins-Monro algorithm can be used, which gets similar asymptotic behavior as (3) (see for instance Gadat and Younes (2007), Kushner and Clark (1978)). With this stochastic method, the updated formula of $\mathbb{P}_n$ becomes:

$$\mathbb{P}_{n+1} = \Pi_{\mathcal{S}}\left[\mathbb{P}_n - \alpha_n \frac{C(\omega_n, .)\epsilon(\omega_n)}{\mathbb{P}_n(.)}\right] \tag{4}$$

where ω_n is any set of p genes sampled with respect to $\mathbb{P}_n$, and $\alpha_n = K/(n+1)$ for any positive constant $K > 0$ is the step of the algorithm. Note that the last expression is always defined since when $\mathbb{P}_n(g) = 0$, we cannot draw this gene in ω_n and $C(\omega_n, g)$ vanishes.

Under mild conditions on the energy $\mathcal{E}$, one can show that this stochastic approximation algorithm converges to a critical point of $\mathcal{E}$. One can also prove the asymptotic normality result:

$$\frac{\mathbb{P}_n - \mathbb{P}_\infty}{\sqrt{\alpha_n}} \to \mathcal{N}(0, V),$$

where the covariance matrix V depends on the energy function $\mathcal{E}$. Further details can be found in (Benveniste et al., 1990).

DOI: 10.2202/1544-6115.1312

2.3 Detailed algorithm

Let $\mathcal{G} = (\delta_1 \dots \delta_{|\mathcal{G}|})$, $\mu \in \mathbb{N}^*$ and η the stopping criterion.

- For $n = 0$ define $\mathbb{P}_0$ as the uniform distribution on $\mathcal{G}$
- While $|\mathbb{P}_{(n+\mu)} - \mathbb{P}_n|_\infty > \eta$:
  - extract ω_n from $\mathcal{G}^p$ with respect to $\mathbb{P}_{n,p} = \mathbb{P}_n^{\otimes p}$
  - construct $\mathbb{A}_{\omega_n}$ and compute $\epsilon(\omega_n)$
  - compute the drift vector $d_n = C(\omega_n, \cdot)\epsilon(\omega_n)/\mathbb{P}_n(\cdot)$
  - update $\mathbb{P}_{n+1} = \Pi_{\mathcal{S}}[\mathbb{P}_n - \alpha_n d_n]$
  - $n = n + 1$

3 Application

We first provide a short description of the two supervised algorithms we apply ofw to: Support Vector Machines (SVM) and Classification And Regression Trees (CART). We next shortly describe other feature selection methods that we compare to our approach.

3.1 Two baseline classifiers are applied to ofw

Support Vector Machines

SVM (Vapnik, 2000) is a binary classifier that attempts to separate the microarrays into $\mathcal{C}_1$ and $\mathcal{C}_2$ by defining an optimal hyperplane between the 2 classes up to a consistency criterion. Linear kernel SVMs are used here because of their good generalization ability compared to more complex kernels.

Classification And Regression Trees

CART (Breiman et al., 1984) is a multi-category classifier that is constructed through a recursive partitioning routine. It builds a classification rule to predict the class label of the microarrays based on the feature information following the Gini criterion. To avoid overfitting, trees are then pruned using a cross validation procedure. Note that CART is naturally unstable: a slight change in the features can lead to a very different construction of the tree.

3.2 Comparisons with existing ranking methods

We briefly present here the several algorithms we performed to compare our OFW approach with. Each of these methods follows the classical framework of feature selection algorithm. A training set is used to compute the rank (or relevancy) of each feature (or gene) and the error of the obtained gene selection is then computed on a test set. Thus, the input of each of these algorithms is simply the training set in our case.

Recursive Feature Elimination

RFE (Guyon et al., 2002) is a feature selection technique exclusively dedicated to SVM. It consists in computing a ranking criterion for all features using the SVM previously computed. Genes with the smallest ranking criterion are then recursively removed (with more than one feature per step for speed reasons). The idea is to construct several stacked feature subsets $\mathcal{G}_m \subset \mathcal{G}_{m-1} \subset \ldots \subset \mathcal{G}_1 = \mathcal{G}$ and find $\mathcal{G}_m$ that is optimal (on the basis of error rates metrics) and that leads to the largest margin of class separation. In this paper we will only focus on the gene ranks that are output from this method and not on the optimal size of the subset so as to compare the different methods. Indeed, all the presented methods do not necessarily give a stopping criterion for an optimal selection size.

l_0 norm SVM

Weston et al. (2003) proposed to minimize the l_0 norm of the normal vector from SVM to provide a way of selecting features and to minimize the training error in one step. As the problem is NP-hard, an approximation of the l_0 norm is proposed. This feature selection method has rarely been used yet in the context of microarray.

Random Forests

RF is a CART aggregation technique. The idea of Breiman (2001) was to introduce two sources of randomness. First with bagging: each unpruned tree is constructed on a bootstrap sample. Second, for each partition building step of the tree, the best variable is chosen among a fixed number of randomly selected variables. Trees are then aggregated by majority vote. There is also an internal importance measure of the variables given by the forest that determines which predictors (*i.e* genes) are the most discriminative. Here we choose the "Mean Decrease Accuracy" measure that consists for each tree in

DOI: 10.2202/1544-6115.1312

randomly permuting the genes values that are not in the bootstrap sample (called "Out-Of-Bag" data) and computing the resulting classification error rate.

Diaz-Uriarte and Alvarez de Andrés (2006) proposed a backward feature selection procedure using RF that has not been applied here as the selection is often extremely small with no redundant genes.

Univariate filter method

One of the aim of this paper is to compare the gene selection using T-statistics to the ones resulting from the multivariate classification methods that were presented above. Note that the False Discovery Rate that controls the number of false positive genes was not applied here as we are selecting a fixed number of genes.

3.3 Public microarray data sets

We present the results obtained on three well known public data sets. *Leukemia* (Golub et al., 1999) compares two different types of leukemia (Acute Myeloid and Acute Lymphoplastic, ALL *vs.* AML) with 3860 genes and 72 microarrays. *Colon* (Alon et al., 1999) was obtained from cancerous or normal colon tissues with 2000 genes and 62 microarrays and *Prostate* (Singh, D. et al., 2002) also compared normal *vs.* cancerous prostate tissues with 102 microarrays and 12600 genes. These data sets will be refered as Leukemia, Colon and Prostate along this paper. We assumed the data sets correctly normalized.

3.4 Error rate assessment

We compared the error rates of all methods on each data set with the e.632+ bootstrap error estimate from (Efron and Tibshirani, 1997) that is adequate for small sample size data sets (Ambroise and MacLachlan, 2002) . The e.632 estimator is defined as $e.632 = .368R + .632B$ where R is the resubstitution error rate and B the ouf-of-bag bootstrap error rate. When the number of genes is much larger than the number of samples, the prediction rule usually overfits (R often equal 0). Efron and Tibshirani proposed the e.632+ estimate

$$e.632+ = (1-w)R + wB$$

with $w = \frac{.632}{1-.368r}$, $\quad r = \frac{B-R}{min(B,\gamma)-R}$, $\quad \gamma = \sum_{i=1}^{2} p_i(1-q_i)$ where r is an overfitting rate and γ the no-information error rate, p_i the proportion of samples of class

$\mathcal{C}_i$, q_i the proportion of samples assigned to class $\mathcal{C}_i$ with the prediction rule and $i = 1, 2$.

Note that e.632+ does not dictate the optimal number of features to select. The error rate estimates that are computed with respect to the number of selected features are only a way to compare the performances of the different methods. Remark at last that each algorithm needs to be learned on each bootstrap sample of the e.632+ bootstrap method to avoid any selection biais (Ambroise and MacLachlan, 2002). Concerning the performance assessment of a T-test selection we used a linear SVM as classifier. We assumed that although SVM is unrelated with this univariate method, it is well appropriate for this two-class problem.

3.5 Computing the efficiency of classification ϵ

The theoretical part showed that the ofw algorithm can be run with any classifier. However, computing the classification efficiency depends on the classifier. For ofw+CART, because of the unstable nature of CART, one needs to aggregate trees as in Breiman (1996) to reduce their variability. For iteration n, we launched B trees on B bootstrap samples on different ω_n^b drawn with respect to $\mathbb{P}_n$, where $b = 1, \ldots, B$. We then defined ϵ as the mean classification error rate on the out-of-bag samples.

No aggregation was needed with SVM, that is known to be very stable, and hence for this case B=1.

3.6 Computational amendements

For ofw+CART a mean gradient was computed that improved the speed of the algorithm

$$G_n = \frac{\sum_{i=1}^{n} \alpha_i \bar{d}_i}{\sum_{i=1}^{n} \alpha_i} \quad \text{with} \quad \bar{d}_i = \sum_{b=1}^{B} \frac{C(\omega_i^b, .)\epsilon(\omega_i^b)}{\mathbb{P}_i(.)}$$

where $\alpha_i = K/(i+1)$, $i = 1..n$ for any positive constant $K > 0$, as defined in equation (4).

DOI: 10.2202/1544-6115.1312

Prostate \ Colon	ofwSVM	RFE	l_0	ofwCART	RF	T-test
ofwSVM	#	14	13	6	0	5
RFE	24	#	27	0	0	0
l_0	21	39	#	1	0	0
ofwCART	4	4	4	#	16	16
RF	6	5	4	17	#	36
T-test	7	5	3	12	31	#

Table 1: Number of genes shared by the several feature selection algorithms on Colon (upper triangle) and Prostate (lower triangle) for a selection of 50 genes.

Leukemia	ofwSVM	RFE	l_0	ofwCART	RF	T-test
ofwSVM	#	16	18	12	15	14
RFE		#	27	10	12	13
l_0			#	8	11	12
ofwCART				#	33	25
RF					#	32
T-test						#

Table 2: Number of genes shared by the several feature selection algorithms on Leukemia for a selection of 50 genes.

Furthermore, to accelerate the computations, the data set Prostate that had a very high classification difficulty was filtered with a very large cut-off T-test p-value (we kept the genes below the p-value 0.1, which corresponded to 3584 remaining genes). Here we made the assumption that most genes are noisy or uninformative and can be removed without affecting the biological study. Indeed, only a very small subset of genes do explain the outcome.

4 Results and discussion

4.1 Numerical results

4.1.1 Comparison of several selections

Table 1 displays the number of shared genes with the different methods when selecting 50 genes on the benchmarks Colon (upper triangular table) and Prostate (lower triangular table).

It first underlined the fact that all gene selections depended on the performed method as there were very few genes that were shared among all methods (less than 36 in Colon and 39 in Prostate). Furthermore, as expected, the methods could be divided in three groups: group 1 and 2 used either the classifier SVM (ofw+SVM, RFE and l_0) or CART (ofw+CART and RF) and group 3 is composed of the method T-test on its own.

Methods in the same group shared an important number of selected genes (for instance at least 13 genes in group 1 and 16 genes in group 2 for Colon). Conversely, the number of genes shared in-between groups was very low (0 to 6 between groups 1 and 2 for Colon). Compared with group 3, more than half of the genes selected with RF were differentially expressed (meaning significant with the T-test) as well as about one third for the genes selected with ofw+CART.

The group 1 did not select many differentially expressed genes (0 to 5 for Colon). The difference is that SVM looks for non redundant genes which lead to a linear separation between the classes C_1 and C_2. These genes are not necessarily differentially expressed. On the other hand, when CART is constructed, it searches genes with the largest difference mean between the two classes. It was hence not surprising to find many differentially expressed genes in group 2.

These latter methods also selected discriminative subsets that were different from the T-test selection. The reason is that groups 1 and 2 take into account interactions between variables, as opposed to filter methods like T-test. The differences between these three groups are less striking in Table 2 on Leukemia as this data set seems more easy for the classification task (see section below). Nevertheless, we can observe that RF and ofwCART shared numerous genes that were also selected with T-Test.

4.1.2 Comparison of the error rate with selection

Figure 1 displays the e.632+ bootstrap error rates obtained with the different methods on the three data sets with respect to the number of selected genes. These graphs first showed the level of classification difficulty of the data sets: for all methods and for a number of selected genes going from 20 to 50, the e.632+ error rates varies from 1 to 6 % on Leukemia (**a**), from 10 to 30% on Colon (**b**), and from 5 to 23% on Prostate (**c**). This variation is even more accentuated as the methods do not have the same performance (Colon, Prostate). Leukemia got similar error rates for all methods as it is known to be relatively easy to classify.

The graphs showed that RF was the most stable and outperformed the

DOI: 10.2202/1544-6115.1312

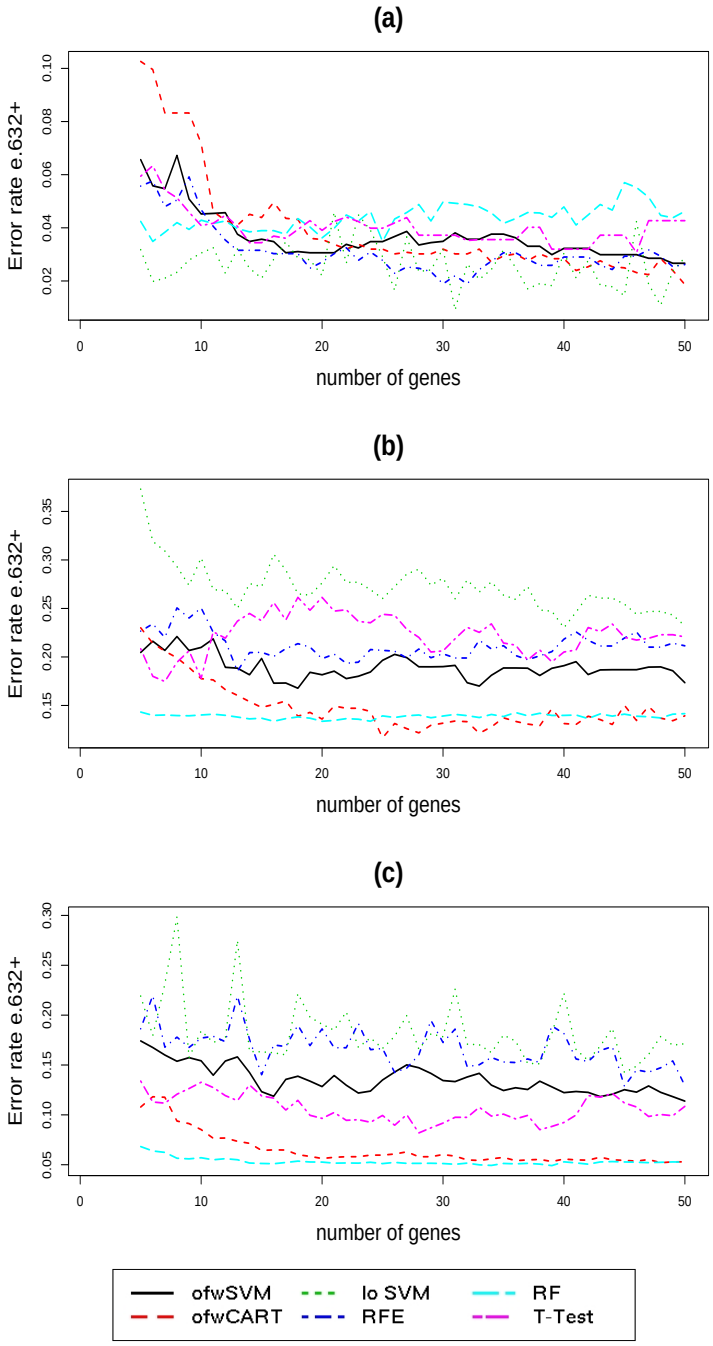

Figure 1: e.632+bootstrap error of several algorithms with respect to the number of genes on Leukemia (**a**) Colon (**b**) and Prostate (**c**).

Criterion \ Method	T-test	RF	ofw CART	ofw SVM	l_0 SVM	RFE
Number of networks	7	4	4	6	3	3
Cancer term frequency in networks	3	2	1	0	1	1
Hematological disease term frequency in networks	0	0	0	0	1	1
Rank of the ontological term in the function list:						
Cancer	9	19	7	17	18	19
Hematological Disease	1	17	9	2	2	13
Number of surface markers	6	20	15	17	9	10
Number of genes associated with the ontological term:						
Leukemia	5	15	15	4	6	4
Myeloid Leukemia	0	3	4	2	0	0
Myeloid leukemia gene name		CD33 SPI1 TOP2B	TOP2B ITGB2 SPI1 CD33	TOP2B ITGB2		
Genes involved in the signaling pathways:						
NfKappaB		TRA@		TRA@	NFKBIA	TRA@
IL4	STAT6			BLVRB		
IL6			IL8			HSPB1
Wnt/Beta Catenin	GNAQ	TCF3	TCF3			
JAK/STAT	STAT6					

Table 3: Analysis of gene selections resulting from several feature selection algorithms on Leukemia data set. Comparisons of gene lists through IPA outputs with several criteria assessing global or specific information.

other methods, except on Leukemia where it performs the worst. This can be explained as the forest is constructed only on the most discriminative variables and is less affected by noisy variables. Hence e.632+ or any error rate computation might not be appropriate to evaluate the performance of RF.

The T-test was not the most efficient as this univariate procedure eliminates noisy genes but does not yield compact non-redundant genes sets. Consequently, genes that are complementary but do not separate the data well are missed.

On the other hand, our two methods were competitive on the more complex data sets Colon and Prostate. On these data sets, ofw+CART gave better performance as CART searches for a non linear separation between features, which a linear SVM cannot perform. These graphs generally showed that a gene selection gives statistical good results when the size of the selection is large enough (greater than 10 genes, depending on the method) but not too large as noisy variables might then enter the selection. It is actually well known that it is impossible to achieve an errorless separation with a single gene. Better results are obtained with a combination of several genes. Note that we did not determine here one optimal gene subset. Only the biological interpretation will give some clue about the relevance of the different selections.

DOI: 10.2202/1544-6115.1312

Method / Criterion	T-test	RF	ofw CART	ofw SVM	l_0 SVM	RFE
Number of networks	4	4	6	4	4	5
Cancer term frequency in networks	1	3	2	1	2	1
Gastrointestinal disease term frequency in networks	0	1	0	0	0	2
Rank of the ontological term in the function list:						
Cancer	11	17	6	4	11	15
Gastrointestinal disease	43	67	49	67	0	22
Tissue development	45	NA	36	2	2	2
Tissue morphology	1	1	37	39	35	26
Skeletal and muscular syst. dev.	3	2	35	40	5	6
Number of genes associated with the ontological term:						
Cancer	11	12	8	12	6	8
Tissue development	2	0	3	6	5	7
Tissue morphology	9	11	8	5	8	6
Skeletal and muscular syst. dev.	12	12	7	7	12	9
Colon Cancer	2	1	0	2	0	1
Colon cancer gene name	CDH3 GUCA2B	CDH3		CDH3 GUCA2B		GUCA2B
Genes involved in the signaling pathways:						
PI3K/AKT			Bcl2	PPP2R5		MEF2C
ERK/MAPK				ETS2 PPP2R5	IL1R2, MEF2	PPP2RC
p38/MAPK						MEF2
Wnt/Beta Catenin	CDH3	CDH3	CSNK2A2	CDH3		PPP2R5C

Table 4: Analysis of gene selections resulting from several feature selection algorithms on Colon data set. Comparisons of gene lists through IPA outputs with several criteria assessing global or specific information.

Method / Criterion	T-test	RF	ofw CART	ofw SVM	l_0 SVM	RFE
Number of networks	6	8	7	7	9	14
Cancer term frequency in networks	3	4	1	3	3	6
Renal and Urological disease term frequency in networks	0	0	0	0	0	0
Rank of the ontological term the in function list:						
Cancer	10	14	6	5	1	2
Renal and Urological Disease	49	0	33	59	0	0
Lipid Metabolism	25	28	27	36	17	15
Number of genes associated with the ontological term:						
Cancer	17	13	12	13	9	13
Prostate Cancer	4	3	3	1	1	4
Prostate cancer gene name	HPN SAT NME1 TGFB3	HPN TGFB3 GSTP1	FOLH1 HPN CLU	FOXO1A	SERPINB5	COX5A HOXC6 PMAIP1 SERPINB5
Genes involved in the signaling pathways:						
C21 steroid hormon metabolism				HSD11B1		
Androgen and Estrogen metabolism				HSD11B1		
Estrogen receptor signaling					CDK7	
Fatty acid metabolism					CYP4F2	CYP4F2
Wnt/Beta Catenin	TGFB3 NME1	TGFB3		WIF1 TLE4	PPP2R5E TLE4	WIF1 PPP2R5E
Pyrimidine or Purine metabolism	GUCY1A3 ATPGV1G1 DPYSL2		NME1	AOX1	RRM1	RRM1
PI3K/AKT				FOXO1A	PPP2R5E	
ERK/MAPK					PPP2R5E PAK1	PPP2R5E PAK1

Table 5: Analysis of gene selections resulting from several feature selection algorithms on Prostate data set. Comparisons of gene lists through IPA outputs with several criteria assessing global or specific information.

4.2 Biological interpretation and discussion

Bioanalysis strategy

In order to elaborate an accurate assessment of the biological relevancy of the various tested methods, we analyzed all lists of 50 selected genes through Ingenuity Pathways Analysis [1] (IPA). IPA was chosen for two main reasons, first for its accuracy: IPA Ontology presents 25 times more classes than Gene Ontology (GO) and 85 high level functions compared to 3 for GO; and second because it supplies a more objective performance estimation compared to manual curating. Hence, with this strategy, we will focus more on global functions associated with a list of genes (*integrative biology*) than on a gene function associated with one gene only. This might allow to identify relevant genes present in a canonical pathway that were not selected with any statistical method.

We explored three outputs from IPA to generate performance indicators of a selected gene list: the *networks* that identify the interactions between the genes, and the most significant *functions* and *signaling pathways* generated by this gene list. This significancy is measured with a p-value of Fisher's exact test determining the probability that each biological function and disease assigned to a gene network or to a gene list was due to chance only. Concerning the canonical pathways, this significancy is furthermore measured by a ratio of the number of genes that map to a given pathway divided by the total number of genes that map to the canonical pathway generated by the gene list. More documentation about IPA can be found online.

The subsequent procedure was followed. First, we uploaded gene identifiers into the IPA application. Each gene identifier was mapped to its corresponding gene object in the Ingenuity Pathways Knowledge Base (IPKB). These genes, called "Focus Genes", were overlaid onto a global molecular network developed from information contained in the IPKB. Networks of these Focus Genes were then algorithmically generated based on their connectivity. Next, the functional analysis of a gene network identified the biological functions and diseases that were the most significant for those given genes. We also took into account the ranks of the most relevant biological functions and the canonical pathways were also considered.

An important remark In this interpretation we do not propose new information for cancer cause as the molecular data set depends entirely on the experimental setting that was chosen by the biologists. The aim of this section

[1] Ingenuity®Systems, www.ingenuity.com

DOI: 10.2202/1544-6115.1312

is simply to check if our statistical results are not biologically aberrant and therefore contain relevant information that would need further experimental proof. The relevant selected genes can be called "predictive" from a statistical point of view (as they are selected on the basis of their predictive power), but from a biological point of view we do not pretend that these genes predict a cancer. The statisticians do hope that the selected genes might be predictive but the biologists can only evaluate the informative characteristics of these genes.

Leukemia data set

The aim of this data set was to select molecular markers distinguishing two leukemia variants arising from lymphoid precursors (Acute Lymphoblastic Leukemia, ALL) or from myeloid precursors (Acute Myeloid Leukemia, AML). Table 3 displays the biological performance estimated for each gene selection method.

In order to check the information quality from a selected gene list, several parameters were defined with various accuracy degrees. The potential abundance of information was first related to the number of networks. The more numerous the generated networks, the more varied the suggested biological clues. One gene list could also be considered as biologically relevant if ontological terms such as "Cancer" or "Hematological Disease" were linked to the networks of interacting genes or found well positionned, according to the p-values functions. We also focused on general leukemia molecular markers and more specifically on AML or ALL markers (Carroll et al., 2006, Pui et al., 2004), as well as surface marker gene families. These latter encode cell surface proteins that would be useful in distinguishing lymphoid from myeloid lineage cells as it was previously demonstrated for the CD33 gene (Drexler, 1987, Malask et al., 2006).

Canonical pathways did not reveal enough relevant differences between the gene lists to compare the methods. Networks generated by IPA were more numerous for the gene list selected by the filter method. It suggests that this method chooses less biologically interconnected genes compared to the wrapper methods. This could be explained by the fact that filter methods disregard the interactions between the features.

When looking for ontological terms, representative of leukemia pathology ("Cancer" and "hematological Disease"), no clear difference arose from any method as they were all well ranked in IPA interacting gene networks or function lists. Surface gene markers found in the networks were mostly selected with ofw+CART, RF and ofw+SVM, suggesting particular biological rele-

vancy for those selected gene lists. Genes linked with "Leukemia" term or more precisely with "Myeloid Leukemia" terms were mostly selected in the lists given by the wrapper methods.

Compared to the wrapper methods, the filter method selected a poor number of general molecular markers linked to the leukemia pathology and surface markers distinguishing AML from ALL. No gene that was linked to Myeloid Leukemia was selected. On the other hand, wrapper methods gave very complementary and relevant gene lists. Three particular methods, ofw+CART, RF and ofw+SVM selected genes that are known to be involved in leukemia pathology *i.e.* TOP2B, ITGB2, SPI1, CD33). This trend was confirmed when we manually curated the gene lists proposed by all methods. ofw+CART, RF and ofw+SVM selected a set of genes involved at different biological level of the leukemia pathology (see for instance this non exhaustive list: CD33, ZYX, CCND3, TOP2B, SPI1, ITGB2, CCNI, NFIC, KPNB1).

To sumarize, we found that the T-test gene selection brought very general cancer-related information and much less information directly related to leukemia pathology than the CART or SVM based wrapper methods. The CART based methods proposed candidates that are linked to Myeloid Leukemia.

Colon data set

The objective of this data set was to select genes distinguishing tumor from normal sample. This is a particularly challenging problem since initial composition of the two types of cells are very different. Indeed, the high composition of tumor richness in epithelial cells and normal tissues in smooth muscle cells produce an important biological parameter that biases cancer-related genes tracking for tumor *vs.* normal cells (Guyon et al., 2002). Biological relevancy of the gene selections was assessed in the same manner as for the Leukemia data set with networks and function lists evaluation (Table 4). We chose ontological terms specific to colon cancer pathology such as "Cancer" and "Gastrointestinal Disease". Ontological terms linked to initial cells composition were also exploited to explain the performances of all methods ("Tissue Morphology", "Skeletal and Muscular System Development and Function"). Specific genes of colon cancer were also taken into account as well as specific signaling pathways.

For this particular data set, the number of networks generated by IPA for any gene selection method was similar. Principal differences arose from ontological functions of those networks. Indeed, rich cancer-related networks were generated from CART-classifiers methods as opposed to poor ones coming from the other methods. For any gene selection, the rank of the ontological

DOI: 10.2202/1544-6115.1312

term "Gastrointestinal Disease" in the function list was surprisingly low (line 5 of Table 4). An explanation of this particular feature could lie under the biological sample composition which is very rich (or too rich) in smooth muscle cells for normal tissue or in epithelial cells for tumor tissue. Interestingly, the ontological terms in IPA function lists "Skeletal and Muscular System Development and Function" or "Tissue Morphology" terms were always on top, lowering the rank of the "Gastrointestinal Disease" term. This observation was also reinforced by the larger number of genes linked with those last functions, comparing to those linked with the "Cancer" term. Therefore, exploitation of functions list analysis does not favor one method against another, as the sample biological composition bias precludes straightforward detection of specific colorectal cancer genes.

When analyzing IPA canonical pathways, Wnt, MAPK and AKT signaling pathways that were (or supposed to be) involved in colorectal cancer, were mainly generated with gene selection involving SVM and CART methods (Oikonomou et al., 2006, Segditsas et al., 2006). Hence, SVM-classifiers (followed by CART-classifiers) seemed to select biologically relevant genes or signaling pathways, even in a data set that has a largely biased gene expression profile.

To summarize, we found that all methods selected genes that were more related to cell composition than to the pathology of interest. Very few colon cancer genes were identified. Despite the important biological biais, the wrapper methods were able to select complementary and relevant genes associated with relevant pathways.

Prostate data set

The identification of gene markers that might help to distinguish tumourous prostate from healthy prostate samples was the main purpose of this third data set. As for colorectal tumor, epithelial content of prostate tumor samples was significantly higher than in normal samples (79 *vs.* 27 %). This results in gene expressions correlated with epithelial content that may preclude cancer-related gene efficient tracking (Singh, D. et al.). Results are displayed in Table 5. We focused this time on the specific ontological terms "Cancer" and "Urological Disease". Prostate cancer specific genes and known deregulated signaling pathways were also used to determine more precisely the relevancy of the different selections.

All lists uploaded into IPA generated the same number of networks (except for RFE that was much higher) and were all linked with the ontological term "Cancer". This term was very well ranked for all function lists whereas, as

observed for the Colon data set, the specific "Urological Disease" term was low ranked. When analysing ontological terms ranked in between, we noted a prevalence for functions involving cell proliferation, regulation of gene expression, lipid metabolism and nucleic acids, *i.e.* biological cell functions that are well known to be involved in prostate cancer disorders (Foley et al., 2004). The number of genes linked with ontological term "Cancer" was the same in any selection. When we focused on specific prostate cancer genes, all gene selection methods brought complementary information. For instance, CART-based methods selected HPN, a gene coding for a transmembrane serine protease involved in colony formation of prostate cancer cell lines (Dhanasekaran et al., 2001). SVM-based methods l_0 norm SVM and RFE selected SERPINB5, a gene coding for a serpin peptidase inhibitor involved in binding of prostate cancer cell lines Tahmatzopoulos et al. (2005).

IPA canonical pathways gave various information depending on the different selections. With the SVM-based methods, we observed signaling pathways involved in prostate cancer pathology (Terry et al., 2006) such as Wnt, MAPK, AKT, pyrimidine and purine signaling pathway. In particular, l_0 norm SVM and RFE selections highlighted the Fatty Acid Metabolism and ofw+SVM the Androgen Signaling Pathway that is actually targeted for prostate cancer therapy (Singh, P. et al., 2002). Hence, SVM-based methods seemed to select here more relevant signaling pathways than the other methods.

To summarize, the T-test and the wrapper methods selected very complementary sets of genes related to prostate cancer in spite of the cell composition bias. All methods were also able to select complementary and relevant genes associated with relevant pathways.

5 Conclusion

The analysis of these three public data sets was performed at two levels. Statistically, we showed that the stochastic algorithm from Gadat and Younes could be applied to microarray data with two classifiers SVM and CART. ofw+CART and ofw+SVM gave excellent results compared to other well known wrapper methods. We also showed that the selected gene lists mostly depended on the chosen classifier.

Biologically, we showed that the relevancy did not only depend on the chosen method but also on the biological sample nature. Indeed, when applying these methods on a simple data set Leukemia, ofw+CART, RF and ofw+SVM proposed very relevant gene lists compared to the others. With a more complex biological matrix like in Colon or Prostate, the expression pattern are

DOI: 10.2202/1544-6115.1312

mixed between constitutive gene expression (*i.e.* expression of a large majority of genes involved in physiological characteristics of a tumor or normal cell) and cancer gene expression. In this setup, we rather observed a global complementarity of the biological information brought by the different selections. However, SVM-based methods seemed to propose interesting signaling pathways for Colon and Prostate data sets.

To summarize, we highlight the fact that the method statistically performing the best prediction does not necessarily give the most interesting biological results. In fact, the application of different methods on the same data set can highlight complementary relationships between the selected genes. Hence, to bring more information, one should not only consider the common features selected between the methods, but also the divergent ones. This means that there is not only one single method that answers a biological question: complementary approaches should be performed to analyze the data.

Availability

The code sources of ofw+SVM (in C++) and ofw+CART (in $\mathbf{R}$[2]) are available on the web site http://www.lsp.ups-tlse.fr/Biopuces/ofw/codesource/. An R package is currently being implemented but can be available upon request to the corresponding author.

References

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, **96**, 6745-6750.

Ambroise, C. and McLachlan, G. (2002) Selection Bias in Gene Extraction in Tumour Classification on Basis of Microarray Gene Expression Data. *Proc. Natl. Acad. Sci. USA*, **99**(10), 6562-6566.

Carroll W.L.,Bhojwani, D., Min, D.J., Moskowitz, N. and Raetz, E.A. (2006) Childhood Acute Lymphoblastic Leukemia in the Age of Genomics. *Pediatric Blood Cancer*, **46**(5), 570-578.

Benveniste, A., Métivier M. and Priouret, P. (1990) Adaptive Algorithms and Stochastic Approximations. (Springer-Verlag).

[2]The Comprehensive R Archive Network, http://cran.r-project.org/

Breiman, L., Friedman, J.H., Olshen, R. A. and Stone, C.J. (1984) Classification and Regression Trees. (Wadsworth, Belmont, CA).

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**, 123-140.

Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-22.

Dhanasekaran, S.M., Barrette, T.R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K.J., Rubin, M.A. and Chinnaiyan, A.M. (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature*, **412**(6849), 822-6.

Diaz-Uriarte, R. and Alvarez de Andrés, S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**(3).

Drexler, H.G. (1987) Classification of acute myeloid leukemias : a comparison of FAB and immunophenotyping. *Leukemia*, **1**(1010), 697-705.

Dudoit, S., Fridlyand, J. and Speed, T. (2000) Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *J. Am. Stat. Assoc.*, **97**, 77-87.

Efron, B. and Tibshirani R.J. (1997) Improvements on cross-validation: the e.632+ bootstrap method. *Journal of American Statistical Association*, **92**, 548-560.

Foley, R., Hollywood, D., and Lawler, M. (2004) Molecular pathology of prostate cancer: the key to identifying new biomarkers of disease. *Endocrine Related Cancer*, **11**(3), 477-488.

Gadat, S. and Younes, L. (2007) A Stochastic Algorithm for Feature Selection in Pattern Recognition. *Journal of Machine Learning Research*, **8**, 509-547.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

Guyon, I. and Weston, J. and Barnhill, S. and Vapnik, V. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389-422.

DOI: 10.2202/1544-6115.1312

Kushner, H. and Clark, D.S. (1978) Stochastic Approximation Method for Constrained and Unconstrained Systems. (Springer-Verlag).

Oikonomou, E. and Pintzas, A. (2006) Cancer genetics of sporadic colorectal cancer: BRAF and PI3KCA mutations, their impact on signaling and novel targeted therapies. *Anticancer Res.*, **26**(2A), 1077-84.

Maslak, P.G., Jurcic, J.G. and Scheinberg, D.A. (2006) Monoclonal antibodies for the treatment of acute myeloid leukemia. *Curr Pharm Biotechnol.*, **7**(5), 343-69.

Pui, C.H., Schrappe, M., Ribeiro, R.C. and Niemeyer, C.M. (2004) Childhood and adolescent lymphoid and myeloid leukemia. *Hematology Am Soc Hematol Educ Program*, **1**, 118-45.

Segditsas, S. and Tomlinson, I. (2006) Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene* **25**, 7531-7537.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander E.S., Loda M., Kantoff P.W., Golub T.R. and Sellers W.R. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **2**, 203-9.

Singh, P., Uzgare, A., Litvinov, I., Denmeade, S.R. and Isaacs, J.T. (2006) Combinatorial androgen receptor targeted therapy for prostate cancer. *Endocr Relat Cancer.*, **3**, 653-666.

Tahmatzopoulos, A., Sheng, S. and Kyprianou, N. (2005) Maspin sensitizes prostate cancer cells to doxazosin-induced apoptosis. *Oncogene*, **24**, 5375-5383.

Terry, S., Yang, X., Chen, M.W., Vacherot,F. and Buttyan, R.(2006) Multifaceted interaction between the androgen and Wnt signaling pathways and the implication for prostate cancer. *J Cell Biochem.*, **99**, 402-410.

Vapnik, V. (2000) The nature of statistical learning theory. (Springer-Verlag).

Weston, J., Elisseeff, A., Schölkopf, B. and Tipping, M. (2003) Use of the Zero-Norm with Linear Models and Kernels Methods. *Journal of Machine Learning Research*, **3**, 1439-1461.