A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis

Cosmin Lazar, Jonatan Taminau, Stijn Meganck, David Steenhoff, Alain Coletta, Colin Molter, Virginie de Schaetzen, Robin Duque, Hugues Bersini, and Ann Nowé

Abstract—A plenitude of feature selection (FS) methods is available in the literature, most of them rising as a need to analyze data of very high dimension, usually hundreds or thousands of variables. Such data sets are now available in various application areas like combinatorial chemistry, text mining, multivariate imaging, or bioinformatics. As a general accepted rule, these methods are grouped in filters, wrappers, and embedded methods. More recently, a new group of methods has been added in the general framework of FS: ensemble techniques. The focus in this survey is on filter feature selection methods for informative feature discovery in gene expression microarray (GEM) analysis, which is also known as differentially expressed genes (*DEGs*) discovery, gene prioritization, or biomarker discovery. We present them in a unified framework, using standardized notations in order to reveal their technical details and to highlight their common characteristics as well as their particularities.

Index Terms—Feature selection, information filters, gene ranking, biomarker discovery, gene prioritization, scoring functions, statistical methods, gene expression data.

1 Introduction

Gene expression microarray (GEM) experiments aim to obtain valuable biological information by collecting biological data from samples (e.g., tissues, cell lines). Recorded GEM data contain gene-wise information across all samples under investigation. In a single experiment, information about thousands of genes is measured and recorded simultaneously. The samples under investigation can be different from many perspectives (e.g., genotype, phenotype, or other biological or clinical relevant annotation). An important research topic in GEM data analysis is the discovery of genes that are able to differentiate between samples originating from different populations or in more general terms, genes which are relevant for a particular target annotation. These genes are called informative genes, biomarkers or *differentially expressed genes (DEGs)*. The discovery of *DEGs* is valuable not only to physicians to diagnose patients but also to pharmaceutical companies aiming to identify genes which can be targeted by drugs. In the last few years, a lot of effort has been put in the development of methodologies for *DEGs* discovery. The problem is still challenging and new algorithms emerge as alternatives to the existing ones. The literature of FS for *DEGs* discovery is abundant. Despite the wide range of approaches proposed to solve this problem, many algorithms share common elements merely differing on details from each other. Our intention is to bring into light an important family of methods (the filters) that are sometimes unfairly discarded to the benefit of more complicated FS techniques. We aim to provide a big picture over filter techniques for *DEGs* discovery in a unified technical framework in order to outline their common points as well as their particularities.

The roadmap of this survey is as follows: in Section 2, the FS problem is described and filter methods are presented in the framework of FS. Section 3 provides the big picture of filter methods in an extended taxonomy inspired from [1]. Two big groups of filters are revealed here: ranking and space search methods. They will be further described in details as follows: Section 4 generally describes the ranking strategy for FS focusing on two main points: the scoring functions used to assign relevance indices to genes and the problem of statistical significance of the estimated scores, while Section 5 is dedicated to the space search strategy. In Section 6, we focus on the evaluation of FS results and we present several evaluation strategies, while in Section 7 we provide the reader with some comments and recommendations which could help in choosing the appropriate methods or for comparison purposes. The last section is dedicated to authors' concluding remarks.

2 Problem Statement

The problem can be stated as follows: let us consider a GEM study where biological data from a population of

- *C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, and A. Nowé are with AI lab, Computational Modeling Group, Department of Computer Science, Vrije Universiteit Brussel, Pleinlaan 2, Brussels 1050, Belgium. E-mail: {vlazar, jtaminau, smeganck, David.Steenhoff, ann.nowe}@vub.ac.be.*
- *A. Coletta, C. Molter, V. de Schaetzen, and R. Duque are with Université Libre de Bruxelles, Building C, 5th floor, Room C5.329, 87 av. Adolphe Buyl, Brussels B-1050, Belgium. E-mail: {alaincoletta, colin.molter}@gmail.com, virgdes@yahoo.com, rduque@vub.ac.be.*
- *H. Bersini is with IRIDIA—Universite Libre de Bruxelles, CP 194/6-50, av. Franklin Roosevelt, Bruxelles 1050, Belgium. E-mail: bersini@ulb.ac.be.*

Manuscript received 29 June 2011; revised 23 Jan. 2012; accepted 1 Feb. 2012; published online 13 Feb. 2012.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2011-06-0163. Digital Object Identifier no. 10.1109/TCBB.2012.33.

samples are collected. The output of the study is recorded as a matrix (called gene expression data matrix) $X^{m\times n} = \{x_{i,j}\}$ containing the expression of m features/genes across n samples, where $x_{i,j}$ is the expression level of gene i in sample j and $m \gg n$. Here we emphasize on the fact that gene expression data matrix X is obtained through a complex process where in the first instance raw, probe-level data are collected. Consequently, gene expression data are derived through a series of preprocessing steps including background correction, log-transformation, normalization, and summarization and further analysis is performed on the preprocessed data. However, a discussion upon these preprocessing methods is beyond the scope of this paper.

Besides gene expression data, metainformation (typically clinical or biological annotations) is also collected during microarray experiments. These annotations usually contain information about the patients but also information about tissue genotype or phenotype, type or time of treatment, etc. Hence, they map either to categorical or to continuous variables. Typical applications on GEM data are *disease prediction*, *disease discovery* [2], or *reconstruction of gene regulatory networks from gene expression data* [3]. Solutions to these problems demand for machine learning techniques such as supervised classification, clustering, and regression. The direct application of these methods on high-dimensional data is usually inefficient [4]. Therefore, it is desirable to select a small subset of features/genes that is discriminative among the subgroups of samples denoted by a target annotation. As mentioned in the introduction, these genes are called *informative genes* or *differentially expressed genes*.

As a generic definition, FS consists in identifying the set of features/genes whose expression levels are indicative of a particular target feature (clinical/biological annotation). In mathematical terms, this problem can be stated as follows: let $X^{m\times n} = \{x_{i,j}\}$ be a matrix containing m genes and n samples originating from different groups denoted by a target annotation (e.g., different phenotypes), $X^{m\times n} = [X_1^{m\times n_1} X_2^{m\times n_2} \dots X_p^{m\times n_p}]$ where each matrix $X_i^{m\times n_i}$ contains samples from the same group and $n_1 + n_2 + \dots n_p = n$. Selecting the most informative genes consists in identifying the subset of genes across the entire population of samples $S^{k\times n} \in X^{m\times n}$, $k \ll m$ which is the most discriminative for the outlined classes. This definition is only valid for classification problems where the groups are clearly identified beforehand (e.g., disease prediction).

Different strategies have been proposed over the last years for feature/gene selection: filter, wrapper, embedded [1], and more recently ensemble techniques [5].

Filter techniques assess the discriminative power of features based only on intrinsic properties of the data. As a general rule, these methods estimate a relevance score and a threshold scheme is used to select the best-scoring features/genes. Filter techniques are not necessarily used to build predictors. As stated in [6], *DEGs* may also be good candidates for genes which can be targeted by drugs. This group of techniques is independent of any classification scheme but under particular conditions they could give the optimal set of features for a given classifier. Saeys et al. [1] also stress on the practical advantages of these methods stating that *"even when the subset of features is not optimal, they may be preferable due to their computational and statistical scalability."*

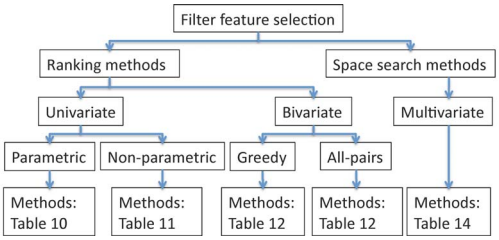


Fig. 1. Proposed taxonomy for filter FS methods.

Wrapper techniques select the most discriminant subset of features by minimizing the prediction error of a particular classifier. These methods are dependent on the classifier being used and they are mainly criticized because of their huge computational demands. More than that, there is no guarantee that the solution provided will be optimal if another classifier is used for prediction.

Embedded techniques represent a different class of methods in the sense that they still allow interactions with the learning algorithm but the computational time is smaller than wrapper methods.

Ensemble techniques represent a relatively new class of methods for FS. They have been proposed to cope with the instability issues observed in many techniques for FS when small perturbations in the training set occur. These methods are based on different subsampling strategies. A particular FS method is run on a number of subsamples and the obtained features/genes are merged into a more stable subset [7].

So far we briefly described the topic of FS but the rest of the paper is entirely dedicated to filter methods for *DEGs* discovery in GEM analysis. Here we stress on the advantage of filters over wrappers or embedded methods which is their independence of classifiers. This particular characteristic of filters avoids all influence of classifier's bias in the FS process.

3 Filter Methods for FS: A Proposed Taxonomy

Building a taxonomy is not a trivial task and moreover, a taxonomy is not unique. Based on the literature reviewed for this paper we propose a taxonomy of filter FS methods for informative genes discovery. As a general observation, two different filter strategies can be identified while surveying the literature. According to the first strategy, one selects features/genes which are top ranked according to some relevance indices estimated with a predefined scoring function. According to the second strategy, features are selected by optimizing a particular cost function which is often defined as a tradeoff between the maximum informativeness and minimum redundancy inside the selected subgroup of features/genes. In the following, we will refer to methods built upon the first strategy as *ranking methods* while those built upon the second strategy will be referred to as *space search methods*.

Our proposed taxonomy (Fig. 1) has many common points with the one presented in [1]. The main difference consists in the fact that on the top level we grouped the filter methods in ranking and space search methods, according to the strategy used to select features. On the

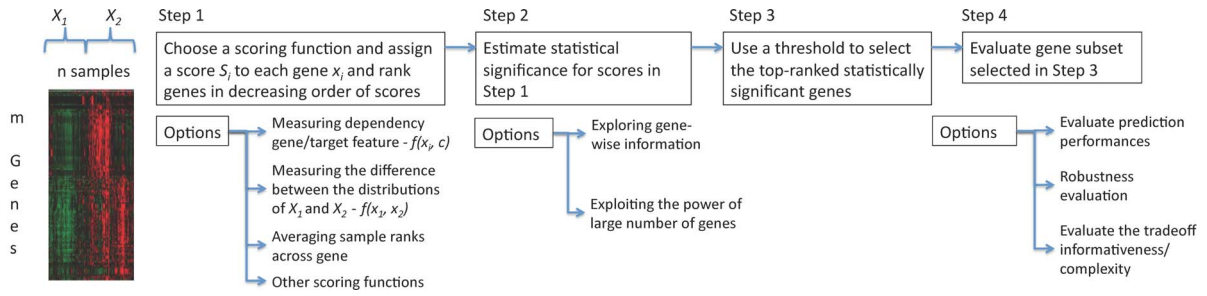


Fig. 2. Illustration of filter ranking methods. Main steps of univariate ranking methods for filter FS. A case study for GEM analysis.

level below, ranking methods are grouped in univariate and bivariate while the space search methods are all multivariate. Subsequently, depending on the parametric assumption used, the univariate methods are split into *parametric* and *nonparametric* while bivariate methods can be *greedy* or *all-pair* methods depending on the strategy used for ranking. The next two sections are dedicated to a synthetic description of the two main classes of filters: ranking and space search methods.

As the biggest part of the reviewed literature for this paper focuses mainly on *DEGs* discovery for disease prediction, the methods designed for classification problems will have a bigger weight compared with those designed for regression, mainly applied to infer networks from gene expression data [8]. We'll briefly point out the distinction between these methods in Section 7.

4 Filter Methods—A Ranking Approach

Most filter methods consider the problem of FS as a ranking problem. The solution is provided by selecting the top scoring features/genes while the rest are discarded. Generally these methods follow a **typical scenario** described below and pictured in Fig. 2.

1. Use a *scoring function* $S(x)$ to quantify the difference in expression between different groups of samples and rank features/genes in decreasing order of the estimated scores. It is supposed that a high score is indicative for a *DEG*.
2. Estimate the *statistical significance* (e.g., p-value, confidence intervals) of the estimated scores.
3. *Select the top ranked features/genes* which are statistically significant as the most informative features/genes (alternatively one could be interested in selecting the top ranked features/genes only as opposed to the top ranked significant ones).
4. *Validate* the selected subset of genes (see Section 5).

In the above-mentioned generic algorithm one can identify two aspects specific to this type of methods which play an important role in identifying informative features/genes: first, the choice of a scoring function to compute the relevance indices (or scores) and second, the assignment of statistical significance to computed scores. They will receive further consideration in order to be able to reveal the main differences between different methods and therefore helping to categorize them.

As an additional remark, the reader should note that ranked lists of features/genes can also be obtained via wrapper/embedded methods not only for filters, e.g., SVM Recursive Feature Elimination (SVMRFE) [9] or Greedy Least Square Regression [10].

Here we also outline the fact that any combination of a scoring function and a statistical significance test designed to quantify the relevance of a feature/gene for a target annotation can be transformed into a ranking method for FS. Since all steps in the generic algorithm described above are independent one from another, the users do have a lot of freedom in the way they wish to perform the selection.

4.1 Scoring Functions—Assigning Relevance Indices to Features

Scoring functions represent the core of ranking methods and they are used to assign a relevance index to each feature/gene. The relevance index actually quantifies the difference in expression (or the informativeness) of a particular feature/gene across the population of samples, relative to a particular target annotation. Various scoring functions are reviewed and categorized here. They cover a wide range of the literature proposed for *DEGs* or biomarkers discovery. The scoring functions are enumerated and categorized according to their syntactic similarities. A similar approach presenting a very comprehensive survey on distance measures between probability density functions has been employed in [11].

Several groups of scoring functions for gene ranking have been identified. In the first group, we gathered scoring functions which estimate an average rank of genes across all samples. Scoring functions from the second group quantify the divergence (or the distance) between the distribution of samples corresponding to different classes associated to a target annotation per feature/gene as a function $f(x_1, x_2)$. The third group contains information theory-based scoring functions while the fourth group measures the degree of association between genes and a target annotation as a function $f(x, c)$ where x and c described in Table 1. The last group gathers a list of miscellaneous scoring functions which cannot be included in the previous four.

The big majority of scoring functions presented here are usually defined to rank single genes but some of them can be easily adapted for pairs or groups of genes. In this section, the notations in Table 1 will be used.

TABLE 1
Notations

$X^{m\times n}$	GEM data set with m genes and n samples
$X_1^{m\times n_1}$, $X_2^{m\times n_2}$	subsets of X denoting samples from two different populations, where $n_1+n_2=n$
x, x_1, x_2	single gene expression across all samples, across samples in X_1 respectively X_2
$\bar{x}$, $\bar{x}_1$, $\bar{x}_2$	mean value of x, x_1 and x_2
σ_x, σ_{x_1}, σ_{x_2}	standard deviation of x, x_1 and x_2
P_x, P_{x_1}, P_{x_2}	probability density function of x, x_1 and x_2
CDF_{x_1}, CDF_{x_2}	cumulative density function of x_1 and x_2
c	class label feature or target annotation
c_1, c_2	labels corresponding to X_1, X_2 respectively
$R_{i,j}$	rank of gene i in the $j-th$ sample
S	relevance index or score associated to a gene
Ω_m, Ω_s, $\lvert\Omega_s\rvert$	the whole set of genes, a subset of genes from Ω_m respectively the number of genes in Ω_s

4.1.1 Ranking Samples Across Features

This group is represented by two scoring functions: rank-sum and rank-product, see Table 2. Supposing x_1 and x_2 are the expression levels of a certain gene in class c_1 and class c_2, respectively, the rank-sum method first combines all the samples in x_1 and x_2 and sorts them in ascending order. Then the ranks are assigned to samples based on that ordering. If k samples have the same value of rank i, then each of them has an average rank given by $i+\frac{k-1}{2}$. If n_1 and n_2 denote the numbers of samples in the smaller and larger group, respectively, then the rank-sum score is computed by summing up the ranks corresponding to samples in c_1, Table 2 first line. For a GEM data set, the rank-product method consists in ordering the genes across all samples in the value ascending order and then for each gene the rank-product score is obtained by taking the geometrical average of the ranks of that gene in all samples.

4.1.2 Measuring the Divergence between the Distributions of Groups of Samples

Another direction toward the identification of informative features/genes is to quantify the difference between the distributions of groups of samples associated to a target annotation. These scoring functions can be generically described as a function $f(x_1,x_2)$ with x_1,x_2 in Table 1. For this purpose, some simple measures rely only on low-order statistics, in particular the first and second moment (mean and variance) of the distribution of expression levels in different groups. This is the simplest way to compare the distributions of two populations and implicitly imposes some more or less realistic assumptions on the distributions of samples in each population (e.g., normal distributed samples). Despite this obvious drawback they are still the most popular scoring functions used to create filters for FS in GEM analysis due to their simplicity. These scoring functions can be grouped in two families: *fold-change family* (Table 3) and *t-test family* (Table 4). A different strategy in comparing the distributions of different populations is to rely on different estimates of the probability density function (*pdf*) or the cumulative density function (*cdf*) of populations but these methods are more expensive computationally. The different families of scoring functions mentioned here will be further presented in this section.

TABLE 2
Rank Score Family

Name	*Metric*	*Ref.*
Wilcoxon rank sum	$S=\sum_{j=1}^{k} R_j$, $k=\min(n_1,n_2)$	[12]
Rank product	$S=(\prod_{j=1}^{n} R_j)^{1/n}$	[13]

TABLE 3
Fold-Change Family

Name	*Metric*	*Ref.*
Fold-change ratio	$S=\frac{\bar{x}_1}{\bar{x}_2}$	[15]
Fold-change difference	$S=\bar{x}_1-\bar{x}_2$	[16]

Fold-change family. Relative indices are assigned to features/genes based only on mean estimates of the expression levels across different groups of samples per gene. According to [14] two forms are encountered for the fold-change scoring functions: fold-change ratio and fold-change difference (Table 3). However, the fold-change difference is less known and usually researchers who mention fold-change in this context actually refer to fold-change ratio. In practice, many packages for GEM analysis typically provide the $\log_2$ of the ratio between the means of group 1 and group 2. The numbers will be either positive or negative preserving the directionality of the expression change.

t-test family. Several forms derived from the ordinary two-sample t-test are used to measure the difference in expression of genes, see Table 4. In the same family, we include the Z-score or the *signal to noise ratio* (SNR) defined as the ratio between the fold-change difference and the standardized square error of a particular gene. These scoring functions make use of both the first and second moments to assign relevance indices to genes.

Bayesian scoring functions. In several studies, the authors have defined scoring functions for informative features discovery in a Bayesian framework. The main motivation behind this is the difficulty in obtaining accurate

TABLE 4
t-Test Family

Name	*Metric*	*Ref.*
Z-score	$S=\frac{\bar{x}_1-\bar{x}_2}{\sigma_x}$	[17]
t-test	$S=\frac{\bar{x}_1-\bar{x}_2}{\sigma_{x_1}+\sigma_{x_2}}$	[2]
Welch t-test	$S=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{\sigma_{x_1}^2}{n_1}+\frac{\sigma_{x_2}^2}{n_2}}}$	[18]
Modified t-test	$S=\frac{\bar{x}_1-\bar{x}_2}{\sigma_x+\sigma_0}$ σ_0 - small positive constant	[19], [20], [21]

TABLE 5
Bayesian Family

Name	*Metric*	*Ref.*
Bayesian t-test	$S = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_p}$, $\sigma_p^2 = \frac{\nu_0\sigma_0^2 + (n-1)\sigma_x^2}{\nu_0 + n - 2}$ ν_0, σ_0 - prior degrees of freedom/variance	[22]
Regularized t-test	$S = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\tilde{\sigma}^2_{x_1}}{n_1} + \frac{\tilde{\sigma}^2_{x_2}}{n_2}}}$ $\tilde{\sigma}^2_{x_1,x_2} = \frac{\nu_0\sigma_0^2 + (n_{1,2}-1)\sigma^2_{x_1,x_2}}{\nu_0 + n_{1,2} - 2}$ ν_0, σ_0 - prior degrees of freedom/variance	[23]
Moderated t-statistics	$S = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\tilde{\sigma}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, $\tilde{\sigma}^2 = \frac{ds^2 + d_0\sigma_0^2}{d + d_0}$, $s^2 = \frac{(n_1-1)\sigma^2_{x_1} + (n_2-1)\sigma^2_{x_2}}{(n_1-1)+(n_2-2)}$ $d = n_1 + n_2 - 2$ d_0 and σ_0 are unknown and must be estimated from the data	[24]
B-statistics	$B = \log A \left[\frac{b + \sigma_x + (\bar{x}_1 - \bar{x}_2)^2}{b + \sigma_x + \frac{(\bar{x}_1 - \bar{x}_2)^2}{1+nh}}\right]^{\nu + \frac{n}{2}}$ $A = \frac{p}{1-p}\frac{1}{\sqrt{1+nh}}$ b and ν - hyperparameters in the inverse gamma prior for the variance h - hyperparameters in the normal prior of the nonzero means p - fixed to sensible values (0.01 or 0.001)	[25]

estimates of the standard deviation of individual genes based on few measurements only. In order to cope with the weak empirical estimation of variance across a single feature/gene, several authors proposed more robust estimations of the variance by adding genes with similar expression values. A list of these scoring functions is presented in Table 5.

PDF-based scoring functions. Scoring functions in this category rely on different estimates of the *pdfs* of populations, from simple histograms to more complex estimators such as the Parzen window estimator [26]. Only few scoring functions based on this idea are used to discover informative features/genes. Here we identified Kolmogorov-Smirnov (K-S) tests [27], Kullback-Leibler divergence [28], or Bhattacharyya distance [29] (Table 6), but the mathematical literature abounds in measures quantifying the distance between *pdfs* revealing new possibilities to look for informative features/genes. We invite the reader to consult [11] for a very comprehensive survey on this topic. Note that the use of these scoring functions for *DEGs* discovery is limited by the low number of samples in GEM experiments which results in unreliable estimates of the *pdf*.

TABLE 6
pdf-Based Scoring Functions

Name	*Metric*	*Ref.*
K-S test	$S = \sup(CDF_{x_1} - CDF_{x_2})$	[27]
KL divergence	$S = \sum_{i=1} P_{x_{1,i}} \log \frac{P_{x_{i,1}}}{P_{x_{2,i}}}$	[28]
Bhattacharyya distance	$S = -\ln \sum_i \sqrt{P_{x_{1i}} P_{x_{2i}}}$	[29]

TABLE 7
Information Theory-Based Scoring Functions

Name	*Metric*	*Ref.*
Info gain	$S = \text{Info}(X) - \text{Info}_x(X)$ $\text{Info}(X) = -\sum_{i=1}^{k} P(c_i, X) \times \log(P(c_i, X))$ $\text{Info}_x(X) = -\sum_{i=1}^{v} \frac{\lvert V_i\rvert}{\lvert X\rvert} \times \text{Info}(V_i)$ k - number of classes v - number of individual values of a gene x V_i - the set of instances whose values in gene x equal x_i $\lvert V_i\rvert$ - number of samples in V_i, $\lvert X\rvert = n$	[30]
Mutual info	$S = \sum_{i=1} \log\left[\frac{P_{x_i}}{P_{x_{1,i}} P_{x_{2,i}}}\right]$	[31], [32]

4.1.3 Information Theory-Based Scoring Functions

These scoring functions rely on different estimates of the information contained both in the target feature c and in the gene expression x. Table 7 presents a list of scoring functions belonging to this group: information gain and the mutual information.

4.1.4 Measuring the Dependency between Features and Target Feature as a Function $f(x, c)$

Scoring functions in this group have the advantage that they allow features/genes ranking when the target annotation is a continuous variable (which is not the case of the previous mentioned scoring functions). They measure the dependency between the gene's expression profile x and the target feature c as a function $f(x, c)$. Pearson's correlation coefficient (PCCs), Table 8, is commonly used for this purpose. Its absolute value equals 1 if x and c are linearly correlated and equals 0 if they are uncorrelated. Note that PCCs is only applied if c is a continuous variable. When c is binary, PCCs comes down to the $Z - score$. A similar measure used for this purpose is Kendall's rank correlation coefficient (KRCCs). A variant of this measure adapted to a two-class problem is proposed in [33].

4.1.5 Other Scoring Functions

A list of scoring functions mentioned in the literature for informative gene discovery which cannot be grouped in the above-mentioned families is presented here. The list presented in Table 9 includes: Area Under ROC Curve (AUC), Area Between the Curve and the Rising diagonal (ABCR), Between-Within class Sum of Squares (BWSS), and Threshold Number of Missclassifications (TNoM). The

TABLE 8
Correlation Gene-Class Label Family

Name	*Metric*	*Ref.*
PCCs	$S = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(c_i - \bar{c})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(c_i - \bar{c})^2}}$	[34]
2-class	$S = \sum_{i \in c_1} \sum_{j \in c_2} h(x_{1,i} - x_{2,j})$	[33]
KRCCs	where $h(x) = \begin{cases} 0, if\ x \leq 0 \\ 1, if\ x > 0 \end{cases}$	

TABLE 9
Other Scoring Functions for Gene Ranking

Name	*Metric*	*Ref.*
AUC	$S = AUC = \sum_{k=1}^{n_0} AUC_k$ n_0 - number of individual values of gene x $AUC_k = \frac{(TPF(h_k,h_{k-1}))(FPF(h_k,h_{k-1}))}{2}$ $TPF(h_k,h_{k-1}) = TPF(h_k) + TPF(h_{k-1})$ $FPF(h_k,h_{k-1}) = FPF(h_k) + FPF(h_{k-1})$ TPF, FPF - true/false positive fraction h_k - unique values of x in decreasing ordered	[35]
ABCR	$S = ABCR = \sum_{k=1}^{n_0} \|AUC_k - A_k\|$ $A_k = \frac{2k-1}{2n_0^2}$, AUC - defined above	[35]
BWSS	$S = BW = \frac{\sum_i \sum_k (c_i=k)(\bar{x}_k - \bar{x})^2}{\sum_i \sum_k (c_i=k)(x_{k,i} - \bar{x}_k)^2}$	[36]
TNoM	$S = TNoM = \min_{d,t} Err(d,t\|x,c)$ $\mathrm{Err}(d,t\|x,c) = \sum_i 1, \{c_i \neq \mathrm{sign}(d(x-t))\}$ d, t - parameters defining the decision rule	[37]

reader is encouraged to consult the associated references in Table 9 for further details about these scoring functions.

4.2 Estimating Statistical Significance for Relevance Indices

Estimating the statistical significance for the relevance indices assigned to each feature/gene has been long addressed in the quest for *DEGs*. It is argued that statistical significance tests quantify the probability that a particular score or relevance index has been obtained by chance. It is common practice that features/genes ranked high in the list according to the relevance index, will be discarded if the computed scores are not statistically significant. There are different ways one can assign statistical significance to a test.[1] Despite many criticisms the most commonly used statistical significance test is the p-value. Many researchers advocate for alternative measures such as confidence intervals, especially due to the fact that p-values only bring evidence against a hypothesis (e.g., the null hypothesis of no "correlation" between features/genes and target annotation) and "confirm" a new hypothesis by rejecting the one which has been tested without bringing any evidence in supporting the new one [38]. Without entering into this debate, it is important to notice that statistical significance tests can be run either by exploring gene-wise information across all samples, either by exploring the large number of features in GEM experiments. Regardless the manner the statistical significance tests are performed, a permutation test is generally employed. It consists of running multiple tests which are identical to the original except that the target feature (or the class label) is permuted differently for each test. An important concept for estimating the statistical significance for *DEGs* discovery is the multiple hypothesis testing which will be described at the end of this section.

4.2.1 Exploring Feature-Wise Information to Asses Statistical Significance

This strategy assumes a large enough number of samples in order to infer upon the statistical significance of computed relevance indices of genes. The statistical significance is estimated for each feature/gene individually based on its intrinsic information.

***p*-values.** In statistics, the p-value is the probability of obtaining a test statistic (in our case a relevance index) at least as extreme as the one that was actually observed. The lower the p-value the more significant the result is (in the sense of statistical significance). Typical cutoff thresholds are set to 0.05 or 0.01 corresponding to a 5 or 1 percent chance that the tested hypothesis is accepted by chance. p-values can be estimated empirically by using a permutation test. However, standard asymptotic methods also exist, reducing substantially the computational time required by permutation tests. These methods rely on the assumption that the test statistic follows a particular distribution and the sample size is sufficiently large. When the sample size is not large enough, asymptotic results may not be valid, with the asymptotic p-values differing substantially from the exact p-values.

4.2.2 Exploiting the Power of Large Number of Features

An alternative strategy to overcome the drawback of the small number of samples in GEM experiments is to take advantage of the large number of features/genes [39]. In order to illustrate this idea we will consider the following: a GEM data set containing gene information about samples originating from two populations c_1 and c_2, and a filter algorithm to search for *DEGs* between c_1 and c_2. Let $S = S_1, \ldots, S_m$ be the relevance indices for all genes, let p_1 be the probability that a gene is discriminating between c_1 and c_2 and $p_0 = 1 - p_1$. Let also $f_1(S)$ be the *pdf* of S for discriminating genes and $f_0(S)$ the *pdf* of S for nondiscriminating genes. Then, we can write

$$f(S) = p_0 f_0(S) + p_1 f_1(S), \tag{1}$$

where $f(S)$ is the mixture of densities of discriminating/nondiscriminating genes.

The usefulness of the model in (1) depends on the estimation of the so called "null distribution" $f_0(S)$. In [40], Efron et al. proposed a method to estimate $f_0(S)$ based on a permutation test. What one is interested in, is the probability that a gene is differentiating between c_1 and c_2, which is p_1. According to [40] there are two strategies to obtain p_1: a bayesian and a frequentist approach. They will be mentioned briefly in the following.

A bayesian framework for estimating statistical significance. In [40], Bayes rule is applied to (1) to estimate the *aposteriori* probability $p_1(S_i)$ that a gene with score S_i is differentially expressed, resulting in

$$p_1(S_i) = 1 - p_0 \frac{f_0(S_i)}{f(S_i)}. \tag{2}$$

This approach can be summarized in four steps:

1. Estimate the relevance indices (scores) S_i.
2. Estimate null relevance indices (null scores) s_i using a permutation test.
3. Estimate the ratio $\frac{f_0(s)}{f(S)}$ based on the densities of S_i and s_i.

1. Here a test consists in verifying whether a feature/gene is informative for a target annotation and it is quantified by a relevance index.

TABLE 10
Univariate Parametric Methods

Name	*Scoring function*	*Statistical significance*	*Ref.*
Fold-change	Fold-change ratio	None	[45]
Regression model	t-test (z-score)	p-values with Bonferonni's correction	[17]
Golub	Welch t-test	p-values	[2]
ANOVA	t-test (z-score)	p-values	[16], [46]
ANOVA in bayesian framework	Bayesian t-test	p-values	[22]
Regularized t-test	Regularized t-test	p-values	[23]
Linear models (LIMMA)	Moderated t-statistics	p-values	[24]
Gene ranking with B statistics	B statistics	p-values	[25]
Gamma model	Similar to B statistics	p-values	[47]

4. Estimate the lower bound for p_1 according to

$$p_1 \geq 1 - \min_S \frac{f(S)}{f_0(S)}. \quad (3)$$

A frequentist framework. This approach relies on direct estimates of f_0 and f and it can be summarized as follows:

1. Estimate the relevance indices (scores) S_i.
2. Estimate null relevance indices (null scores) s_i using a permutation test.
3. Expected null relevance indices are computed according to

$$\bar{s}_i(b) = \frac{1}{B}\sum_{b=1}^{B} s_i(b), \quad (4)$$

where B is the total number of permutations.
4. Plot points $(\bar{s}_i, S_i)$.
5. For several threshold values t, estimate the number of true and false positives (TP, respectively, FP).
6. (Optional) Estimate the false discovery rate (FDR) for increasing values of t.

4.2.3 Multiple Hypothesis Testing Approach

The study of Dudoit et al. [41] was the first work describing the multiple hypothesis testing for GEM experiments in a statistical framework. In the context of *DEGs* discovery, multiple hypothesis testing is seen as *simultaneously testing for each gene the null hypothesis of no association between the expression level and the responses or target features* [41]. According to them, any test can result in two type of errors: false positive or Type I errors and false negative or Type II errors. Multiple hypothesis testing procedures aim to provide statistically significant results by controlling the incidence rate of these errors. In other words, provide a way of setting appropriate thresholds in declaring a result statistically significant. The most popular methods for multiple hypothesis testing focus on controlling Type I error rate. This is done by imposing a certain threshold α for the Type I error rate and then applying a method to produce a list of rejected hypothesis until the error rate is less than or equal with the specified threshold. Well-known methods for multiple hypothesis testing are as follows.

p-value with Bonferroni correction is an improved version of the classical p-value and consists in increasing the statistical threshold for declaring a gene significant by dividing the desired significance with the number of statistical tests performed [17].

False discovery rate (FDR) is a recent alternative for significance testing and has been proposed as an extension of the concept of p-values [42]. The FDR is defined as $FDR = E[\frac{F}{G}]$, where F is the number of false positive genes and G is the number of genes found as being significant. In order to overcome the situations where FDR is not defined (when $G = 0$), Storey [43] proposed a modified version of the FDR called positive false discovery rate (pFDR) defined as $pFDR = E[\frac{F}{G}|(G > 0)]$.

A less accurate alternative to the FDR for significance testing is the family-wise error rate (FWER) which is defined as the probability of at least one truly insignificant feature to be called significant.

q-value is an extension of FDR which has been proposed to answer the need of assigning a statistical significance score to each gene in the same way that the p-value does [44]. The q-value is defined as being the minimum pFDR at which a test may be called significant. The reader should be aware that the q-value can be defined either in terms of the original statistics or in terms of the p-values, see [43].

4.3 Ranking Methods for FS—Examples

In this section, we discuss and review ranking methods for FS by extending the taxonomy presented in Fig. 1.

4.3.1 Univariate Methods

According to [1], univariate methods for FS can be either parametric or nonparametric. Here, we provide a brief description of both groups.

Parametric methods. These methods rely on some more or less explicit assumption that the data are drawn from a given probability distribution. The scoring functions used to measure the difference in expression between groups of samples for each gene provide meaningful results only if this assumption holds. In particular, many researchers state that the t-test can be used to identify *DEGs* only if the data in each class are drawn from some normal distribution with mean μ and standard deviation σ. Candidates for this class of methods described are in Table 10.

TABLE 11
Univariate Nonparametric Methods

	Name	*Scoring function*	*Statistical significance*	*Ref.*
Model free	Rank-sum	Rank-sum	p-values	[12]
	Rank-products	Rank-product	q-values	[48]
	Between-within classes sum of squares	BWSS	none	[49]
	Relative entropy	Relative entropy	p-value	[28]
	Threshold number of miss-classifications	TNoM	p-value	[37]
	ABCR	ABCR	p-value	[35]
Random permutations	Significance Analysis of Microarrays (SAM)	Modified t-test	Observed relative difference vs. Expected relative difference	[19]
	Empirical Bayes Analysis (EBA)	Modified t-test	Empirical Bayesian inference	[21]
	Park	Kendall's rank CC	p-values	[33]
	Mixture Model Method (MMM)	Modified t-test	Likelihood ratio test (LRT)	[50]

Nonparametric methods. These methods assume by definition that the data are drawn from some unknown distribution. The scoring functions used to quantify the difference in expression between classes rely either on some estimates of the *pdfs* or on averaged ranks of genes or samples. Obviously, these methods have a higher generalization power but for most of them (especially those relying on estimates of the *pdfs*), the computational cost is higher.

In [1], univariate nonparametric filter techniques are split in two groups: pure model-free methods and methods based on random permutation associated to parametric tests. Pure model free methods use nonparametric scoring functions to assign a relevance index to each gene and then the statistical relevance of that index is estimated in terms of either p-value, FDR or q-value. Methods based on random permutations associated with a parametric test take advantage on the large number of genes/features in order to find genes/features which present significant changes in expression. In a first instance, they make use of a parametric scoring function to assign a relevance index to each gene and then employ a nonparametric statistical significance test to check for *DEGs*. The nonparametric significance test consists in comparing the distribution of relevance indices of genes estimated in the previous step and the null distribution of the test statistic (or relevance index). The null distribution of the test statistic is usually estimated using a permutation test. Table 11 lists the most well-known methods from this class.

4.3.2 Bivariate Ranking Methods

Ranking pairs of genes according to their discrimination power between two or more conditions can be performed either using a "greedy strategy" or "all pair strategy."

Greedy strategies. Methods in this group first rank all genes by individual ranking (using one of the criteria employed by univariate ranking methods); subsequently the highest scoring gene g_i is paired with the gene g_j that gives the highest gene pair score. After the first pair has been selected, the next highest ranked gene remaining g_s is paired with the gene g_r that maximizes the pair score, and so on. In [51], a greedy gene pair ranking method has been proposed where initially the t-test was employed to first rank genes individually while the pair score measures how well the pair in combination distinguishes between two populations. Concretely, the gene pair score is the t-test of the projected coordinates of each experiment on the diagonal linear discriminant (DLD) axis, using only these two genes. For further details we invite the reader to consult [51].

All pairs strategies. Unlike greedy pairs methods, all pairs strategies examine all possible gene pairs by computing the pair score for all pairs. The pairs are then ranked by pair score, and the gene ranking list is compiled by selecting nonoverlapping pairs, and selecting highest scoring pairs first. This method is computationally very expensive. A list of bivariate gene ranking methods is presented in Table 12.

5 Filter Methods—Space Search Approach

The second direction to create filters for FS is to adopt an optimization strategy which will come up with the most informative and least redundant subset of features among the whole set. This strategy implies three main steps described as follows:

1. Define a cost function to optimize.
2. Use an optimization algorithm to find the subgroup of features which optimizes the cost function.
3. Validate the selected subset of genes.

Wrappers and embedded methods also make use of "space search" strategies to select features, but as mentioned in Section 2 they are built around a classifier. The main difference between filter space search methods and wrappers/embedded methods is that the cost function is different: for filters the cost function is independent on any output of the classifier while for wrappers/embedded methods the cost function is in general the classifier's accuracy itself.

TABLE 12
Pair-Wise Ranking Methods

	Name	*Scoring function*	*Statistical significance*	*Ref.*
Pair-wise methods	Greedy t-test	t-test	none	[51]
	All pairs t-test	t-test	none	[51]
	Top-scoring pairs	ML voting rule	p-value	[52]
	Uncorrelated Shrunken Centroid (USC)	Modified t-test and pair-wise correlation	none	[53]

TABLE 13
Objective Functions for Space Search FS Methods

Name	*Metric*	*Ref.*
Group correlation coefficient	$GC(\Omega_s, c) = \frac{k r_{\Omega_s c}}{\sqrt{k+k(k-1) r_{\Omega_s \Omega_s}}}$, $k = \lvert\Omega_s\rvert$ $r_{\Omega_s c}$, $r_{\Omega_s \Omega_s}$ - the average feature to class/feature to feature correlation	[55]
Mutual information difference	$MID = \max_{i \in \Omega_m} I(x_i, c) - \frac{1}{\lvert\Omega_s\rvert} \sum_j I(x_i, x_j)$	[54]
Mutual information quotient	$MIQ = \max_{i \in \Omega_m} \frac{I(x_i, c)}{\frac{1}{\lvert\Omega_s\rvert} \sum_j I(x_i, x_j)}$ $I(x_i, c)$ - the mutual information between gene $x_i \in \Omega_m$ and class label c $I(x_i, x_j)$ - the mutual information between gene $x_i \in \Omega_m$ and gene $x_j \in \Omega_s$	[54]
Conditional entropy	$\Delta_{\Omega_s} = \sum_i P(x_i) D(P(c\lvert x_i) \Vert P(c\lvert x_j))$, $x_i \in \Omega_m$, $x_j \in \Omega_s$ $D(P(c\lvert x_i) \Vert P(c\lvert x_j))$ - the KL divergence between $P(c\lvert x_i)$ and $P(c\lvert x_j)$	[56]

Methods using space search strategies are not as numerous as the ranking methods and they are less popular, especially due to the optimization step which is often computationally expensive. Following we will briefly refer to Steps 1 and 2 in the generic algorithm mentioned above.

5.1 Objective Functions

As previously mentioned, these methods make use of an objective functions defined as a tradeoff between the maximum informativeness of the selected features/genes and their number (or minimum redundancy). Table 13 lists the objective functions we identified through the literature. As the search space might be too large for an exhaustive search, typically heuristic search algorithms are used for the optimization [54], [55].

5.2 A List of Existing Space Search Filter Methods

These methods are far less numerous than the ranking methods. They are all multivariate in the sense that, in order to identify the optimum subset of features they take into account not only the correlation between the features and the target annotation but also feature-feature correlation. This can result in better classification accuracy than the ranking methods but on the other hand, from the biomarkers discovery point of view, they are prone to filter out informative genes which might be of potential interest for biologists. Table 14 presents a list of optimization filter methods for informative genes discovery as identified in the literature.

6 On the Evaluation of Filter Methods

The evaluation of the selected subgroup of features also called *signature* is a mandatory step common to all FS methods. As we mentioned in the introduction, the selection of features/genes in GEM analysis is mainly performed for two reasons: class prediction/discovery and biomarkers identification. If the goal is class prediction/discovery, the evaluation is performed using some classifier dependent performance indices which are described further in this section. If the goal is the identification of informative features which are potentially useful for further investigations (biomarkers discovery), then the classification performances are ignored and the selected genes are evaluated individually by estimating the statistical significance of their relevance score. An important aspect on the evaluation of FS methods is the robustness or the stability of the signature defined as the variation in the FS results due to small changes in the data set [57] or as the variation resulted when different methods are used on the same data set. The changes in the data set can be considered either at the instance level (e.g., by removing or adding samples) or at the feature level (e.g., by adding noise to features) [57].

6.1 Evaluating the Prediction Power of the Signature

Class prediction is an important reason one is interested in building and using FS in gene expression analysis. In this context, the evaluation of FS methods is performed with respect to the performances of classifiers when the input data are described by the subset of selected features/genes. If two filter methods are compared, the same classifier is (or should be) used to estimate the accuracy or the prediction power of those signatures. Different methods to evaluate classifiers can be employed in this context: ROC analysis, accuracy, precision-recall curves. Note that wrapper and embedded methods guarantee accurate results only if the classifier used for FS is also used to predict new samples while filters can be used with a broader range of classifiers.

Received Operating Characteristic (ROC) analysis. ROC analysis is mainly used to measure and visualize the performances of classifiers. ROC graphs are built by plotting on the X axis the *false positive rate* (FPR), while the Y axis stands for the *true positive rate* (TPR). The values of FPR (the number of negatives incorrectly classified) and TPR (the

TABLE 14
Space Search Filter Methods

	Name	*Objective function*	*Ref.*
Exploring high order gene interactions	Correlation based FS (CFS)	Group Correlation Coefficient	[55]
	Minimum Redundancy Maximum Relevance (MRMR)	Mutual information difference/quotient	[54]
	Markov Blanket Filtering (MBF)	Conditional entropy	[56]

number of positives correctly classified) represent the output of a classifier and they are defined by $FPR = \frac{NNIC}{TNN}$ and $TPR = \frac{NPCC}{TNP}$, where $NNIC$ denotes the number of negatives incorrectly classified, $NPCC$ the number of positives correctly classified while TNN and TNP denote the total number of negatives and positives, respectively.

The terms *sensitivity* and *specificity* are often associated with the ROC curves ($sensitivity = TPR$ and $specificity = 1 - FPR$). For more comprehensive texts on ROC analysis, we invite the reader to consult [58].

Note that *Precision-Recall Curves* (PRC) can be used as an alternative to ROC curves for problems with unbalanced data between classes [59]. PRC space is built by plotting on X axis the $Recall$ which is the same as TPR, while Y axis stands for $Precision$ which measures the fraction of samples classified as positive that are truly positive.

Prediction power or accuracy. In contrast to ROC analysis, the prediction power or the accuracy is designed to quantify the performances of classifiers in a number. The classification accuracy is typically expressed in percentage as $accuracy = \frac{TNCC}{TNS} \times 100$, where $TNCC$ stands for the total number of correctly classified samples while TNS denotes the total number of samples. Alternatively one can use the error rate which is defined by $Err(\%) = 1 - accuracy$.

If the samples are unbalanced between the different classes, the prediction power or the accuracy of a classifier is less informative. To overcome this drawback, *balanced accuracy* defined as the mean value of sensitivity and specificity should be used instead [60].

6.2 Classifier Independent Evaluation of the Signature

It is well known that two different classifiers may report different results for the same input data. In order to remove as much as possible the classifier's influence on the evaluation process one solution is to evaluate the signature with respect to several classifiers, but this strategy demands extra computational effort. One way to avoid this inconvenience is to use classifier independent measures. These measures take into account only the intrinsic content of the selected subset of features and the target feature.

Group correlation coefficient is a classifier independent tool used for the evaluation of FS methods [61]. It is actually a tradeoff between the overall goodness of fit (computed as the average correlation between the subset of features/genes and the class label) and its redundancy (defined as the intracorrelation of the signature). The group correlation coefficient is defined in Table 13.

6.3 Evaluating the Robustness of the Signature

A different evaluation strategy in the context of FS is represented by the robustness or stability tests. We stress on the fact that stability tests are never used as a single evaluation measure and they should always be combined with some predictive indices. Several studies conducted for *DEGs* discovery show that many FS methods are highly dependent on the training set of samples [62], [63] resulting in lists of features which are unstable under variations in the training population. In this context, evaluating the robustness of the list becomes mandatory in declaring it as being relevant with respect to a target feature.

According to [64], the variability of the list depends on two aspects: first is the use of different scoring functions to select features/genes while the second one is due to the use of the same scoring function under slight variations in the data set. Stability tests should be performed to check for variability in the list originating from both sources.

As a common rule, stability tests compare two lists of a fixed number of features/genes by quantifying the size of the intersection between the two lists. A well-known method is the *percentage of overlapping genes* (POG). Several studies make use of the *POG* index to evaluate the reproducibility of the results for FS methods [65], [66]. It consists in comparing the lists of the top k most discriminating genes found by different methods and computing the percentage of genes found in all lists. Another example is *Correspondence At the Top* (CAT) [67] which is a visualization method that represents the proportion of overlap of the top p features/genes versus p. One pitfall of these two methods is the fact that the list of features must be of equal size while in practice one might be interested in comparing lists of different size. One method able to cope with this drawback is the overlap score described in [68].

As previously mentioned, stability tests are used for testing the variability of lists of features/genes both while using different scoring functions and while introducing perturbations in the data set. Stability tests under perturbations in the data set are generally performed via a subsampling-based strategy. A number of k subsamples of different size are drawn from the entire population of samples. FS is then performed on each of the k subsamples and a measure of robustness (e.g., *POG index*) is computed. In [57], the overall stability is defined in general terms as the average over all pair-wise comparisons between different signatures

$$R_{tot} = \frac{2\sum_{i=1}^{k}\sum_{j=i+1}^{k} R(f_i, f_j)}{k(k-1)}, \tag{5}$$

where f_i is the outcome of the FS method applied to subsample i and $R(f_i, f_j)$ is a similarity measure between f_i and f_j. For filter methods, this similarity measure could be the Spearman rank correlation coefficient. For a more detailed explanation on the stability tests for FS we invite the reader to consult [64].

7 Final Comments and Recommendations

As we mentioned in the end of Section 3, most filter methods presented in this survey are designed for supervised classification problems. However, several methods (those making use of scoring functions able to deal with continuous target annotations) can be used for regression problems as well, see Table 15.

From a practitioner point of view, the choice of a method could be a very difficult task without a solid experience in the field. Here, we provide some basic guidelines which hopefully will help the practitioner in choosing the appropriate method for his application. We focus on two types of recommendations: one concerning the choice of the scoring function and one for the choice of the statistical significance test (this is only valid for ranking methods). In the choice of the scoring function the following aspects should be considered: its complexity, the minimum number

TABLE 15
Summary of Scoring Functions for Filter FS: The Complexity is Proportional with the Number of Parameters in the Scoring Function; for the Minimum Number of Samples Required, One Star Denotes That the Scoring Function Has Been Proposed to Cope with Few Samples Data Sets; Class Label Feature May Be of Three Types: Binary (B), Multivalue (M) or Continuous (C)

Family	*Complexity*	*No. of samples*	*Class label variable type*			*Distribution of samples population*
			B	*M*	*C*	
Correlation based	*	**	+	+	+	Any
Ranking samples across features	*	**	+	-	-	Any
Fold change family	*	**	+	-	-	Normal
T-test family	*	**	+	-	-	Normal
Bayesian family	***	*	+	-	-	Different parametric assumptions
PDF based family	***	***	+	+	+	Any
Information theory family	**	**	+	+	+	Any
Objective functions for space search	**	**	+	+	+	Any
Others	**	**	+	+	-	Any

Some scoring functions are valid under some parametric assumptions on the population of samples which restricts their use in the case where the population of samples do not follows the assumptions.

of samples needed to obtain accurate results, the parametric assumptions and the variable type of the target annotation (binary, multivalued, continuous), see Table 15. In general, the practitioner will choose the scoring function according to the best tradeoff among these parameters as follows: it would be preferred the simplest scoring function, with the slightest parametric assumptions, according to the number of samples available and to the variable type of the target annotation. Concerning the statistical significance tests, the practitioner should be aware that more accurate results in terms of false discovery rates (features/genes declared informative but which are actually not) are obtained using more elaborate methods such as the multiple hypothesis testing approach but this will increase the overall complexity of the method. However, it is very often the case that simpler statistical significance tests (e.g., p-values) will provide comparable results. Nevertheless, for a trustful inference from the results, the practitioner should compare the outcome of several methods both in terms of prediction power and stability as mentioned in Section 6. The most frequently used methods for *DEGs* discovery seems to be Significance Analysis of Microarrays (SAM), Analysis of Variance (ANOVA), Empirical Bayes t-statistic, the Welch t-statistic, Fold-Change or the Rank Product which are commonly used in comparative studies [69].

Table 15 shows a brief comparison between scoring functions used to identify informative features/genes from GEM data in terms of complexity, minimum number of samples required, variable type of the target annotation and parametric assumptions. We assigned scores (stars from 1 to 3) to quantify the complexity of the different families of scoring functions according to the number of parameters in their formulation (one star denotes the simplest scoring functions) as well as the minimum number of samples required (one star denotes that the scoring function provides trustful results only with few samples). As one could be aware of, most of them are dedicated to binary (B) annotations. Multivalued (M) or continuous (C) annotations are also available in clinical annotations but these are more difficult to handle. The way one can deal with multivalued annotations is to generalize the methods designed for binary annotations by using a one-against-all strategy. Continuous annotations can be handled using PCCs or similar scoring functions (for regression problems). Parametric scoring functions (e.g., t-tests, fold change, etc.) provide meaningful results only if the parametric assumptions hold, while nonparametric scoring functions can be used in a more general framework, for any distributions or where the distribution of samples is unknown.

8 Conclusion

The paper aims to provide a comprehensible and as complete as possible survey on filter methods for FS. The literature surveyed covers exclusively the FS methodology for *DEGs* discovery from GEM data. This paper does not aim in any way to provide numerical evaluations of filter methods in terms of which one is the best, but to gather as much as possible domain knowledge about this particular topic. This upcoming task is a natural continuation of this work and comparative studies can easily be performed based on this material.[2]

The literature dedicated to filter methods for informative genes discovery is very vast and the differences between existing methods are often very subtle which can be easily seen from the multitude of related scoring functions used to propose different techniques for individual feature/genes ranking. This often renders the conceptual comparison of methods not so obvious.

In the presentation of the surveyed literature, we have been guided by a top-bottom strategy which is partially illustrated in the taxonomy in Fig. 1. From the definition of the problem we first situated the filter methods in the context of filter selection among the other groups of FS methods: wrapper, embedded, and ensemble. Consequently, we aimed to identify the most general characteristics proper to filter methods. Here, one has many choices but we decided to pursue our survey following two main strategies adopted by

2. Here we provide some additional information aiming to guide the readers interested in comparative studies of these techniques. Many filter methods for FS are available in Bioconductor R package and some also exist in Bioinformatics Toolbox in Matlab. For testing purposes, curated GEM data are available through the InSilicoDb R/Bioconductor package [70], developed inside the InSilico project (http://insilico.ulb.ac.be).

researchers to develop filter methods for FS: ranking and space search methods. Each one of these groups has been further presented in a synthetic manner by reviewing its own specific aspects. For the ranking methods we reviewed and categorized the scoring functions and we also presented the methods for estimating statistical significance while for the space search methods we presented the most used optimization functions. We completed the top-bottom strategy by resuming the different filter methods for FS, mentioning for each one of them the scoring function/statistical significance for ranking methods, respectively, objective function for space search methods.

In order to have a complete picture on the topic we mentioned the most common validation techniques which equally apply to all FS methods and in the end we formulated some guidelines aiming to help novel practitioners in choosing the appropriate method for their applications. We also provided a conceptual comparison between scoring functions for filters in Table 15.

ACKNOWLEDGMENTS

The authors would like to thank the Brussels Institute for Research and Innovation (INNOVIRIS) who funded this research and also the referees for their constructive comments.

REFERENCES

[1] Y. Saeys, I. Inza, and P. Larrañaga, "A Review of Feature Selection Techniques in Bioinformatics," *Bioinformatics,* vol. 23, no. 19, pp. 2507-2517, 2007.

[2] T.R. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science,* vol. 286, no. 5439, pp. 531-537, 1999.

[3] C.A. Penfold and D.L. Wild, "How to Infer Gene Networks from Expression Profiles, Revisited," *Interface Focus,* vol. 1, no. 6, pp. 857-870, 2011.

[4] R.L. Somorjai, B. Dolenko, and R. Baumgartner, "Class Prediction and Discovery Using Gene Microarray and Proteomics Mass Spectroscopy Data: Curses, Caveats, Cautions," *Bioinformatics,* vol. 19, no. 12, pp. 1484-1491, 2003.

[5] P. Yang et al., "A Review of Ensemble Methods in Bioinformatics," *Current Bioinformatics,* vol. 5, no. 4, pp. 296-308, 2010.

[6] I. Guyon, "An Introduction to Variable and Feature Selection," *J. Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[7] A.-C. Haury, P. Gestraud, and J.-P. Vert, "The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures," *PLoS ONE,* vol. 6, no. 12, p. e28210, 2011.

[8] M. Bansal et al., "How to Infer Gene Networks from Expression Profiles," *Moleculer Systems Biology,* vol. 3, p. 78, 2007.

[9] I. Guyon et al., "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning,* vol. 46, nos. 1-3, pp. 389-422, 2002.

[10] T. Zhang, "On the Consistency of Feature Selection Using Greedy Least Squares Regression," *J. Machine Learning Research,* vol. 10, pp. 555-568, 2009.

[11] S.-H. Cha, "Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions," *Int'l J. Math. Models and Methods in Applied Sciences,* vol. 1, no. 4, pp. 300-307, 2007.

[12] L. Deng et al., "A Rank Sum Test Method for Informative Gene Discovery," *Proc. 10th ACM SIGKDD Int'l Conf. Knowldge Discovery and Data Mining,* pp. 410-419, 2004.

[13] R. Breitling et al., "Rank Products: A Simple, Yet Powerful, New Method to Detect Differentially Regulated Genes in Replicated Microarray Experiments," *FEBS Letters,* vol. 573, nos. 1-3, pp. 83-92, 2004.

[14] D. Witten and R. Tibshirani, "A Comparison of Fold-Change and the t-Statistic for Microarray Data Analysis," technical report, Stanford Univ., 2007.

[15] H. Tao et al., "Functional Genomics: Expression Analysis of Escherichia Coli Growing on Minimal and Rich Media," *J. Bacteriology,* vol. 181, pp. 6425-6440, 1999.

[16] M.K. Kerr, M. Martin, and G.A. Churchill, "Analysis of Variance for Gene Expression Microarray Data," *J. Computational Biology,* vol. 7, no. 6, pp. 819-837, 2000.

[17] J.G. Thomas et al., "An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles," *Genome Research,* vol. 11, no. 7, pp. 1227-1236, 2001.

[18] S. Dudoit et al., "Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments," *Statistica Sinica,* vol. 12, pp. 111-139, 2002.

[19] V.G. Tusher, R. Tibshirani, and G. Chu, "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," *Proc. Nat'l Academy of Sciences USA,* vol. 98, no. 9, pp. 5116-5121, 2001.

[20] R. Tibshirani et al., "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proc. Nat'l Academy of Sciences USA,* vol. 99, no. 10, pp. 6567-6572, 2002.

[21] B. Efron et al., "Empirical Bayes Analysis of a Microarray Experiment," *J. Am. Statistical Assoc.,* vol. 96, no. 456, pp. 1151-1160, 2001.

[22] A.D. Long et al., "Improved Statistical Inference from DNA Microarray Data Using Analysis of Variance and A Bayesian Statistical Framework," *J. Biological Chemistry,* vol. 276, no. 23, pp. 19937-19944, 2001.

[23] P. Baldi and A.D. Long, "A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-Test and Statistical Inferences of Gene Changes," *Bioinformatics,* vol. 17, no. 6, pp. 509-519, 2001.

[24] G.K. Smyth, "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments," *Statistical Applications in Genetics and Moleculer Biology,* vol. 3, no. 1, pp. 1-25, 2004.

[25] I. Lönnstedt and T. Speed, "Replicated Microarray Data," *Statistica Sinica,* vol. 12, p. 31, 2001.

[26] E. Parzen, "On Estimation of a Probability Density Function and Mode," *The Annals of Math. Statistics,* vol. 33, no. 3, pp. 1065-1076, 1962.

[27] A. Wilinski, S. Osowski, and K. Siwek, "Gene Selection for Cancer Classification through Ensemble of Methods," *Proc. Ninth Int'l Conf. Adaptive and Natural Computing Algorithms (ICANNGA '09),* pp. 507-516, 2009.

[28] X. Yan et al., "Detecting Differentially Expressed Genes by Relative Entropy," *J. Theoretical Biology,* vol. 234, no. 3, pp. 395-402, 2005.

[29] J.-G. Zhang and H.-W. Deng, "Gene Selection for Classification of Microarray Data Based on the Bayes Error," *BMC Bioinformatics,* vol. 8, no. 1, article 370, 2007.

[30] L.-Y. Chuang et al., "A Two-Stage Feature Selection Method for Gene Expression Data," *OMICS: J. Integrative Biology,* vol. 13, pp. 127-137, 2009.

[31] R. Steuer et al., "The Mutual Information: Detecting and Evaluating Dependencies Between Variables," *Bioinformatics,* vol. 18, suppl. 2, pp. S23-S240, 2002.

[32] X. Liu, A. Krishnan, and A. Mondry, "An Entropy-Based Gene Selection Method for Cancer Classification Using Microarray Data," *BMC Bioinformatics,* vol. 6, article 76, 2005.

[33] B.M. Park PJ and M. Pagano, "A Nonparametric Scoring Algorithm for Identifying Informative Genes from Microarray Data," *Proc. Pacific Symp. Biocomputing,* pp. 52-63, 2001.

[34] L.J. van 't Veer et al., "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer," *Nature,* vol. 415, no. 6871, pp. 530-536, 2002.

[35] S. Parodi, V. Pistoia, and M. Muselli, "Not Proper Roc Curves as New Tool for the Analysis of Differentially Expressed Genes in Microarray Experiments," *BMC Bioinformatics,* vol. 9, no. 1, article 410, 2008.

[36] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *J. Am. Statistical Assoc.,* vol. 97, no. 457, pp. 77-87, 2002.

[37] A. Ben-Dor et al., "Tissue Classification with Gene Expression Profiles," *J. Computational Biology,* vol. 7, pp. 559-583, 2000.

[38] J. Cohen, "The Earth is Round ($p < .05$)," *Am. Psychologist,* vol. 38, pp. 997-1003, 1994.

[39] W. Pan, J. Lin, and C.T. Le, "A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data," *Functional and Integrative Genomics,* vol. 3, no. 3, pp. 117-124, 2003.
[40] B. Efron et al., "Microarrays and Their Use in a Comparative Experiment," technical report, Dept. of Statistics, Stanford Univ., 2000.
[41] S. Dudoit, J.P. Shaffer, and J.C. Boldrick, "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science,* vol. 18, no. 1, pp. 71-103, 2003.
[42] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. Royal Statistical Soc. Series B (Methodological),* vol. 57, no. 1, pp. 289-300, 1995.
[43] D. Storey, "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value," *Annals of Statistics,* vol. 31, pp. 2013-2035, 2003.
[44] J.D. Storey, "A Direct Approach to False Discovery Rates," *J. Royal Statistics Soc.: Series B,* vol. 64, no. 3, pp. 479-498, 2002.
[45] J. DeRisi, V. Iyer, and P. Brown, "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale," *Science,* vol. 278, no. 5338, pp. 680-686, 1997.
[46] S. Draghici et al., "Noise Sampling Method: An ANOVA Approach Allowing Robust Selection of Differentially Regulated Genes Measured by DNA Microarrays," *Bioinformatics,* vol. 19, no. 11, pp. 1348-1359, 2003.
[47] M.A. Newton et al., "On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes from Microarray Data," *J. Computational Biology,* vol. 8, pp. 37-52, 2001.
[48] R. Breitling et al., "Rank Products: A Simple, Yet Powerful, New Method to Detect Differentially Regulated Genes in Replicated Microarray Experiments," *FEBS Letters,* vol. 573, nos. 1-3, pp. 83-92, 2004.
[49] S. Dudoit, J. Fridlyand, and T. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *J. Am. Statistical Assoc.,* vol. 97, no. 457, pp. 77-87, 2002.
[50] W. Pan, "On the Use of Permutation in and the Performance of a Class of Nonparametric Methods to Detect Differential Gene Expression," *Bioinformatics,* vol. 19, no. 11, pp. 1333-13340, 2003.
[51] T. Bø and I. Jonassen, "New Feature Subset Selection Procedures for Classification of Expression Profiles," *Genome Biology,* vol. 4, no. 4, pp. research0017.1-research0017.11, 2002.
[52] D. Geman et al., "Classifying Gene Expression Profiles from Pairwise mrna Comparisons," *Statistical Applications in Genetics and Moleculer Biology,* vol. 3, pp. 1-19, 2004.
[53] K. Yeung and R. Bumgarner, "Multiclass Classification of Microarray Data with Repeated Measurements: Application to Cancer," *Genome Biology,* vol. 4, no. 12, p. R83, 2003.
[54] C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *J. Bioinformatics and Computational Biology,* pp. 185-205, 2005.
[55] Y. Wang et al., "Gene Selection from Microarray Data for Cancer Classification—A Machine Learning Approach," *Computational Biology and Chemistry,* vol. 29, no. 1, pp. 37-46, 2005.
[56] E.P. Xing, M.I. Jordan, and R.M. Karp, "Feature Selection for High-Dimensional Genomic Microarray Data," *Proc. 18th Int'l Conf. Machine Learning (ICML '01),* pp. 601-608, 2001.
[57] Y. Saeys, T. Abeel, and Y. Peer, "Robust Feature Selection Using Ensemble Feature Selection Techniques," *Proc. European Conf. Machine Learning and Knowledge Discovery in Databases,* pp. 313-325, 2008.
[58] T. Fawcett, "Roc Graphs: Notes and Practical Considerations for Researchers," technical report, 2004.
[59] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," *Proc. 23rd Int'l Conf. Machine Learning,* pp. 233-240, 2006.
[60] R. Powers, M. Goldszmidt, and I. Cohen, "Short Term Performance Forecasting in Enterprise Systems," *Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD '05),* pp. 801-807, 2005.
[61] A. Ben-Dor and Z. Yakhini, "Clustering Gene Expression Patterns," *Proc. Third Ann. Int'l Conf. Computational Moleculer Biology (RECOMB '99),* pp. 33-42, 1999.
[62] L. Ein-Dor et al., "Outcome Signature Genes in Breast Cancer: Is There a Unique Set?" *Bioinformatics,* vol. 21, no. 2, pp. 171-178, 2005.
[63] S. Michiels, S. Koscielny, and C. Hill, "Prediction of Cancer Outcome with Microarrays: A Multiple Random Validation Strategy," *Lancet,* vol. 365, no. 9458, pp. 488-492, 2005.
[64] A.-L. Boulesteix and M. Slawski, "Stability and Aggregation of Ranked Gene Lists," *Briefings in Bioinformatics,* vol. 10, no. 5, pp. 556-568, 2009.
[65] K. Kadota, Y. Nakai, and K. Shimizu, "Ranking Differentially Expressed Genes from Affymetrix Gene Expression Data: Methods with Reproducibility, Sensitivity, and Specificity," *Algorithms for Moleculer Biology,* vol. 4, p. 7, 2009.
[66] M. Zhang et al., "Evaluating Reproducibility of Differential Expression Discoveries in Microarray Studies by Considering Correlated Molecular Changes," *Bioinformatics,* vol. 25, no. 13, pp. 1662-1668, 2009.
[67] R.A. Irizarry et al., "Multiple-Laboratory Comparison of Microarray Platforms," *Nature Methods,* vol. 2, no. 5, pp. 345-350, 2005.
[68] X. Yang et al., "Similarities of Ordered Gene Lists," *J. Bioinformatics and Computational Biology,* vol. 4, no. 3, pp. 693-708, 2006.
[69] I. Jeffery, D. Higgins, and A. Culhane, "Comparison and Evaluation of Methods for Generating Differentially Expressed Gene Lists from Microarray Data," *BMC Bioinformatics,* vol. 7, no. 1, article 359, 2006.
[70] J. Taminau et al., "inSilicoDb: An R/Bioconductor Package for Accessing Human Affymetrix Expert-Curated Data sets from GEO," *Bioinformatics,* vol. 27, pp. 3204-3205, 2011.

Cosmin Lazar received the PhD degree in informatics, automatics and signal processing from the University of Reims Champagne Ardenne, France. He is currently working with CoMo Lab at Vrije Universiteit Brussel (VUB) as a postdoctoral researcher. His research interests include data mining, supervised/unsupervised learning, feature selection/extraction, blind source separation and their application in the analysis of GEM data, multi/hyperspectral images, and time series.

Jonatan Taminau received the advanced master's degree in bioinformatics. He is currently working toward the PhD degree at the Vrije Universiteit Brussel on the topic of large-scale analysis of microarray data.

Stijn Meganck received the PhD degree from the Vrije Universiteit Brussels (VUB), Belgium in 2008. Since then he has been working as a postdoctoral researcher at two research groups at the VUB: AI and ETRO. His main research interests include bioinformatics, probabilistic graphical models, and causality.

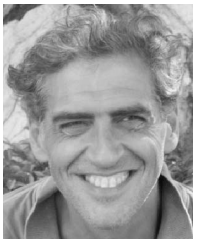
David Steenhoff received the master's degree in Sciences of Industrial Engineering in electronics and ICT in 2008. This was facilitated by the Erasmushogeschool Brussel, Vrije Universiteit Brussel and the Hanoi University of Technology. His research interests include machine learning and data mining applied in microarray gene expression analysis and hyperspectral imaging.

Alain Coletta received the graduate degree from the Université Libre de Bruxelles and received the PhD degree from Manchester University, Faculty of Engineering ad Physical Sciences, Advanced Interfaces Group.

Colin Molter received the PhD degree in artificial intelligence from the Université Libre de Bruxelles (ULB), Belgium in 2005. After receiving the PhD degree, he started the postdoctoral in computational neuroscience at the RIKEN-Brain Science Institute first and the Ecole Polytechnique Fédérale de Lausanne next. Recently, he became interested in genetics and started working on the inSilico project in Bruxelles.

Virginie de Schaetzen completed the medical studies from Universite Catholique de Louvain-la Neuve. She did the dermatology residency in the Hospital St Louis in Paris, the CHU of Montpellier, and the CHU of Liege in Belgium. She started working as a biocurator for the IRIDIA project in June 2010.

Robin Duque received the graduate degree in 2008 as master in sciences of industrial engineering in electromechanics followed by a master in management in 2009, both from Vrije Universiteit Brussel (VUB). Currently he is working as programming engineer/developer at IRIDIA laboratory, Universit Libre de Bruxelles (ULB).



Hugues Bersini received the MS degree in 1983 and the PhD degree in engineering in 1989 both from Université Libre de Bruxelles (ULB). He is now heading the IRIDIA laboratory (the AI laboratory of ULB) with Marco Dorigo. Since 1992, he has been an assistant professor at ULB and has now become full professor, teaching computer science, programming, and AI. Over the last 20 years, he has published about 250 papers on his research work which covers the domains of cognitive sciences, AI for process control, connectionism, fuzzy control, lazy learning for modeling and control, reinforcement learning, biological networks, the use of neural nets for medical applications, frustration in complex systems, chaos, computational chemistry, object-oriented technologies, immune engineering, and epistemology.

Ann Nowé received the MS degree from Universiteit Gent, Belgium, in 1987, where she studied mathematics with a minor in computer science, and the PhD degree from Vrije Universiteit Brussels (VUB), Belgium, in collaboration with Queen Mary and Westfield College, University of London, United Kingdom, in 1994. Currently, she is a full professor at the VUB and cohead of the Computational Modeling Lab. Her research interests include machine learning, including multiagent reinforcement learning, and bioinformatics.