A mathematical model for the validation of gene selection methods

Marco Muselli, *Member, IEEE,* Alberto Bertoni, Marco Frasca, Alessandro Beghini, Francesca Ruffino, and Giorgio Valentini

Abstract

Gene selection methods aim at determining biologically relevant subsets of genes in DNA microarray experiments. However, their assessment and validation represent a major difficulty since the subset of biologically relevant genes is usually unknown. To solve this problem a novel procedure for generating biologically plausible synthetic gene expression data is proposed. It is based on a proper mathematical model representing gene expression signatures and expression profiles through Boolean threshold functions. The results show that the proposed procedure can be successfully adopted to analyze the quality of statistical and machine learning-based gene selection algorithms.

Index Terms

Gene selection, feature selection, mathematical models, gene expression, Boolean functions.

1 INTRODUCTION

The identification of relevant influence factors on genomic data is based on the integration of multiple information sources, such as single nucleotide polymorphisms [1], array comparative genomic hybridization [2] and gene expression data [3], or a priori knowledge on pathways [4].

- *Marco Muselli is with IEIIT, Istituto di Elettronica, di Ingegneria dell'Informazione e delle Telecomunicazioni, Consiglio Nazionale delle Ricerche, Via De Marini 6, Genova, Italy,*
  E-mail: marco.muselli@ieiit.cnr.it
- *Alberto Bertoni, Marco Frasca, Francesca Ruffino, and Giorgio Valentini are with DSI, Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Via Comelico 39, Milano, Italy,*
  E-mail: {bertoni,frasca,ruffino,valentini}@dsi.unimi.it
- *Alessandro Beghini is with Dipartimento di Biologia e Genetica per le Scienze Mediche, Università degli Studi di Milano, Via Viotti 5, Milano, Italy*
  E-mail: alessandro.beghini@unimi.it

Considering the relevant impact of DNA microarray analysis on several fields of molecular biology and medicine [5], [6], [7], in this paper we focus on the computational aspects related to the analysis of gene expression data.

In this specific context, three classes of problems have been mainly studied within the community of bioinformaticians for answering three basic questions [8]:

1) Class prediction: how the functional state of an example can be determined through the expression level of its genes?
2) Gene selection: how the genes correlated to the functional state in exam can be identified?
3) Class discovery: what are the groups (clusters) of co-expressed and functionally correlated genes/examples?

In parallel with the development of algorithms for these problems, methods for testing the validity of their performances have been proposed. While effective validation procedures exist for the analysis of class prediction methods [9], [10], [11] and new techniques for the assessment of clustering solutions have been proposed [12], [13], [14], the evaluation of gene selection methods still remains a major problem.

From a computational standpoint gene selection is essentially equivalent to feature selection, which is known to be NP-hard [15]. Consequently, the task of finding an optimal feature set is unfeasible, even when the number of features is largely lower than the usual number of gene-expression values. Moreover, the set of biologically "relevant" genes that characterizes a given expression profile is usually unknown or only partially known: as a consequence, also from a biological standpoint the evaluation of the real effectiveness of gene selection methods is very difficult and in most cases unfeasible.

Finally, in [16], the authors showed that we need very large samples of gene expression data to select with high confidence the most relevant genes, and this is unfeasible with real data, where only relatively small samples are available instead.

Summarizing, these arguments show that there are strong impediments to evaluate feature selection algorithms using real gene expression data.

Most of the proposed approaches in literature indirectly evaluate gene selection methods by assessing the accuracy of the classifiers tested on the selected subsets of genes. According to this approach, a rigorous unbiased protocol to perform a fair comparison of gene selection techniques is proposed in [17] and an extensive application of the unbiased protocol to frequently used multivariate and univariate gene selection methods is provided in [18]. This approach is useful to correctly identify subsets of genes related to a specific phenotype, but is more focused on the evaluation of the performance of predictors of disease state, because the assessment is based on the estimation of the generalization error of the classifier, rather than on the direct estimation of the genes related to a specific disease.

By following a different approach, the problem of the assessment of gene selection methods has been considered by using massively parallel computation to find out optimal subsets of genes for a given real gene expression data set [19]. With this approach subsets of optimal genes are computed by exhaustive search and stored in databases, and made available for evaluating the performance of a given gene

selection method; the resulting genetic test bed has been proven useful and effective to evaluate gene selection methods, but, as observed by the authors, due to the intrinsic combinatorial nature of feature selection problems, only 70 genes previously established in [20] are considered, far below from the thousands of genes typical of a microarray.

A network generator that creates synthetic transcriptional regulatory networks and produces simulated gene expression data, modeled by equations based on Michaelis-Menten and Hill kinetics, has been developed, but its primary goal consists in testing and optimizing network inference methods [21]. Other artificial models have been previously proposed in the literature, but they did not explicitly consider the biological characteristics of gene expression data [22], [23], [24], resulting in a drastic simplification of the model.

We propose a different approach, by considering and developing an artificial model to generate data for the assessment of gene selection methods. Considering that available gene expression data are too complex and the available sample size not sufficient to infer their underlying distribution, our aim does not consist in reproducing true existing gene expression data. Our goal is limited to the generation of data that are biologically plausible and useful for the validation of gene selection methods, without pretending to recover the unknown distribution of a specific gene expression data set.

To this end we propose a mathematical model that describes the relationships between the expression levels of the genes of an artificial example and its phenotype, by taking into account the biological characteristics of gene expression data, according to the main results in the literature. Even if our goal does not consist in the simulation of a specific instance of gene expression data, we evaluated whether our model is also able to generate data similar to real ones. To this aim we applied a statistical test to assess the similarity between artificial and real data.

In this paper we modify and extend a preliminary model presented in [25]. More precisely, we introduce an explicit formal definition of expression signature and expression profile, using Boolean threshold functions and m-of-n expressions [26]. Moreover, we extend the stochastic component of the model and provide a mathematical characterization of the concept of gene relevance, and a modified version of the algorithmic procedure to generate synthetic gene expression data. In the Supplementary Information (http://homes.dsi.unimi.it/$\sim$valenti/SW/genemodel) we discuss in detail several characteristics of the proposed model, not considered in the previous work, such as the relationships between the relevance of variables representing genes and the compactness of the mathematical representation of expression profiles. We introduce also a new statistical analysis of the model, as well as new experiments to assess the performance of gene selection methods.

2 Biological characteristics of gene expression data

Several results published in the bio-medical and bioinformatics literature point out the main structures underlying gene expression data. Their analysis allows to derive a collection of specific characteristics,

which must be satisfied by an artificial model so as to produce biologically plausible data.

2.1 Profiles and expression signatures

In the bio-medical literature sets of biologically relevant and differentially expressed genes are named *expression signatures* [27], [28], [29], [30], [31]. To our knowledge, this term has been firstly introduced in [27] to characterize patterns found by gene expression profiling. More precisely, it refers to a group of genes coordinately expressed in a given set of specimens and in a specific physiological or pathophysiological condition.

The correlation among the mRNA levels of the genes is due to the underlying regulatory system, by which subsets of transcription factors (TFs) and binding sites may be directly or indirectly shared by the genes belonging to the same expression signature. It is worth noting that co-expression is in general not synonymous with co-regulation: co-regulation means having the same TF(s) regulating all the genes, while co–expression implies that several different TFs can be involved in the regulation, each one on a subset of the co-expressed genes [32].

Hence, a gene expression signature indicates a cluster of coordinately expressed genes, whose coordination reveals the fact that they may participate to the same biological process. Indeed, they are usually named by either the cell type in which their component genes are expressed, or by the biological process in which their component genes are known to function.

From this standpoint the overall *expression profile* of a specimen can be interpreted as a collection of gene expression signatures that reveal different biological features of the analyzed sample [27].

Expression signatures have been mainly discovered and analyzed in gene expression profiles of diseases. For instance, the expression profiling of B-cell malignancies through hierarchical clustering revealed expression signatures related to cell-proliferation, lymph-nodes, T-cells, Germinal Center B-cells (GCB) and others [27]. Independent Component Analysis performed on gene expression data from ovarian cancer tissues found gene expression signatures representing potential pathophysiological processes in ovarian tissue samples [28]. Expression profiling of rhabdomyosarcoma (RMS), the most common soft tissue sarcoma in children, identified two signatures associated with metastatic RMS, responsible for most of the fatal outcome of this disease [31], while two way hierarchical clustering analysis identified several expression signatures expressed in different types of bladder carcinoma [29].

Expression signatures have been also identified in species other than humans and in contexts not related to tumoral differentiation. For instance, comparative functional genomics based on shared patterns of regulations across orthologous genes identified shared expression signatures of aging in orthologous genes of *D. melanogaster* and *C. elegans* [30].

Since *expression profiles* and *expression signatures* are well-established biological structures that characterize gene expression data, they can be employed as the corner stones of our artificial model. To this aim, in the next subsection the main properties of gene expression signatures will be analyzed and discussed.

2.2 Characteristics of gene expression signatures

Differential expression and co-expression. Differential expression analysis of single genes, even if useful to identify specific genes involved in biological processes [33], cannot capture the complexity of tightly regulated processes, crucial for the proper functioning of a cell. Correlations among gene expression levels have been observed [34], [27], reflecting the fact that in most biological processes genes are co-regulated. As recently observed, not all the changes in co-regulation are manifested by up or down regulation of individual genes, and we need to explicitly consider interactions among genes to discover patterns in the data [35]. This corresponds to examine sets of co-regulated genes, i.e. expression signatures, to reveal functional relationships among genes.

Gene expression signatures as a whole rather than single genes contain predictive information. Many times is the signature taken as a whole that seems to contain predictive information for a biologically meaningful identification of tissue samples. For instance, it was found an expression signature of 8 upregulated and 9 downregulated genes associated with metastasis in different types of adenocarcinoma: none of these genes represents a marker, but it is the signature as a whole that represents a "collective marker" of tumor metastasis [36].

In other works it has been shown that in some cases relevant differences are subtle at the level of individual genes but coordinate in gene expression groups [35].

Genes may belong to different gene expression signatures at the same time and gene expression profiles may share common signatures. Each gene may be influenced by several transcription factors, each of which affects several genes [28], and many underlying conditions in a given sample may concur to define a gene expression signature (e.g. tumorigenesis, angiogenesis, apoptosis) [37]. Moreover, it has been shown that gene expression signatures may be shared and partially expressed in different gene expression profiles [27], [36], [38]. Hence, gene expression profiles of functionally different tissues may share some expression signatures, differing only for a subset of them. These expression signatures may be also partially represented (that is, not all the genes belonging to the expression signature are over-expressed or under-expressed), reflecting functional alterations in diseased patients.

Expression signatures can characterize subclasses of diseases. Several other works showed that a bio-molecular characterization of tumors can discover different subtypes of malignancies, not detectable with traditional morphological and histopathological features (see e.g. [27], [39]). For instance, an expression signature of 153 genes can be used to correctly classify hepatocellular carcinoma (HCC) intra-hepatic metastasis from metastatic-free HCC [38]. This expression signature, that embeds high predictive information, has been shown to be independent of tumor size, tumor encapsulation and patient age, but very similar to that of their corresponding metastases.

2.3 Modeling issues

In the light of the characteristics of gene expression signatures described in the previous section, we can identify the following main issues, which must be taken into account in the construction of a biologically plausible artificial model for gene expression data:

1) Expression profiles (*EP*) may be characterized as a set of gene expression signatures (*ES*), which uniquely determines a *functional group* of samples. Thus, the model should allow us to define expression profiles in terms of expression signatures, ensuring a large flexibility with respect to the number and the kind of genes composing the synthetic expression signatures.
2) Expression signatures are interpreted in the literature as a set of modulated genes (up-regulated or down-regulated) with respect to a particular condition. Accordingly, in the model, each expression signature should be defined as a set of up-regulated or down-regulated genes, that represent genes with expression levels above or below a given threshold. We say that a signature is *active* if its genes are coordinately modulated.
3) The number of genes within an expression signature usually vary from few units to few hundreds and genes may belong to different signatures at the same time. Accordingly, the model should permit to select the number of elements for each expression signature and to assign the same gene to different signatures.
4) Apart from technical variation (that in principle should be detected and reduced by proper design and implementation of bio-technological experiments and suitable pre-processing procedures [40]), gene expression is biologically variable also within functional classes (conditions) [41]. Moreover, gene expression variation among individuals may introduce variation into expression signatures. Hence, the model should permit to introduce flexibility in the number of genes that can be underexpressed or overexpressed, as well as to introduce individual variability within a functional group.
5) Some signatures may be only partially expressed within a particular expression profile. Accordingly, the model should be sufficiently flexible to allow different ways of constructing an expression profile. For instance, it must provide for signatures that may or may not be expressed, as well as for "mandatory" signatures, whose activation is necessary for a given expression profile.

3 METHODS

3.1 A mathematical model for gene expression

On the basis of the biological analysis presented in Sec. 2 and, in particular, starting from the concepts of expression profile, expression signature and gene modulation, we aim to develop a model describing the relationship between the functional state or phenotype of a tissue and the expression level of its genes.

To this end let us consider:

- A set $A = \{g_1, \ldots, g_r\}$ containing all the genes whose expression level will be analyzed; the vector $\boldsymbol{x} = [x_1, \ldots, x_r] \in \mathbb{R}^r$ is a real vector where x_k represents the expression level of the gene g_k;
- c functions $f_i : \mathbb{R}^r \to \{0, 1\}, i = 1, \ldots, c$, where f_i represents the i^{th} of c possible functional states collected in the set $S = \{s_1, \ldots, s_c\}$; we have that $f_i(\boldsymbol{x})$ assumes the value 1 if the tissue whose gene expression level is represented by the vector $\boldsymbol{x}$ is in the functional state s_i and 0 otherwise;
- A random noise N.

By following the concept of gene modulation introduced in the previous section, we suppose that, for each gene g_i belonging to A, there exists a modulation threshold t_i such that g_i is *overexpressed (OE)* if the value x_i of its expression exceeds t_i, or alternatively is *underexpressed (UE)* if $x_i < -t_i$.

A gene is considered to be *modulated* if it is OE or, alternatively, UE. In particular, we suppose that a function $\Psi : S \times A \to \{\text{OE}, \text{UE}\}$ exists that determines, for each functional state, if a gene is modulated when OE or when UE. Thus, $\Psi(s_i, g_j) = \text{OE}$ means that the gene g_j, relatively to the functional state s_i, is modulated when overexpressed.

By starting from Ψ and by considering the modulation thresholds t_j, we can define, for each functional state $s_i \in S$, a mapping $\boldsymbol{\beta}_i : \mathbb{R}^r \to \{0, 1\}^r$ that returns, for each gene g_j, the value 1 if that gene is modulated and 0 otherwise.

Hence, $\beta_i(x_1, \ldots, x_r) = (z_1, \ldots, z_r)$, where

$$z_j = \begin{cases} 1 & \text{if } \Psi(s_i, g_j) = \text{OE and } x_i > t_i \\ & \text{or} \\ & \Psi(s_i, g_j) = \text{UE and } x_i < -t_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

A biologically plausible assumption for the model implies that the output is uniquely determined by the modulation (present or not) of the r genes and does not depend on their specific expression values. Then, each function f_i can be written as $f_i(\boldsymbol{x}) = \varphi_i(\boldsymbol{\beta}_i(\boldsymbol{x}))$, where φ_i is a suitable Boolean function defined on binary strings in $\{0, 1\}^r$. Consequently, once the mappings $\boldsymbol{\beta}_i$ are completely described, the components f_i of our model are uniquely determined by the construction of the Boolean functions φ_i.

Without loss of generality, we consider a unique functional state s represented by the function $f(\boldsymbol{x}) = \varphi(\beta(\boldsymbol{x}))$. By taking into account the biological specifications presented in Sec. 2.3, we will show that the function φ can be represented by means of the biologically motivated concepts of expression signature and expression profile.

Formally, we introduce the following

Definition 1: Given the set of Boolean variables $\{z_1, \ldots, z_r\}$ associated with the genes $\{g_1, \ldots, g_r\}$, an *expression signature* (*ES*) is a couple

$$S = < \{z_{j_1}, \ldots, z_{j_l}\}, q >$$

where $\{z_{j_1}, \ldots, z_{j_l}\} \subset \{z_1, \ldots, z_r\}$ and $q \leq l$ is a positive integer.

An expression signature S defines in a natural way the Boolean threshold function $I_S(z_1, \ldots, z_r) = \theta(\sum_{k=1}^{l} z_{j_k} - q)$, where $\theta(x)$ assumes value 1 if $x \geq 0$, and 0 otherwise. We say that the expression signature S is *active* if $I_S(\boldsymbol{z}) = 1$, i.e. if at least q components of $\{z_{j_1}, \ldots, z_{j_l}\}$ have value 1 and *not active* otherwise. The Boolean function $\theta(\sum_{k=1}^{l} z_{j_k} - q)$ is sometimes called m-of-n expression [26].

Suppose, for example, that $S =< \{z_1, z_2, z_3\}, 2 >$ and $r = 3$. Then, $I_S(\boldsymbol{z}) = 0$, i.e. S is not active, for $\boldsymbol{z} = (z_1, z_2, z_3) = (1, 0, 0)$, while $I_S(\boldsymbol{z}) = 1$, i.e. S is active, for $\boldsymbol{z} = (1, 1, 0)$ or $\boldsymbol{z} = (1, 0, 1)$.

In a similar way we give a definition of expression profile.

Definition 2: An *expression profile* (*EP*) is a couple

$$P =< \{S_1, \ldots, S_h\}, s >$$

where $S_1, \ldots, S_h$ are h ESs and $s \leq h$ is a positive integer.

Thus, the function $I_P(\boldsymbol{z}) = \theta(\sum_{k=1}^{h} I_{S_k}(\boldsymbol{z}) - s)$ can be defined and the expression profile P is considered *active* if at least s of h expression signatures are active.

In order to simplify the notation, instead of $< \{z_{j_1}, \ldots, z_{j_l}\}, q >$ we will write $(z_{j_1}, \ldots, z_{j_l})_q$ that will be also used to denote the function $\theta(\sum_{k=1}^{l} z_{j_k} - q)$.

Extending this notation to expression profiles, the function φ representing P can be written as:

$$\varphi(\boldsymbol{z}) = [(z_{j_{11}}, \ldots, z_{j_{1l_1}})_{q_1}, \ldots, (z_{j_{h1}}, \ldots, z_{j_{hl_h}})_{q_h}]_s \tag{2}$$

Finally, if we combine the representation of the function φ defined in equation (2) and the mapping $\boldsymbol{\beta}$ defined in equation (1), the function $f(\boldsymbol{x}) = \varphi(\boldsymbol{\beta}(\boldsymbol{x}))$ can be written in an explicit form. It is sufficient to replace the components of the vector $\boldsymbol{z}$ belonging to the expression profile with the mapping $\boldsymbol{\beta}$ that determines their values. For example, if the gene g_i is modulated when overexpressed we can replace z_i with $x_i > t_i$, otherwise we replace z_i with $x_i < -t_i$. As an example, suppose that

$$\varphi(\boldsymbol{z}) = [(z_1, z_2, z_3)_2, (z_4)_1, (z_5)_1]_2 \tag{3}$$

and that $\boldsymbol{\beta}$ is defined as follows

$$z_1 = \beta_1(\boldsymbol{x}) = \begin{cases} 1 \text{ if } x_1 > 2 \\ 0 \text{ otherwise} \end{cases}$$

$$z_2 = \beta_2(\boldsymbol{x}) = \begin{cases} 1 \text{ if } x_2 < -3 \\ 0 \text{ otherwise} \end{cases}$$

$$z_3 = \beta_3(\boldsymbol{x}) = \begin{cases} 1 \text{ if } x_3 > 1 \\ 0 \text{ otherwise} \end{cases}$$

$$z_4 = \beta_4(\boldsymbol{x}) = \begin{cases} 1 \text{ if } x_4 > 3 \\ 0 \text{ otherwise} \end{cases}$$

$$z_5 = \beta_5(\boldsymbol{x}) = \begin{cases} 1 \text{ if } x_5 < -1 \\ 0 \text{ otherwise} \end{cases}$$

Then, $f(\boldsymbol{x}) = \varphi(\boldsymbol{\beta}(\boldsymbol{x}))$ can be written as:

$$f(\boldsymbol{x}) = [(x_1 > 2, x_2 < -3, x_3 > 1)_2, \\ (x_4 > 3)_1, (x_5 < -1)_1]_2 \tag{4}$$

3.2 Mathematical characterization of the concept of gene relevance

In this section we show that the set of relevant genes, i.e. those genes whose expression can contribute to change the functional state of a tissue, may be characterized through positive Boolean functions, defined as follows.

Definition 3: A Boolean function $\varphi : \{0,1\}^r \rightarrow \{0,1\}$ is called *positive* if and only if $\boldsymbol{u} \leq \boldsymbol{z}$ implies $\varphi(\boldsymbol{u}) \leq \varphi(\boldsymbol{z})$ for all $\boldsymbol{u}, \boldsymbol{z} \in \{0,1\}^r$.

Moreover, we prove that each positive Boolean function can be described by using only its relevant variables; hence, an EP can be represented by employing only the subset of relevant genes.

The following result characterizes the class of Boolean functions that can be represented by expression profiles.

Fact 1: The class of the positive Boolean functions coincides with the class of the Boolean functions representable through an expression profile.

The advantage of using the expression profile representation instead of the classical PDNF (Positive Disjunctive Normal Form) representation is discussed in the Supplementary Information.

Then, the following definitions allow to introduce the mathematical concept of gene relevance.

Definition 4: For $\boldsymbol{z} \in \{0,1\}^r$ and $\varphi : \{0,1\}^r \rightarrow \{0,1\}$ we say that the j^{th} variable z_j is *pivotal* if flipping the value of z_j changes the value of φ.
Formally, if we define

$$\sigma_j(z_1, \ldots, z_{j-1}, z_j, z_{j+1}, \ldots, z_r) \\ = (z_1, \ldots, z_{j-1}, 1 - z_j, z_{j+1}, \ldots, z_r)$$

we can say that the j^{th} variable is pivotal if $\varphi(\sigma_j(\boldsymbol{z})) \neq \varphi(\boldsymbol{z})$.

Definition 5 ([42]): The *influence* of the j^{th} variable z_j on a Boolean function φ, denoted with $I_j(\varphi)$, is the probability that z_j is pivotal, i.e.

$$I_j(\varphi) = \mathbb{P}\{\varphi(\sigma_j(\boldsymbol{z})) \neq \varphi(\boldsymbol{z})\}$$

where $z_1, \ldots, z_r$ are supposed to be independent, and $\mathbb{P}\{z_j = 1\} = \mathbb{P}\{z_j = 0\} = 1/2$ for every $j = 1, \ldots, r$.

We call *relevant* the variables z_i whose influence $I_i(\varphi)$ is strictly positive; the other variables are called *irrelevant*. From a biological standpoint the set of relevant genes are those differentially expressed that are significant for the determination of a given phenotype.

Note that the concept of gene (variable) relevance has a further notable biological meaning. As a matter of fact, the biological idea that each functional state could be represented through an expression profile

containing all and only the genes relevant for that state is formalized by the following mathematical result:

Fact 2: Every positive Boolean function $\varphi : \{0,1\}^r \rightarrow \{0,1\}$ can be represented through an expression profile that contains all and only relevant variables (genes).

This fact is discussed in the Supplementary Information.

3.3 Simulation of gene expression data

By adopting the mathematical model introduced in Sec. 3.1, a procedure for building artificial gene expression data to evaluate feature selection methods can be derived.

Every example is given by a pair $(\boldsymbol{x}, y)$, where $\boldsymbol{x}$ is a real-valued vector, whose components represent the gene expression levels, and the output y can vary into a set of c different values, each one denoting the class, i.e. a phenotype of the corresponding sample.

In the following, for the sake of simplicity, we consider the case $c = 2$.

The relationship between the gene expression values of an example and its functional state is realized by using the model described in Sec. 3.1.

We introduce only the main logical steps underlying the procedure to generate artificial gene expression data; for more details see the Supplementary Information.

I. *Model synthesis*. Two deterministic functions $f_1 = \varphi_1 \circ \beta_1$ and $f_2 = \varphi_2 \circ \beta_2 : \mathbb{R}^r \rightarrow \{0,1\}$, where r is the total number of virtual genes, are defined. φ_1 and φ_2 are specified by expression profiles, as in eq. (2), while β_1 and β_2 are defined as in eq. (1). The parameters necessary for the construction of the two functions (i.e. the number of ESs, the values q_i and s, the values of the thresholds (Sec. 3.1)) are determined by the user.

II. *Virtual gene expression data generation*. In the second step n r-dimensional real vectors are generated so that the first n_1 examples belong to the class 1, i.e. verify $f_1(\boldsymbol{x}) = 1$ and $f_2(\boldsymbol{x}) = 0$, while the remaining $n_2 = n - n_1$ vectors belong to the class 2, i.e. verify $f_1(\boldsymbol{x}) = 0$ and $f_2(\boldsymbol{x}) = 1$. The examples are built so that the components corresponding to genes that do not belong to the EPs are drawn according to normal distributions with mean and variance chosen by user. On the other hand, the components corresponding to relevant genes are chosen to ensure that the number of modulated genes is sufficient to make the EP verified. More details can be found in the Supplementary Information.

III. *Noise addition*. In the last step the vectors $\boldsymbol{x}$ and the labels y are perturbed according to specific probability distributions, by simulating both the errors in data acquisition and the uncertainty in the attribution of the correct class to the tissues. In particular, a parameter p, corresponding to the probability that an example is assigned to the wrong state, is added. In this way if $f_1(\boldsymbol{x}) = 1$, i.e. if $\boldsymbol{x}$ belongs to the class 1, then $y = 1$, with probability $1 - p$ and $y = 2$ with probability p. Similarly, if the example $\boldsymbol{x}$ verifies $f_2(\boldsymbol{x}) = 1$, i.e. it belongs to the class 2, we assign it to the class 1 with probability p.

Finally, possible errors in data acquisition are simulated by adding to each component of the examples a noise value generated according to a normal distribution with mean zero and variance chosen by the user.

Note that for most of the algorithmic procedure parameters reasonable default values are provided, but users can supply their own parameter values to obtain specific expression profiles.

4 RESULTS AND DISCUSSION

In this section we propose a non parametric statistical test to assess the adherence of the artificial data to the actual gene expression datasets to be analyzed. This test have been successively applied to four real cases, after properly tuning the parameters of the model. Then we present two procedures to compare and evaluate several gene selection methods using both real and artificial data sets.

4.1 A non-parametric test to compare distributions

To this end, we employed the non-parametric test proposed in [43], which is based on the comparison of distribution functions. In particular we applied the *Diggle test* using the distribution of the t-statistics of the genes. This approach is motivated by two basic reasons:

1) the high dimensionality of the data prevents a direct comparison of the tables;
2) the t-statistics preserves the relevant information on the genes differentially expressed in the analysis of gene expression data.

The test has been implemented as follows. A real dataset has been considered and, for each gene g_i, with $i = 1, \ldots, r$, the following value t_i is calculated

$$t_i = \frac{\bar{x}_i^1 - \bar{x}_i^2}{\sqrt{\frac{s_i^1}{n^1} + \frac{s_i^2}{n^2}}}$$

where $\bar{x}_i^1$, $\bar{x}_i^2$ and s_i^1, s_i^2 are the sample mean and variance of the i^{th} gene belonging respectively to the class 1 and 2 (that is to the two different functional states), and n^1, n^2 are the number of examples belonging to the two classes. We denote with $\hat{G}_0$ the empirical cumulative distribution function (cdf) of the values t_i obtained from real gene expression data:

$$\hat{G}_0(t) = \frac{|\{t_i : i = 1, \ldots, m, t_i \leq t\}|}{r}$$

Then, ν artificial datasets having the same dimension of the real one are created by using the same set of input parameters. For each dataset the corresponding cdf $\hat{G}_j$, for $j = 1, \ldots, \nu$, is obtained.

Denote with

$$\overline{G}_j(t) = \frac{1}{\nu} \sum_{k \neq j} \hat{G}_k(t) \qquad j = 0, \ldots, \nu$$

the mean of the cdfs when the j^{th} empirical distribution is excluded and with

$$u_j = \int (\hat{G}_j(t) - \overline{G}_j(t))^2 dt \qquad j = 0, \ldots, \nu$$

a measure of the difference between the jth cdf and the mean of the ν remaining cdfs.

The null hypothesis (no difference between the t distribution of real and synthetic data) can be restated as follows: u_0 is drawn from the same distribution of $u_1, \ldots, u_\nu$. More precisely, we can sort in decreasing order the components of the vector $(u_0, \ldots, u_\nu)$ and denote with $\mathrm{rank}(u_0)$ the position of the term u_0 in this sorted list. It can be seen that the *p-value* amounts to $\frac{\nu+1-rank(u_0)}{\nu+1}$; thus, we can reject the null hypothesis at $\alpha = \frac{\nu+1-i}{\nu+1}$ significance level if and only if $rank(u_0) \geq i$, for $0 \leq i < \nu$.

4.2 Test Results

In this section we apply the Diggle test to verify whether the proposed model can simulate four real gene expression data published in the literature, and available on the web:

- *Colon cancer* consists of 62 tissues, 22 healthy and 40 struck by colon cancer [44]. For each tissue 2000 genes have been considered.
- *Leukemia* collects 72 samples, subdivided in 47 ALL (Acute Lymphoblastic Leukemia) and and 25 AML (Acute Myeloid Leukemia) cases of cancer of the white blood cells [39]. Each sample is represented by the expression level of 7129 genes.
- *DLBCL-FL* consists of 77 samples, 58 Diffuse Large B-cell Lymphoma (DLBCL) and 19 Follicular Lymphoma (FL) specimens, with 6285 analyzed genes for each sample [45].
- *DLBCL-outcome*: all specimens refer to DLBCL patients. The 58 patients are subdivided according to the clinical outcome: 32 are cured and 26 fatal/refractory patients for whom the therapy was unsuccessful [45]. Each sample is represented by the expression level of 6285 genes.

Filtering and gene normalization of the above gene expression data have been performed according to the procedures described in [44], [39], [45].

We set the number of artificial datasets to $\nu = 1000$ and chose the input parameters to simulate the above data sets by looking at the expression signatures found in [44], [39], [45], by data inspection, and by heuristic optimization (the parameters values are available in the Supplementary Information).

By applying the Diggle test, for all the considered data sets we cannot reject the null hypothesis at 0.01 significance level: in other words we cannot reject the hypothesis that u_0, obtained from real data, comes from the same distribution of $u_1, \ldots, u_{1000}$, obtained from the synthetic data. Hence, according to the Diggle test, there is no significant statistical difference between the t-statistic distributions in real and simulated synthetic data.

In order to strengthen the validity of our model we applied the Diggle test to 5 artificial datasets, generated according to simple multivariate Gaussian distributions (with different means and covariance matrices, see Supplementary Information for more details). For each data set, whose dimension is the

same of the Colon cancer data set, 1000 instances have been realized. The results reveal that, for all the artificial data sets, the value u_0 exceeds the values u_j ($j \in \{1, \ldots, 1000\}$) of an order of magnitude, and according to the Diggle test, we can reject the null hypothesis that real and artificial data are drawn from the same distribution with very high confidence.

4.3 Comparison of the behaviour of gene selection methods on real and synthetic data

In this section we study whether genes selected using real and corresponding simulated gene expression data are comparable, by applying different gene selection methods. For our experiments we used the real and simulated gene expression data described in Sec. 4.2. We chose different statistical, linear regression and machine learning based methods: t-test, Wilcoxon rank sum test, variable selection through Lasso [46], Elastic net [47] and random forests [48]. We investigated whether the percentage of genes jointly selected by each pair of gene selection methods is comparable in both real and synthetic data.

Fig. 1 shows the results obtained with *Colon* and *Leukemia* data sets. Each point represents the percentage of genes jointly selected by a specific pair of gene selection methods with synthetic (abscissa) and real (ordinate) gene expression data. Each of the 10 different pairwise comparisons has been repeated 5 times for each data set, thus resulting in 50 points for each graph. The compared percentages are relative to the first 200 (Fig. 1 a) and c)) and 400 (Fig. 1 b) and d)) top ranked genes.

Note that all the points lie straight along the bisector: this means that, independently of the considered pair of gene selection methods, we obtain approximately the same percentage of commonly selected genes with the real and the corresponding simulated artificial gene expression data. These results show that the behavior of gene selection methods is similar in both situations and consequently we can use data simulated through our model to analyze the performance of gene selection methods. More detailed results relative to all the considered data sets (comprising the two *DLBCL*s) are available in the Supplementary Information.

4.4 Experimental assessment and comparison of feature selection methods

We propose a simple procedure for evaluating and comparing feature selection methods, although other different approaches can be applied by using the synthetic data obtained with our mathematical model. Suppose we want to compare and evaluate H feature selection methods F_i, with $i = 1, \ldots, H$, on a single artificial dataset A created by using the model described in Sec. 3.

Denote with R the set containing all the relevant genes for the two states of interest included in the two distinct expression profiles. If r is the total number of genes, we have $|R| = k \leq r$. The result produced by each feature selection method F_i consists of a list L_{ik} containing the k more relevant genes identified by the method F_i. We can evaluate the performance of the method F_i as the fraction p_i of overlapping between the set of the true relevant genes R and the set L_{ik} of the genes identified by the method F_i.

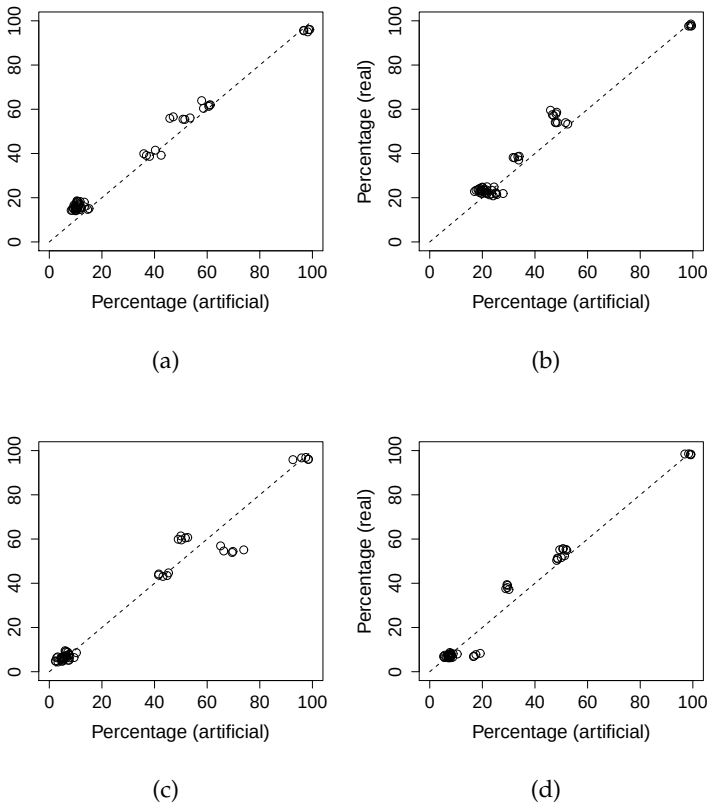

Fig. 1. Comparison of the percentage of common genes obtained with synthetic (abscissa) and real (ordinate) data for all the pairwise combinations of gene selection methods. *Colon*: a) 200 and b) 400 top ranked genes. *Leukemia*: c) 200 and d) 400 top ranked genes.

$$p_i = \frac{|R \cap L_{ik}|}{k} \ , \ i = 1, \ldots, H \tag{5}$$

We can compare the performances of the H methods by ordering the p_i values, for $i = 1, \ldots, H$. In particular, the higher is the value p_i, the more effective is the corresponding method F_i.

Since this comparison, based on a single data set, may lead to unreliable conclusions, the procedure can be repeated for a fixed number T of different instances A_j, with $j = 1, \ldots, T$, obtained from the same mathematical model. Thus, for each method F_i, we obtain a vector $\boldsymbol{p}_i$ whose j^{th} component p_{ij}, $j = 1, \ldots, T$, is the overlapping between the relevant genes of the j^{th} dataset A_j, and the set of the first $k_j = |R_j|$ genes obtained by applying the method F_i on A_j. By using the same approach, gene selection methods can be compared also starting from synthetic data obtained with different input parameters.

As an example, we compare two feature selection techniques, the Golub method [39] and SVM-RFE [49], by using 100 artificial datasets obtained with the parameters satisfying the Diggle test (Sec. 4.2) for

the Colon dataset. To compare the two methods we simply count how many times the Golub method achieves a better performance with respect to SVM-RFE, i.e. how many times the p_is computed according to eq. (5) are larger for the Golub technique with respect to SVM-RFE. In this way, we obtain that the Golub method outperforms SVM-RFE on $57/100$ instances of synthetic data. The mean and the variance of the p_i computed across the 100 gene selection experiments are respectively 0.81 ± 0.01 and 0.71 ± 0.08 for Golub and SVM-RFE method (see Supplementary Information for full details on the results). Both results confirm that, with synthetic Colon-like data, the simple Golub univariate statistical method works better than the more complex SVM-RFE approach.

4.5 Limitations and relevance of the proposed method

Recently, new bio-technologies have made several sources of data available for gene selection and for other relevant problems in functional genomics: new array types (e.g. exon or SNP arrays) [50], [51], transcription factors, pathway or proteomics data [52], [4], till to ultra high-throughput sequencing technologies [53] have opened new perspectives for the analysis of genomic data, and methods for the integration of multiple sources of data represent a key issue in functional genomics [54], [55]. From this standpoint, searching for influence factors affecting gene expression should be realized by integrating multiple sources of data to achieve more reliable results and to take into account of factors not detectable using only a single source of data.

Nevertheless, gene expression analysis through DNA microarrays represents a mature bio-technology with relevant applications in molecular biology and medicine, as witnessed by the wide literature available on these topics and in particular by very recent publications on newly proposed gene selection methods for DNA microarray data analysis [56], [57], [58]. In this context, our proposed method has been conceived to compare the performance of existing gene selection methods based on DNA microarray data analysis. It is worth noting that the mathematical modeling of profiles and expression signatures cannot be directly applied to identify the ”most relevant genes” or unknown influence factors affecting their expression, but it can be applied to simulate biologically plausible gene expression data, in order to provide benchmarks to assess the performance of gene selection methods, exploiting the fact that the most relevant synthetic genes are known in advance.

In principle, other mathematical models can be designed to simulate different sources of biomolecular data, and in perspective different typologies of synthetic data could be integrated to provide a synthetic system to simulate complex heterogeneous biomolecular data, but this is far beyond the scope of this work.

5 Conclusion

In this paper we proposed an artificial model for validating gene selection methods. Besides the biological motivations underlying our model, we showed from a statistical standpoint that we may obtain artificial

data reasonably close to real gene expression data. As a consequence, we may generate biologically plausible virtual gene expression data that may be easily used to evaluate gene selection methods, since, in this case, we know in advance the set of "relevant" genes. On the basis of the mathematical model, we proposed an algorithmic procedure to generate artificial gene expression data, and we showed how to apply the algorithm to the analysis of the performance of statistical and machine learning based gene selection methods. We argue that our proposed mathematical model and the derived algorithmic procedure to generate gene expression data can be applied to perform a large scale experimental analysis of state-of-the-art gene selection methods.

ACKNOWLEDGMENT

We thank the Editor and the anonymous reviewers for their comments and suggestions. This work has been developed in the context of *CIMAINA* Center of Excellence and has been also partially supported by the Italian MIUR project "Laboratory of Interdisciplinary Technologies in Bioinformatics (LITBIO)", and by the PASCAL2 Network of Excellence under EC grant no. 216886.

REFERENCES

[1] A. Syvanen, "Accessing genetic variation: genotyping single nucleotide polymorphisms." *Nature Review Genetics*, vol. 2, no. 18, pp. 930–942, 2001.

[2] M. Shinawi and S. Cheung, "The array CGH and its clinical applications," *Drug Discovery Today*, vol. 13, no. 17-18, pp. 760–770, 2008.

[3] D. Lockhart and E. Winzeler, "Genomics, gene expression and DNA arrays," *Nature*, vol. 405, pp. 827–836, 2000.

[4] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, "Kegg for linking genomes to life and the environment," *Nucleic Acids Research*, vol. 36, pp. D480–D484, 2008.

[5] D. Allison, X. Cui, G. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus." *Nat Rev Genet.*, vol. 7, no. 1, pp. 55–65, 2006.

[6] S. Wang and Q. Cheng, "Microarray analysis in drug discovery and clinical applications." *Methods Mol Biol.*, vol. 316, pp. 49–65, 2006.

[7] Z. Lee, "An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer," *Artificial Intelligence in Medicine*, vol. 42, no. 1, pp. 81–93, 2008.

[8] J. Dopazo, "Functional interpretation of microarray experiments," *OMICS*, vol. 3, no. 10, 2006.

[9] U. Braga-Neto and E. Dougherty, "Is cross-validation valid for small-sample microarray classification?" *Bioinformatics*, vol. 20, pp. 374–380, 2004.

[10] W. Fu, R. Carroll, and S. Wang, "Estimating misclassification error with small samples via bootstrap cross-validation," *Bioinformatics*, vol. 21, no. 9, pp. 1979–1986, 2005.

[11] A. Molinaro, R. Simon, and R. Pfeiffer, "Prediction error estimation: a comparison of resampling methods," *Bioinformatics*, vol. 21, no. 15, pp. 3301–3307, 2005.

[12] S. Dudoit and J. Fridlyand, "A prediction-based resampling method for estimating the number of clusters in a dataset," *Genome Biology*, vol. 3, no. 7, pp. 1–21, 2002.

[13] T. Lange, V. Roth, M. Braun, and J. Buhmann, "Stability-based validation of clustering solutions," *Neural Computation*, vol. 16, pp. 1299–1323, 2004.

[14] G. Valentini, "Mosclust: a software library for discovering significant structures in bio-molecular data," *Bioinformatics*, vol. 23, no. 3, pp. 387–389, 2007.

[15] E. Amaldi and V. Kann, "On the approximation of minimizing non zero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, pp. 237–260, 1998.

[16] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 5923–5928, 2006.

[17] L. Wessels, M. Reinders, A. Hart, C. Veenman, H. Dai, Y. He, and v. LJ., "A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets," *Bioinformatics*, vol. 21, no. 19, pp. 3755–62, 2005.

[18] C. Lai, M. Reinders, L. van't Veer, and L. Wessels, "A protocol for building and evaluating predictors of disease state based on microarray data," *BMC Bioinformatics*, vol. 7, no. 235, 2006.

[19] A. Choudary, M. Brun, J. Hua, J. Lowey, E. Suh, and E. Dougherty, "Genetic test bed for feature selection," *Bioinformatics*, vol. 22, no. 7, pp. 837–842, 2006.

[20] van de Vijver, M.J. et al., "A gene-expression signature as a predictor of survival in breast cancer," *N. Engl. J. Med.*, vol. 374, pp. 1999–2009, 2002.

[21] T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal, "Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms," *BMC Bioinformatics*, vol. 7, 2006.

[22] J. Weston, A. Elisseeff, B. Scholkopf, and M. Tipping, "Use of the zero-norm with linear models and kernels methods," *Journal of Machine Learning Research*, vol. 3, pp. 1439–1461, 2003.

[23] S. Dudoit and J. Fridlyand, "Bagging to improve the accuracy of a clustering procedure," *Bioinformatics*, vol. 19, no. 9, pp. 1090–1099, 2003.

[24] J. Koo, I. Sohn, K. Sujong, and J. Won Lee, "Structured polychotomous machine diagnosis of multiple cancer types using gene expression," *Bioinformatics*, vol. 22, no. 8, pp. 950–958, 2006.

[25] F. Ruffino, M. Muselli, and G. Valentini, "Gene expression modelling through positive boolean functions," *International Journal of Approximate Reasoning*, vol. 47, no. 1, pp. 97–108, 2008.

[26] G. Towell and J. Shavlik, "Extracting refined rules from knowledge-based neural networks," *Machine Learning*, vol. 131, pp. 71–101, 1993.

[27] A. Alizadeh, M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, J. Powell, L. Yang, G. Marti, T. Moore, J. Hudson, L. Lu, D. Lewis, R. Tibshirani, G. Sherlock, W. Chan, T. Greiner, D. Weisenburger, J. Armitage, R. Warnke, R. Levy, W. Wilson, M. Grever, J. Byrd, D. Botstein, P. Brown, and L. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.

[28] A. Martoglio, J. Miskin, S. Smith, and D. MacKay, "A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer," *Bioinformatics*, vol. 18, no. 12, pp. 1617–1624, 2002.

[29] L. Dyrskjøt, T. Thykjaer, M. Kruhøffer, J. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, and T. Ørntoft, "Identifying distinct classes of bladder carcinoma using microarrays," *Nature Genetics*, vol. 33, no. jan., pp. 90–96, 2003.

[30] S. McCarroll, C. Murphy, S. Zou, S. Pletcher, C. Chin, Y. Jan, C. Kenyon, C. Bargmann, and H. Li, "Comparing genomic expression patterns across species identifies shared transcriptional profile in aging," *Nature Genetics*, vol. 36, no. 2, pp. 197–204, 2004.

[31] Y. Yu, J. Khan, C. Khanna, L. Helman, P. Meltzer, and G. Merlino, "Expression profiling identifies the cytoskeletal organizer ezrin and the developmental homoprotein Six-1 as key metastatic regulators." *Nature Medicine*, vol. 10, no. 2, pp. 175–181, 2004.

[32] M. Tompa *et al.*, "Assessing computational tools for the discovery of transcription factor binding sites," *Nature Biotechnology*, vol. 1, no. 23, pp. 137–44, 2005.

[33] X. Cui and G. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biology*, vol. 4, no. 4, 2003.

[34] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *PNAS*, vol. 95, no. 25, pp. 14 863–14 868, 1998.

[35] D. Kotska and R. Spang, "Finding disease specific alterations in the co–expression of genes," *Bioinformatics*, vol. 20, no. suppl.1, pp. i194–i199, 2004.

[36] S. Ramaswamy, K. Ross, E. Lander, and T. Golub, "A molecular signature of metastasis in primary solid tumors," *Nature Genetics*, vol. 33, pp. 49–54, 2003.

[37] J. Ihmels and N. Bergmann, S.and Barkai, "Defining transcription modules using large-scale gene expression data," *Bioinformatics*, vol. 20, no. 13, pp. 1993–2003, 2004.

[38] Q. Ye, L. Qin, M. Forgues, P. He, J. Kim, A. Peng, R. Simon, Y. Li, A. Robles, Y. Chen, Z. Ma, Z. Wu, S. Ye, Y. Liu, Z. Tang, and X. Wang, "Predicting hepatitis b virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning," *Nature Medicine*, vol. 9, no. 4, pp. 416–423, 2003.

[39] T. Golub *et al.*, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, pp. 531–537, 1999.

[40] J. Chen, R. Delongchamp, C. Tsai, H. Hsueh, F. Sisatare, K. Thompson, V. Deasi, and J. Fuscoe, "Analysis of variance components in gene expression data," *Bioinformatics*, vol. 20, no. 9, pp. 1436–1446, 2004.

[41] V. Cheung, L. Conlin, T. Weber, M. Arcaro, K. Jen, M. Morley, and R. Spielman, "Natural variation in human gene expression assessed in lymphoblastoid cells," *Nature Genetics*, vol. 33, no. 3, pp. 422–425, 2003.

[42] M. Ben-Or and N. Linial, "Collective coin flipping," in *Randomness and Computation*. New York: Academic Press, 1990, pp. 91–115.

[43] P. Diggle, *Statistical analysis of spatial point patterns*. London: Academic Press, 1983.

[44] Alon, U. et al., "Broad patterns of gene expressions revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *PNAS*, vol. 96, pp. 6745–6750, 1999.

[45] M. Shipp, K. Ross, P. Tamayo, A. Weng, J. Kutok, R. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. Pinkus, T. Ray, M. Koval, K. Last, A. Norton, T. Lister, J. Mesirov, D. Neuberg, E. Lander, J. Aster, and T. Golub, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning." *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[46] M. Young-Park and T. Hastie, "L1-regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society B*, vol. 69, no. 4, pp. 659–677, 2007.

[47] H. Zhou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society B*, vol. 67, no. 2, pp. 301–320, 2007.

[48] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[49] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1/3, pp. 389–422, 2002.

[50] K. Kapur, Y. Xing, Z. Ouyang, and W. Wong, "Exon array assessment of gene expression," *Genome Biology*, vol. 8, no. R82, 2007.

[51] T. LaFramboise, "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances," *Nucleic Acids Research*, vol. 37, no. 13, pp. 4181–4193, 2009.

[52] V. Matys *et al.*, "TRANSFAC: transcriptional regulation, from patterns to profiles." *Nucleic Acids Res.*, vol. 31, no. 1, pp. 374–378, 2003.

[53] J. Bullard, E. Purdom, K. Hansen, and S. Dudoit, ""evaluation of statistical methods for normalization and differential expression in mrna-seq experiments"," U.C. Berkeley Division of Biostatistics Working Paper Series, Tech. Rep. Working Paper 247, 2009.

[54] W. Noble and A. Ben-Hur, "Integrating information for protein function prediction," in *Bioinformatics - From Genomes to Therapies*, T. Lengauer, Ed. Wiley-VCH, 2007, vol. 3, pp. 1297–1314.

[55] J. Dopazo, "Formulating and testing hypotheses in functional genomics," *Artificial Intelligence in Medicine*, vol. 45, no. 2-3, pp. 97–107, 2009.

[56] C. De Mol, S. Mosci, M. Traskine, and A. Verri, "A regularized method for selecting nested groups of relevant genes from microarray data,," *Journal of Computational Biology*, vol. 16, pp. 1–14, 2009.

[57] J. Shaik and M. Yeasin, "Fuzzy-adaptive-subspace-iteration-based two-way clustering of microarray data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 2, pp. 244–259, 2009.

[58] S. Niijima and Y. Okuno, "Laplacian linear discriminant analysis approach to unsupervised feature selection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 4, pp. 605–614, 2009.