Available online at www.sciencedirect.com
ScienceDirect
International Journal of Approximate Reasoning
47 (2008) 97–108
INTERNATIONAL JOURNAL OF APPROXIMATE REASONING
www.elsevier.com/locate/ijar

Gene expression modeling through positive Boolean functions

Francesca Ruffino [a], Marco Muselli [b], Giorgio Valentini [a,*]

[a] *DSI, Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano, Via Comelico 39, Milano, Italy*
[b] *IEIIT, Istituto di Elettronica, Ingegneria dell'Informazione e delle Telecomunicazioni, Consiglio Nazionale delle Ricerche, Genova, Italy*

Received 20 April 2006; received in revised form 5 September 2006; accepted 15 March 2007
Available online 11 April 2007

Abstract

In the framework of gene expression data analysis, the selection of biologically relevant sets of genes and the discovery of new subclasses of diseases at bio-molecular level represent two significant problems. Unfortunately, in both cases the correct solution is usually unknown and the evaluation of the performance of gene selection and clustering methods is difficult and in many cases unfeasible. A natural approach to this complex issue consists in developing an artificial model for the generation of biologically plausible gene expression data, thus allowing to know in advance the set of relevant genes and the functional classes involved in the problem.

In this work we propose a mathematical model, based on positive Boolean functions, for the generation of synthetic gene expression data. Despite its simplicity, this model is sufficiently rich to take account of the specific peculiarities of gene expression, including the biological variability, viewed as a sort of random source. As an applicative example, we also provide some data simulations and numerical experiments for the analysis of the performances of gene selection methods.

Keywords: Gene expression modeling; Gene selection; Gene expression data clustering; Positive Boolean functions; DNA microarrays

1. Introduction

DNA microarrays provide the gene expression level for thousands of genes pertaining to a given tissue, thus allowing to understand mechanisms regulating biological processes, such as the onset of a disease or the effects of a drug [2]. To this end, supervised and unsupervised machine learning and statistical methods have been largely applied to the analysis of gene expression data [14,15,20,23].

In some situations the quality of the solution offered by a given technique can be easily evaluated; this is the case of pattern recognition problems, where the accuracy of a classifier can be measured through cross-validation or hold-out estimation. In other problems the performance of a statistic or learning method cannot be assessed since the correct solution is not available, even in a subset of cases.

For instance, several statistic and machine learning techniques [10,11,18] have been proposed in the literature to face with the important problem of *gene selection*, where the subset of genes involved in a biological

* Corresponding author. Tel.: +39 2503 16225; fax: +39 2503 16373.
E-mail address: valentini@dsi.unimi.it (G. Valentini).

0888-613X/$ - see front matter
doi:10.1016/j.ijar.2007.03.010

process of interest is to be determined from a collection of microarray experiments. Unfortunately, the entire set of genes involved in a specific biological process is usually unknown or only partially known. Consequently, the evaluation of the real effectiveness of gene selection methods is very difficult and in many cases unfeasible.

Other important problems, such as the discovery of new subclasses of diseases detected at bio-molecular level may be formalized as unsupervised clustering problems [1,19]. However, besides the fact that unsupervised clustering is in general an ill-posed problem, in this case no a priori solutions are known in advance, as the "real" bio-molecular classes are usually unknown.

To provide some kind of performance evaluation, several models have been proposed to produce synthetic gene expression data for classification, clustering and gene selection problems [6,24]. Even if in principle they may be helpful to test gene selection methods, their main limitation consists in a drastic simplification of the model, which is not sufficiently rich to take into account the peculiarities of gene expression data.

In this paper we propose a new biologically motivated mathematical model capable to describe the relationships between the expression levels of the genes of a virtual tissue and its functional state. In this way it is possible to design an artificial system for a genome-wide synthesis of gene expression data. In particular, the randomness due to biological variability and measurement errors is gathered in a specific term, whereas it is shown that the deterministic part of the model can be implemented by a positive Boolean function acting on relevant genes.

Furthermore, a convenient manner of writing this kind of functions consists in employing *m-of-n expressions*, which are able to capture the main biological characteristics of gene expression, while maintaining a sufficient simplicity. Numerical experiments show how to apply the proposed model to the analysis of the performances of largely used statistical and machine learning gene selection methods.

The structure of the paper is as follows: Section 2 analyze in detail the biological characteristics of gene expression data that must be taken into account in the development of an artificial model for the generation of virtual microarray experiments. The proposed model based on positive Boolean functions is described in Section 3, whereas in Section 4 numerical experiments show how to apply the proposed model to the performance analysis of gene selection methods. Section 5 reports some conclusions.

2. Biological characteristics of gene expression data

Many important 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 gene expression levels.

2.1. Profiles and expression signatures

The main goal of gene selection methods consists in finding sets of genes significantly related to a specific functional state (e.g. diseased vs. healthy). In the bio-molecular literature sets of biologically relevant and differentially expressed genes are named *expression signatures* [1,7,16,17,26]. This term has been firstly introduced by Alizadeh et al. [1] to characterize gene expression patterns found by gene expression profiling. More precisely this term 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 the same set of transcription factors and binding sites may be directly or indirectly shared by the genes belonging to the same expression signature. Hence, a gene expression signature indicates a cluster of coordinately expressed genes, whose coordination reveals the fact that they participate to the same biological process (and hence they are controlled by the same set of regulation factors). 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 patient can be interpreted as a collection of gene expression signatures that reveal different biological features of the analyzed sample [1].

Expression signatures has 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 [1].

Independent Component Analysis performed on gene expression data from ovarian cancer tissues found gene expression signatures representing potential pathophysiological processes in ovarian tissue samples [16]. 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 [26], while two way hierarchical clustering analysis identified several expression signatures expressed in different types of bladder carcinoma [7].

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* [17].

Since *expression profiles* and *expression signatures* seem to be 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

2.2.1. Differential expression and co-expression

Differential expression analysis of single genes, even if it may be useful to identify specific genes involved in biological processes [5], cannot capture the complexity of tightly regulated processes, crucial for the proper functioning of a cell.

Correlations among gene expression levels have been observed [1,8], 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 [13]. This corresponds to examine sets of co-regulated genes, i.e. expression signatures, to reveal functional relationships among genes.

2.2.2. 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 [21].

In other works [13,21] it has been shown that in some cases relevant differences are subtle at the level of individual genes but coordinate in gene expression groups.

2.2.3. Genes may belong to different gene expression signatures at the same time

Many genes may be involved in a number of distinct behaviors, depending on the specific conditions of the tissue. From this standpoint they may belong to different expression signatures [9]. Indeed, each gene may be influenced by several transcription factors, each of which affects several genes [16]. Moreover, many underlying conditions in a given sample may concur to define a gene expression signature (e.g. tumorigenesis, angiogenesis, apoptosis) [12].

2.2.4. Expression signatures may be independent of clinical parameters

An expression signature of 153 genes can be used to correctly classify hepatocellular carcinoma (HCC) intra-hepatic metastasis from metastatic-free HCC [25]. 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.

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. [1,10]).

2.2.5. Different gene expression profiles may share signatures and may differ only for few signatures

It has been shown that gene expression signatures may be shared and partially expressed in different gene expression profiles [1,21,25].

For instance, it has been shown that Diffuse Large B-Cell Lymphoma (DLBCL) subgroups (GCB-like and activated B-like DLBCL) share most of the expression signatures but differ mainly for two signatures (GCB and activated B-cell signatures), partially expressed respectively in germinal center B-cell and activated peripheral blood B cell [1].

Moreover, hierarchical clustering, in the space of a 128 genes signature of metastatic adenocarcinoma nodules of diverse origin, showed two clusters of primary tumors that were highly correlated with metastatic ones: this fact, together with a differential overall survival in primary adenocarcinoma tumors, showed that the considered gene expression signature is present in a subpopulation of primary tumors [21].

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 expressed (that is, not all the genes belonging to the expression signature are over-expressed or under-expressed), reflecting functional alterations in diseased patients.

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 may be characterized as a set of gene expression signatures, 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 coexpressed genes; these genes may be overexpressed or underexpressed with respect to a particular condition. Accordingly, in the model, each expression signature should be defined as a set of overexpressed or underexpressed genes, that is genes with expression levels above or below a given threshold. The model should define a signature *active* if its genes are coordinately over(under)expressed.
(3) Expression signatures may be defined either by the overall available knowledge about bio-molecular processes (e.g. by Gene Ontology categories) or may be discovered through statistical and machine learning methods. Hence, the model should permit to define arbitrary signatures, in order to face with a large range of applications in different biological contexts.
(4) Genes may belong to different signatures at the same time. Consequently, the model should allow to assign the same gene to different signatures.
(5) The number of genes within an expression signature usually vary from few units to few hundreds. Accordingly, the model should permit to select within this range the number of elements for each gene expression signature.
(6) Apart from technical variation (that in principle should be detected and canceled by proper design and implementation of bio-technological experiments and suitable pre-processing procedures [3]), gene expression is biologically variable also within functional classes (conditions) [4]. Thus, the model should reproduce the variation of gene expression data, which may be simulated by sampling from a predefined distribution. Our preliminary analysis showed that gene expression values are close to be normally distributed.
(7) Not always expression signatures show large variations of gene expression levels: some signatures may present modest but coordinate variations. Consequently, the model should be sufficiently flexible to allow small variations of coexpressed genes, and to this end it should include tunable parameters of the gene distributions.
(8) Not all the genes within a signature may be expressed in all the samples. 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.

(9) Different expression profiles may differ only for few signatures, i.e. different functional groups may share the same (or very similar) expression signatures. This situation must be permitted by the artificial model when developing expression signatures for different functional states.

(10) 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 functional state.

3. The mathematical model

On the basis of the biological analysis presented in Section 2 and, in particular, starting from the concepts of expression profile, expression signature and gene modulation, we propose a mathematical model describing the relationship between the expression levels of genes and functional state of a tissue. Our model will receive in input a set of values representing the gene expression levels of a tissue and will return in output the value 1 if the tissue is in the functional state of interest and 0 otherwise.

Since in a real situation, due to both biological variability and possible measurement errors occurring in DNA-microarray experiments, a deterministic relationship between gene expression values and the functional state of the tissue does not exist, the model will be composed by a deterministic part described through a function $f : \mathbb{R}^m \to \{0,1\}$ and by a random term e corresponding to the probability that a tissue is assigned to the wrong state. If we denote with y the output of the model and with $\boldsymbol{x}$ the input vector we will have

$$y = \begin{cases} f(\boldsymbol{x}) & \text{with probability } 1-e \\ 1-f(\boldsymbol{x}) & \text{with probability } e \end{cases}$$

To define the model function f let us introduce the input set $A = \{g_1, \ldots, g_m\}$, given by the collection of the total number m of analyzed genes, and the real vector $\boldsymbol{x} = (x_1, \ldots, x_m)$ including the expression levels of the m genes belonging to A.

Suppose that, for each gene g_i belonging to A, a modulation threshold t_i exists so that we can assert that the gene g_i is *overexpressed* if the value x_i of its expression exceeds t_i and underexpressed if $x_i < -t_i$. More precisely, we say that a gene is *modulated* when it is overexpressed or underexpressed with respect to a given functional state.

Therefore, it is possible to define a mapping $\boldsymbol{\beta} : \mathbb{R}^m \to \{0,1\}^m$ that depends on the modulation thresholds t_i and returns for each gene the value 1 if that gene is modulated and 0 otherwise.

$$z_i = \beta_i(\boldsymbol{x}) = \begin{cases} 1 & \text{if } g_i \text{ is modulated (i.e. if } x_i > t_i \text{ or } x_i < -t_i) \\ 0 & \text{if } g_i \text{ is not modulated} \end{cases} \tag{1}$$

Suppose the output is uniquely determined by the state (modulated or not) of the m genes and does not depend on their specific expression values. Then, the function f can be written as $f(\boldsymbol{x}) = \varphi(\boldsymbol{\beta}(\boldsymbol{x}))$, where φ is a Boolean function defined on binary strings in $\{0,1\}^m$. Consequently, once the mapping $\boldsymbol{\beta}$ is completely described, the deterministic component f of our model is uniquely determined by the construction of the Boolean function φ.

On the input set $\{0,1\}^m$, having cardinality 2^m, we consider the standard partial ordering $(\{0,1\}^m, \leqslant)$, i.e. for any pair $\boldsymbol{u}, \boldsymbol{z} \in \{0,1\}^m$ we have $\boldsymbol{u} \leqslant \boldsymbol{z}$ if and only if $u_i \vee z_i = z_i$ for every $i \in \{1, \ldots, m\}$, where $\vee$ denotes the logical OR operator. A Boolean function $\varphi : \{0,1\}^m \to \{0,1\}$ will be called *positive* if and only if $\boldsymbol{u} \leqslant \boldsymbol{z}$ implies $\varphi(\boldsymbol{u}) \leqslant \varphi(\boldsymbol{z})$ for all $\boldsymbol{u}, \boldsymbol{z} \in \{0,1\}^m$.

Consider the truth table of a positive Boolean function $\varphi : \{0,1\}^m \to \{0,1\}$. Denote with p_i the fraction of input vectors $\boldsymbol{z}$ with output 1 having the ith component $z_i = 1$ and with p the fraction of patterns $\boldsymbol{z} \in \{0,1\}^m$ with output 1 out of the total 2^m.

$$p_i = \frac{\sum_{\boldsymbol{z} \in \{0,1\}^m, z_i = 1} \varphi(\boldsymbol{z})}{2^{m-1}}, \quad p = \frac{\sum_{\boldsymbol{z} \in \{0,1\}^m} \varphi(\boldsymbol{z})}{2^m} \tag{2}$$

By using the definition of positive Boolean function it can be easily seen that $p_i \geqslant p$. Thus, we can denote with R_φ the set of the indexes of the input components for which the strict inequality holds, i.e.

$$R_\varphi = \{i \in \{1, \ldots, m\} : p_i > p,\}$$

and with $k = \mid R_\varphi \mid$ its cardinality (number of elements).

In a similar way we define

$$I_\varphi = \{i \in \{1, \ldots, m\} : p_i = p\}$$

I_φ will obviously have cardinality $m - k$ and will include the indexes of the components of the input vector not relevant for the output. Hence, the function φ can be equivalently defined by a truth table with only k inputs; this new description leaves fractions p and p_i unchanged.

As an example, consider the truth table in Table 1a for the positive Boolean function φ_1. From a direct inspection, using the definitions (2), we obtain:

$$p = \frac{3}{8}, \quad p_1 = \frac{1}{2}, \quad p_2 = \frac{1}{2}, \quad p_3 = \frac{3}{4}, \quad p_4 = \frac{3}{8}$$

Since p_1, p_2 and p_3 are strictly greater than p, whereas $p_4 = p$, the same function can be described by a truth table with three inputs, reported in Table 1b. As we can note, the values of p, p_1, p_2 and p_3 remain unchanged.

It will be shown in this section that the adoption of a positive Boolean function φ for the construction of the model allows to satisfy the biological requirements outlined in Section 2. In particular, if gene g_i (corresponding to the ith input z_i) is defined to be *relevant* when $p_i > p$, i.e. when $i \in R_\varphi$, from a biological point of view, the set R_φ determines the expression profile of the functional state described by φ, since the presence in it of indexes of modulated genes increases the probability that the corresponding tissue is in the considered functional state. The elements having indexes in I_φ, on the contrary, are irrelevant for the determination of the output value for φ and, consequently, are associated with genes not belonging to the expression profile.

Consider a positive Boolean function and restrict its domain to include only the relevant k input variables. If $D_1(f)$ is the collection of the input vectors $\boldsymbol{z}$ for which the output value of φ is 1

$$D_1(f) = \{\boldsymbol{z} \in \{0,1\}^k : \varphi(\boldsymbol{z}) = 1\}$$

and $P(\boldsymbol{z})$ is the set of the indexes of the components of $\boldsymbol{z}$ with value 1

$$P(\boldsymbol{z}) = \{i \in \{1, \ldots, k\} : z_i = 1\}$$

we have

$$\varphi(z_1, \ldots, z_k) = \bigvee_{\boldsymbol{z} \in D_1} \bigwedge_{P(\boldsymbol{z})} z_i \tag{3}$$

where $\wedge_{P(\boldsymbol{z})}$ is the logical product (AND) among all the components of $\boldsymbol{z}$ with indexes in $P(\boldsymbol{z})$, while $\bigvee_{D_1}$ is the logical sum (OR) on the input vectors in D_1.

Table 1
(Panel a) Truth table of the positive Boolean function φ_1 and (panel b) truth table equivalent to (panel a)

Panel (a)										Panel (b)			
$\boldsymbol{z}$				$\varphi_1(\boldsymbol{z})$	$\boldsymbol{z}$				$\varphi_1(\boldsymbol{z})$	$\boldsymbol{z}$			$\varphi_1(\boldsymbol{z})$
z_1	z_2	z_3	z_4		z_1	z_2	z_3	z_4		z_1	z_2	z_3	
0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	1	0	0	1	0	0	0	1	0
0	0	1	0	0	1	0	1	0	1	0	1	0	0
0	0	1	1	0	1	0	1	1	1	0	1	1	1
0	1	0	0	0	1	1	0	0	0	1	0	0	0
0	1	0	1	0	1	1	0	1	0	1	0	1	1
0	1	1	0	1	1	1	1	0	1	1	1	0	0
0	1	1	1	1	1	1	1	1	1	1	1	1	1

In this way every positive Boolean function can be expressed through a logical sum ($\bigvee$) of logical products ($\bigwedge$) of their inputs. As an example, consider the truth table in Table 1b: expression (3) for φ_1 has the form

$$\varphi_1(z_1, z_2, z_3) = (z_2 \wedge z_3) \vee (z_1 \wedge z_3) \vee (z_1 \wedge z_2 \wedge z_3) = (z_2 \wedge z_3) \vee (z_1 \wedge z_3) \tag{4}$$

An alternative way of representing a positive Boolean function can be derived by extending the concept of *m-of-n expression* defined in [22]. To this aim we introduce the following:

Definition 1. If

$$G(q) = \{z_{j_1}, \ldots, z_{j_l}, j_r \neq j_s \text{ if } r \neq s\}$$

is a set composed by l distinct components of the generic vector $\boldsymbol{z} \in \{0,1\}^k$ and q is a positive integer with $q \leqslant l$, we say that $G(q)$ is *active* if at least q of its components have value 1.

Suppose, for example, that $k = 4$ and $G(2) = \{z_1, z_2, z_3\}$. Then, $G(2)$ is not active for $\boldsymbol{z} = (1,0,0,1)$, while $G(2)$ is active for $\boldsymbol{z} = (1,1,0,0)$ or $\boldsymbol{z} = (1,0,1,0)$.

Definition 2. If $G_1(q_1), \ldots, G_h(q_h)$ are defined as above, with $|\, G_i(q_i) \,| = l_i$, $q_i \leqslant l_i$, and p, h are positive integers with $p \leqslant h$, the m-of-n expression of a positive Boolean function $\varphi : \{0,1\}^k \to \{0,1\}$ is given by the following representation:

$$\varphi(z_1, \ldots, z_k) = \begin{cases} 1 & \text{if at least } p \text{ of the } h \text{ sets } G_1(q_1), \ldots, G_h(q_h) \text{ are } \textit{active} \\ 0 & \text{otherwise} \end{cases}$$

It can be shown that

Theorem 3. *A positive Boolean function $\varphi : \{0,1\}^k \to \{0,1\}$ can always be written in the form of an m-of-n expression.*

Proof. Denote with $h = |D_1|$ the cardinality of the set D_1 and with $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{|D_1|}$ its elements. Then, the theorem is proved by setting $q_i = |P(\boldsymbol{z}_i)|$, $G_i(q_i) = \{z_j : j \in P(\boldsymbol{z}_i)\}$, for every $i \in \{1 \ldots, h\}$, and $p = 1$. In this way the m-of-n expression of φ is equivalent to the AND-OR expression in (3). □

According to the proof of Theorem 3, the function φ_1 of Table 1b can be put in the form of an m-of-n expression by taking the following two sets of components:

$$G_1(2) = \{z_1, z_3\}, \quad G_2(2) = \{z_2, z_3\} \tag{5}$$

each of them gives rise to a logical product in expression (4) since every set is active when all its components has value 1, i.e. when the corresponding logical product gives output 1. Then, by taking $p = 1$, the logical OR in (4) is obtained.

In general, by denoting

$$G_i(q_i) = (z_{j_{i,1}}, \ldots, z_{j_{i,l_i}})_{q_i}$$

we can represent φ as follows:

$$\varphi(\boldsymbol{z}) = [(z_{j_{1,1}}, \ldots, z_{j_{1,l_1}})_{q_1}, \ldots, (z_{j_{h,1}}, \ldots, z_{j_{h,l_h}})_{q_h}]_p \tag{6}$$

As an example, from (5) we obtain:

$$\varphi_1(\boldsymbol{z}) = [(z_1, z_3)_2, (z_2, z_3)_2]_1$$

However, this representation of φ_1 as an m-of-n expression is not unique; the same function can also be obtained by

$$\varphi_1(\boldsymbol{z}) = [(z_1, z_2)_1, (z_3)_1]_2$$

As a matter of fact, the resulting truth table, presented in Table 2, is equivalent to that reported in Table 1b.

The following example shows how, when the dimension k of the input domain is large, m-of-n expressions can provide a more compact description of positive Boolean functions with respect to (3).

Let $\varphi_2 : \{0,1\}^5 \to \{0,1\}$ be the positive Boolean function having the truth table in Table 3. By using (3) we obtain:

$$\varphi_2(z_1,z_2,z_3,z_4,z_5) = (z_4 \wedge z_5) \vee (z_1 \wedge z_2 \wedge z_4) \vee (z_1 \wedge z_2 \wedge z_5) \vee (z_1 \wedge z_3 \wedge z_4) \vee (z_1 \wedge z_3 \wedge z_5) \vee (z_2 \wedge z_3 \wedge z_4) \vee (z_2 \wedge z_3 \wedge z_5)$$

However, it can be easily seen that the m-of-n expression

$$\varphi_2(\boldsymbol{z}) = [(z_1,z_2,z_3)_2,(z_4)_1,(z_5)_1]_2 \tag{7}$$

leads to the same positive Boolean function. Because of this property, m-of-n expressions will be employed to derive a compact representation for positive Boolean functions.

Finally, by combining the representation of the function φ defined in (6) and the mapping $\boldsymbol{\beta}$ defined in (1), the model 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 sets $G_i(q_i)$ with the mapping $\boldsymbol{\beta}$ that determines their values. For example, if z_i assume value 1 when $x_i > t_i$, 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 in $f_2(\boldsymbol{x}) = \varphi_2(\boldsymbol{\beta}(\boldsymbol{x}))$, the mapping $\boldsymbol{\beta}$ has the following form:

$$z_1 = \beta_1(\boldsymbol{x}) = \begin{cases} 1 \text{ if } x_1 > 2, \\ 0 \text{ otherwise}, \end{cases} \quad z_2 = \beta_2(\boldsymbol{x}) = \begin{cases} 1 \text{ if } x_2 < -3 \\ 0 \text{ otherwise} \end{cases}$$

$$z_3 = \beta_3(\boldsymbol{x}) = \begin{cases} 1 \text{ if } x_3 > 1, \\ 0 \text{ otherwise}, \end{cases} \quad z_4 = \beta_4(\boldsymbol{x}) = \begin{cases} 1 \text{ if } x_4 > 3 \\ 0 \text{ otherwise} \end{cases}$$

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

Table 2
Truth table equivalent to Table 1b

$\boldsymbol{z}$			G_1	G_2	$\varphi_1(\boldsymbol{z})$
z_1	z_2	z_3			
0	0	0	Not active	Not active	0
0	0	1	Not active	Active	0
0	1	0	Active	Not active	0
0	1	1	Active	Active	1
1	0	0	Active	Not active	0
1	0	1	Active	Active	1
1	1	0	Active	Not active	0
1	1	1	Active	Active	1

Table 3
Truth table of $\varphi_2 : \{0,1\}^5 \to \{0,1\}$

$\boldsymbol{z}$					φ_2	$\boldsymbol{z}$					φ_2	$\boldsymbol{z}$					φ_2
z_1	z_2	z_3	z_4	z_5		z_1	z_2	z_3	z_4	z_5		z_1	z_2	z_3	z_4	z_5	
0	0	0	0	0	0	0	1	0	1	1	1	1	0	1	1	0	1
0	0	0	0	1	0	0	1	1	0	0	0	1	0	1	1	1	1
0	0	0	1	0	0	0	1	1	0	1	1	1	1	0	0	0	0
0	0	0	1	1	1	0	1	1	1	0	1	1	1	0	0	1	1
0	0	1	0	0	0	0	1	1	1	1	1	1	1	0	1	0	1
0	0	1	0	1	0	1	0	0	0	0	0	1	1	0	1	1	1
0	0	1	1	0	0	1	0	0	0	1	0	1	1	1	0	0	0
0	0	1	1	1	1	1	0	0	1	0	0	1	1	1	0	1	1
0	1	0	0	0	0	1	0	0	1	1	1	1	1	1	1	0	1
0	1	0	0	1	0	1	0	1	0	0	0	1	1	1	1	1	1
0	1	0	1	0	0	1	0	1	0	1	1						

Then, by extending the representation (7), f_2 can be written as follows:

$$f_2(\boldsymbol{x}) = [(x_1 > 2, x_2 < -3, x_3 > 1)_2, (x_4 > 3)_1, (x_5 < -1)_1]_2 \tag{8}$$

In this way, when a vector $\boldsymbol{x}$ is presented to the model, we can immediately know if $f_2(\boldsymbol{x}) = 1$. In addition, if we interpret each set $G_i(q_i)$ as an expression signature, it is easy to see that the proposed model implements the biological specifications presented in Section 2:

- The expression profile is defined in terms of expression signatures;
- Each expression signature is defined as a set of underexpressed or overexpressed genes, that is genes with gene expression levels above or below a given threshold;
- Genes may belong to different expression signatures at the same time;
- By choosing a value of q lower than the cardinality of the sets $G(q)$, not all the genes belonging to the expression signature have to be modulated to make $G(q)$ active. In a similar way, by taking a value for p less than h, not all the expression signatures have to be active to induce the output value 1.

4. An application to the evaluation of gene selection methods

The model proposed in the previous section can be employed to evaluate the performance of gene selection methods in determining the correct set of relevant genes when analyzing a collection of examples derived from synthetic microarray experiments, each of which is associated with a virtual tissue. Every example is given by a pair $(\boldsymbol{x}, y)$, where $\boldsymbol{x}$ is a real-valued input vector whose components represent the gene expression levels for the corresponding tissue.

The output y can vary into a set of c different values, each one denoting the class which the associated tissue belongs to. In this way situations where the analyzed tissue belongs to one of c different possible classes are simulated; this corresponds to consider c different functional states, one for each output class. The case $c = 2$, where the output y can assume the values 1 and -1, will be examined henceforth; a generalization of the analysis to higher values of c is straightforward.

The mathematical model developed in the previous section can be adopted to describe each of the two functional states. Two subsequent phases have been devised: in the first one the two functions f_1 and f_2, related to the two different functional states, are built, whereas in the second one the gene expression levels of n virtual tissues are generated.

As described in the previous section, randomness inherent the determination of the functional state can be collected into a real parameter e, so that with probability $1 - e$ each virtual tissue belonging to the output class 1 (resp. -1) has gene expression levels forming a vector $\boldsymbol{x}$ verifying $f_1(\boldsymbol{x}) = 1$ (resp. $f_2(\boldsymbol{x}) = 1$). If the classes are mutually exclusive (as it is usually the case), it should be guaranteed that each tissue belongs to only one functional state, i.e. if $\boldsymbol{x}$ is the associated input vector only one model provides the output 1.

The collection of virtual tissues generated by the model can be collected into a matrix X, where each row corresponds to a tissue and each column to a gene. Then, a final column Y representing the class of each tissue is added. Feature selection and clustering methods can be applied to $Z = [X, Y]$ and X respectively. However, since both the rule determining the membership of a tissue to a class and the relationship among the virtual genes are completely known, these methods can be directly tested and their performances can be easily evaluated.

As an example, we compare two feature selection methods, the technique proposed by Golub et al. in [10] (a simple variation of the classic t-test) and the SVM-RFE procedure [11], on two different collections of examples built by adopting the model described in the previous section. The evaluation of the performances of the two methods has been performed by counting how many relevant genes, actually belonging to the expression profile, are found.

The first dataset X_1 is composed by 100 artificial tissues, 60 belonging to the first class and 40 in the second class, with 6000 virtual genes. The expression profiles of the two functional states, represented by the functions f_1 and f_2, contain 144 genes in total.

The m-of-n expression of f_1 has been built by using the mathematical model described in the previous section with parameters:

- $h = 5$;
- $l_1 = 17, l_2 = 20, l_3 = 10, l_4 = 11, l_5 = 16$;
- $q_1 = 7, q_2 = 8, q_3 = 4, q_4 = 5, q_5 = 7$;
- $p = 3$;

while the values of the parameters for the function f_2 are the following:

- $h = 6$;
- $l_1 = 14, l_2 = 12, l_3 = 13, l_4 = 11, l_5 = 11, l_6 = 10$;
- $q_1 = 7, q_2 = 6, q_3 = 7, q_4 = 6, q_5 = 6, q_6 = 5$;
- $p = 4$;

For both the functional states the parameter e has been fixed to 0.1.

Both the Golub's method and SVM-RFE have been applied to the complete dataset $Z_1 = [X_1, Y_1]$, being Y_1 the vector containing the labels y of the class of each tissue $\boldsymbol{x}$ ($y = 1$ if $f_1(\boldsymbol{x}) = 1$ or $y = -1$ if $f_2(\boldsymbol{x}) = 1$). Every gene selection method assigns a rank value to each of the 6000 genes: the higher is the rank the more relevant is the corresponding gene. The first 144 genes with greater rank values are then compared with the 144 genes actually belonging to the two expression profiles.

If we denote with G_{144} and S_{144} the set of the 144 most relevant genes selected by Golub's method and by SVM-RFE, respectively, we can evaluate the intersections between G_{144} or S_{144} and the set M_{144} of the genes included in the two expression profiles. The greater is the size of the intersection, the better is the performance of the gene selection method. A relative measure of this term is given by the fraction P_G (resp. P_S) of relevant genes contained in G_{144} (resp. R_{144}).

The results show that

$$P_G = \frac{|G_{144} \cap M_{144}|}{|M_{144}|} = \frac{132}{144} = 0.92$$

and

$$P_S = \frac{|S_{144} \cap M_{144}|}{|M_{144}|} = \frac{24}{144} = 0.17$$

having denoted with $|A|$ the cardinality (number of elements) of the set A. The comparison between the values of P_G and P_S shows that in this artificial dataset the behavior of the Golub's method is significantly better than that of SVM-RFE. In particular, the former is able to retrieve most (92%) of the relevant genes.

The application of the same approach to a second artificial dataset may help to understand if this result has a more general validity. To this aim a new data matrix $Z_2 = [X_2, Y_2]$ has been generated, where X_2 contains 80 virtual tissues (50 belonging to the first class and 30 to the second class) and 2500 virtual genes. The parameters for the construction of the m-of-n expression f_1 for the first functional state are

- $h = 5$;
- $l_1 = 13, l_2 = 17, l_3 = 10, l_4 = 17, l_5 = 10$;
- $q_1 = 6, q_2 = 7, q_3 = 4, q_4 = 7, q_5 = 4$;
- $p = 5$;

while the model f_2 for the second functional state is generated starting from the following parameters:

- $h = 6$;
- $l_1 = 12, l_2 = 15, l_3 = 12, l_4 = 10, l_5 = 12, l_6 = 10$;
- $q_1 = 5, q_2 = 6, q_3 = 5, q_4 = 4, q_5 = 5, q_6 = 4$;
- $p = 6$;

The value of the parameter e has been fixed to 0.05.

Since, in this case, the total number of genes belonging to the two expression profiles is 133, we consider the sets G_{133} and S_{133} obtained by applying the Golub's method and SVM-RFE, respectively, to the dataset Z_2 and by taking the 133 genes with highest rank for both methods. In this way, we can again compute the quantities P_G and P_S, given by the fraction of relevant genes included in G_{133} and S_{133}:

$$P_G = \frac{|G_{133} \cap M_{133}|}{|M_{133}|} = \frac{124}{133} = 0.93$$

while

$$P_S = \frac{|S_{133} \cap M_{133}|}{|M_{133}|} = \frac{39}{133} = 0.29$$

M_{133} is the set of the relevant genes adopted for the construction of the m-of-n expressions of f_1 and f_2. As one can note, also in this case the Golub's method achieves by far the best performance.

5. Conclusions

An artificial model for the generation of biologically plausible gene expression data, to be adopted in the evaluation of gene selection and clustering methods, has been proposed. Starting from the concepts of gene expression signature and gene expression profile, whose properties can be derived by publications in the bio-medical and bioinformatics literature, we have obtained a list of requirements that must be fulfilled by the artificial model to guarantee a sufficient degree of similarity between virtual and real gene expression data.

A mathematical model, composed by a random term and by a positive Boolean function φ, has been shown to satisfy the required specifications. The adoption of a particular form, called m-of-n expression, for the function φ allows to significantly simplify the generation process of the model, emphasizing the mathematical counterparts of gene expression signature and gene expression profile.

An application of the proposed artificial model in evaluating the performances of two gene selection techniques, Golub's method [10] and SVM-RFE [11], has been also presented. The analysis of two artificial datasets, where the collection of relevant genes is considerably smaller than the whole set of genes characterizing the virtual tissue, has permitted to derive that the Golub's method performs significantly better than SVM-RFE, being able to retrieve more than 90% of the relevant genes.

Acknowledgements

This work was partially supported by the Italian MIUR projects "Laboratory of Interdisciplinary Technologies in Bioinformatics (LITBIO)" and has been developed in the context of *CIMAINA* Center of Excellence. We thank the reviewers for their comments to our paper.

References

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

[2] P. Baldi, G.W. Hatfield, DNA Microarrays and Gene Expression, Cambridge University Press, Cambridge, UK, 2002.

[3] J.J. Chen, R. Delongchamp, C. Tsai, H. Hsueh, F. Sisatare, K. Thompson, V. Deasi, J. Fuscoe, Analysis of variance components in gene expression data, Bioinformatics 20 (9) (2004) 1436–1446.

[4] V. Cheung, L. Conlin, T. Weber, M. Arcaro, K. Jen, M. Morley, R. Spielman, Natural variation in human gene expression assessed in lymphoblastoid cells, Nature Genetics 33 (3) (2003) 422–425.

[5] X. Cui, G. Churchill, Statistical tests for differential expression in cDNA microarray experiments, Genome Biology 4 (4) (2003).

[6] S. Dudoit, J. Fridlyand, Bagging to improve the accuracy of a clustering procedure, Bioinformatics 19 (9) (2003) 1090–1099.

[7] L. Dyrskjøt, T. Thykjaer, M. Kruhøffer, J. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, T. Ørntoft, Identifying distinct classes of bladder carcinoma using microarrays, Nature Genetics 33 (2003) 90–96.

[8] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstei, Cluster analysis and display of genome-wide expression patterns, PNAS 95 (25) (1998) 14863–14868.

[9] P. Gasch, M. Eisen, Exploring the conditional regulation of yeast gene expression through fuzzy k-means clustering, Genome Biology 3 (11) (2002).
[10] T.R. Golub et al., Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science 286 (1999) 531–537.
[11] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning 46 (2002) 389–422.
[12] J. Ihmels, S. Bergmann, N. Barkai, Defining transcription modules using large-scale gene expression data, Bioinformatics 20 (13) (2004) 1993–2003.
[13] D. Kotska, R. Spang, Finding disease specific alterations in the co–expression of genes, Bioinformatics 20 (2004) i194–i199.
[14] T. Li, C. Zhang, M. Ogihara, A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, Bioinformatics 20 (2004) 2429–2437.
[15] Y. Lu, J. Han, Cancer classification using gene expression data, Information Systems 28 (2003) 243–268.
[16] A. Martoglio, J. Miskin, S. Smith, D. MacKay, A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer, Bioinformatics 18 (12) (2002) 1617–1624.
[17] S.A. McCarroll, C. Murphy, S. Zou, S. Pletcher, C. Chin, Y. Jan, C. Kenyon, C. Bargmann, H. Li, Comparing genomic expression patterns across species identifies shared transcriptional profile in aging, Nature Genetics 36 (2) (2004) 197–204.
[18] M. Muselli, Gene selection through Switched Neural Networks, in: NETTAB-2003, Workshop on Bioinformatics for Microarrays, Bologna, Italy, 2003.
[19] M.D. Onken, L.A. Worley, J.P. Ehlers, J.W. Harbour, Gene expression profiling in uveal melanoma reveals two molecular classes and predicts metastatic death, Cancer Research 64 (2004) 7205–7209.
[20] J. Quackenbush, Computational analysis of microarray data, Nature Reviews Genetics 2 (6) (2001) 418–427.
[21] S. Ramaswamy, K. Ross, E. Lander, T. Golub, A molecular signature of metastasis in primary solid tumors, Nature Genetics 33 (2003) 49–54.
[22] G. Towell, J. Shavlik, Extracting Refined Rules from Knowledge-Based Neural Networks, Machine Learning 131 (1993) 71–101.
[23] G. Valentini, M. Muselli, F. Ruffino, Cancer recognition with bagged ensembles of Support Vector Machines, Neurocomputing 56C (2004) 461–466.
[24] J. Weston, A. Elisseeff, B. Scholkopf, M. Tipping, Use of the zero-norm with linear models and kernels methods, Journal of Machine Learning Research 3 (2003) 1439–1461.
[25] 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, X. Wang, Predicting hepatitis b virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning, Nature Medicine 9 (4) (2003) 416–423.
[26] Y. Yu, J. Khan, C. Khanna, L. Helman, P. Meltzer, G. Merlino, Expression profiling identifies the cytoskeletal organizer ezrin and the developmental homoprotein Six-1 as key metastatic regulators, Nature Medicine 10 (2) (2004) 175–181.