journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Detecting reliable gene interactions by a hierarchy of Bayesian network classifiers

Rubén Armañanzas[a,*], *Iñaki Inza*[a], *Pedro Larrañaga*[b]

[a] *Department of Computer Science and Artificial Intelligence, University of the Basque Country, Paseo Manuel Lardizabal 1, 20018 Donostia-San Sebastián, Gipuzkoa, Spain*

[b] *Department of Artificial Intelligence, Technical University of Madrid, 28660 Boadilla del Monte, Madrid, Spain*

ARTICLE INFO

Article history:
Received 5 September 2007
Received in revised form
8 February 2008
Accepted 28 February 2008

Keywords:
Bayesian network classifiers
Robust arc identification
Gene interactions
DNA microarrays
Knowledge discovery

ABSTRACT

The main purpose of a gene interaction network is to map the relationships of the genes that are out of sight when a genomic study is tackled. DNA microarrays allow the measure of gene expression of thousands of genes at the same time. These data constitute the numeric seed for the induction of the gene networks. In this paper, we propose a new approach to build gene networks by means of Bayesian classifiers, variable selection and bootstrap resampling. The interactions induced by the Bayesian classifiers are based both on the expression levels and on the phenotype information of the supervised variable. Feature selection and bootstrap resampling add reliability and robustness to the overall process removing the false positive findings. The consensus among all the induced models produces a hierarchy of dependences and, thus, of variables. Biologists can define the depth level of the model hierarchy so the set of interactions and genes involved can vary from a sparse to a dense set. Experimental results show how these networks perform well on classification tasks. The biological validation matches previous biological findings and opens new hypothesis for future studies.

1. Introduction

Gene networks or gene interaction networks [1] are currently a topic under heavy research in the computational biology field. High throughput biological devices such as DNA microarrays have reduced the gap between the traditional medicine and what is known nowadays as biomedicine. But in this context, not only the proof of a certain gene activity is necessary, but also the investigation of how a set of genes interact among them is crucial for the understanding of different complex diseases.

However, there is still a tendency to analyse gene expression data only from a pure numeric point of view, that is, to look for the smallest and most accurate set of genes that are able to distinguish between two or more phenotypes [2–5]. This analysis strategy still falls into the problems related with the low number of instances in a typical genomic study, broadly known as the *curse of the dimensionality* [6]. As the DNA microarray devices begin to be less expensive, the amount of available data will allow to overcome these problems such as, for instance, the overfit effect [7,8] on the microarray studies.

Apart from these studies, computational techniques have proven their capacity to help physicians to analyse the gene activities of complex diseases. In order to understand such complex relations, many approaches have been gone on stage. From pure Bayesian networks [9–11] to statistical validations by multiple random simulation [12], new graphical models to match gene interactions [13,14] or biological validation of previously reported interactions [15,16] have been approached. The main corpus of all these works is to assume that a gene behaves as a random variable of an unknown probabilistic dis-

* *Corresponding author*. Tel.: +34 943018070; fax: +34 943015590.
E-mail addresses: ruben@si.ehu.es (R. Armañanzas), inza@si.ehu.es (I. Inza), pedro.larranaga@fi.upm.es (P. Larrañaga).

0169-2607/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.cmpb.2008.02.010

tribution. Over that distribution, the regulatory interactions between the genes are expected to produce corresponding probabilistic dependences within their expression levels [17].

In this framework, the majority of the works just look for differentially expressed genes to build their models. However, few of them are explicitly focussed on the statistical information that the comparison of different sample types contributes. The conditional probabilities learnt through the phenotype statistical distribution in the database will be used to report interactions among genes not only based on their individual expression levels, but also on their behaviour through the different conditions. This fact involves the addition of the probabilistic relationship that associates the sample class (or phenotype) with each relevant gene or feature under the study, that is, a supervised-class experimental design [18,19]. Our new proposal belongs to these supervised studies, stressing the search of robust results by means of a hierarchy of supervised Bayesian classifiers.

Based on the frequency of appearance of each arc within an induced pool of Bayesian classifiers, our approach assigns confidence levels to those arcs. Depending on the confidence level fixed by a biologist or physician (hereafter *expert*), the final model can vary from a very simple structure including a small set of dependences to a deep forest-like one with hundreds of them. This property allows to retrieve a hierarchy of autoinclusive models: from the simplest and most reliable one with only one interaction to the most complex one that includes all the detected interactions. These hierarchical networks are computed by means of a set of tools well-suited for the biological characteristics, taken from the machine learning and statistics fields:

- The estimations produced by stratified sampling with replacement, known as *non-parametric bootstrap*, are cautious. The ratio of false positives in the features induced with this procedure is very low [20]. This fact is significantly important when dealing with biological data in which the number of samples is still very low.
- A small set of genes gathers most of the information in an entire microarray. A feature selection procedure must be applied to reduce the dimensionality from thousands to only hundreds of candidate genes [21].
- No *a priori* biological information is used by the Bayesian classifiers, only the phenotype distribution is considered. Therefore, no previous biological premise will bias the final models.
- Consensus conclusions in the analysis of microarray data have already demonstrated good results [22–24]. When seeking for robust gene interactions, finding a parsimonious set of both genes and dependences, which have a high degree of confidence on the basis of the data, guarantees a low number of false positives in the final network.

Specifically, our approach combines a resampling method with an inner feature selection technique and a Bayesian k-dependence classifier [25] to obtain a gene interaction network formed by arcs which surpass a certain confidence level. The expert can fix the complexity threshold of the relationships among the genes in the output network so it can be used as a tool to unveil or corroborate biological hypothesis.

The use of Bayesian classifiers to tackle this task implies that, first, the statistical dependences among the genes can reveal real interactions among them. Secondly, the gene interactions not only describe relationships among solely genes, but also describe different biological behaviours based on the phenotype distribution of each gene's expression. Similar studies with the same aim [9,11,26] make use of the classical *score + search* Bayesian learning scheme and focus their attention on partially directed models. Our method returns directed acyclic models with directed edges and it can be configured with both different variable set selections and Bayesian classifier inductors.

Because of this flexibility, the approach can also be seen as a consensus feature selection if the expert is only interested in the genes or variables connected by the arcs of the output model. Therefore, two different biological validations can be performed: the discussion of the selected genes' relevance and the discussion of the relations reported among them. According to this idea, the reliability of the results collected in this work is also discussed in both ways: from a pure classification and from a biological point of view.

The remainder of this paper is organized as follows. In Section 2 a brief introduction to the basics of Bayesian classifiers is presented, with more emphasis on the k-dependence Bayesian classifier [25]. The main corpus of our new proposal is presented in Section 3, while results over different gene expression datasets are gathered with an extended numeric and biological discussion in Section 4. Finally, conclusions on the work are discussed in Section 5.

2. Bayesian classifiers

A classifier can be seen as a function that assigns labels to observations,

$$\gamma : (x_1, \ldots, x_n) \to \{1, 2, \ldots, m\},$$

where $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{R}^n$ conforms the observation and $\{1, 2, \ldots, m\}$ are the range of possible values for the class variable. The main assumption is the existence of an unknown underlying probability joint distribution where the observations come from

$$p(x_1, \ldots, x_n, c) = p(c|x_1, \ldots, x_n)p(x_1, \ldots, x_n) = p(x_1, \ldots, x_n|c)p(c). \tag{1}$$

In practice, this joint probability distribution $p(x_1, \ldots, x_n, c)$ can be estimated from a random sample,

$$\{(\boldsymbol{x}^{(1)}, c^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, c^{(N)})\},$$

extracted from the true joint probability distribution.

The usual classification scheme assumes a 0/1 loss function for which the cost of a false positive is the same as the cost of a false negative. In this scenario, the Bayes classifier assigns to the observation $\boldsymbol{x} = (x_1, \ldots, x_n)$ the class with higher a posteriori probability [6]:

$$\gamma(\boldsymbol{x}) = arg\max_c p(c|x_1, \ldots, x_n). \tag{2}$$

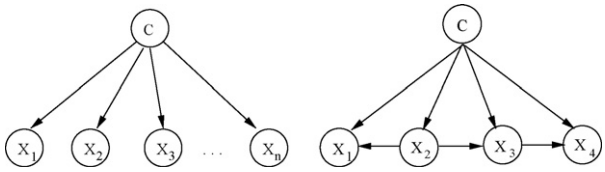

Fig. 1 – Graphical structures of a naïve Bayes classification model (left) with *n* predictive variables and a possible tree augmented naïve Bayes (right) with four predictive variables.

2.1. *Naïve Bayes and tree augmented naïve Bayes classifiers*

The naïve Bayes (NB) classifier [27] uses the Bayes theorem in conjunction with the conditional independence hypothesis. The naïve Bayes paradigm is thus based on two conditions over the features or predictive variables and the class or variable to predict:

- the class variable C can only take one of its *m* possible values $c_1, \ldots, c_m$;
- if the class value is known, the knowledge of some predictive variables is irrelevant to the rest of features. This condition can be mathematically expressed as

$$p(x_1, \ldots, x_n|c) = \prod_{i=1}^{n} p(x_i|c). \quad (3)$$

Therefore, the search for the most probable class value, c^*, once all the $(x_1, \ldots, x_n)$ features are known, can be reduced in the naïve Bayes paradigm to look for

$$c^* = arg\max_c p(c) \prod_{i=1}^{n} p(x_i|c). \quad (4)$$

However, the conditional independence assumption of this paradigm is a very restrictive condition that could ignore relations among the predictive variables. So as to overcome this limitation, the naïve Bayes paradigm can be augmented by embedding a tree-like dependence structure among its predictive variables. This more complex classifier is formally known as tree augmented naïve Bayes [28] or TAN classifier.

In order to induce a TAN classifier, a tree structure among the predictive variables is firstly built, and, then, the class node is related to all the variables. The technique to build the tree is based on the mutual information conditioned to the class variable,

$$I(X, Y|C) = \sum_{i=1}^{v} \sum_{j=1}^{w} \sum_{r=1}^{m} p(x_i, y_j, c_r) \log \frac{p(x_i, y_j|c_r)}{p(x_i|c_r)p(y_j|c_r)}, \quad (5)$$

being X and Y two discrete predictive variables and C the class variable. The complete TAN induction algorithm is based on a tree building algorithm [28] in conjunction with the Kruskall algorithm, and, it can be reviewed in [29]. An example of both naïve Bayes and TAN structures is also shown in Fig. 1.

Step 1. For each predictive variable X_i, $i = 1, \ldots, n$, compute the mutual information with respect to the class variable C, $I(X_i, C)$

Step 2. For each pair of predictive variables, compute the mutual information conditioned to the class, $I(X_i, X_j|C)$, with $i < j$ and $i, j = 1, \ldots, n$

Step 3. Initialize to empty the list of used variables $\aleph$

Step 4. Initialize the Bayesian network classifier to build, BN, to a single node, the one corresponding to the C variable

Step 5. Repeat until $\aleph$ includes all the variables

  Step 5.1. Choose among the variables not included in $\aleph$, that variable X_{max} with highest mutual information with respect to C

  Step 5.2. Add X_{max} into BN

  Step 5.3. Add an arc from C to X_{max} in BN

  Step 5.4. Add $m = \min(|\aleph|, k)$ arcs from the m different variables X_j of $\aleph$ that have the highest values for $I(X_{max}, X_j|C)$

  Step 5.5. Add X_{max} into $\aleph$

Step 6. Compute the conditional probabilities needed to specify the Bayesian network classifier BN

Fig. 2 – kDB algorithm pseudocode [25].

2.2. *k-Dependence Bayesian classifier*

Sahami [25] presents an algorithm called *k*-dependence Bayesian classifier (kDB) that allows to go through the wide spectrum from the naïve Bayes to a complete Bayesian network. The algorithm has its basis in a naïve Bayes structure that allows each predictive variable to have a maximum number of *k* parent variables (excluding the class one).

The simple naïve Bayes classifier corresponds to the 0-dependence Bayesian classifier, the TAN model would be the 1-dependence and the complete Bayesian classifier–structure where there is no independence–would correspond to a $(n-1)$-dependence Bayesian classifier. The kDB induction pseudocode is presented in Fig. 2.

The main idea of this algorithm is to extend the algorithm proposed by Friedman et al. [28] allowing a variable to have a number of parents, excluding the class variable C, bounded by k. This *k* parameter will allow the expert to vary the sparsity degree of the results, focusing on single iteration or in more complex ones. An example of a kDB structure is shown in Fig. 3.

Sahami also introduces a modification in Step 5.4 of the algorithm. The variant, named kDB-θ, do not consider all the possible parent's set bounded by the *k* value, it only includes those dependences which surpass a given threshold θ within the conditional mutual information $I(X_{max}, X_j|C)$.

3. Induction of reliable Bayesian networks

3.1. *Robust arc identification*

The disposal of a low number of instances forces every kind of machine learning technique to look for robustness in its

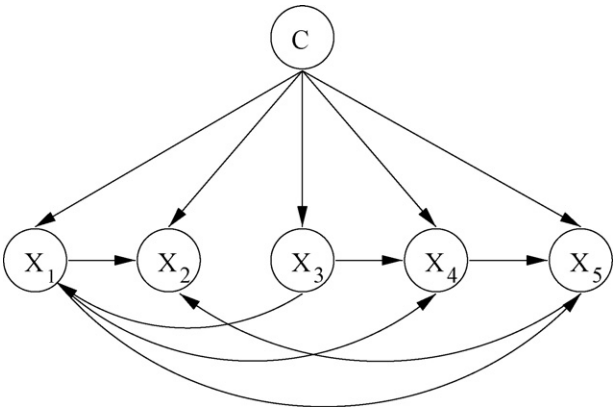

Fig. 3 – Example of a *k*-dependence Bayesian classifier structure.

results. In the microarray context and with this purpose, we propose the combination of two widely known techniques: a *stratified bootstrap* resampling [30] and a feature subset selection [21].

The bootstrap approach was first introduced by Efron [30]. It is based on sampling intermediate databases from the original one. These databases are conformed by instances randomly selected from the original dataset with replacement. The proportion between classes in the original dataset is maintained in each resampled dataset, which is known as stratified bootstrap. This bootstrap scheme is known as non-parametrical bootstrap [20] due to the fact that it needs no external parameter to adjust or compute. On domains where the number of cases is low, the bootstrap scheme is widely used to analyse these data [31].

After the stratified sampling of the dataset, an intermediate feature subset selection step is undertaken. Throughout this step, we look for the most relevant features in each different resampled dataset; datasets that can show differences among them due to the stochastic nature of the bootstrap resampling. The relevant feature selection constitutes a running parameter to be chosen by the researcher. Feature selection methods that return sets of variables rather than individual relevancies are recommended in this step (e.g. correlation feature selection [32], see Section 4.2).

Subsequently, a *k*-dependence Bayesian classifier [25] is induced for each resampled dataset reduced to the found relevant features. On the basis of all the induced *k*DB graphical structures, the confidence of each configured arc between a pair of variables is computed as the relative frequency of its presence in the *B*-induced classification models. Fig. 4 shows the proposed algorithm.

In a *k*-dependence Bayesian classifier model, all the nodes of its structure graph conditionally depend on the class node. These common dependences will not be taken into account: our aim is to find repeated dependency structures among the predictive variables, as well as to identify which variables are reported by those dependences.

Step 1. Repeat B times

Step 1.1. Stratified randomly sample N instances with replacement from the original dataset

Step 1.2. Select an optimal feature subset and reduce the sampled dataset to only those selected features

Step 1.3. Run the induction algorithm on the new reduced dataset, learning a *k*DB classification model

Step 2. Compute the confidence level of each arc as the relative frequency of its presence among all the B induced models

Fig. 4 – Robust arc identification algorithm.

3.2. *Bayesian networks with high-confidence dependences*

Let l_{ij} be the arc from variable X_i to variable X_j. On the basis of the robust arc identification algorithm presented in the previous section, we can define a_{ijr} as

$$a_{ijr} = \begin{cases} 1 & \text{if } l_{ij} \text{ is present in the } r\text{th-induced graph,} \\ 0 & \text{otherwise.} \end{cases}$$

The number of occurrences of a certain arc l_{ij} over the B induced classifiers can be expressed as

$$o_{ij} = \sum_{r=1}^{B} a_{ijr}. \tag{6}$$

From now on, each arc l_{ij} will be associated with its correspondent number of occurrences, o_{ij}. The set of arcs L that have been configured at least once over all the models can be expressed as

$$L = \{l_{ij} | o_{ij} \geq 1\}. \tag{7}$$

Let t be the confidence threshold or reliability level, that is, the number of times that sets the confidence border of the features for an in-depth study. In our case, the set of arcs from L that overcome the threshold t, hereafter known as the set of t-reliability dependences, L_t, is then defined as

$$L_t = \{l_{ij} \in L | o_{ij} \geq t\}. \tag{8}$$

Analogously, the set of variables included in a set of t-reliability dependences L_t, $S(L_t)$, is defined as

$$S(L_t) = \{\mathbf{X}_t \subseteq \{X_1, \ldots, X_n\} | \forall X_i \in \mathbf{X}_t \, \exists X_j \in \mathbf{X}_t \text{ s.t. } l_{ij} \in L_t \vee l_{ij} \in L_t\}. \tag{9}$$

According to L_t and $S(L_t)$, it is possible to build a probabilistic graphical model G_t of t-reliability dependences. In this model, we can find cycles between two variables due to the inclusion of the same arc, but in opposite directions. In such cases, we only take into account the dependence that shows the larger number of occurrences.

Changing the reliability level t, we can build a hierarchy of models, from an empty model to a model that includes

almost all the found arcs. The simplest model corresponds to a reliability level of $t = \max\{o_{ij}\}$, $i, j \in \{1, \ldots, n\}$, when this maximum is unique, L_t only comprises a single dependence and G_t includes two variables and one link between them. At the limit, when $t = 1$, almost every dependence is included; only those that are removed to avoid cycles are not included. In this way, when the value of t varies, the autoinclusion property between all the models is verified, reporting a hierarchy of graphical model structures that can be profoundly analysed:

$$G_{\max\{o_{ij}\}} \subseteq \cdots \subseteq G_t \subseteq \cdots \subseteq G_0. \quad (10)$$

Finally, once a t level is set, the structure of the model G_t can be retrieved and then the parameters obtained from the dataset [33]. The autoinclusive property adds a new characteristic to this gene interaction network: the capability to study how the sets of dependences and variables evolve step-by-step throughout all the models. The $G_{\max\{o_{ij}\}}$ model will presumably include just two variables and an arc between them. Since we decrease the threshold, more variables and arcs will appear in the general model. Thus, it is possible for an expert to control the study deepness and to isolate findings that could comprise a future work hypothesis. In the biodata mining field, the control over the false positives is of crucial interest. So, work hypothesis based on high-confidence thresholds is presumed to be far from a statistic artifact.

4. Experimental results

4.1. Microarray databases

The proposed method is tested using four array sets. Three of them are well-known microarray benchmark sets and have been widely used for this purpose. The latter one corresponds to a regional cancer research alliance [34] and was produced taking into account different colorectal cancer samples. Details about these supervised datasets are as follows:

- *Colon* [35]—This array set comes from a colon gene expression study of 62 samples – 40 tumoral and 22 non-tumoral – with 1989 features from the original 2000 (removing 11 Affymetrix microarray control sequences). Feature intensity values of each microarray are scaled into an average intensity value of 50.
- *Leukemia* [36]—Leukemia dataset is composed of 72 samples in two classes of leukemias: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). From the 7070 original features, only those with 75% presence value in the raw data are included in this study, that is, 1161 features. The two phenotypes are distributed as 47 (ALL patients) and 25 (AML patients) samples per class. Features have been scaled using the factors provided by its authors.
- *Lymphoma* [37]—One of the very first works in high-throughput microarray technology was the analysis of different cells coming from a variety of lymphoma tumors. The set is originally composed of 96 samples and 4026 probes were measured. There are 9 diagnosis classes corresponding with different lymphocyte cell types with cardinalities 46, 2, 2, 10, 6, 6, 9, 4 and 11, respectively.
- *CRC* [34]—This array set comes from a local project under research. It is composed of 65 Agilent microarrays with samples coming from colorectal cancer patients that underwent surgery. There are two classes of samples: those extracted from the cancer polypos and, those extracted from normal colon mucosa epithelium. Searching for possible genetic markers, the experimental design comprised the hybridization of each sample against a reference pool consisting of the non tumoral samples. This way, a total of 32 arrays were hybridized comparing tumoral samples and the pool, and 33 comparing each non tumoral sample with the pool. After an exhaustive process of preprocessing to ensure the data quality – imputation of lost values, the reading quality of the probes, data smoothing and intraclass variability filtering [38]– the set comprises a total of 8104 probes or variables.

Bayesian classifiers can only deal with discrete variables, so a discretization process of the original continuous data is approached. On the basis of its biological activity, it is assumed that a gene can only be in a few different numbers of activity states. As a general criterion in microarray analysis [9,39], this number of states is three: an upregulated, a downregulated and a baseline or null activity. Following this idea, we consider the equal width [40] discretization in three different bins as the most appropriate method to parse the continuous values into discrete states.

4.2. Running parameters

The methodological proposal introduced in Section 3 includes a set of running parameters to be fixed, principally the feature subset selection, a boundary for the maximum number of parents k for the k-dependence Bayesian classifier and the number of times that the boostrap loop is performed. Moreover, and especially in the microarray context, all these parameters are expected to set a scenario in which the running time could be affordable.

Correlation feature selection [32] or CFS has been widely used in this bioinformatics context, reporting good results both in time and in relevant genes [41,42]. CFS belongs to the filter multivariate subset selection techniques [43] and it addresses two fundamental issues by a heuristic function: avoid redundancy and irrelevancy in the selected subset of features. It is able to identify a set of features highly correlated with the phenotype distribution keeping the redundancy among them minimum. Making use of the uncertainty coefficient [44], it defines the heuristic function that guides the search in the space of all possible subset combinations of features. Since this search is NP-hard [21,45], the search strategy is configured in a classical forward greedy hill-climbing search that starts from an empty set of features and incrementally adds new features until the heuristic function is no longer improved. This search strategy also guarantees that the cardinality of the output subsets is not of a high dimension.

Once the dataset is reduced by the CFS, the Bayesian classifier to be learnt is a kDB with a k value of 4. This value allows the graphical models to be both flexible and not sparse when inducing the structures of dependences. Moreover, it implies a sufficient value so none of the possible relevant dependences

could be outside the models. It has been tested experimentally that the kDB-θ does not add any improvement to this design within the microarray domain.

Finally, the proposed algorithm in Section 3.1 is repeated a thousand times, that is, the bootstrap parameter B is set to a value of 1000. This way, we search for arcs that occur a number of times that can be widely considered as reliable.

The complexity order of the full algorithm configured with these parameters can be estimated as the product of the bootstrap parameter B times the computational cost of the feature subset selection and the kDB structure induction. Computing the kDB network structure requires $O(n^2Nmv^2)$, where n is the number of variables, N is the number of cases, m is the number of phenotypes and v is the maximum number of discrete values a predictor variable may take (three in our case). The computational cost of the CFS step is not closed; it depends on the search strategy and on the database's characteristics. In the worst case and by using forward greedy search, the computational cost can be expressed in function of an α parameter that relates the number of original features n and the number of selected features s ($s \approxeq \alpha \cdot n$). For each bootstrap iteration the value of α changes, but we will only consider its maximum value for all the B iterations. In such cases, the total number of operations for a CFS run is delimited by the polynomial expression $\alpha n^3 + (N - \alpha)n^2 - Nn$. In short, the result of the joint algorithm is asymptotically of $\Theta(B\alpha n^3)$ order and the time for computing the conditional probability tables, when the structure is used as a classifier, linearly depends on the number of variables and dependences included when setting the reliability threshold.

4.3. *Graphical outputs*

Table 1 presents a summary of the numeric results provided for each microarray set. Column $|S(L_1)|$ shows the number of probes that are selected at least once from the original set. The next column, $\overline{|L_1|}$, reflects the average number of arcs configured through all the induced classification models—removing those that create cycles among them. Lastly, column Arc collects the most times configured probabilistic relationship for each array set; within each set, the reported arc is included in a total of Max t models out of a thousand ones.

For the *Colon* array set, the total number of variables in $S(L_1)$ selected represents 31% of the original set. The variables not included can be safely discarded for the subsequent knowledge discovery process. Moreover, and taking into account the arcs configured at least a hundred times (threshold $t = 100$), we can radically reduce this number to only 13 variables. Fig. 5 shows the graphical structure compounded by all the arcs included in at least a hundred of the models (shaded nodes match variables without parents apart from the class one). On each arc, the number of times that arc has been included is displayed. Moreover, the graphical thickness of each arc is proportional to each arc's weight. This way it is possible to study the relevance of each dependence and the variables involved within at a glance.

As for the *Leukemia* dataset, Fig. 6 reflects the dependences found at least in a hundred runs out of the total thousand ones. The most configured arc is included in a total of 205 models (D49400_at → U46751_at) which implies that both variables have been jointly selected by the feature selection algorithm and linked by the kDB induction algorithm in such a number of resamplings. From the 1161 original variables, in Fig. 6 only 14 are collected, showing the potential of this technique as a feature subset selection approach.

Regarding the *Lymphoma* array set, the fact that there is no significant reduction in the number of selected variables is interesting. Almost 93% of all the original 4026 variables, a total of 3710 ones, are selected at least one time throughout the experiments. This high number of variables explains the high-average number of arcs configured, more than 180 arcs per model. Both effects come from the fact that this array set is very complex in its phenotype separability: nine classes distributed throughout only 96 samples. With such a low number of instances per class, the conditional statistics evaluated for the classification models make them very dispersed.

Lastly, the results on the *CRC* array set fit with results of the first two array sets. The reduction range in the number of variables is similar, 21.26% variables are selected, and the number of arcs is consistent with this reduction. For this set, the high number of times that the arc TCF3 → ENC1 is included is noticeable, see Fig. 7 for a complete view. At least 80% of all the feature selection runs selected both variables, and, within all those runs, 799 out of 1000 models induced this arc. This fact entails a very high degree of confidence to this dependency. In Section 4.5 the biological background of this finding and others in this array set are discussed in detail, showing a clear correspondence between the statistical models and the biological findings.

4.4. *Clasification accuracy*

Although the priority of this work is to present and apply a new knowledge discovery method, a reliable set of dependences can also be used in a pure classification application. For this purpose, firstly, the expert has to fix a certain value for the dependency threshold t to return the set of variables and arcs which surpass that level, obtaining a single model. This way, the complexity of the models can be tuned, assessing the study's scope, variables or aims. After that, the class node is included in the model, adding arcs from it to the rest of the variables. This way the graphical structure is completed and the correspondent conditional probabilities are computed by their maximum likelihood estimators. Fig. 8 represents the model structures for the *Lymphoma* and *CRC* array sets for thresholds $t = 300$ and 400, respectively, that is, each model contains the probabilistic relationships that have been jointly selected and configured 300 or 400 times at least.

As the confidence threshold falls, the sparsity degree of the models decreases and, thus, the number of variables to be evaluated increases. Therefore, it is of interest to study how the classification models evolve from the very simplest to the most dense ones. In order to analyse this effect, an evaluation of the classification accuracy of each model is performed. Due to the number of models to be evaluated, the total runs and the required computing time for the whole process, a five-fold cross-validation method is used to estimate the final classification accuracy. This estimation scheme was proven to be well suited for the microarray context [46,47], guaranteeing a fair and not overfitted accuracy percentage. For each fold, the

Table 1 – Run statistics for each microarray set

	Features	$\lvert S(L_1)\rvert$	$\overline{\lvert L_1\rvert}$	Max t	Arc
Colon	1989	617	10.67	317	M76378 → J02854
Leukemia	1162	587	15.96	205	D49400_at → U46751_at
Lymphoma	4027	3710	180.64	321	g4012X → g1171X
CRC	8104	1723	35.19	799	TCF3 → ENC1

First column shows the total number of variables for each dataset. Second column, $\lvert S(L_1)\rvert$, gathers the total number of variables selected by the feature selection throughout the B bootstrap iterations. Column $\overline{\lvert L_1\rvert}$ shows the average number of arcs induced by the models and max t the number of times that the most configured dependence (the referred in column *Arc*) is included.

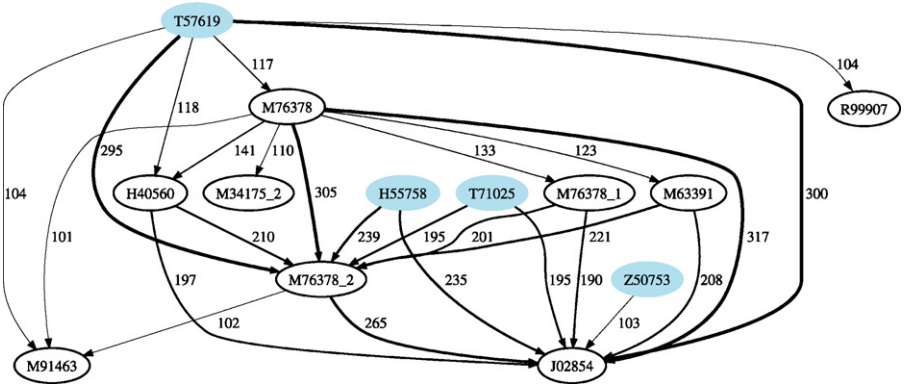

Fig. 5 – Graphical structure of the high-reliable dependences network for the *Colon* dataset and a t value of 100.

run parameters are equal to the ones used in Section 4.2: a thousand bootstrap loops, CFS as multivariate filter method and a value of 4 for the kDB classifiers.

Table 2 gathers, for each array set and for each fold, the number of selected variables, the total number of arcs induced in all the models, the number of times the most often retrieved dependence is recovered, and the maximum average accuracy achieved. Notice that the accuracies shown are jointly evaluated for a fixed confidence threshold. As a visual tool to study the tendency in classification, we have collected for each threshold the number of variables, arcs, mean accuracies and standard deviation in a single plot (see Fig. 9). This figure could be useful to decide to which degree of complexity a biologist is willing to analyse, taking into account the number of variables, arcs and the accuracy level that the model is able to reach.

Inspecting these results shows that there is no direct relationship between the number of arcs/variables and the

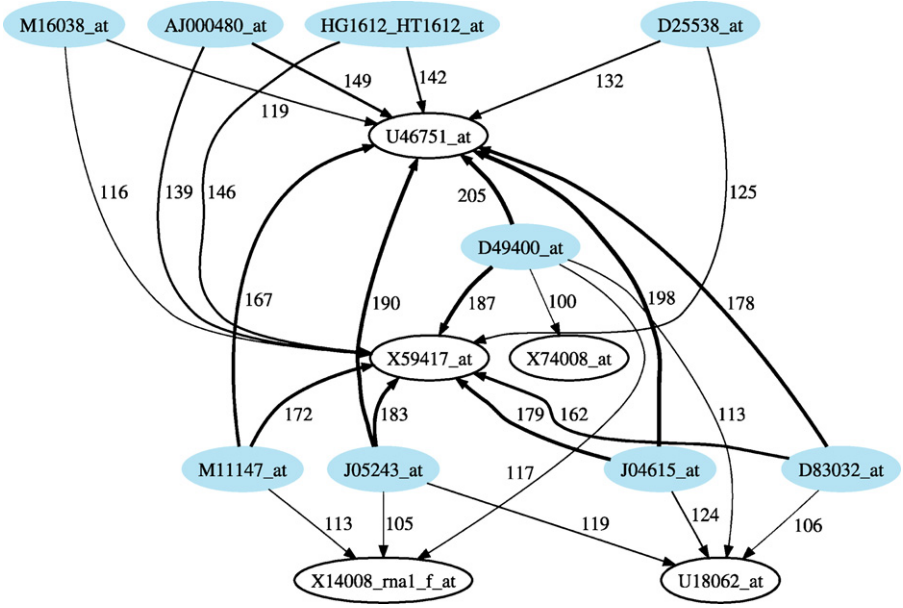

Fig. 6 – Graphical structure of the high-reliable dependences network for the *Leukemia* dataset and a t value of 100.

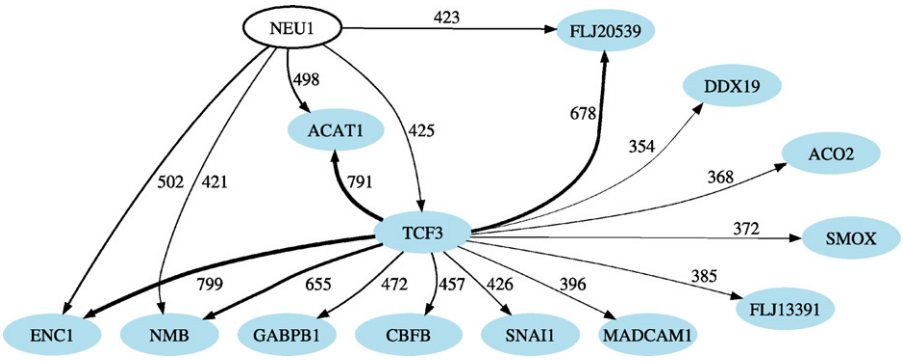

Fig. 7 – Graphical structure of the high-reliable dependences network for the CRC dataset and a t value of 350.

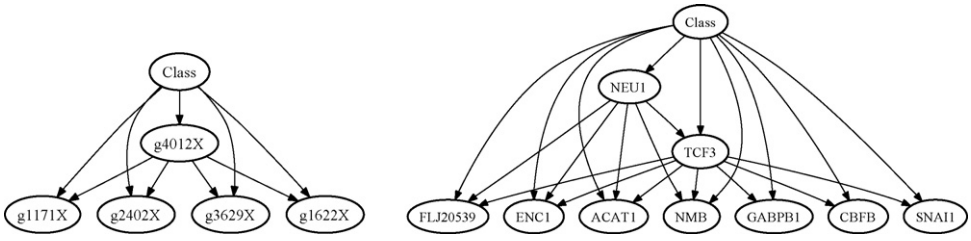

Fig. 8 – Examples of the graphical structures of the network classifiers configured from the high-confidence dependences set. For the *Lymphoma* array set, the threshold is set at 300 (left); for the CRC array set, the corresponding threshold is set at 400 (right).

model's accuracy. Fig. 9 illustrates how, despite the addition of new arcs and thus more variables, there is no guarantee that the accuracies of a more complex model would be higher than the ones from a simpler model. There is a nuclear set of variables/arcs that are able to work out a high degree of the classification separability: more complex models do not necessarily correspond with higher accurate models. For instance, in the *Leukemia* set results, at a confidence level of $t = 200$, four variables with four arcs correctly predict 70% of the samples; in the CRC domain, at a level of $t = 310$, the estimation of the accuracy with only four variables and three arcs achieves a mean value of 96%. This fact corroborates other studies regarding gene expression classification based on a reduced number of genes [4,12,48].

Table 2 – Details about the number of variables and arcs for each cross-validation fold

	$Train_1$	$Train_2$	$Train_3$	$Train_4$	$Train_5$	Mean	S.D.
Colon (1989 vars)							
$\|S(L_1)\|$	461	652	636	668	513	586	92.92
$\overline{\|L_1\|}$	6.43	11.56	10.24	12.85	7.38	9.69	2.73
Max t	352	267	411	265	336	326.2	61.65
Max acc. ($t = 264$)	76.92	92.31	83.33	100	66.67	83.85	13.00
Leukemia (1162 vars)							
$\|S(L_1)\|$	545	489	492	413	534	494.6	51.93
$\overline{\|L_1\|}$	15.00	11.02	12.52	8.71	12.05	11.86	2.29
Max t	209	241	271	217	284	251.25	33.43
Max acc. ($t = 88$)	86.67	60.0	85.71	85.71	64.29	76.48	13.18
Lymphoma (4027 vars)							
$\|S(L_1)\|$	2511	2495	2434	2501	2505	2489.2	31.40
$\overline{\|L_1\|}$	70.28	75.30	72.04	76.90	85.69	76.04	6.00
Max t	462	395	343	454	259	382.6	84.26
Max acc. ($t = 99$)	70	84.21	94.74	89.47	89.47	85.58	9.47
CRC (8104 vars)							
$\|S(L_1)\|$	1223	1205	1188	1073	952	1128.2	114.62
$\overline{\|L_1\|}$	23.55	25.63	23.65	19.08	17.54	21.89	3.41
Max t	689	380	433	439	323	452.8	140.11
Max acc. ($t = 89$)	100	100	100	100	83.33	96.67	7.45

The cardinality of the highest configured arc is included.

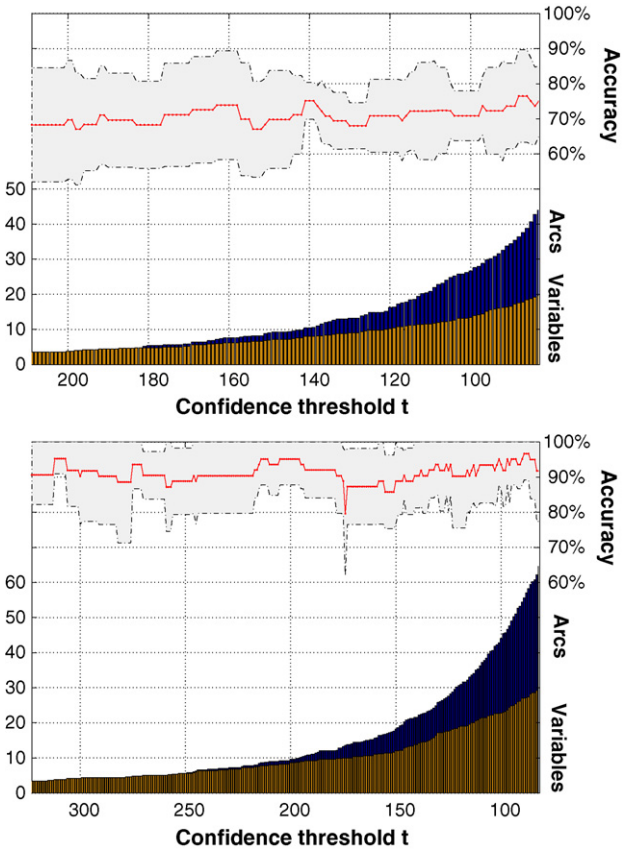

Fig. 9 – Estimated accuracy tendency over the *Leukemia* (up) and CRC (down) array sets. Mean accuracies are presented with their associated standard deviation for each confidence threshold, as well as the number of variables and edges included for that threshold.

The low number of instances in the test set of each fold forces the mean accuracy to have a high level of standard deviation. Thus, accuracy percentages for each array set do not improve the state-of-the-art error rates, but clearly show that recovered high-confidence structures are also able to clear up a significant piece of the phenotype information. All these genes and dependences can be of great interest to reveal new underlying biological knowledge.

4.5. *Biological discussion*

Three out of the four array sets used in the experimentation part of this work, namely *Colon*, *Leukemia* and *Lymphoma*, were published in the late Nineties. From that time, many works in the field have analysed these data and published different sets of relevant genes and interactions. Considering its originality and possible future applications, we will focus our attention in the biological importance of the fourth array set results, the *CRC*, which has been meticulously processed from the very initial production stages to the preprocess and final data retrieval [38].

The graphical dependency structure reported in Fig. 8 for the *CRC* set gathers a total of nine genes given a t threshold of 400. From all of them, TCF3 results in a kernel gene that shows dependences with all of the rest of the genes. This finding perfectly matches the biological function of TCF3, which is the transcription factor 3 or E2A immunoglobulin enhancer binding factors E12/E47. TCF3 coordinately regulates the expression of genes involved in cell survival, cell cycle progression, lipid metabolism, stress response, and lymphoid maturation [49].

In the downstream dependences we find the gene FLJ20539, also known as GBP. Lee et al. [50] describes a physiological regulation of [beta]-catenin stability by TCF3 and CK1epsilon. Moreover, another of the reported genes, CBFB encodes a protein that belongs to the beta subunit of a heterodimeric core-binding transcription factor belonging to the PEBP2/CBF transcription factor family which master-regulates a host of genes specific to hematopoiesis [51](e.g. RUNX1) and osteogenesis (e.g. RUNX2). The expression of CBFB is down regulated in a significant portion of gastric cancer cases, which may be involved in gastric carcinogenesis [52]. In addition, several studies suggest that lack of RUNX3 function is causally related to the genesis and progression of human gastric cancer, but potential roles of other members of the RUNX family genes have not yet been reported. Furthermore, CBFB, the gene encoding the cofactor of RUNX1, -2, -3, was also downregulated in significant fraction (32%, $p < 0.05$). The percentage of downregulation of RUNX1, RUNX3 and CBFB increases as the cancer stage progresses. All these findings and relationships constitute a serious biological hypothesis between the activity of CBFB and the gastric or colorectal carcinogenesis.

From its part, SNAI1 gene is present in activated mesenchymal cells indicating its relevance in the communication between tumor and stroma and this fact suggests that it can promote the conversion of carcinoma cells to stromal cells [53]. Its expression in colorectal tumors is also associated with downregulation of E-cadherin (CDH1) and vitamin D receptor gene products [54]. The work by Takahashi et al. [55] demonstrated that inhibition in SNAI1 is directly induced by TCF3. In mice, a human ortholog of human TCF3 is reported as a direct sequence-specific activator of negative vitamin D response element [56], which clearly supports the SNAI1 findings in colorectal tumors.

Other dependence found by our method shows a relation between NEU1 and the transcription factor TCF3. The protein encoded by NEU1 encodes the lysosomal enzyme, which cleaves terminal sialic acid residues from substrates such as glycoproteins and glycolipids. Upregulation of the NEU1 expression is important for the primary function of macrophages and there is a link between NEU1 and the cellular immune response [57]: data show that the differentiation of monocytes into macrophages is associated with the specific upregulation of the enzyme activity of NEU1 [58]. Greenbaum et al. [59] reported that TCF3 is a negative regulator of a set of genes involved in the development of B lymphocytes, thus, showing the link between TCF3 and NEU1.

Table 3 gathers the summary of the dependences that have been previously reported by biological works. Notice that the confidence levels for these arcs are very high ($t = 400$), which corroborates the reliability of these results. From the rest of these genes, two of them are directly related with colorectal cancers: ENC1 and NMB. Ectodermal-neural cortex 1 (ENC1) belongs to the p53-induced gene set and it is also known

Table 3 – High-confidence ($o_{ij} \geq 400$) interactions reported by both our method and also by the biological literature for the CRC array set

Dependence	Confidence level	Reference
TCF3 → FLJ20539	$o_{ij} = 678$	[50]
TCF3 → CBFB	$o_{ij} = 457$	[51]
TCF3 → SNAI1	$o_{ij} = 426$	[55,56]
NEU1 → TCF3	$o_{ij} = 425$	[59]

as PIG10 gene [60]. The influence of this PIG with colorectal cancer was firstly published by [61]. NMB (neuromedin B) is associated with eating behaviours and obesity [62]; NMB and its receptor are coexpressed by proliferating cells in which they act in an autocrine fashion with similar and modest potency in both normal and malignant colonic epithelial cells [63].

The last two genes are not yet related to the cancer field. Acetyl-coenzyme A acetyltransferase 1 (ACAT1) is associated with the alpha-methylacetoaceticaciduria disorder, an inborn error of isoleucine catabolism characterized by urinary excretion of 2-methyl-3-hydroxybutyric acid, 2-methylacetoacetic acid, tiglylglycine, and butanone. The length of ACAT1 is approximately of 27 kb and contains 12 exons. Due to this high dimension, many mutations have been found for this gene [64] and many of them are under study [65]. Lastly, GABPB1 (GA-binding protein transcription factor, beta subunit) stimulates transcription of target genes. The encoded protein may be involved in activation of cytochrome oxidase expression and nuclear control of mitochondrial function. Biologists have identified multiple transcript variants encoding distinct isoforms of the protein. All of this suggests a general purpose compound that may be found in many biological processes.

5. Conclusions

Throughout this work a new approach to identify gene interactions has been proposed based on the consensus of Bayesian networks learnt from a pool of bootstrap samples. A major feature of our proposal is the possibility to set confidence levels in order to rely only on interactions highly supported by the expression data. It offers to the expert a broad range of probabilistic dependences to be studied, depending on the available time and laboratory resources.

A gene interaction network represents information in a richer way than univariate lists of genes. It describes groups of closely connected genes, unveiling biological knowledge or work hypothesis for both biologists and physicians. Hypothesis driven studies can converge with this data driven technique: it opens the possibility to study how a given gene or dependence interacts with the rest of the genes included in a study. This way, a beforehand hypothesis could be corroborated by a 'blind' data mining approach.

Our approach is grouped into the supervised classification methods because it makes use of the phenotype class variable. Bayesian classifiers induce their structure by means of class-conditional probabilities, therefore, studies that compare control against illness samples are feasible targets for this technique. The conjunction of a triplet of well-known machine learning procedures (a stratified boostrap, a feature selection and a Bayesian k-dependence classifier) assures a robust set of results, and, even more importantly, a low number of false positives. A hierarchy of structures is computed, allowing the user to set a threshold in the frequency of appearance of each arc in the pool of bootstrap models. The hierarchy reports for this given threshold both a set of dependences and a set of variables; therefore, it also constitutes a variable subset selector.

Reported results in this work have shown the potentiality of the induced models in a pure classification task. Reduced sets of dependences/variables are able to achieve a competitive degree of accuracy when performing a class-discrimination procedure, corroborating previous statements in the microarray analysis field.

Besides, the biological analysis of the results for the novel CRC array set has proven a flawless correspondence between our method's findings and the evidences found in the biological state-of-the-art. As important as the accomplishment of previous hypothesis is the pointing out of new research targets. This knowledge discovery application brings into focus a new set of tools to help understand complex diseases that show relationships of different degrees among the involved genes.

Conflict of interest statement

None declared.

Acknowledgments

This work has been partially supported by the Etortek, Saiotek and Research Groups 2007-2012 (IT-242-07) programs (Basque Government), TIN2005-03824 and Consolider Ingenio 2010-CSD2007-00018 projects (Spanish Ministry of Education and Science) and COMBIOMED network in computational biomedicine (Carlos III Health Institute). R. Armañanzas is supported by the Basque Government grant AE-BFI-05/430.

REFERENCES

[1] N. Friedman, Inferring cellular networks using probabilistic graphical models, Science 303 (5659) (2004) 799–805.
[2] G. Bontempi, A blocking strategy to improve gene selection for classification of gene expression data, IEEE/ACM Trans. Comput. Biol. Bioinform. 4 (2) (2007) 293–300.
[3] T. Lin, R. Liu, C. Chen, Y. Chao, S. Chen, Pattern classification in DNA microarray data of multiple tumor types, Pattern Recogn. 39 (2006) 2426–2438.
[4] L. Wang, F. Chu, W. Xie, Accurate cancer classification using expressions of very few genes, IEEE/ACM Trans. Comput. Biol. Bioinform. 4 (1) (2007) 40–53.
[5] K. Yang, Z. Cai, J. Li, G. Lin, A stable gene selection in microarray data analysis, BMC Bioinform. 7 (1) (2006)228.
[6] R. Duda, P. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, 1973.
[7] U.M. Braga-Neto, E.R. Dougherty, Is cross-validation valid for small-sample microarray classification? Bioinformatics 20 (3) (2004) 374–380.

[8] S. Michiels, S. Koscielny, C. Hill, Prediction of cancer outcome with microarrays: a multiple random validation strategy, Lancet 365 (2005) 488–492.
[9] N. Friedman, M. Linial, I. Nachman, D. Pe'er, Using Bayesian networks to analyze expression data, J. Comput. Biol. 7 (2000) 601–620.
[10] J.M. Peña, J. Björkegren, J. Tegnér, Growing Bayesian network models of gene networks from seed genes, Bioinformatics 21 (Suppl. 2) (2005) ii224–ii229.
[11] D. Pe'er, A. Regev, G. Elidan, N. Friedman, Inferring subnetworks from perturbed expression profiles, Bioinformatics 17 (2001) S215–S224.
[12] S.G. Baker, B.S. Kramer, Identifying genes that contribute most to good classification in microarrays, BMC Bioinform. 7 (2006) 407.
[13] I. Shmulevich, I. Gluhovsky, R. Hashimoto, E.R. Dougherty, W. Zhang, Steady-state analysis of genetic regulatory networks modeled by probabilistic Boolean networks, Comp. Funct. Genom. 4 (2003) 601–608.
[14] J. Wang, L.W. Cheung, J. Delabie, New probabilistic graphical models for genetic regulatory networks studies, J. Biomed. Inform. 38 (2005) 443–455.
[15] A.J. Hartemink, D.K. Gifford, T.S. Jaakkola, R.A. Young, Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks, in: Proceedings of the Pacific Symposium on Biocomputing, vol. 6, 2001, pp. 422–433.
[16] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, J. Vert, Classification of microarray data using gene networks, BMC Bioinform. 8 (1) (2007) 35.
[17] D. Pe'er, A. Tanay, A. Regev, Minreg: a scalable algorithm for learning parsimonious regulatory networks in yeast and mammals, J. Mach. Learn. Res. 7 (2006) 167–189.
[18] P. Larrañaga, J.A. Lozano, J.M. Peña, I. Inza (guest editors), Special issue on probabilistic graphical models for classification, Mach. Learn. 59 (3) (2005).
[19] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J.A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, V. Robles, Machine learning in bioinformatics, Brief. Bioinform. 7 (1) (2006) 86–112.
[20] N. Friedman, M. Goldsmidt, A. Wyner, Data analysis with Bayesian networks: a bootstrap approach, in: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, 1999, pp. 196–205.
[21] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[22] W. Li, Y. Yang, How many genes are needed for a discriminant microarray data analysis? in: S.M. Lin, K.F. Johnson (Eds.), Methods of Microarray Data Analysis: Papers from CAMDA'00, Kluwer Academic, Boston, 2002, pp. 137–150.
[23] S. Monti, P. Tamayo, J. Mesirov, T. Golub, Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data, Mach. Learn. 52 (1–2) (2003) 91–118.
[24] S. Swift, A. Tucker, V. Vinciotti, N. Martin, C. Orengo, X. Liu, P. Kellam, Consensus clustering and functional interpretation of gene-expression data, Genome Biol. 5 (11) (2004) R94.1–R94.16.
[25] M. Sahami, Learning limited dependence Bayesian classifiers, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 335–338.
[26] X. Zhou, X. Wang, R. Pal, I. Ivanov, M. Bittner, E.R. Dougherty, A Bayesian connectivity-based approach to constructing probabilistic gene regulatory networks, Bioinformatics 20 (17) (2004) 2918–2927.
[27] M. Minsky, Steps toward artificial intelligence, Trans. Inst. Radio Eng. 49 (1961) 8–30.
[28] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (2) (1997) 131–164.
[29] R. Blanco, I. Inza, M. Merino, J. Quiroga, P. Larrañaga, Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS, J. Biomed. Inform. 38 (5) (2005) 376–388.
[30] B. Efron, Bootstrap methods: another look at the jackknife, Ann. Stat. 7 (1979) 1–26.
[31] J.L. Simon, Resampling: The New Statistics, Resampling Stats, 1997.
[32] M.A. Hall, L.A. Smith, Feature subset selection: a correlation based filter approach, in: Proceedings of the Fourth International Conference on Neural Information Processing and Intelligent Information Systems, Dunedin, 1997, pp. 855–858.
[33] D. Heckerman, D. Geiger, D.M. Chickering, Learning Bayesian networks: the combination of knowledge and statistical data, Mach. Learn. 20 (1995) 197–243.
[34] A. García, A. Freije, R. Armañanzas, I. Inza, Z. Ispizua, P. Heredia, P. Larrañaga, G. López-Vivanco, T. Suárez, M. Betanzos, Simultaneous search of genomic and proteomic biomarkers in human colorectal cancer, in: Genomes to Systems Conference, Manchester, 2006.
[35] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. U.S.A. 96 (12) (1999) 6745–6750.
[36] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531–537.
[37] A.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, L.M. Staudt Jr., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511.
[38] A. García, A. Freije, R. Armañanzas, I. Inza, Z. Ispizua, P. Heredia, P. Larrañaga, G. LópezVivanco, T. Suárez, M. Betanzos, Gene expression model for the classification of human colorectal cancer and potential CRC biomarkers search, in: Drug Discovery Technology, London, 2007.
[39] H.C. Causton, J. Quackenbush, A. Brazma, Microarray Gene Expression Data Analysis: A Beginner's Guide, Blackwell Publishers, Malden, 2003.
[40] R. Kerber, Chimerge: discretization for numeric attributes, in: National Conference on Artificial Intelligence, 1992, pp. 123–128.
[41] M.A. Hall, L.A. Smith, Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper, in: Proceedings of the Florida Artificial Intelligence Research Symposium, Orlando, 1999, pp. 235–239.
[42] Y. Wang, I.V. Tetko, M.A. Hall, E. Frank, A. Facius, K.F.X. Mayer, H.W. Mewes, Gene selection from microarray data for cancer classification—a machine learning approach, Comput. Biol. Chem. 29 (2005) 37–46.
[43] M. Ben-Bassat, Use of distance measures, information measures and error bounds in feature evaluation, in: P.R. Krishnaiah, L.N. Kanal (Eds.), Handbook of Statistics, vol. 2, North-Holland Publishing Company, 1982, pp. 773–791.
[44] M.A. Hall, Correlation-Based Feature Subset Selection for Machine Learning, PhD Thesis, Department of Computer Science, University of Waikato, 1999.

[45] M.R. Garey, D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W.H. Freeman, San Francisco, 1979.
[46] R.R. Bouckaert, E. Frank, Evaluating the replicability of significance tests for comparing learning algorithms, in: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining—PAKDD, Sydney, 2004, pp. 3–12.
[47] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, Bioinformatics 21 (5) (2005) 631–643.
[48] 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 (15) (2004) 2429–2437.
[49] R. Schwartz, I. Engel, M. Fallahi-Sichani, H.T. Petrie, C. Murre, Gene expression patterns define novel roles for E47 in cell cycle progression, cytokine-mediated signaling, and T lineage development, Proc. Natl. Acad. Sci. U.S.A. 103 (26) (2006) 9976–9981.
[50] E. Lee, A. Salic, M.W. Kirschner, Physiological regulation of beta-catenin stability by Tcf3 and CK1epsilon, J. Cell Biol. 154 (5) (2001) 983–993.
[51] R. Bayly, L. Chuen, R.A. Currie, B.D. Hyndman, R. Casselman, G.A. Blobel, D.P. LeBrun, E2A-PBX1 interacts directly with the KIX domain of CBP/p300 in the induction of proliferation in primary hematopoietic cells, J. Biol. Chem. 279 (53) (2004) 55362–55371.
[52] C. Sakakura, A. Hagiwara, K. Miyagawa, S. Nakashima, T. Yoshikawa, S. Kin, Y. Nakase, K. Ito, H. Yamagishi, S. Yazumi, T. Chiba, Y. Ito, Frequent downregulation of the runt domain transcription factors RUNX1, RUNX3 and their cofactor CBFB in gastric cancer, Int. J. Cancer 113 (2) (2005) 221–228.
[53] C. Francí, M. Takkunen, N. Dave, F. Alameda, S. Gómez, R. Rodríguez, M. Escrivà, B. Montserrat-Sentís, T. Baró, M. Garrido, F. Bonilla, I. Virtanen, A. García de Herreros, Expression of SNAIL protein in tumor–stroma interface, Oncogene 25 (37) (2006) 5134–5144.
[54] C. Peña, J.M. García, J. Silva, V. García, R. Rodríguez, I. Alonso, I. Millán, C. Salas, A. García de Herreros, A. Muñoz, F. Bonilla, E-cadherin and vitamin D receptor regulation by SNAIL and ZEB1 in colon cancer: clinicopathological correlations, Hum. Mol. Genet. 14 (22) (2005) 3361–3370.
[55] E. Takahashi, N. Funato, N. Higashihori, Y. Hata, T. Gridley, M. Nakamura, Snail regulates p21(WAF/CIP1) expression in cooperation with E2A and Twist, Biochem. Biophys. Res. Commun. 325 (4) (2004) 1136–1144.
[56] A. Murayama, M.S. Kim, J. Yanagisawa, K. Takeyama, S. Kato, Transrepression by a liganded nuclear receptor via a bHLH activator through co-regulator switching, EMBO J. 23 (7) (2004) 1598–1608.
[57] F. Liang, V. Seyrantepe, K. Landry, R. Ahmad, A. Ahmad, N.M. Stamatos, A.V. Pshezhetsky, Monocyte differentiation up-regulates the expression of the lysosomal sialidase, Neu1, and triggers its targeting to the plasma membrane via major histocompatibility complex class II-positive compartments, J. Biol. Chem. 281 (37) (2006) 27526–27538.
[58] N.M. Stamatos, F. Liang, X. Nan, K. Landry, A.S. Cross, L.X. Wang, A.V. Pshezhetsky, Differential expression of endogenous sialidases of human monocytes during cellular differentiation into macrophages, J. Feder. Eur. Biochem. Soc. 272 (10) (2005) 2445–2456.
[59] S. Greenbaum, A.S. Lazorchak, Y. Zhuang, Differential functions for the transcription factor E2A in positive and negative gene regulation in pre-B lymphocytes, J. Biol. Chem. 279 (43) (2004) 45028–45035.
[60] K. Polyak, Y. Xia, J.L. Zweier, K.W. Kinzler, B. Vogelstein, A model for p53-induced apoptosis, Nature 389 (6648) (1997) 300–305.
[61] M. Fujita, Y. Furukawa, T. Tsunoda, T. Tanaka, M. Ogawa, Y. Nakamura, Up-regulation of the ectodermal-neural cortex 1 (ENC1) gene, a downstream target of the beta-catenin/T-cell factor complex, in colorectal carcinomas, Cancer Res. 61 (21) (2001) 7722–7726.
[62] L. Bouchard, V. Drapeau, V. Provencher, S. Lemieux, Y. Chagnon, T. Rice, D.C. Rao, M.C. Vohl, A. Tremblay, C. Bouchard, L. Pèrusse, Neuromedin beta: a strong candidate gene linking eating behaviors and susceptibility to obesity, Am. J. Clin. Nutr. 80 (6) (2004) 1478–1486.
[63] D. Matusiak, S. Glover, R. Nathaniel, K. Matkowskyj, J. Yang, R.V. Benya, Neuromedin B and its receptor are mitogens in both normal and malignant epithelial cells lining the colon, Am. J. Physiol.: Gastrointest. Liver Physiol. 288 (4) (2005) G718–G728.
[64] T. Fukao, N. Matsuo, G.X. Zhang, R. Urasawa, T. Kubo, Y. Kohno, N. Kondo, Single base substitutions at the initiator codon in the mitochondrial acetoacetyl-CoA thiolase (ACAT1/T2) gene result in production of varying amounts of wild-type T2 polypeptide, Hum. Mutat. 21 (6) (2003) 587–592.
[65] G. Zhang, T. Fukao, S. Sakurai, K. Yamada, K. Michael Gibson, N. Kondo, Identification of Alu-mediated, large deletion-spanning exons 2–4 in a patient with mitochondrial acetoacetyl-CoA thiolase deficiency, Mol. Genet. Metab. 89 (3) (2006) 222–226.