BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 21 2007, pages 2866–2872
doi:10.1093/bioinformatics/btm429

*Gene expression*

# Feature construction from synergic pairs to improve microarray-based classification

Blaise Hanczar[1,*,†], Jean-Daniel Zucker[1,2,3,4,†], Corneliu Henegar[2,3,4] and Lorenza Saitta[5]

[1]Laboratoire d'Informatique Médicale et Bioinformatique (Lim & Bio), Université Paris 13, 93017 Bobigny, [2]Université Paris Descarte, F-75006, [3]Université Pierre et Marie Curie - Paris 6, Centre de recherche des Cordeliers, UMR S 872, [4]INSERM, U872, Paris, F-75006, France and [5]Dipartimento di Informatica, Universitá del Piemonte Orientale, 15100 Alessandria, Italy

Received on February 5, 2007; revised on July 6, 2007; accepted on August 18, 2007

Advance Access publication October 9, 2007

Associate Editor: Joaquin Dopazo

## ABSTRACT

**Motivation:** Microarray experiments that allow simultaneous expression profiling of thousands of genes in various conditions (tissues, cells or time) generate data whose analysis raises difficult problems. In particular, there is a vast disproportion between the number of attributes (tens of thousands) and the number of examples (several tens). Dimension reduction is therefore a key step before applying classification approaches. Many methods have been proposed to this purpose, but only a few of them considered a direct quantification of transcriptional interactions. We describe and experimentally validate a new dimension reduction and feature construction method, which assesses interactions between expression profiles to improve microarray-based classification accuracy.

**Results:** Our approach relies on a mutual information measure that exposes some elementary constituents of the information contained in a pair of gene expression profiles. We show that their analysis implies a term that represents the information of the interaction between the two genes. The principle of our method, called *FeatKNN*, is to exploit the information provided by highly synergic gene pairs to improve classification accuracy. First, a heuristic search selects the most informative gene pairs. Then, for each selected pair, a new feature, representing the classification margin of a KNN classifier in the gene pairs space, is constructed. We show experimentally that the interactional information has a degree of significance comparable to that of the gene expression profiles considered separately. Our method has been tested with different classifiers and yielded significant improvements in accuracy on several public microarray databases. Moreover, a synthetic assessment of the biological significance of the concept of synergic gene pairs suggested its ability to uncover relevant mechanisms underlying interactions among various cellular processes.

**Contact:** hanczar_blaise@yahoo.fr

**Supplementary information:** Complementary results can be found on the companion website at http://featknn.nutriomique.org

## 1 INTRODUCTION

Most cellular processes need to accommodate concomitantly various types of solicitations, related either to the specificities of a particular cellular state, or to variations of the parameters of the intra- or the extracellular environments. Complex regulatory mechanisms are therefore integrating internal demands, environmental fluctuations as well as various extracellular signals (e.g. growth factors, mediators, hormones, other autocrine and paracrine signals, etc.), and initiate specific adaptive processes to maintain the metabolic homeostasis of the cell and to assure its systemic role in the organism. In this article, we propose an original approach designed to capture synergic interactions between cellular processes from the information encoded in the gene expression profiles, to improve the accuracy of the classification of microarray experiments. We investigate the feasibility and the potential advantages of considering gene information interactions in the very phase of dimension reduction. We show that pairs of genes with a high discrimination power need not include genes that are both individually discriminant. Therefore, a feature reduction method that does not consider interactions explicitly is likely to miss such useful pairs. Figure 1 illustrates this situation. Genes Hsa.1221 and Hsa.9025 are used to discriminate between two classes (control subjects and patients affected by colon cancer), represented by black and white dots, respectively. Expression of gene Hsa.1221 is very useful for discrimination, as it appears from the presence of a large proportion of white dots between the values $-0.5$ and 2. This is not the case of gene Hsa.9025: the values of the samples in both classes are spread over the whole range of expression levels. A standard feature selection method is likely to consider only the first gene as relevant for the classification. However, as Fig. 1 suggests, the association of these two genes may improve significantly the discrimination between the two classes.

We devised an original approach that computes the mutual information contained in the gene expression profiles to identify gene pairs showing the strongest synergies, and then used them to improve the accuracy of microarray experiments classification. We point out that such synergic interactions capture a biological information which is contextually relevant.

*To whom correspondence should be addressed.
†The authors wish it to be known that in their opinion the first two authors should be regarded as First Authors.

 

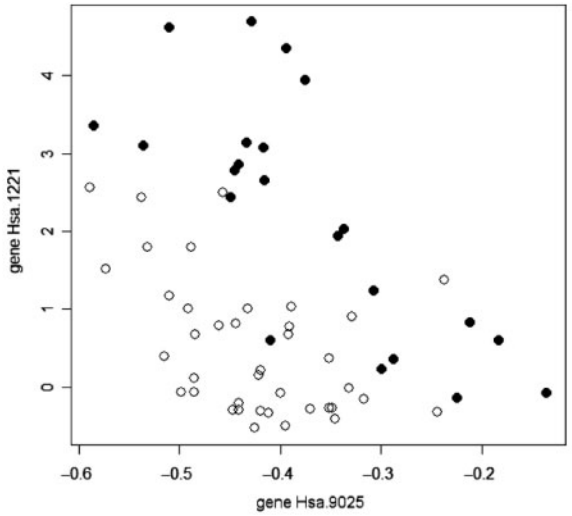


**Fig. 1.** Example of synergy between two genes. The plot shows the expressions of genes Hsa.9025 and Hsa.1221 from the colon cancer dataset. White dots represent sick patients and black dots normal controls. The association of the two genes clearly distinguishes the two conditions.

## 2 RELATED WORK

There is a vast amount of work on gene reduction methods to improve microarray data classification. Widely used, the *scoring* approaches take an individual perspective by computing for each gene a relevance score, depending on how well the gene distinguishes the examples of different classes. A good review of this kind of approaches is provided by Ben-Dor (Ben-Dor *et al*., 2000). These methods are useful for microarray data because they are fast (linear complexity with the number of dimensions). However, they can only evaluate the relevance of genes with respect to the class, but cannot discover redundancy and basic interactions among genes. For this reason, the most competitive methods are multivariate ones that rely on groups of genes (i.e. the selection of a specific gene depends on the others), instead of considering each gene individually. Researchers assume generally that a good gene subset is one that contains genes that are highly correlated with the class, yet uncorrelated among them. Based on this idea, several selection methods have been developed. For example, MRMR (Maximum Relevance, Minimum Redundancy) (Ding and Peng, 2003) uses mutual information to select genes with maximum relevance and minimal redundancy. ProGene (Hanczar *et al*., 2003) reduces redundancy by building new features from subsets of similar genes. In a recent work, Dai *et al*. (2006) have carried out an extensive study to compare three reduction methods, including partial least square (PLS). PLS builds new features corresponding to components that maximize the covariance between the variables and the class.

Other types of approaches aim to improve microarray classification by integrating available a priori biological knowledge about gene interactions. In this category, Rapaport *et al*. (2007) have recently proposed an original method, which integrates KEGG metabolic interaction networks into a spectral decomposition of gene expression profiles to derive a classification algorithm. Whenever available, this type of data integration should be used to improve any of the gene selection approaches mentioned above.

Most of the available dimension-reduction methods do not take into account explicitly the interactions among genes, although some proposals of using pairwise gene interactions do exist. Bo and Jonassen (2002) evaluate the gene pairs by computing the projected coordinates of each example on the axis of the diagonal linear discriminant in the gene-pair space. The score is the two sample $t$-statistic on the projected points. Geman *et al*. (2004) do not use the expression value, but the expression rank of the genes. The pair score is computed from the probability that the expression rank of the first gene of the pair is higher than that of the second one, in each class. Their experimental results confirm the claim that class prediction can be improved using pairs of genes. In this article, we propose to identify strongly interacting genes and systematically exploit these pairs of synergies to improve the classification accuracy and the biological significance of the results.

## 3 DECOMPOSITION OF GENE PAIR INFORMATION

Let us first recall some definitions concerning mutual information (Shannon, 1948). The entropy $H(X)$ of a variable $X$, which can take $m$ values $\{x_1, \ldots, x_m\}$, each value $x_i$ with a probability $p(X = x_i)$, is defined as follows:

$$H(X) = -\sum_{i=1}^{m} p(X = x_i) \log\ p(X = x_i)$$

The following propriety about entropy holds for any pair of stochastic variables $X$ and $Y$:

$$H(X, Y) = H(Y, X) = H(X|Y) + H(Y)$$

The mutual information $I(X,Y)$ is a measure of the dependency between two variables $X$ and $Y$:

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y, X)$$

The following propriety about information holds for any pair of stochastic variables $X$ and $Y$:

$$0 \leq I(X, Y) \leq \inf(H(X), H(Y))$$

Unlike the second order mutual information, the third order mutual information can be either positive or negative. The mutual information between three variables $(X, Y, Z)$ is defined as follows (Matsuda, 2000):

$$I(X, Y, Z) = -H(X, Y, Z) + H(X, Y) + H(X, Z) + H(Y, Z) - H(X) - H(Y) - H(Z)$$

In the case of microarray-based classification, the mutual information between a gene $G_i$ and a class $C$ represents the information that the gene provides to classify. The higher the mutual information, the more informative the gene. For reasons of tractability and simplicity, expression levels are often discretized. The most straightforward and widely used approach relies on a histogram-based technique (Butte and Kohane, 2000). The data is partitioned into equal-width discrete bins, and the equations above may be used. The mutual information $I(G_i, C)$ can be expressed as follows:

$$I(G_i, C) = H(C) - H(C|G_i)$$

According to the above formula, the mutual information between a gene $G_i$ and the class $C$ can be seen as the reduction of the class entropy caused by the knowledge of $G_i$. In the same way, we define the mutual information between the class $C$ and a pair of gene $\{G_iG_j\}$ formed by the two genes $G_i$ and $G_j$:

$$I(G_iG_j, C) = I(G_i, C) + I(G_j, C) - I(G_i, G_j, C) \quad (1)$$

Formula (1) can be proved by the following argument. The left-hand side of (1) can also be written:

$$I(G_iG_j, C) = H(C) - H(C|G_i, G_j) = \\ = H(C) - H(G_i, G_j, C) + H(G_i, G_j)$$

By introducing the terms $H(G_i, C)$, $H(G_j, C)$, $H(G_i)$, $H(G_j)$ and $H(C)$, we obtain:

$$\begin{aligned} I(G_iG_j, C) = &-H(G_i, C) + H(G_i) + H(C) \\ &-H(G_j, C) + H(G_j) + H(C) \\ &-H(G_i, G_j, C) + H(G_i, G_j) + H(G_i, C) \\ &+H(G_j, C) - H(G_i) - H(G_j) - H(C) \\ &= I(G_i, C) + I(G_j, C) - I(G_i, G_j, C) \end{aligned}$$

Formula (1) shows that the information of a gene pair is the sum of the information of the first gene, the information of the second gene, and the third order mutual information between $G_i$, $G_j$ and $C$. This last term represents the information provided to the classification by the association of the two genes. We call this term the *interaction*, which can be either positive or negative. In the case of a positive interaction, the information of the gene pair is lower than the sum of the information of the two genes. In this case part of the information provided by the genes is similar, and therefore we may speak of *redundancy* between the two genes. On the contrary, when the interaction is negative the information of the gene pair is higher than the sum of the information of the two genes, which means that the association of the two genes provides new information. We speak then of *synergy* between the genes (Jakulin and Bratko, 2003). In the example of Figure 1, the information contained in the expression of gene Hsa.1221 is 0.20, whereas it is 0.03 for gene Hsa.9025, and −0.27 for their synergic interaction. The information of the pair formed by these two genes is $0.50 = 0.20 + 0.03 - (-0.27)$.

It should be underlined that the mutual information is computed from the probability distributions of the gene expression (Steuer *et al.*, 2002). However, because the real probabilities are unknown, as they are only estimated from limited data, we have conducted a set of experiments (details of these experiments can be found on the companion website), which shows that the mutual information is accurate enough to identify the most informative pairs of genes.

## 4 REDUCING DIMENSIONALITY USING SYNERGIES

In this section, we will describe a new dimension reduction method, called *FeatKNN*, based on the use of synergic pairs of genes. The most informative gene pairs are identified using a sequential forward search (SFS) procedure. Then, for each gene pair, a new feature, which summarizes the information contained in the pair, is constructed.

### 4.1 The search for the most informative pairs of genes

The naive approach to finding informative pairs of genes consists in computing the mutual information with respect to the class of all the $N(N-1)/2$ different pairs, where $N$ is the number of genes, and then selecting the $p$ best ones. However, this approach has a complexity $O(N^2)$; in the context of microarray data, where the number of genes is of the order of several thousands, this solution is often computationally infeasible. But, even in the case where the complexity may not be a problem, there are two more reasons why this simple approach might be unsuitable. The first is that we would like to select pairs that not only provide high information for the classification, but which provide an information superior to the one brought by the single genes (i.e. the genes should interact negatively). In fact, only those genes that satisfy this property are interesting in the context of this work, which is based on the assumption (grounded on biological findings) that synergic gene interaction is important for improving classification.

The second reason is that correlated attributes usually provide duplicated (redundant) information; in learning, it is well known that it is preferable to exploit as diverse and independent sources of information as possible. Preliminary experiments showed that in some rare cases a dataset may contain an exceptionally informative gene, such that it forms a 'good' pair when coupled with a large number of the other genes. This situation is undesirable, and we want to avoid it.

In order to face the above problems, we propose in this article a simple search algorithm, guided by a powerful heuristics, which allows the $p$ most informative pairs of genes to be found with a complexity $O(pN)$. The search proceeds as follows: in the beginning, the mutual information $I(G_i, C)$ between each single gene $G_i$ and the class $C$ is computed, and the most informative gene $G_{i^*}$ is selected. Then, the mutual information $I(G_{i^*}G_j, C)$, between the class and each pair of genes which include the gene $G_{i^*}$, is computed. The gene pair $(G_{i^*}G_{j^*})$ that maximizes $I(G_{i^*}G_j, C)$ is selected, and the genes $G_{i^*}$ and $G_{j^*}$ are removed from the list of genes to be analyzed. This procedure is iterated $p$ times to obtain $p$ pairs of genes. The deletion of the selected genes from the list of the available ones is motivated by the goal of eliminating the redundant pairs mentioned earlier.

### 4.2 Feature construction from gene pairs

For each of the $p$ informative gene pairs $(G_iG_j)$, a new feature $A_{i,j}$ is constructed. The idea is the following one: the higher the difference between the densities of the classes around a point, the higher the probability that this point belongs to the higher density class. The value of the new feature is the difference between the local densities of the classes.

More precisely, let $E = \{e_1, \ldots, e_M\}$ be a set of $M$ instances, each belonging to one of the two classes $\{C_a, C_b\}$. From an informative pair of genes $(G_iG_j)$ we construct a new feature $A_{i,j}$ as follows: the instances in $E$ are projected onto the two-dimensional space defined by the expressions of $G_1$ and $G_2$. We have to define the value $A_{i,j}(x)$ of the feature $A_{i,j}$ for every

point $x$ of this space. The probabilities $p_a(x)$ and $p_b(x)$ at point $x$ belongs either to class $C_a$ or to class $C_b$, respectively, can be approximated by using the $k$-nearest neighbors of $x : p_a(x) \sim n_a(x)/k$ and $p_b(x) \sim 1 - p_a(x)$, where $n_a(x)$ is the number of points belonging to class $C_a$ among the $k$-nearest neighbors of $x$. The value of the new feature $A_{i,j}$ at the point $x$ is the difference between $p_a(x)$ and $p_b(x)$:

$$A(x) = p_a(x) - p_b(x) \sim \frac{n_a(x)}{k} - (1 - \frac{n_a(x)}{k}) = -1 + 2\frac{n_a(x)}{k}$$

The values of the new feature are between $-1$ and $+1$. When an instance is close to intances belonging to class $C_a$ (respectively, $C_b$), the feature tends to $+1$ (respectively, $-1$).

It should be underlined that the definition of this new feature is the same as the *margin* defined by Shapire for the voting methods in machine learning (Schapire *et al.*, 1997). The new feature, that we have defined, represents the classification margin of the $k$ nearest neighbor classifier in the space of the gene pairs.

## 5 RESULTS AND DISCUSSION

In order to test the effectiveness of the proposed method, an experimental study was designed and set up to answer the following questions: is our selection heuristic adapted to find informative pairs of genes? What is the amount of information contained in the interaction compared to that of the individual genes? Is our feature construction method effective for synthesizing the information contained in a pair of genes? Does FeatKNN improve classification accuracy?

Six public datasets are used in these experiments, their characteristics are described in Table 1.

### 5.1 Identification of the most informative pairs of genes

Our method is based on the identification of the most informative gene pairs that is performed by SFS procedure. The choice of this SFS has been based on the assumption that the most informative pairs of genes include at least one of the most informative gene. To validate this assumption, we examined the rank of the genes forming the best pairs. We defined a ranking of the genes based on their mutual information with the class. The gene with the highest mutual information has rank 1, and the one with the lowest mutual information has the lowest rank. In the same way, we compute the mutual information between the class and all gene pairs, and we defined a ranking of the gene pairs. Notice that here we do not select the best pairs using the SFS procedure, but we compute the mutual information for all exclusive pairs of genes. A figure on the companion website shows the average rank of the two genes forming the most informative gene pairs. It shows that the first genes of the best pairs were among the top informative genes, while the second genes have a much higher rank. For example, the two genes forming the 50 best pairs have on average rank 58 and 209, respectively. The same results were observed on the other six datasets which are described in the Table 1.

The observation of the values of the mutual information of the genes and pairs of genes leads to the conclusion that all the best pairs include, on average, a highly informative gene. This observation validates our assumption for microarray data. Also, this suggests a positive answer (from an empirical point of view) to the second question regarding the ability of our heuristic to find highly informative gene pairs. It also indicates the usefulness of this heuristic choice for selecting pairs, which are formed starting from the gene with the best rank. It should be underlined that standard feature selection methods, disregarding interactions, may miss many useful pairs, as it was already said in the introduction.

### 5.2 Analysis of the most informative pairs

Our assumption is that the explicit account of the synergic interaction between genes may improve classification accuracy. To validate this assumption we computed, for each dataset, the information of all genes and all gene pairs; both genes and gene pairs have been ordered according to increasing rank. Figure 2 shows the decomposition of the information of the 100 best gene pairs of the six datasets. It can be seen that around 40% of

**Table 1.** Description of the datasets

| Dataset name | Number of Genes | Number of Samples | Class $C_a$ | Class $C_b$ |
|---|---|---|---|---|
| Leukemia | 7129 | 72 | 47 | 25 |
| Colon cancer | 2000 | 62 | 40 | 22 |
| Prostate cancer | 12 600 | 102 | 52 | 50 |
| SRBCT | 6567 | 63 | 43 | 20 |
| Lung cancer | 3588 | 43 | 22 | 21 |
| Breast cancer | 7129 | 49 | 25 | 24 |

It shows the data type, the number of genes measured and the number of samples contained in each class.

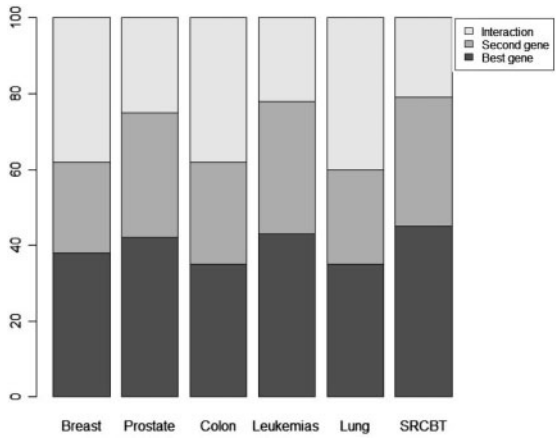


**Fig. 2.** Decomposition of the information contained in the best gene pairs. The black part shows the amount of information of the most informative genes. The dark grey part shows the amount of information of the second genes. The light grey part shows the amount of information of the interaction.

the information provided by the pairs of genes resides in the interaction of their components. For example, in the 100 most informative gene pairs of the breast cancer dataset, 38% of the information is provided by the best gene, 24% by the second gene and 38% by their interaction.

### 5.3 Information obtained by feature construction

The aim of feature construction is to synthesize the information contained in the genes and their interactions. In order to measure the effectiveness of our feature construction method, we empirically compared the information contained in the most informative gene pairs $(G_i, G_j)$ and in the associated newly constructed features $A_{i,j}$. For the $p = 100$ best gene pairs of the six datasets described above, the mutual information of the newly constructed attribute is on average in 90% of the cases higher than the information of the best gene in the pair. These results suggest that our feature construction method is effective for synthesizing the information contained in a gene pair, thus answering our third question.

### 5.4 Classification accuracy

To measure the impact of our dimension reduction method on the classification, we examined classification accuracy on the six datasets. The cross-validation estimator is the most commonly used error estimation method in microarray-based classification. We have used the 10-times 10-fold cross-validation procedure in our experiment to measure the generalization error. However, Braga-Neto and Dougherty (2004) have shown that this estimator is not the most appropriate one for a small instance sample like the ones available in microarrays analysis. Cross-validation has a high variance and therefore bootstrap estimators are preferred, in particular the 0.632 estimator (Efron, 1983). The 0.632 bootstrap estimator is a weighted sum of the empirical error and the out-of-bag bootstrap error. We have also used the 0.632 bootstrap estimator in our experiment to complete the results obtained by cross-validation. Total 100 bootstrap iterations were performed. It should be noted that for the evaluation procedure (both cross-validation and bootstrap) the test samples were not used in the dimension reduction and classifier design. Thus, we avoided the problem of selection bias pointed out by Ambroise and McLachlan (2002) and Reunanen (2003). It should be underlined that the number of features used in classifier design is a meta-parameter, whose value is chosen by an internal cross-validation procedure. A figure representing the study design can be found on the companion website (Fig. 1).

In our experiments, each dimension reduction methods is associated to three different classification algorithms: the support vector machines (SVM), $k$-nearest neighbors (KNN) and diagonal linear discriminant (DLD). We have chosen these algorithms because they are among the most efficient for microarray data classification. Dudoit *et al*. (2002) have pointed out the excellent results of the simplest methods like KNN and DLD. Furey *et al*. (2000) and Lee *et al*. (2005) have published a comparative study of the classification methods for microarray data, and they concluded that the SVMs are the best model. All of the experiments have been performed with the statistical environment R, the numbers of neighbors $k$ for KNN was 3 and the SVM has been implemented using the package 'e1071' with a radial kernel.

We have compared FeatKNN to other methods that are widely used in the literature and reach good performances. These dimension reduction methods are the following:

- *All genes*: all the genes are used.
- *Single MI*: the genes with the highest mutual information with respect to the class are selected. This method is commonly used in the literature (Ben-Dor *et al*., 2000; Wang *et al*., 2005).
- *Pair MI*: the gene pairs with the highest mutual information with the class are selected (i.e. FeatKNN without the feature construction step). This method is tested to show the importance of feature construction.
- *BO*: the gene pair-based method developed by Bo and Jonassen (2002).
- *Geman*: the gene pair selection method used by Geman *et al*. (2004) in their TSP classifier.
- *PLS*: new features are constructed as the components which maximize the covariance between the class and the variables (Dai *et al*., 2006).

In this article, we focus on the results obtained by 0.632 bootstrap; the results by cross-validation can be found on the companion website, and leads to the same conclusions. Table 2 reports the classification error rates for different reduction methods. We used a paired Wilcoxon test to compare the results. The detail of the $P$-values of significance can be found in the companion website.

It is not surprising to see that dimension reduction methods improve classification performance considerably. In all cases the performances reported in the column 'All genes', are worse than the others. The columns 'Single MI' and 'Pair MI' represent the results obtained when we select the genes (respectively, gene pairs) having the highest mutual information with respect to the class. We see that the methods using single genes (column 'Single MI') and pairs of genes (column 'Pair MI') obtain similar results. We have shown that the gene pairs were more informative than single genes. This may suggest that the information contained in the interaction between the genes composing the pairs is not well exploited by the classification algorithms, and therefore much of the information computed during the pair-selection phase is lost. This phenomenon is avoided in FeatKNN, thanks to the feature construction step. The new features constructed by FeatKNN synthesize the information contained in the genes and their interactions, which explain the better results. FeatKNN outperforms Geman's method in all datasets with the three classifiers. Bo's method is competitive, especially with the DLD classifier: 9 times out of 18 Bo's results are as good as FeatKNN's ones, and it outperforms FeatKNN on the SRBCT dataset with a DLD classifier. FeatKNN is statistically significantly (95% level) better than all other methods but PLS. PLS results are almost as good as FeatKNN. It is the only method that is not significantly worse than FeatkNN.

**Table 2.** Classification results on six public datasets

| Classifier | Data | All gene | Single MI | Pair MI | FeatKNN | Bo | Geman | PLS |
|---|---|---|---|---|---|---|---|---|
| SVM | Leukemia | 12.3 ± 1.1 | 4.3 ± 1 | 4.8 ± 1.2 | 2.8 ± 1.0 | 3.9 ± 0.9 | 6.1 ± 1.1 | **2.4** ± 1.3 |
| | Colon cancer | 17.5 ± 1.1 | 12.5 ± 1.3 | 11.8 ± 1.0 | **10.7** ± 1.1 | 13.9 ± 1.3 | 14.6 ± 1.0 | 11.1 ± 1 |
| | Prostate cancer | 9.5 ± 1.1 | 6.1 ± 1.0 | 6 ± 1.0 | 6.0 ± 0.8 | **5.6** ± 0.9 | 6 ± 0.7 | 6.2 ± 0.6 |
| | SRBCT | 7.6 ± 0.5 | 2.1 ± 0.4 | 3.8 ± 0.3 | **0.2** ± 0.2 | **0.2** ± 0.2 | 0.7 ± 0.3 | 1.7 ± 0.5 |
| | Lung cancer | 24.9 ± 1.1 | 21.7 ± 0.8 | 21 ± 0.9 | **19.5** ± 1.0 | 21.5 ± 1.1 | 21 ± 1.0 | 20.7 ± 1.3 |
| | Breast cancer | 14.6 ± 0.8 | 11.4 ± 1.1 | 11.2 ± 0.9 | **8.7** ± 1.0 | 11.4 ± 1.1 | 11.2 ± 1.0 | 9.7 ± 1 |
| KNN | Leukemia | 8.4 ± 1.1 | 6.1 ± 0.9 | 6.2 ± 1.0 | 5.0 ± 1.2 | **4.6** ± 1.2 | 6.3 ± 1.1 | 5.4 ± 0.9 |
| | colon cancer | 20.0 ± 1.2 | 14.9 ± 1.2 | 14.4 ± 1.0 | 12.8 ± 0.9 | 15.9 ± 1.0 | 16 ± 1.1 | **12.4** ± 1.1 |
| | Prostate cancer | 20.2 ± 1.0 | 17.8 ± 0.8 | 18 ± 0.9 | **8.1** ± 0.7 | 8.7 ± 0.8 | 9.8 ± 1.0 | 8.5 ± 0.9 |
| | SRBCT | 11.6 ± 0.7 | 1.1 ± 0.2 | 1 ± 0.5 | **0.1** ± 0.1 | **0.1** ± 0.1 | **0.1** ± 0.1 | 1.3 ± 0.4 |
| | Lung cancer | 35 ± 0.7 | 29.1 ± 1.0 | 28.2 ± 0.9 | **20.8** ± 1.0 | 23.7 ± 1.3 | 24.2 ± 1.0 | 21.7 ± 1.4 |
| | Breast cancer | 20.8 ± 0.7 | 14.3 ± 0.9 | 13.5 ± 1.0 | 9.0 ± 1.0 | 12 ± 1.1 | 13.1 ± 1.0 | **8.4** ± 1 |
| DLD | Leukemia | 11.5 ± 1.2 | 4.8 ± 1.2 | 4.8 ± 1.0 | 3.8 ± 1.1 | 4.1 ± 1.0 | 5.0 ± 1.1 | **2.7** ± 0.8 |
| | colon cancer | 19.5 ± 1.4 | 15.7 ± 1.2 | 15.4 ± 1.0 | **12.5** ± 1.0 | 14.4 ± 1.1 | 15 ± 1.3 | 12.9 ± 1.1 |
| | Prostate cancer | 37.5 ± 1.0 | 10.5 ± 0.8 | 10.1 ± 0.9 | 7.6 ± 0.9 | **7.3** ± 1.0 | 8 ± 0.7 | **7.3** ± 1 |
| | SRBCT | 5.4 ± 0.7 | 0.8 ± 0.2 | 0.5 ± 0.2 | 0.7 ± 0.2 | 0.2 ± 0.1 | **0.1** ± 0.1 | 2.5 ± 0.6 |
| | Lung cancer | 25.4 ± 1.2 | 21.6 ± 0.8 | 22.1 ± 0.9 | 20.6 ± 1.0 | 20.3 ± 0.8 | **20.2** ± 1.0 | **20.2** ± 1.1 |
| | Breast cancer | 14.9 ± 0.6 | 10.7 ± 1.0 | 10.9 ± 1.0 | **9.1** ± 1.1 | 9.3 ± 0.9 | 10.1 ± 0.9 | 9.6 ± 1 |

All errors are estimated using the 0.632 bootstrap estimators.
Boldfaced values highlight the best results.

## 5.5 A biological interpretation of the concept of synergic transcript pairs

The exploration of the biological significance that may be enfolded in the concept of synergic transcript pairs had to consider two distinct aspects, one more particular related to the analyzed clinical situations, and another more general regarding the type of biological interactions that may explain the synergic behavior exhibited by the genes belonging to the same pair. To answer these questions, we started by separating the two components of gene pairs and then we carried out a discriminative functional profiling of the two resulting lists of genes (i.e. a *first pairs component* list and a *second pairs component* list). Based on the functional assignments provided by the Gene Ontology (GO) consortium (http://www.geneontology.org), and by the NCBI genomic repository (http://www.ncbi.nlm.nih.gov), an automated annotation procedure combined with a gene set enrichment analysis allowed to identify biological themes significantly overrepresented in each of the two lists of genes.

Figure 3 shows overrepresented biological themes characterizing each of the two components of the first 100 most informative pairs extracted from the colon cancer dataset, which resulted from microarray experiments performed to compare expression profiles of tumor and normal colon tissues. The functional profiles depicted in Figure 3 seem to indicate highly distinct biological assignments for the two components of gene pairs. Thus, while the first pair component seems to be related mostly to intracellular processes located either in the nucleus (i.e. *nucleus*, *nuclear part*) or in the cytoplasm (i.e. *intracellular part*, *intracellular non-membrane-bound organelle*, *cytoskeletal part*), the second pair component appears to be involved exclusively in cell membrane-related processes (i.e. *plasma membrane*, *membrane part*, *intracellular membrane-bound organelle*). Moreover, these findings seem to be well supported by the molecular functions assigned to the

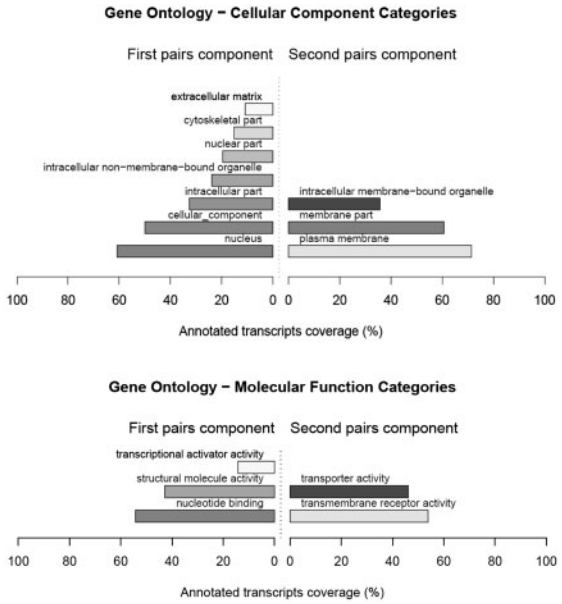


**Fig. 3.** Overrepresented biological themes discriminating the two components of the first 100 most informative synergic transcript pairs extracted from the colon cancer dataset (see text for details).

translation products of these genes, which were found to be related to *nucleotide binding*, *structural molecule activity* and *transcriptional activator activity* for the first pairs component, and to *transmembrane receptor activity* and *transporter activity* for the second pairs component.

Considering the experimental framework from which this dataset resulted, these functional profiles seem to indicate that the biological themes that best distinguish tumoral from normal colon cells concern essentially the *nuclear transcriptional control* of a large panel of intracellular processes (i.e. cell differentiation, proliferation, metabolism, apoptosis, extracellular matrix production, etc.) on one side, and the *cell communication and extracellular signaling* modulated by cell membrane structures (i.e. involved in processes as focal adhesions, cell attachment and migration, growth factor receptor expression and signaling, etc.) on the other.

These findings are in total agreement with the most up-to-date understanding of the tumoral biology. Indeed, it is well acknowledged that tumor cell survival is dictated by both internal properties of the cell, such as status of components of the apoptotic machinery, and its extracellular environment, such as extracellular matrix and growth factor receptor expression and signaling (Dennis and Kastan, 1998) that modulate apoptosis regulation. Apoptotic anomalies represent a major distinction between tumor and normal cells, as they allow tumor cells to avoid programmed cellular death, and confer them the capacity to proliferate indefinitely. Some of these anomalies, which involve the nuclear compartment, may result in an inactivation of key tumor suppressor genes (TSGs), acknowledged as being central to the development of all forms of human cancer (Rhee *et al.*, 2002). This inactivation is induced by a combination between epigenetic silencing and promoter hypermethylation of various TSGs. On the other side, the importance of extracellular survival signals as key regulators of apoptosis is now being recognized by the ability of growth factors (GFs), GF receptors (GFRs) and GFR signaling to promote cellular survival. Indeed, recent evidence suggests that a number of abnormal cell membrane constituents may disrupt apoptosis regulation in tumor cells by pathologically amplifying the anti-apoptotic effect of normal extracellular survival signals (Leask and Abraham, 2006). All these evidences suggest that the concept of synergic gene pairs not only uncovered a noteworthy functional behavior singularizing tumor cells, but also captured the synergic aspect of the interaction between underlying biological mechanisms.

## 6 CONCLUSION

In this article, we have presented a dimension reduction procedure for microarray data oriented towards improving classification performance. This method is based on the idea that the information provided by the interaction between genes cannot be ignored in the feature selection phase. We have proposed a decomposition of the information contained in the gene pairs. Although it is natural to quantify information from genes and interactions from the computation of mutual information, this simple reduction does not necessarily improve performance. Therefore, we have developed a feature construction method that forces learning algorithms to take into account pairs with a high level of information. The usefulness of this approach was experimentally assessed on six datasets and yielded a significant improvement in performance. Moreover, a synthetic assessment of the biological significance of the concept of synergic gene pairs suggested its ability to uncover relevant mechanisms underlying interactions among various cellular processes.

*Conflict of Interest*: none declared.

## REFERENCES


Ambroise,C. and McLachlan,G. (2002) Selection bias in gene extraction on the basis of microarray gene expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.

Ben-Dor,A. *et al.* (2000) Scoring genes for relevance. Technical report AGL-2000-13, Agilent Technologies. Institute of Computer Science, Hebrew University, Jerusalem.

Bo,T. and Jonassen,I. (2002) New feature subset selection procedures for classification of expression profiles. *Genome Biology*, **3**, research0017.1–research0017.11.

Braga-Neto,U. and Dougherty,E. (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, **20**, 374–380.

Butte,A.J. and Kohane,I.S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, 418–429. Using Smart Source Parsing.

Dai,J. *et al.* (2006) Dimension reduction for classification with gene expression microarray data. *Stat. Appl. Genet. Mol. Biol.*, **5**, article 6.

Dennis,P.A. and Kastan,M.B. (1998) Cellular survival pathways and resistance to cancer therapy. *Drug Resist. Updat.*, **1**, 301–309.

Ding,C. and Peng,H. (2003) In *Proceedings of the IEEE Computer Society Conference on Bioinformatics*. Stanford, CA, USA, pp. 523–529.

Dudoit,S. *et al.* (2002) Comparison of discrimination methods for classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Efron,B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc.*, **78**, 316–331.

Furey,T. *et al.* (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Geman,D. *et al.* (2004) Classifying gene expression profiles from pairwise mRNA comparisons. *Stat. Appl. Genet. Mol. Biol.*, **3**, article 19.

Hanczar,B. *et al.* (2003) Improving classification of microarray data using prototype-based feature selection. *SIGKDD Explor.*, **5**, 23–30.

Jakulin,A. and Bratko,I. (2003) Analyzing attribute dependencies. In *Proceedings A of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pp. 229–240.

Leask,A. and Abraham,D.J. (2006) All in the CCN family: essential matricellular signaling modulators emerge from the bunker. *J. Cell. Sci*, **119**, 4803–4810.

Lee,J. *et al.* (2005) An extensive comparison of recent classification tools applied to microarray data. *Comput. Stat. Data Analy.*, **48**, 869–885.

Matsuda,H. (2000) Physical nature of higher-order mutual information: intrinsic correlations and frustration. *Phys. Rev. E*, **62**, 3096–3102.

Rapaport,F. *et al.* (2007) Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**.

Reunanen,J. (2003) Overfitting in making comparisons between variable selection methods. *J. Mach. Learn. Res.*, **3**, 1371–1382.

Rhee,I. *et al.* (2002) DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature*, **416**, 552–556.

Schapire,R.E. *et al.* (1997) Boosting the margin: a new explanation for the effectiveness of voting methods. In *Proceedings 14th International Conference on Machine Learning*. Morgan Kaufmann, Nashville, TN, USA, pp. 322–330.

Shannon,E. (1948) A mathematical theory of communication. *Bell Sys. Tech. J.*, **27**, 623–656.

Steuer,R. *et al.* (2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, **18**, 231–240.

Wang,Y. *et al.* (2005) Gene selection from microarray data for cancer classification – a machine learning approach. *Comput. Biol. Chem.*, **29**, 37–46.