Comparing and combining feature estimation methods for the analysis of microarray data

Antoine Cornuéjols[1,3], Christine Froidevaux[2] and Jérémie Mary[1]

[1] Équipe TAO
[2] Équipe Bioinformatique
Laboratoire de Recherche en Informatique, CNRS UMR 8623
Bâtiment 490, Université Paris-Sud, 91405 - Orsay Cedex
[3] Institut d'Informatique d'Entreprise,
18, allée Jean Rostand, 91025 - Evry Cedex
antoine@lri.fr et http://www.lri.fr/∼antoine

Abstract: *Microarray data are expected to provide important clues about the role of the genome in the cell organization and behavior. However, the parameters of interest are difficult to reliably estimate with only a small number of array samples and poor sample distributions of gene expression levels. A number of statistical and hypothesis testing methods have been adapted in recent years and brought to bear on this problem, but the determination of which and how many genes are involved in a particular biological process remains a stumbling block.*
In this paper, we propose a kind of meta-approach that makes use of several methods for gene selection. We show how to evaluate their independence and how to combine their results in order both to determine the most likely number of relevant genes and to select them.
We illustrate this approach on a microarray data set devoted to the biological detection of low radiation doses, and show how it greatly improves on previously reported results.

Keywords: Microarray data analysis, Feature selection, combination of methods.

1 Introduction

One essential issue commonly encountered in the analysis of microarray data is to decide which and how many genes should be selected for further study because they are likely to be involved in the tested biological phenomenon. In this setting, each reported gene expression level can be seen as a feature describing an experiment. By measuring expression levels associated with two kinds of tissue (e.g. tumor or non-tumor) or two kinds of condition (e.g. irradiated or not irradiated), one obtains labeled data sets that can be used to build diagnostic classifiers, or more generally to help understand the underlying genetic processes at play. Unfortunately, the number of replicates in these experiments is usually severely limited, in the order of a few tens as compared to several thousands of genes. In this state of affairs, it is illusory to directly use automatic classification systems to identify the relevant genes. Too many spurious regularities may put forward features that look like perfect predictors of the class on the data set but are really uncorrelated with it. This is why one relies instead on feature selection methods in order to detect the likely informative genes.

Numerous methods have been proposed in recent years for gene selection (see [3,6,9,11,14,16,17]). Most of them assume that each gene expression level is in some way directly correlated to the class, and that the expression level obeys simple statistical models (e.g. normal distribution). Equipped with these simplifying

assumptions, these methods are used to evaluate the relevance of each gene. In addition, one often relies on hypothesis testing methods to set a threshold that separates the good candidate genes from the other ones.

However, the poor quality of the data together with their scarcity, render the estimation of the diverse parameters of the model quite unreliable. There is therefore a recurrent concern about the minimum sample size of the data set that would allow one to have confidence in the results (see [5,10]) and the question about which genes and how many should be selected is still a daily burning issue in biology labs. Additionally, one is confronted with the choice of a feature selection method among the many that exist. Why one should prefer one method over another is often a difficult question, especially when the data do not completely satisfy the requirements of orthodox statistics.

While nothing can replace the knowledge of the true nature of the data and, therefore, of the matching best feature selection method, in the absence of such information, an alternative might be to combine the use of several methods in the hope of benefiting from the qualities of each one. This is, however, more easily said than done because, in contrast to supervised learning, one cannot evaluate directly the worth of a feature selection method from the training set. In this paper, we show first how it is possible to estimate the information recovered by a feature selection method, second, how to measure the correlation between two feature selection methods, and third, how to combine the use of two methods to obtain more precise information about the number of relevant genes and about their identity. Notice that the approach demonstrated here for pairs of methods can be generalized to more than two methods.

Section 2 of the paper provides an overview of feature selection methods and presents in more details three among them: SAM [16], ANOVA and RELIEF [7,15]. Section 3 shows how the determination of the relevant genes and of their number is usually done. This is illustrated on a microarrray data set pertaining to the detection of biological effects of low doses of radiation [12]. Next, the section 4 starts by a discussion about the notion of correlation of different selection techniques and how it can be measured. A method for combining two methods using a maximum likelihood approach is then presented in section 5, together with its application to the data set introduced in section 3. Finally, section 6 sums up the main lines of the approach and describes vista offered by this research.

2 Approaches and methods for feature selection

Feature selection methods aim at identifying features that are useful for classification purposes. Each pattern is described by a set of d features (e.g. genes) and belongs to a class (generally, in bio-informatics, one of two conditions, e.g. tumor or non-tumor). The training set provides examples of patterns together with their, supposedly, true class. The problem is to identify the features that are the more informative with respect to the classification of known, and, more importantly, yet to be observed, patterns. In addition, one can be interested in a minimal set of features that allow the prediction of the class, or in discovering all the features involved, even if they are redundant. The latter case is more representative of the concerns in microarray data analysis.

It is important to notice that features can be informative about the class independently of each other (called linear correlation), or in combination (higher order correlation). Evidently, higher order correlations are more difficult to discover than linear ones, and usually require more data. Feature selection methods in bio-informatics are thus generally targeted at linear correlations between the attributes and the class.

Three broad classes of approaches exist for feature selection: the *embedded*, *wrapper* and *filter* methods [2,4,8]. The first one consists in directly using the learning system on the training data in the hope that the features useful for classification will naturally be selected or highlighted in some way by the system. For instance, a decision tree classifier automatically provides these features in the decision nodes of the induced tree. Unfortunately, such an approach is bound to report false discoveries in the case of very few data points

(also called patterns or samples) compared to the number of features. The so-called *embedded methods* are therefore not feasible in microarray data analysis.

The *wrapper* methods assess subsets of variables according to their usefulness to a given predictor. Given a classifier (e.g. a neural network) and a set of features $\mathcal{F}$, a wrapper method searches the space of subsets of $\mathcal{F}$, using cross-validation to compare the performance of the trained classifier on each subset. Intuitively, wrapper methods have the advantage to select feature subsets that are well-tailored to maximize the final classifier system performance, which is the overall goal. However, while the over-optimistic character of this approach has recently been put into question [1,17], its main disadvantage is the huge computation time required to effectively explore the set of subsets of $\mathcal{F}$.

For this reason, one usually resorts to the *filter* methods. These are often considered as a preprocessing step, independent of the choice of the predictor. Usually, they consider each feature independently and evaluate their correlation with the class of the sample, or, in other words, their ability to discriminate between the classes. Under certain independence and orthogonality assumptions, the variables thus detected as informative may be optimal with respect to a given predictor. But, feature ranking is not necessarily used to build predictors. In microarray data analysis, it is commonly used as a way to evaluate the degree to which each gene is involved in the biological process under study. One significant advantage of this approach is its computational efficiency since it only requires the calculation of d scores, and then a sorting operation.

In the rest of this paper, we focus on filter methods as methods for ranking genes.

2.1 Filter methods

We note l the number of patterns $\{\mathbf{x}_k, y_k\}$ defined on a space of d dimensions or features (e.g. the genes) $x_{k,i}$ ($i = 1, ..., d$) and its class y_k. Filter methods use an evaluation function, or scoring function, $S(i)$ computed from the values $x_{k,i}$ and $y_k, k = 1, ..., l$ to rank the features.

In the context of microarray data analysis, two families of methods in particular have been used: methods based on statistical hypothesis testing and methods based on class separability measures.

The *statistical hypothesis testing* framework examines each feature separately and investigate whether the value it take for the different classes differ significantly. This is expressed as deciding between two options:

H_1 : The values of the feature differ significantly depending on the class
H_0 : The values of the feature do not differ significantly

H_0 is known as the *null hypothesis* and H_1 as the alternative hypothesis. In order to make a decision, one has to make statistical assumptions about the process generating the data. The most common assumption associated with the null hypothesis is that the data distribution follows a normal law defined by a mean μ and a standard deviation σ. The test can then be rephrased as deciding between the hypotheses:

$H_1 : E[x_i] \neq \mu(i)$
$H_0 : E[x_i] = \mu(i)$

To this end, the following test statistics is defined: $q(i) = \frac{\bar{x}(i) - \mu(i)}{\sigma(i)/\sqrt{l}}$ where $\bar{x}(i)$ is the observed mean of the $x_{k,i}$ ($k = 1, ..., l$) values for feature i, and $\mu(i)$ is its supposed mean under hypothesis H_0.

When the standard deviation $\sigma(i)$ is supposed to be known, $q(i)$ approximately follows a $\mathcal{N}(0, 1)$ probability distribution, otherwise, if the standard deviation is unknown, $q(i)$ is defined in terms of an estimate $\hat{\sigma}(i)$, and

it follows a t-distribution with $l-1$ degrees of freedom[4]. In both cases, using tabulated confidence intervals, it is easy to decide between H_0 ($q(i)$ falls in the acceptance interval) and H_1 (it does not).

More generally, one can use the value of $q(i)$ to evaluate the relevance of a feature i. The more $q(i)$ is remote from 0, the more the feature is unlikely to be ruled by the null hypothesis, and, therefore, the more it is likely to be relevant.

A *second scoring technique* involves a measure of the class separability given each feature. Again, there are methods relying on statistical assumption about the data distribution. This is the case, for instance, of divergence measures like the Kullback-Leibler distance. The non-parametric methods, on the other hand, do not use statistical assumptions. The RELIEF feature selection algorithm is an example of these methods.

In the following, we describe three methods commonly used in microarray data analysis: SAM [16] and ANOVA which are parametric, and the *Bio*RELIEF algorithm which is non-parametric.

2.2 ANOVA

Analysis of variance (ANOVA) can be used to evaluate the correlation of each feature to the class. Its principle relies on a comparison of the variances of the values of each feature when the class of the data points is taken into account, and when it is not. If these variances are significantly different, this indicates that the feature is informative about the class. ANOVA is a parametric procedure that relies on the assumption that the feature values are normally distributed. The F-test is used in evaluating the relevance of each attribute separately.

2.3 SAM

Significance Analysis of Microarrays (SAM) has been introduced by [16] as a statistical method adapted specifically for microarrays. Like ANOVA, it belongs to the family of statistical parametric tests and it relies on the t-distribution and the t-test mentioned earlier in section 2. The score function is defined as:

$$S(i) = \frac{\bar{x}_{C_1}(i) - \bar{x}_{C_2}(i)}{\sigma(i) + s_0}$$

where i stands for a given gene, C_1, and C_2 for the two classes, $\bar{x}_{C_1}(i)$ and $\bar{x}_{C_1}(i)$ for the average levels of expression for gene i in class C_1 and C_2, respectively. $\sigma(i)$ is the standard deviation of the expression measurements on the data. s_0 is a normalization factor that tends to penalize genes which have a high $S(i)$ thanks to a low variance. It is also used as an implicit threshold for the false decovery rate (FDR).

2.4 *Bio*-RELIEF

*Bio*RELIEF, developed by the authors for bio-informatics purposes[5], is a variant of the RELIEF system [7,15]. It is a feature estimation method that evaluates the features according to their apparent correlation with the class to be predicted. The score of a feature is a function of the variation of its value within each class compared to the variation between classes. However, it does not rely on statistical assumptions about the data distribution. In addition, while the score is evaluated for each feature independently, the score function uses the distance

[4] With a slight modification when the available number of samples in each class is not the same. One must then make a test involving the F-distribution.

[5] Available at: `http://www.lri.fr/ chris/bioinfo/BioRelief`

between patterns in the whole feature space, and that tends to highlight the correlated features. *Bio*RELIEF is thus well-fitted to microarrray data analysis, whereas, it is less suitable for the discovery of non redundant sets of features. More details about the *Bio*RELIEF algorithm can be found in [11].

2.5 Three questions in genes selection

The study of microarray data, characterized by very few replicates compared to the number of genes and a low signal to noise ratio, induces three challenges. First, is there any useful information in the data? Second, given a feature estimation method, how to determine a threshold for deciding which genes are likely to be relevant and deserve further examination. This decision is often rather arbitrary or is based on informed guess that comes from information that is foreign to data analysis standards. Unfortunately, the selected genes generally incorporate false positives. The third question, then, relates to the determination of the most likely ratio of true positives to the selected genes.

We illustrate these three questions on a microarray data analysis task related to the biological detection of low radiation. We show how these questions were answered in a previous study. Sections 4 and 5 will then presents a new technique to solve them.

3 A case study: biological detection of low radiation

In a work reported in [11,12], the microarray technology was used to measure the effects of low doses of radiation. To analyze these effects, the expression level of most of the yeast genes (*S. cerevisia*) in cells grown with (I) and without (NI) low doses of irradiation was monitored after an exposition of 20 hours. The relative expression level of each gene was estimated with glass slide microarrays spotted with 6135 denatured DNA sequences corresponding to all of the open reading frames (ORFs) of *S. cerevisiae*. These measured intensities were then normalized to suppress the many experimental biases. The training data thus obtained consisted of 12 non treated cultures (class NI) and 6 treated ones (class I). Two feature selection methods were used to study the microarray data: ANOVA and *Bio*RELIEF.

The first step was to assess the reality of an effect of low radiation doses on the genomes of the cultures. This was done by comparing the scores obtained for each gene on the training data with the scores that would be obtained if there was no effect (null hypothesis). This null hypothesis was implemented by computing the mean score of each gene (and its variance) when the class of the cultures were randomly permuted (2000 permutations of the 12 NI and the 6 I labels were used to compute these averages). Figure 1 shows the two curves thus obtained using *Bio*RELIEF, one for the null hypothesis, surrounded by the 95% confidence intervals, and the curve of the scores for the real labels. It is clear that the null hypothesis is very unlikely, and that, therefore, there is an effect of low radiation exposure on the genomes of *S. cerevisiae*.

The second question concerns the number of genes one must retain. This is generally related to a tradeoff between the number of false positives one is ready to accommodate and the number of true positives that one risks to miss. One solution is to chose *a priori* a number n of genes to retain, using an educated guess. The second approach is to chose a threshold based on confidence intervals. For instance one can choose to select all genes that have less than a 5% probability to be ruled by the null hypothesis. A third solution consists in controlling in some way the tradeoff.

In this manner, using the curve on figure 1, on can choose several thresholds yielding different ratio of true positives to false positives in the selected pool of genes. For instance, if one selects the score 0.5 as a threshold, it is observed that 35 genes have a higher score with the true labeled data, whereas none on average reach that

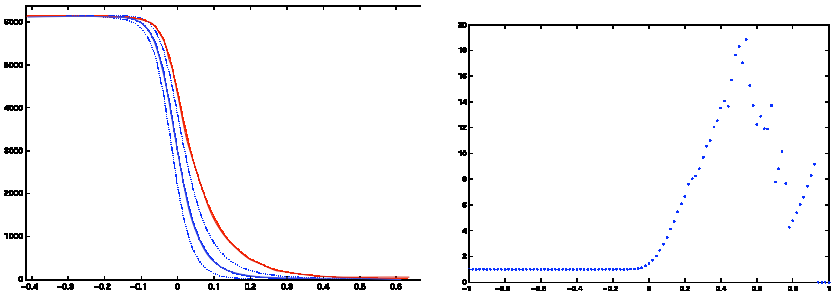

Figure 1. Left: these curves show the number of genes (on the y-axis) that have a score above the value indicated on the x-axis (scores computed with *Bio*RELIEF). The higher the score, the less genes have a score greater than or equal to it. The null hypothesis curve is significantly lower than the curve obtained for the true labeled data. **Right**: Curve of the ratio of the scores with the true labeled data and the mean score under the null hypothesis.

score under the null hypothesis, that is on the average of random permutations of the labels. The 35 genes thus selected are therefore likely to be true positives. But there exists another interesting quantity that may help set the tradeoff.

Figure 1 on the right shows the ratio between the score values obtained for the true labeled data, and the values obtained under the null hypothesis. In a way, this is akin to a signal to noise ratio. This curve presents a sharp increase for score values starting around 0.1. For a threshold of 0.3, the ratio is approximately equals to 9, and the number of selected genes is 171. This means that one can expect approximately $171/9$, or 19, selected genes that are false positives out of the 171 selected ones.

One problem with these approaches is that they try to optimize, under some preference criterion, the tradeoff between the number of false positives and the number of true positives, but they do not give the likely total number of true positives, a quantity of central interest for the biologists. We next describe a way to compute that number.

4 Measuring the correlation of feature selection methods

Confronted with uncertain results from one method, it is tempting to try to check them with results from another method. In the case of the study on low doses of radiation, this was done by comparing the outputs of the SAM, ANOVA and *Bio*RELIEF methods. The idea was to take the top-ranked n (top_n) genes from each method and to measure the size of their intersection.

The intersection of the top_{500} from SAM and ANOVA is equal to 409, and much higher that the intersection obtained with ANOVA and *Bio*RELIEF, which is equal to 281. Which conclusions should be drawn from these figures? Do they mean that, since SAM and ANOVA do seem to agree better than ANOVA and *Bio*RELIEF, they should be trusted more? How to interpret these intersection sizes?

The intersection can result from three causes:

– *Randomness*. Any two subsets randomly drawn from a given finite set of elements may have a non empty intersection. In fact, the distribution probability over the size of this intersection can be computed from the

hypergeometric law:

$$H(d,n,k) \;=\; \frac{\binom{n}{k}\cdot\binom{d-n}{n-k}}{\binom{d}{n}} \tag{1}$$

where d is the total number of elements (genes), n the number of elements in each subset (the n genes top-ranked by each method), and k the size of the observed intersection (the number of genes found in both top$_n$).
For instance, in the case of the intersection of two subsets of 500 elements randomly drawn from a set of 6135 elements, the most likely intersection size is 40 ($H(6135,500,40) = 0.069$), while the probability for the case $k \geq 281$ is less than 5.16×10^{-199}. In other words, the intersection size of 281 observed for ANOVA and *Bio*RELIEF is extremely unlikely to have happen by chance alone.

– *A priori correlation* between the methods. If someone used two times the very same feature selection method, he/she should not be surprised to get an intersection size of n for two top$_n$ rankings. This could be entirely explained by the fact that, no matter the data, the "two" methods will always completely agree on their rankings. There is certainly a full spectrum of *a priori* correlation between methods, and part of the obtained intersection size must be attributable to this.
– *The information (regularities) in the data* that both methods are able to extract and agree on, beside their a priori alignment, is the quantity of interest. This is that part of the intersection, these informative genes, that we would like to identify.

As was mentioned, it is easy to compute the intersection size that can be expected from randomness alone. The application to the low radiation microarray data was reported in [11,12]. We now show how to evaluate the *a priori* correlation between two feature selection methods.

One way to measure the correlation of two ranking methods M_1 and M_2 is to compute the expected size of the intersection of their top$_n$ if the data were random.

$$\langle M1, M_2\rangle \;=\; \mathsf{E}_{\mathcal{D}}(|\mathtt{top}_n(M_1) \cap \mathtt{top}_n(M_2)|) \tag{2}$$

where $|.|$ is used to denote the cardinality of a set, and top$_n(M)$ denotes the n genes top-ranked by method M. The expectation is defined over a distribution $\mathcal{D}$ of the data.

To ensure that the bias introduced by the data distribution is the same for random data sets and for the true data set, a simple solution is to take the expectation over the original data set with random permutation of the labels. Short of being able to analytically compute this distribution, it is possible to get an empirical estimate of it from the measured intersection size on a number of random data sets obtained in that way.

Table 1 reports the results obtained for various values of n for the ANOVA and the *Bio*RELIEF methods, both for random data sets (mean value and standard-deviation), and for the true low radiation microarray data set.

n	100	200	300	400	500	600	700	800	900	1000
$\mu_{\mathcal{H}_0}$	21.2	54.2	93.2	135.4	180.3	226.9	276.3	326.2	378.9	432.5
$\sigma_{\mathcal{H}_0}$	8.0	16.9	24.5	32.3	41.8	50.3	57.7	64.1	71.3	78.0
k	**37**	**93**	**149**	**210**	**281**	**339**	**406**	**470**	**535**	**605**

Table 1. Intersection of two top$_n$ from ANOVA and *Bio*RELIEF for various values of n, under the null hypothesis ($\mu_{\mathcal{H}_0}$ and $\sigma_{\mathcal{H}_0}$) and observed for the true data set (k).

In figure 2, the curves corresponding to the observed intersection, the a priori correlation and the random intersection are shown. The x-axis stands for the size n of the top$_n$, while the y-axis stands for the ratio of the intersection sizes to n (e.g. for $n = 500$, the observed intersection size is equal to 281, or $.562 \times 500$, hence the value .562).

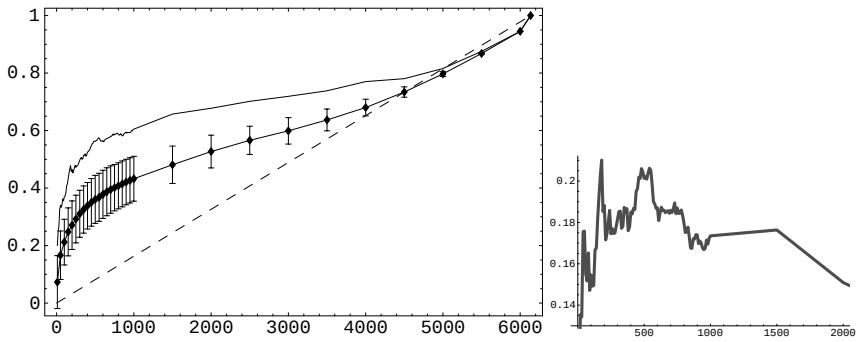

Figure 2. The x-axis stands for the number n of top-ranked features (e.g. genes) considered. The y-axis stands for the ratio of the intersection size to n. **Left**: (Top curve) the intersection size for the true data. (Center curve): the mean intersection size due to the a priori correlation between ANOVA and *Bio*RELIEF (with some standard deviation bars). (Lower curve): the intersection size explainable by randomness alone. **Right**: Curve of the relative difference, with respect to n, of the observed intersection size k and the intersection size $\sigma_{\mathcal{H}_0}$ due to a priori correlation between ANOVA and *Bio*RELIEF. The curve focuses on the beginning of the curve, for $n < 2000$, since it is the more interesting part.

Equally instructive is the curve of figure 2 (right) showing the difference between the observed intersection size k for the true data and the expected intersection size $\mu_{\mathcal{H}_0}$.

Two conclusions are warranted from these figures. First, there is indeed some specific information in the true data, since the observed intersection size, k, is much higher than the expected value $\mu_{\mathcal{H}_0}$ (in fact, k is generally more than two standard-deviation away from the expected mean value). Second, it is possible to determine the best n value, the one for which the measured intersection size on the true data exceeds the most the expected mean value. Indeed, their relative difference is maximal for $n \approx 180$ and $n \approx 540$. This suggests that it is best to consider the top$_{180}$ or the top$_{540}$ ranked genes by ANOVA on one hand and by *Bio*RELIEF on the other hand because they should contain the largest number of genes corresponding to information that is specific to the data, and not explainable by randomness or by a priori correlation between the methods.

The next section shows how one could draw more information about the biological phenomenon under study from the observed intersection size and the measured a priori correlation.

5 Combining feature selection methods

It is possible to propose a parametric generative model governing the intersection size distribution k observed on the true data.

Let us suppose that d is the total number of features or genes, p is the number of (biologically) relevant features (the ones that we wish to identify or, at least, that we would like to count), and n the number of top-

ranked features by both feature selection methods. Additionally, let us assume that both methods are equally able to draw m relevant features from the existing p[6]. Figure 3 depicts the corresponding situation.

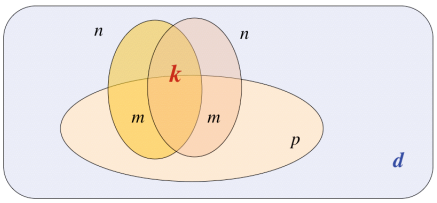

Figure 3. The sets involved in the generative model of the intersection size k.

The probability distribution of the intersection size k can then be computed from the following formula:

$$p(\cap = k|d,p,n,m,\mu_{\mathcal{H}_0}) = \frac{\binom{p}{m}\binom{d-p}{n-m}\sum_{k^+=2m-p}^{m}\binom{m}{k^+}\binom{p-m}{m-k^+}\binom{n-m}{k-k^+}\binom{d-n-(p-m)}{n-m-(k-k^+)}}{\binom{d}{n}\cdot\binom{d}{n}} \Big/ C(\mu_{\mathcal{H}_0}) \tag{3}$$

This seemingly complicated expression computes the number of ways one can get an intersection size of k given d, p, n, m and the a priori correlation size $\mu_{\mathcal{H}_0}$ divided by the total number of ways one can get two drawings of n features among d. The denominator $C(\mu_{\mathcal{H}_0})$ stands for a normalization factor associated with the a priori correlation of the methods and is computed in the same way as the numerator. k^+ stands for the part of the intersection size k that correspond to relevant features. In the ideal case, every feature in the intersection would be a relevant one.

From empirical measurements on the available data using two different feature selection methods, one can get values for d, n, k and $\mu_{\mathcal{H}_0}$. It is then possible, using a *maximum likelihood principle* to compute the most likely values for the quantities p and m. In this way, one can estimate the likely total number p of relevant features among the d features, and to estimate as well the likely number of relevant features identified by the methods among the n features they top-rank.

For instance, in the case of the low radiation doses data, the maximum likelihood principle, applied with $d = 6135, n = 500, \mu_{\mathcal{H}_0} = 181$, plugged in the formula yields $p = 420 \pm 20$ and $m = 340 \pm 20$ as the most likely numbers of total relevant genes and of the relevant genes among the top$_{500}$ ranked by both methods.

Notice however that, while these values seem reasonable, they result from the simplifying assumption that ANOVA and *Bio*RELIEF were equally good on these data, i.e. that both returned the same number of relevant genes in their n top-ranked genes. Furthermore, even if a combination of methods can provide more precise information about the data than a single one, it cannot make up for their scarcity. Consequently, there remains a rather large uncertainty about the estimated values. Nevertheless, there is now a firmer and less arbitrary basis for the determination of the number of relevant genes as well as for assessing the number of relevant genes that are recalled in the n top-ranked genes from each method.

[6] It is straightforward to generalize our discussion to two methods drawing different numbers of features n_1 and n_2 and identifying different numbers of relevant features m_1 and m_2. For lack of space, we provide the simplest formula in this paper, the one corresponding to two supposedly equally powerful methods.

6 Conclusion

The recurrent questions in microarray data analysis include (i) whether there exists a measurable effect of the experiment on the genome, (ii) the determination of the total number of genes involved if any, and (iii) their identity. The usual approach is to use one feature selection technique and to rely on risky statistical assumptions or on educated guesses to set a False Discovery Rate threshold.

Yang et al. in [18] recently presented a method that allows one to take into account two (or, potentially, more) different genes evaluation methods. For this, they propose to represent each gene as a point in space where each dimension stands for the evaluation by one method. They then compare each point (gene) with an extreme point corresponding to a (possibly virtual) gene that would have the highest evaluation by each method. This extreme point thus defines an axis in space that represents the correlation trend of the data points. From the observation of the points that markedly differ from this axis, they are thus able to lower the rank of the genes with discordant measurements. This work represents a step toward the combination and synthesis of information coming from different evaluation methods. However, it is more aimed at pointing out genes that do not exhibit the same overall correlation between measurements as the group all together, rather than aimed at extracting more information from the data by combining the different view points of ranking methods. In addition, their approach does not allow to measure the correlation between the methods in terms of the information they extract from the data.

In this paper, we proposed instead a method that takes advantage of the information provided by a combination of feature selection techniques to sharpen our estimates about the number of relevant genes and about their identity.

By focusing on the intersection of the top-ranked genes by several techniques, we have developed a new method to assess the degree of correlation between feature selection techniques. This permits to evaluate the significance of the intersection size. A high intersection size is significant only up to the point that the feature selection techniques used are reasonably a priori uncorrelated.

We have also shown how, from the observed intersection size, and the one expected from a priori correlation, it is possible to get more accurate estimates of the total number of relevant genes, and of the ratio of true positives returned by each method in their top_n ranked genes. A further analysis, not reported here for lack of space, is able to estimate the number of true positives within the intersection of two top_n sets of genes.

We think that this study points to a very promising new approach where one can benefit from the combination of several techniques. Our work is currently directed at (i) evaluating the approach on other sets of microarray data, (ii) using other feature selection techniques beside ANOVA and *Bio*RELIEF, and (iii) at investigating the mathematical properties of our proposed correlation measure between ranking methods. We strongly believe that the approach presented here could be applied in other problems as well where elements are ranked and where different ranking methods exist.

References

[1] C. Ambroise and G. J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, *Proceedings of the National Academy of Sciences, USA*, 99(10)6562-6566, 2002.

[2] A. Blum and P. Langley, Selection of relevant features and examples in machine learning, *Artificial Intelligence journal*,97:245-271, 1997.

[3] S. Dudoit and Y. H. Yang and M. Callow and T. Speed, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Technical Report, Stanford University, No.578, August 2000.

[4] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *Journal of Machine Learning Research*,3:1157-1182, 2003.

[5] D. Hwang and W. A. Schmitt and G. Stephanopoulos and G. Stephanopoulos, Determination of minimum sample size and discriminatory expression patterns in microarray data, *Bioinformatics*,18(9):1184-1193, 2002.

[6] K. Kerr and M. Martin and G. Churchill, Analysis of variance for gene expression microarray data, *J. of Comp. Biol.*, 7(6):818-837, 2000.

[7] K. Kira and L. Rendell, A practical approach to feature selection, *Int. Conf. on Machine Learning (ICML-92)*, Morgan Kaufmann, 249-256, 1992.

[8] R. Kohavi and G. John, Wrappers for feature subset selection, *Artificial Intelligence journal*, 273-324, 1997.

[9] W. Li and I. Grosse, Gene selection criterion for discriminant microarray data analysis based on extreme value distributions, *RECOMB'03*, ACM Press, 2003.

[10] W. Li and Y. Yang, How many genes are needed for a discriminant microarray analysis ?, *LANL e-print physics/0104029*, 2001.

[11] J. Mary and G. Mercier and J.-P. Comet and A. Cornuéjols and C. Froidevaux and M. Dutreix, An attribute estimation technique for the analysis of microarray data, *Proceedings of the Dieppe School on Modelling and Simulation of Biological processes in the Context of Genomics*, Amar, Ph. and Képès, F. and Norris, V. and Tracqui, P. (Eds), Publisher Frontier Group, 69-77, 2003.

[12] G. Mercier and N. Berthault and J. Mary and A. Antoniadis and J.-P. Comet and A. Cornuéjols and C. Froidevaux and M. Dutreix, Biological detection of low radiation by combining results of two analysis methods, *Nucleic Acids Research (NAR)*, 32(1):1-8, 2004.

[13] A. Ng, On feature selection: Learning with exponentially many irrelevant features as training examples, *Int. Conf. on Machine Learning (ICML-98)*, 1998.

[14] W. Pan, On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression, *Bioinformatics*, 19(11):1333-1340, 2003.

[15] M. Robnik-Sikonja and I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, *Machine Learning Journal*, 53(23), 2003.

[16] V. G. Tusher and R. Tibshirani and G. Chu, Significance analysis of microarrays applied to the ionizong radiation response, *Proceedings of the National Academy of Sciences, USA*, 98(9):5116-5121, 2001.

[17] E. Xing and M. Jordan and R. Karp, Feature selection for high-dimensional genomic microarray data, *Int. Conf. on Machine Learning (ICML-01)*, 2001.

[18] Y. H. Yang and Y. Xiao and M. Segal, Identifying differentially expressed genes from microarray experiments via statistic synthesis, *Bioinformatics*, 21(7):1084-1093, 2005.