JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Number 6, 2004

Pp. 1090–1109

Blind Source Separation and the Analysis of Microarray Data

P. CHIAPPETTA, M.C. ROUBAUD, and B. TORRÉSANI

ABSTRACT

We develop an approach for the exploratory analysis of gene expression data, based upon blind source separation techniques. This approach exploits higher-order statistics to identify a linear model for (logarithms of) expression profiles, described as linear combinations of "independent sources." As a result, it yields "elementary expression patterns" (the "sources"), which may be interpreted as potential regulation pathways. Further analysis of the so-obtained sources show that they are generally characterized by a small number of specific coexpressed or antiexpressed genes. In addition, the projections of the expression profiles onto the estimated sources often provides significant clustering of conditions. The algorithm relies on a large number of runs of "independent component analysis" with random initializations, followed by a search of "consensus sources." It then provides estimates for independent sources, together with an assessment of their robustness. The results obtained on two datasets (namely, breast cancer data and *Bacillus subtilis* sulfur metabolism data) show that some of the obtained gene families correspond to well known families of coregulated genes, which validates the proposed approach.

Key words: gene expression data, blind source separation, independent component analysis, coregulated genes.

1. INTRODUCTION

TODAYS MICROARRAY EXPERIMENTS yield large volumes of data and raise numerous statistical problems. Some of them are directly related to the experimental protocol and correction of the various biases (for example, normalization, image analysis, background correction) while others address data analysis and interpretational issues. This paper is devoted to the statistical analysis of microarray data using "blind source separation techniques," aiming at identifying "elementary" independent expression patterns, which may be thought of as potential candidates for regulation pathways.

Besides methods for studying differentially expressed genes in microarray data (see, for example, Dudoit *et al.* [2002]), many different approaches have been proposed for various goals including, among others, condition and/or gene clustering and condition discrimination. We shall deal in this paper with an approach based on a *linear modeling* of logarithms of expression data, as opposed to clustering-based approaches.

Laboratoire d'Analyse, Topologie et Probabilités, Centre de Mathématiques et Informatique, Université de Provence, France.

Clustering-based methods (including hierarchical methods such as UPGMA, or agglomerative approaches such as K-means, SOM, etc.) have been very popular, because they are fairly easy to use and require (at first sight) little tuning effort.[1] However, clustering is not always completely adapted to microarray datasets: though genes involved in the same biological process are likely to exhibit somewhat similar expression patterns and thus be correctly clustered by most clustering algorithms, some may be significantly involved in several biological processes and could therefore be naturally clustered with several different gene groups. Also, hierarchical clustering methods do not provide any simple way to find genes or gene groups with opposite expression patterns. In addition, microarray data often yield "unregulated" genes, whose expression profile does not contain much information, and prior filtering of such genes which are not regulated is generally necessary before applying clustering.

Linear models attempt to describe expression data as linear combinations of elementary "modes," to be interpreted as "expression patterns." An elementary mode takes the form of a "fake" microarray, i.e., a set of gene expression data. After such elementary modes have been identified, conditions may be compared to them, which yields useful informations in terms of condition classification. Also, the distribution of gene expression levels in a given source generally features a small number of significantly overexpressed or underexpressed genes, which kind of "govern" the source. Those genes generally form very biologically coherent groups and may be interpreted in terms of regulatory pathways.

A criterion is necessary to extract these elementary modes from datasets. Quite often (this is also true for clustering techniques), one relies on tools (for example, distances or dissimilarities) closely connected to second-order statistics (covariance, correlation). This is in particular the case of PCA- (principal component analysis) or SVD- (singular value decomposition) based approaches (see, for example, Alter *et al.* [2003], Ghosh [2002], and Wall *et al.* [2001]), or clustering methods (see, for example, Eisen *et al.* [1998], Ben-Dor *et al.* [1999], and Peterson [2002]) using correlation or Euclidean distances. However, higher-order statistics (for example, moments of higher order) contain significant complementary information. This is the case in particular as soon as the statistical distribution of data differs significantly from normal distributions, which turns out to happen quite often in microarray data. Indeed, some particular genes may happen to be significantly overexpressed in some specific conditions (and underexpressed in some others), which yields "heavy tail" distributions. Therefore, it makes sense to try to exploit such higher order statistics for the analysis of expression data; this turns out to yield information which are complementary to the information provided by first and second order moments.

In this paper, we report on an approach based on signal-processing techniques known as blind source separation methods, which amount to estimating linear mixtures of *statistically independent* modes from observations, which are therefore assumed to originate from a linear combination of independent, non-Gaussian *sources*. The estimation of such independent components is based on techniques which go under the name of *independent component analysis* (ICA for short; see Cardoso [1998] and Hyvärinen [1999] for reviews). Our approach models logarithms of expression profiles (by expression profile, we shall mean a set of expression levels for a given condition and a fixed gene set) as linear combinations of "elementary" sources (logarithms of profiles) which are statistically independent. The estimation of these independent sources, and of the corresponding mixing coefficients is performed using an algorithm called FastICA developed by Hyvärinen (1999). The rows of the mixing matrix represent the coefficients of the projection of the conditions onto the estimated sources. As such, they provide useful information in terms of discrimination or clustering of conditions. Further analysis shows that the estimated sources often exhibit a certain number of overexpressed and underexpressed genes, which may be used for further analysis of the data under consideration.

Since all ICA implementations we know of have to face the classical problem of convergence to local optima (a problem already alluded to by Liebermeister [2002], who mentioned that his results were "quite reproducible"), our approach relies on multiple ICA runs (with random initializations), followed by a search of "consensus independent sources," which yields extremely stable and robust estimates for the sources, as well as indications relative to their stability.

We illustrate our approach on a couple of case studies, using two significantly different datasets. The first one is a dataset of breast cancer data provided by the TAGC team (CIML Marseille). As a result, the source

[1]This is not totally true, since all clustering strategies require prior decisions regarding, for instance, the choice of distances or dissimilarities, the agglomeration strategies, and others which may be subject to arguments.

separation produces a number of sources which may be given a clear biological interpretation. In particular, three of the estimated sources may be put in correspondence with already-known facts in breast cancer microarray data, which had been observed previously in the dataset under study. We also obtain other independent sources, featuring families of over- or underexpressed genes with good biological coherence, which had not been reported before. The second dataset consists of *Bacillus subtilis* gene expression data obtained by Sekowska *et al.* (2000) for testing differences in gene expression when *Bacillus subtilis* is grown under different sulfur sources (methionine or methylthioribose). Blind source separation identified sources related to the main factors of variation in the experiment. Our results confirm the results of Sekowska *et al.* (2000); namely, the link between arginin metabolism and sulfur metabolism, and the role of the late competence operons. Blind source separation turns out to be able to identify the corresponding component, without any a priori information on experimental conditions (unlike the ANOVA-based analysis of Sekowska *et al.* [2000]). The analysis also points out families of genes related to mobility, which did not appear on previous studies.

All together, our results show that blind source separation techniques are a promising approach for exploratory microarray data analysis. In the two examples considered here, they allowed us to identify groups of genes with a good biological coherence. This shows that, even though the proposed method is *not* a clustering technique, it may be used to identify (possibly intersecting) classes of genes.

Note. After a first version of this work was completed (Chiappetta *et al.*, 2002b), we became aware of the works of Liebermeister (2002) and Hori *et al.* (2001), who have developed strategies close to ours for microarray analysis. The work reported here presents similarities with Liebermeister (2002). It also includes an additional "consensus source" search algorithm which yields finer estimates for the sources, as well as indications relative to their "credibility" (which is an important point since the search algorithms are likely to yield local optima rather than global ones). Also, the datasets we have considered are different from those studied by these authors, and we obtain for these results which were not reported in the original publications.

2. BLIND SOURCE SEPARATION AND INDEPENDENT COMPONENT ANALYSIS

Blind source separation is a recently introduced technique which originates from the signal-processing literature. The main idea is to disentangle statistically independent signals (sometimes called "linear modes," see Liebermeister [2002]) which have been linearly mixed. The approach we propose is to use such models of linear mixtures of independent sources to model logarithms of expression profiles and corresponding ICA algorithms to estimate the sources. Even though it is unlikely that such simple models can describe accurately complete microarray datasets, one may hope to be able to identify a few significant sources and put them in correspondence with biologically relevant features.

The general model goes as follows. Assume that we are given a family of vectors $\mathbf{S}^1, \mathbf{S}^2, \ldots \mathbf{S}^M$ (each being a vector of dimension denoted by I), termed "sources," and a family of observations $\mathbf{Y}^1, \mathbf{Y}^2, \ldots \mathbf{Y}^N$ (vectors of the same dimension), obtained by a linear "mixing" of the sources, in the form

$$\mathbf{Y}^n = \sum_{m=1}^{M} A_m^n \mathbf{S}^m, \quad n = 1, \ldots N. \tag{1}$$

Componentwise, we also write

$$y_i^n = \sum_{m=1}^{M} A_m^n s_i^m, \quad n = 1, \ldots N, \quad i = 1, \ldots I.$$

Here, $\mathbb{A} = \{A_m^n, m = 1, \ldots M, n = 1, \ldots N\}$ is an $M \times N$ matrix, called the *mixing matrix*. Throughout this paper, we shall assume that $N \geq M$ (in the practical examples to be discussed below, N shall be much larger than M). The blind source separation problem amounts to "inverting" the mixing, i.e., estimating the mixing matrix and the sources from the observations. To be able to solve this problem uniquely, further

assumptions have to be made. The sources are therefore modeled as random vectors over some probability space and assumed to be (statistically) independent. This independence assumption is the main ingredient of the estimation method.

Estimation of the parameters of a linear model is often performed via a principal component analysis: the diagonalization of the covariance (or the correlation) matrix of the expression data indeed yields a representation such as (1). In such a situation, the decorrelated "sources" $\mathbf{S}^m$ are pairwise orthogonal.

However, taking into account the distribution of (logs of) expression data (in particular, the existence of very large values of expression data), the principal component analysis is not necessarily the most appropriate answer. In particular, it does not take into account the information contained in higher-order moments. Independent component analysis (ICA) aims at performing such an analysis by seeking the sources $\mathbf{S}^m$ which are maximally independent. As a result, ICA produces estimates for the sources of the form

$$\hat{\mathbf{S}}^m = \sum_{n=1}^{N} B_n^m \mathbf{Y}^n \tag{2}$$

where $\mathbb{B} = \{B_n^m\}$ is a "disentangling matrix." In other words, the probability distribution of the unmixed vectors $\hat{\mathbf{S}}^1, \ldots \hat{\mathbf{S}}^M$ equals (or, in practice, is as close as possible to) the product distribution of the marginal distributions of $\hat{\mathbf{S}}^1, \ldots$ and $\hat{\mathbf{S}}^M$.

The mutual information provides an appropriate way of studying the departure from independence. For the sake of simplicity, we limit our discussion to continuous random variables to which a probability density may be associated (discrete variables are treated similarly, replacing integrals with sums). In this context, the mutual information essentially provides an average measure of the logarithm of the ratio between the joint probability density and the product of the marginal densities: given M (continuous) random variables with joint density ρ and marginal densities $f_1, \ldots f_M$, one defines

$$I[\rho] = \int \rho(x_1, \ldots, x_M) \log_2 \left(\frac{\rho(x_1, \ldots, x_M)}{\prod_{m=1}^{M} f_m(x_m)} \right) dx_1 \ldots dx_M.$$

An elementary calculation then shows that

$$I[\rho] = -H[\rho] + \sum_{m=1}^{M} H[f_m], \tag{3}$$

where $H[\rho]$ and $H[f_m]$ denote the Shannon entropies associated with the density ρ and the density f_m, respectively,

$$H[\rho] = -\int \rho(x_1, \ldots x_M) \log_2(\rho(x_1, \ldots x_M))\, dx_1, \ldots dx_M, \tag{4}$$

$$H[f_m] = -\int f_m(x) \log_2(f_m(x))\, dx, \quad m = 1, \ldots M, \tag{5}$$

and measure the information content of the corresponding probability distribution (the interested reader may refer to Shannon [1949] and the last chapters of Renyi [1970] for a detailed discussion of the mathematical properties of Shannon's entropy and interpretations).

Ideally, finding the independent sources may be performed by seeking the disentangling matrix $\mathbb{B}$ which minimizes the mutual information. However, estimating entropies or mutual information turns out to be quite difficult from the statistical point of view (see, for example, Beirlant *et al.* [1997] and references therein). Actually, it may be shown (see, for example, Cardoso [1998] and Hyvärinen [1999]) that minimizing the mutual information under the constraint that the estimated sources are uncorrelated is equivalent

to maximizing the entropy, i.e., finding (uncorrelated) directions in the data space in which the distribution of projections is *maximally non-Gaussian*.[2] This remark is the key point for the existence of efficient algorithms which yield estimates for independent sources, given a dataset.

Blind source separation has become an increasingly active field during the last decade, and many computer softwares have been contributed by different groups. A number of them are available at the ICA Central website.[3] In this work, we have mainly followed the approach and the algorithm proposed by Hyvärinen and Oja (2000), which are based upon the following strategy. Let Φ be any (nonquadratic) function, and let $\mathbf{V}$ denote an N-dimensional random vector; one seeks uncorrelated directions (given by unit vectors $\mathbf{w} \in \mathbb{R}^N$) which are maximally non-Gaussian in the following sense:

$$\sup_{\mathbf{w}, \|\mathbf{w}\|=1} (\mathbb{E}\{\Phi(\mathbf{w} \cdot \mathbf{V})\} - \mathbb{E}\{\Phi(\Gamma)\})^2 \tag{6}$$

under the constraint

$$\mathbb{E}\{(\mathbf{w} \cdot \mathbf{V})^2\} = 1. \tag{7}$$

Here, $\mathbb{E}\{X\}$ denotes the expectation of a random variable X, $\|\mathbf{v}\|$ denotes the Euclidean norm of a vector $\mathbf{v}$, $\mathbf{w} \cdot \mathbf{V}$ is the inner product of the vectors $\mathbf{w}$ and $\mathbf{V}$, and Γ is a reference $\mathcal{N}(0, 1)$ random variable.

Many different choices for the non-Gaussianity criterion are possible, the most popular one being probably the Kurtosis, given by the function

$$\Phi_{\text{kurt}}(u) = u^4,$$

(used by Hori *et al.* [2001]) for which $\mathbb{E}\{\Phi(\Gamma)\} = 3$. Other choices suggested by Hyvärinen and Oja (2000) include

$$\Phi_{\tanh}(u) = \log\cosh(u), \qquad \Phi_{\text{Gauss}}(u) = 1 - e^{-u^2/2},$$

each choice emphasizing different types of departure from Gaussianity (Liebermeister [2002] used the Gaussian). Rescaled and normalized versions of these three choices are displayed in Fig. 1. One of their main differences is the importance they give to large values.

Remark. The influence of the criterion Φ has been discussed by several authors (see, for example, Hyvärinen [1999]). In statistical terms, different criteria yield different estimators for the independent directions. For example, it is clear from Fig. 1 that they give variable importance to large values (quartic behavior for the Kurtosis, versus linear behavior for the $\log\cosh$). The choice of Φ, together with the nature of the distribution of the (unknown) source, influence the variance of the estimator. Hyvärinen (1999) proposes criteria for choosing the "optimal" criterion for a given distribution. In practice, the nature of the distributions being generally unknown in advance, "general purpose" criteria are used, at least in a first "exploratory" stage.

3. APPLICATION TO EXPRESSION DATA

We are concerned with the problem of analyzing and interpreting gene expression data. Our starting point is an array

$$\mathbb{X} = \{X_g^c, g = 1, \ldots N_g,\ c = 1, \ldots N_c\}, \tag{8}$$

where X_g^c denotes the measured expression level for the gene g in the condition (chip) c. We assume that the data under consideration have been appropriately normalized (normalization issues will therefore not

[2]The heuristic is that mixing has a tendency to "Gaussianize," and the mathematical reason is that the probability distribution which maximizes the Shannon entropy under the constraint of unit variance is the normal distribution.

[3]See *http://www.tsi.enst.fr/icacentral/index.html*.

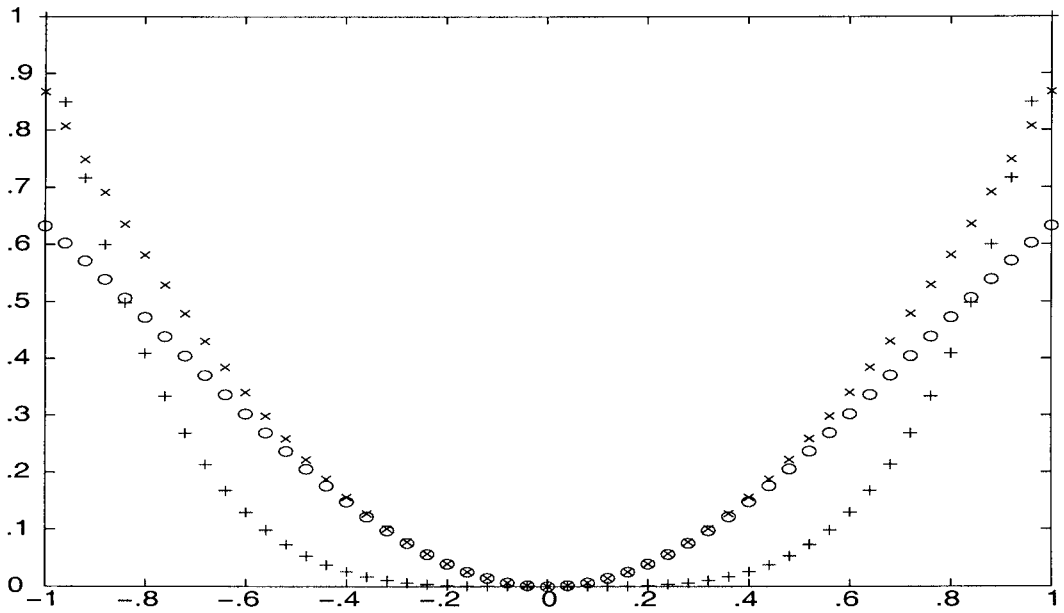

FIG. 1. Three different choices for the function Φ characterizing the non-Gaussianity criterion: rescaled versions of u^4 ("+"), $\log\cosh$ ("o"), and negative Gaussian ("x").

be discussed here as such). We shall denote by $\mathbf{X}^c$ (respectively, $\mathbf{X}_g$) the column (respectively, row) vectors corresponding to conditions (respectively, genes). Typical values for N_g and N_c are of the order of a few thousands and between fifty and two hundred, respectively.

3.1. Data preprocessing

The preprocessing of data is a standard but quite important aspect of microarray data analysis. There is quite a general consensus in the literature on applying logarithmic corrections to the data, for several reasons. A first justification is that some effects under study are likely to have a multiplicative behavior, which becomes linear after being log transformed. A more "data analysis" oriented one is that in the data under consideration, one often observes an extremely large amount of small values X_g^c, together with a few very large values; most data analysis techniques are strongly affected by such unbalanced distributions, which may be corrected by a logarithmic transformation of the type

$$X_g^c \hookrightarrow \log X_g^c.$$

Other standard choices include square root and hyperbolic tangent. More sophisticated approaches (see, for example, Durbin *et al.* [2002] for a transformation which takes into account noise models) may also be used, which, however, requires prior knowledge or estimates of the noise distribution.

However, for the datasets we have used in our study, such information was hardly available or usable, so that we had to stick to simple logarithmic corrections. In addition, for the breast cancer dataset, the data at hand feature a large number of zero or very small values of X_g^c (presumably resulting from a background noise removal and/or thresholding in the data acquisition protocol) for which the singular behavior of the log function at the origin may cause problems.

To overcome this shortcoming, we suggest "regularizing" the small values by adding a small "noise" ϵ_g^c. The usual practice is to use a small, suitably chosen constant for ϵ_g^c, or to make it condition dependent (i.e., to introduce a dependence with respect to c only) to account for normalization differences between chips. However, when the number of zero or small values is large, this introduces a large number of values equal to $\log\epsilon$, at the risk of introducing a systematic bias in the analysis. For that reason, we suggest using pseudo-random values for ϵ_g^c, with an appropriately chosen standard deviation. This, of course,

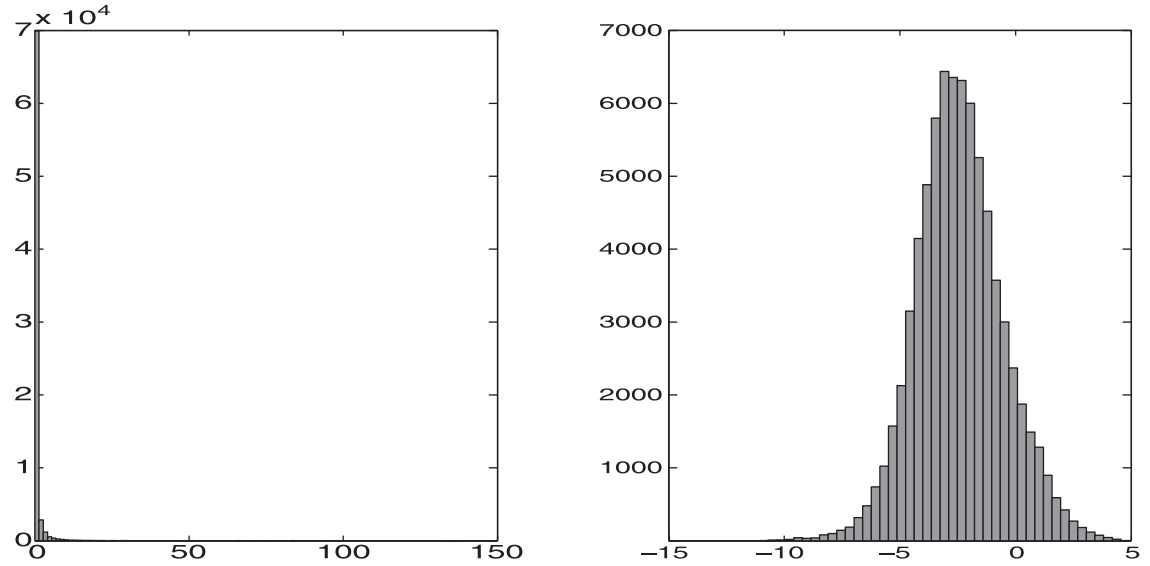

FIG. 2. Histograms of expression data (**left**) and of their "corrected logarithms" (**right**), in the case of breast cancer microarray data.

modifies significantly the quantitative information contained in small values, but preserves the qualitative information (small values remain small, instead of being thrown away) and does not introduce arbitrary bias. More precisely, we set

$$Y_g^c = \log(\tilde{X}_g^c) = \begin{cases} \log(X_g^c), & \text{if } X_g^c \neq 0 \\ \log(\epsilon_g^c), & \text{if } X_g^c = 0, \end{cases} \tag{9}$$

where the (positive valued) random perturbation is chosen as follows: for a fixed condition c, the mean and standard deviation of ϵ_g^c are adapted to the smallest observed value $\min_g(X_g^c)$ in the condition c. In the examples discussed in the present article, we have taken the ϵ_g^c's uniformly distributed between 0 and some constant (an example is shown in Fig. 2 below), the latter being proportional to the smallest observed value in the condition c. Even though such a choice may appear arbitrary, further numerical tests performed using different distributions did not change significantly the results.

Remark. Such a preprocessing turns out to be relevant when the number of zero values is large and no prior information on the noise is available. Otherwise, different approaches may be preferred. Let us, however, point out an advantage of our procedure, in terms of test of robustness. Performing the normalization in that way, and the subsequent statistical analysis, several times using different seeds for the generation of the regularizing numbers ϵ provides a simple way to test the robustness of the procedure. The results presented below turn out to be quite robust.

3.2. Independent component analysis

We denote by $\mathbf{Y}^c = \log \widetilde{\mathbf{X}}^c$ the corrected logarithms of expression levels and start from a model of the form

$$\mathbf{Y}^c = \sum_{m=1}^{M} A_m^c \mathbf{S}^m, \tag{10}$$

where the $\mathbf{S}^m$ are independent sources and $\mathbb{A}$ is the mixing matrix, to be estimated from the data.

In independent component analysis, the number of parameters to estimate turns out to be quite large, as discussed for instance by Cardoso (1998), and it is advantageous to reduce it. In order to focus on

higher-order moments, we first "whiten" (or "sphere") the data, using a principal component analysis: from the covariance matrix

$$\mathbb{C} = \mathbf{Y}'\mathbf{Y}$$

(where the prime symbol denotes matrix transposition) the principal axes are computed, and the variance along them is normalized to unity. This has the effect of "factoring out" the intrinsic variability of genes and attenuating the possible impact of outliers. Directions in which the variance is too small may also be factored out, which results in a reduction of the dimensionality of the data. By an abuse of notation, we still denote by $\mathbf{Y}$ the data after sphering (and dimension reduction when it is indeed performed).

Using the FastICA algorithm (see Hyvärinen and Oja [2000]), one obtains estimates for the independent sources (whose number has to be chosen in advance, but the choice may be refined afterwards, see below) and for the mixing matrix. The algorithm is essentially based upon a numerical optimization of the chosen criterion (see Equation (6)), using a Newton-type iterative method. An estimated source (i.e., a linear combination of the expression profiles $\mathbf{Y}^c$) takes the form of a vector of "fake" (logarithms of) expression levels and may be thought of as representing an "elementary regulation pattern," or "independent linear mode" (see Liebermeister [2002]), to be confirmed by a subsequent analysis.

3.3. Search for consensus sources and corresponding mixing matrices

Unlike principal component analysis, which is based only on linear algebra techniques, ICA requires searching the maxima of a target function in a large-dimensional configuration space. Therefore, one often encounters difficulties with local maxima in which most algorithms may get stuck, and the result may be sensitive to initialization. This is the case of the FastICA algorithm we use, even though we could observe empirically on all experiments we made that several interesting sources were strikingly stable (similar observation was also made by Liebermeister [2002]).

In addition, the results obtained from an ICA algorithm are not "ordered": if N_s sources are looked for, N_s sources are obtained, without any indication regarding their significance. This problem was addressed by Liebermeister (2002), who proposed to rank the estimated sources according to a contrast function, accounting for the percentage of total variation they carry.

To overcome these difficulties, we use the following procedure. The independent source estimation is run several times (say, 100 times), with different random initializations, and "consensus sources" are recorded: namely, sources which are obtained (to a certain controlled approximation) with a frequency larger than a certain threshold are conserved, and their frequencies of appearance are recorded and used as "credibility indices." As a result, one is led to a (variable, data-driven) number of average consensus sources $\bar{\mathbf{S}}^1, \ldots \bar{\mathbf{S}}^{N_s}$ (the algorithm for consensus sources is described in the appendix).

Finally, the corresponding $N_c \times N_s$ consensus mixing matrix $\mathbb{A}$ is computed by solving for

$$\min_{\mathbb{A}} \left\| \mathbf{Y}^c - \sum_{m=1}^{N_s} A_m^c \bar{\mathbf{S}}^m \right\|^2, \quad c = 1 \ldots N_c, \tag{11}$$

which implies the matrix equation $(\bar{\mathbf{S}}^n)'\mathbf{Y}^c = \sum_{m=1}^{N_s} A_m^c (\bar{\mathbf{S}}^n)'\bar{\mathbf{S}}^m$ (the "prime" standing for matrix or vector transposition). As long as the average sources are linearly independent (which should be the case if N_s is not too large, and must be checked for), this yields the following "pseudo-inverse" solution

$$A_m^c = \sum_{n=1}^{N_s} V_m^n (\bar{\mathbf{S}}^n)'\mathbf{Y}^c, \quad c = 1, \ldots N_c,\ m = 1, \ldots N_s, \tag{12}$$

where $\mathbb{V} = \{V_m^n, n, m = 1, \ldots N_s\}$ is the inverse of the $N_s \times N_s$ matrix $\mathbb{U}$ of the scalar product of the sources ($U_{mm'} = (\bar{\mathbf{S}}^m)'\bar{\mathbf{S}}^{m'}$).

Remark. Let us stress that the "linear mixture of independent modes" model on which our approach is based upon is quite speculative, and to be considered cautiously. Therefore, the "credibility" index

obtained as a result of the consensus sources search, though it does not provide any indication on the biological significance of the result (as well as the criterion used by Liebermeister [2002] for ranking sources, which is tied to the chosen non-Gaussianity criterion), is still a useful tool: when an independent source is obtained more than 90% of the times, with various random initializations, this may be considered a significant result.

3.4. Interpretation of ICA results

As a result, the blind source separation approach yields "pseudo" expression profiles, to be interpreted in more detail. A first step of the analysis is the study of the mixing matrix $\mathbb{A}$. For a fixed source, say source m, the coefficients A_m^c represent the projection of condition c on source m, or the "importance" of source m in condition c. If one believes in the "linear mixture of independent sources" model, and accepts identifying a source with a regulation pathway in first approximation, the coefficients A_m^c would allow one to assert to which extent the source m was (positively or negatively) "active" in condition c.

The distribution of the values of $\{A_m^c, c = 1, \ldots N_c\}$ is often interesting and may reveal specific features of the dataset. Particularly interesting is the situation where the distribution of mixing coefficients for a given source exhibit a bimodal or multimodal behavior. This indicates that the source under consideration has a good discriminating power between two or more different classes of conditions. However, let us point out that even though bimodal distributions yield spectacular results, good discrimination may also be obtained without such a behavior, as we shall see on the breast cancer dataset below.

A second step in the interpretation of ICA results is to analyze carefully the behavior of specific genes in different sources. It generally happens that a given independent source is characterized by a number of significantly overexpressed (or underexpressed) genes. Putting such genes into correspondence with conditions, or clinical data, may happen to be extremely informative. More precisely, we proceed as follows. For each estimated consensus source $\bar{\mathbf{S}}^m$, we pick the genes whose expression level in the considered source exceeds—in absolute value—some critical value z: $|\bar{S}_g^c| \geq z$. For the sake of simplicity (since departure from Gaussianity is the criterion on which the approach is based), z is chosen to be the critical value for the normal distribution corresponding to some fixed risk, for example, 0.1% or 0.01%. As we shall see in the case studies below, very coherent groups of genes may be obtained in this way.

3.5. Miscellaneous comments

Before turning to the discussion of test results, a few comments are necessary. It is important to realize that the approach proposed here is still far from an automatic procedure for expression data analysis, in several respects. We list here a few of them.

- A first point concerns the choice of the non-Gaussianity criterion. Several choices are possible. In addition, it has been shown that for a given (non-Gaussian) distribution, there exists a choice for the non-Gaussianity criterion which is optimal in the sense that it minimizes the variance of the estimator. The distributions of the sources being generally unknown, we chose to stick to general purpose criteria, namely the log cosh criterion, which does not give too much importance to large values. However, there is probably room for improvement at this point.
- The second point concerns the algorithm itself. The FastICA algorithm on which our approach is based has two main drawbacks for application to microarray data: the number of sources has to be chosen in advance, and the algorithm itself often yields estimates which correspond to local optima of the target function. Even though preliminary work showed some form of stability, we developed a consensus-source-search algorithm for postprocessing the ICA results. Not only does this procedure yield completely stable estimates for the sources, but it also provides a credibility estimate for them (see the appendix for mode details). Therefore, the number of retained sources may be made completely data driven, by limiting oneself to the sources whose credibility index exceeds some fixed tolerance. Nevertheless, it still makes sense to study alternatives to FastICA-type fixed-point algorithms.

Despite these remarks, our results definitely show that source separation techniques perform quite well in a variety of situations. In addition, the relative simplicity of the search algorithm makes it possible to perform quite a large number of ICA runs in a reasonable amount of time. To give a rough idea, the whole

process, involving 100 ICA runs and the corresponding consensus source search on the *B. subtilis* dataset described below (16 conditions, 16 sources, approximately 4,000 genes), took less than five minutes on a Pentium 4 processor (1.5 GHz).

4. TEST RESULTS

The source separation method has been applied to several datasets, including breast cancer data discussed by Bertucci *et al.* (2000, 2002) and *Bacillus subtilis* sulfur metabolism data (see Sekowska *et al.* [2000]). We now discuss results obtained with these two significantly different datasets.

4.1. Breast cancer data

There already exists a significant amount of literature on the study of breast cancer through microarrays, using different technologies (see, for example, Gruvberger *et al.* [2001], Perou *et al.* [1999], and West *et al.* [2001]).[4] These references were often concerned with the search for classes of genes of particular clinical interest, or on class prediction. We focus here on a dataset provided by the TAGC team of the Centre d'Immunologie de Marseille-Luminy (CIML), which has been analyzed by Bertucci *et al.* (2000, 2002). The data consist of nylon microarray data in which PCR products from cDNA were arranged on a nylon membrane and hybridized with a radioactive probe. They form a set of 1,045 genes and 67 microarrays (corresponding to 55 breast cancer tumors and 12 cell lines). The proposed approach has been applied to the complete $1{,}045 \times 67$ dataset, as well as the reduced $1{,}045 \times 55$ dataset (with the cell lines removed). Notice that no prior filtering of nonregulated genes was necessary, since one of the goals of the method is to be able to automatically identify significant gene families. In what follows, we discuss the results obtained using the blind source separation technique and compare them with those of Bertucci *et al.* (2002). Let us point out that PCA did not provide results that could be directly interpretable. We also refer to results we obtained by Chiappetta *et al.* (2002a) using clustering techniques.

4.1.1. The complete dataset. The corrected logarithms of normalized expression data have been computed as described above (in the dataset under consideration, the number of vanishing expression values was quite large, so that the random correction was needed). The distributions of normalized expression data and their corrected logarithms are shown in Fig. 2, from which it may be seen that the global distribution of corrected logarithms is bell shaped. We stress that very small values are likely to originate from random restoration of logs of vanishing values (which would have been small anyway). We have checked by running several times the log correction (with different seeds for the random number generator) that such very small values do not affect the results presented here.

A principal component analysis (performed on either the covariance matrix or correlation matrix) didn't yield any significant result, either in the gene space or in the conditions space.

The blind source separation algorithm (involving 100 ICA runs, followed by the consensus sources estimation) was run on the complete dataset (including breast tumor cells and cell lines), using the $\Phi = \log\cosh$ non-Gaussianity criterion (after PCA-based dimension reduction from 67 to 30). Twenty independent sources were estimated, and consensus sources were searched for from 100 ICA runs. The distributions of the mixing matrix coefficients for the 12 most significant sources (whose credibility exceeds 60%) are displayed in Fig. 3 (we used kernel estimators for the pdfs, with a Gaussian kernel, whose bandwidth was set to a fifth of the standard deviation; the values are plotted under the pdf plot). The pdfs essentially feature three main types of departure from Gaussianity:

- Bimodal distributions (in particular, the sources **2**, **9**, and to some extent **4**). Such a situation is particularly interesting, as it expresses the fact that the source under consideration has a significant discriminating power between conditions. We shall come back to those cases in more details below.

[4] A rather large account of the existing references on the subject may be found at the web site *www.clarkelabs.georgetown.edu/BreastStudies.html*.

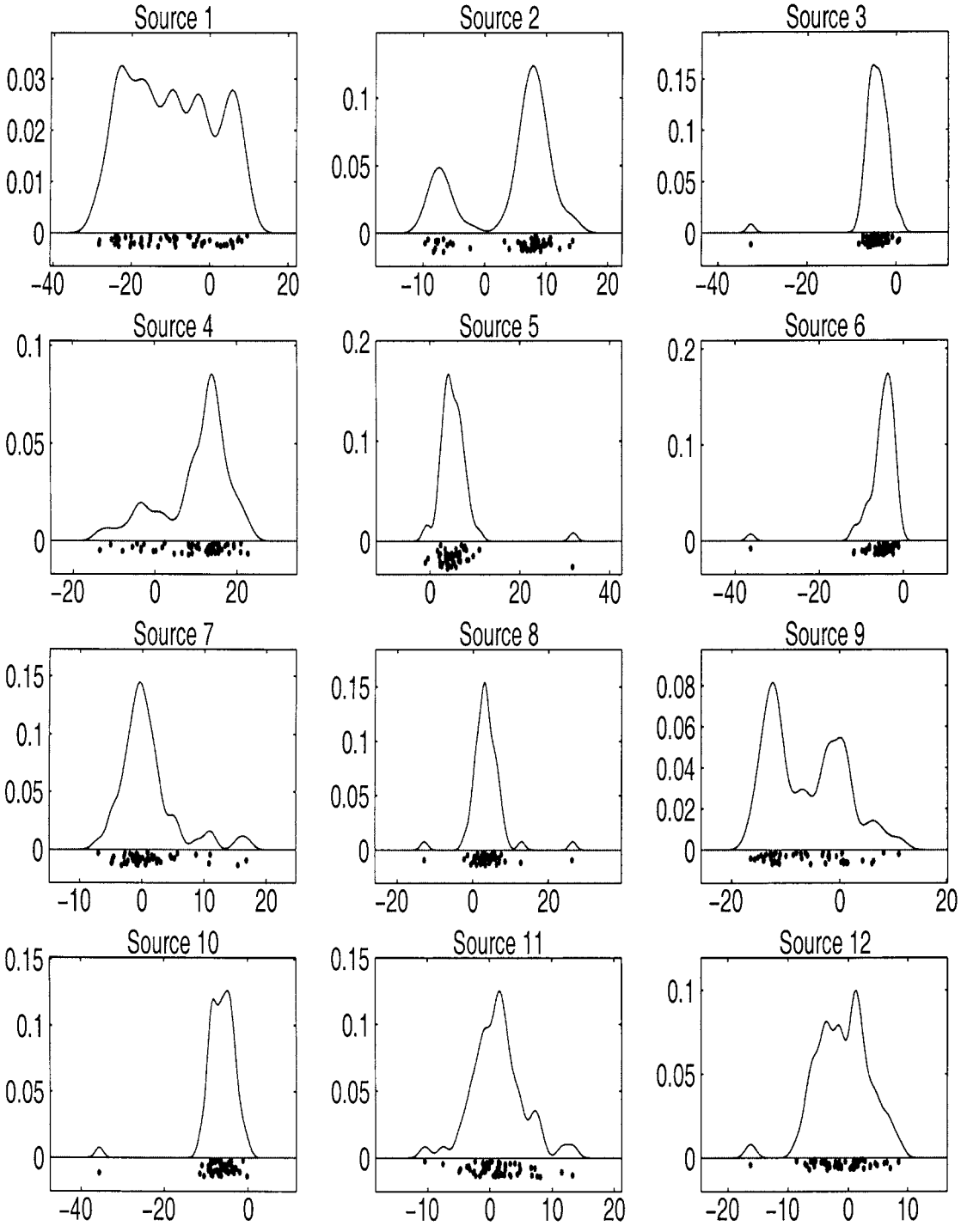

FIG. 3. Breast cancer data: pdf of estimated mixing matrix coefficients for the 12 "most credible" consensus sources (complete dataset).

- Distributions with one or a few very large values (for example, sources **2, 5, 6, 7, 10**, and **12**). Such situations indicate that a given condition played a prominent role in the selection of the source. This fact may originate from the possible presence of outliers, or stress a particular difference between one of a few conditions with respect to others.
- Distributions close to uniform; here the consensus source **1**. (Let us recall that the consensus sources have been ordered by decreasing "credibility.") Among the consensus sources which were not retained, some also exhibited a manifestly asymmetric distribution.

We now focus more closely on a few specific sources and examine the genes which are significantly overexpressed or underexpressed in the corresponding source. Besides sources which were essentially characterized by a large (or small) value for one particular condition (which were here quite difficult to interpret, and which we chose to disregard in this paper), the following four consensus sources appear to be particularly interesting, for various reasons.

Immune response genes (source 1). The stablest source (credibility 100%) yields a "uniform-like" pdf for the mixing matrix coefficients A_1^c. However, interestingly enough, the conditions c with largest values of A_1^c all correspond to cell lines and generate the right "hill" in the upper left plot of Fig. 3. Even though there is no gap between cell lines and tumor cells in that particular source, the separation nevertheless appears clearly.

The distribution of gene expression levels in that source (not shown here) exhibits a significant asymmetry. The genes significantly overexpressed in this source form a group closely connected to immune response. Namely, the following genes are significantly overexpressed in the source: immunoglobulin genes IGHM, IGHA1, IGKV1D-8, IGL, the GATA transcription factors GATA1, GATA2, GATA4, GATA6, and others among which are IL2RG (two clones), CSF1, NFYB, SUI1, RELA, NCOA3, SILV, FOS, CD79A, MYB, TNFRSF7, NFKB1. As a consequence of the asymmetry of the gene expression levels distribution, no significant family of genes antiexpressed with those was found. These results are coherent with those obtained by Bertucci *et al.* (2002), where a significant overexpression of this group of genes in the tissues was mentioned (a result similar to the one obtained here, even though no clear gap appears in the histogram). Let us also mention that the results are coherent with those we obtained by Chiappetta *et al.* (2002a), with a completely different approach.

Prolactin receptor and CIDE A (source 2). The credibility index of that source was found to equal 100%. The pdf of mixing matrix coefficients A_2^c for the consensus source **2** in Fig. 3 exhibits a spectacularly bimodal distribution, with a significant gap between the two components. This shows that the source under consideration has a high discriminating power.

The distribution of gene expression levels appears again to be quite asymmetric. A closer examination of the expression profile of the consensus source **2** shows that it is characterized by a significant overexpression of a certain number of specific genes, including (we quote the most significant) PRLR (prolactin receptor), CIDE A (cell death activator), CDKN3 (cyclin-dependent kinase inhibitor 3), TC21, CDH15, CNTFR, KLF1, and a splice variant of BCL2 (some of these genes have been shown to be associated to chemotherapy resistance).

Our results confirm results previously obtained by Bertucci and coworkers (Bertucci *et al.*, 2000, 2002), who also showed that this set of genes has a high predictive power in terms of patient survival. High expression of CIDE A, PRLR, and a few other genes has been shown to be associated with poor prognosis.

Stromal source (source 4). The consensus source **4** was again found very stable (credibility: 96%). The mixing matrix coefficients A_4^c exhibit a bimodal distribution. A closer examination of the coefficients shows that the small values of A_4^c actually correspond to the cell lines, which are well separated from the tumor cells by this particular component, although no clear gap appears (unlike the case of source 2).

Again, the distribution of gene expression levels is quite asymmetric, and features a number of significantly overexpressed genes, among which the collagen genes (COL6A1, COL1A1), IGF2 (insulin-like growth factor 2), the group of matrix metalloproteinases (MMP2, MMP3, MMP11, MMP13), SPP1, CDH11, IGF2 and IGHA1. This group was already mentioned by Bertucci *et al.* (2002) (identified as the "stromal cluster"), and its ability to discriminate between cell lines and tissues was observed. One may also notice the existence of a small complementary group, involving in particular VNN2 (vanin 2), the T-cell gene CD3G, CEACAM1, and the gene ATR involved in the growth of tumor cells.

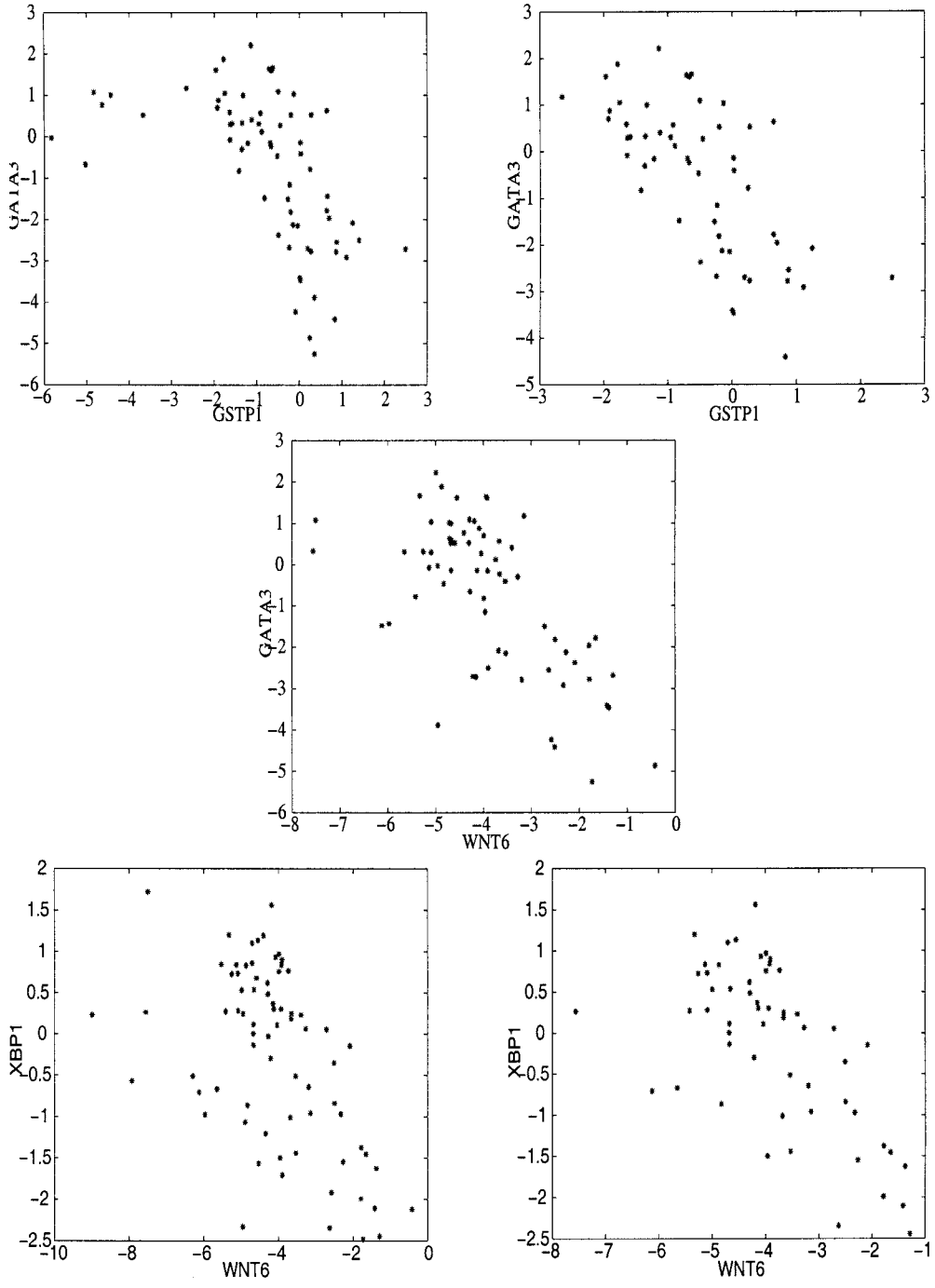

FIG. 4. Logarithms of expression levels. **Top row:** GSTP1 versus GATA3; **left:** complete dataset; **right:** breast cancer cells only. **Middle row:** WNT6 versus GATA3. **Bottom row:** WNT6 versus XBP1; **left:** complete dataset; **right:** breast cancer cells only.

TABLE 1. VALUES OF THE PEARSON CORRELATION COEFFICIENT FOR SIGNIFICANT GENES OF CONSENSUS SOURCE **9**

	GSTP1	*WNT6*	*GATA3*	*ESR1*	*MYB*	*XBP1*	*CRAPB2*
GSTP1	1.0000	0.3808	−0.6744	−0.5596	−0.4041	−0.4881	−0.3996
WNT6		1.0000	−0.6419	−0.4834	−0.4704	−0.6309	−0.3188
GATA3			1.0000	0.8351	0.6183	0.7156	0.5552
ESR1				1.0000	0.5071	0.6392	0.4162
MYB					1.0000	0.5760	0.2166
XBP1						1.0000	0.3645
CRABP2							1.0000

ESR1 and GATA3 versus GSTP1 (source 9). The consensus source **9** (credibility: 70%) is also extremely interesting. The histogram of mixing matrix coefficients in Fig. 3 exhibits a significantly bimodal behavior (although no clear gap appears in the values).

The distribution of gene expression levels in that particular source does not present an asymmetric shape, but is rather characterized by two groups of genes, which appear to be antiregulated. The significantly overexpressed genes are GATA3 (GATA binding protein 3, two clones), ESR1 (estrogen receptor), KRT19 (keratin 19, two clones), and MUC1, MYB, XBP1, CRABP2, IGFBP1, a family of genes which had already been reported to possess a great importance for prognosis.

These genes go together with a family of systematically underexpressed genes, in particular two clones of GSTP1 (glutathione S-transferase Pi 1), MAGEA3, WNT6 (wingless-related MMTV integration site 6 protein), EEF1G and CDH3. As may be seen in Fig. 4 (top left), significant underexpression (respectively, overexpression) of GSTP1 goes together with overexpression (respectively, underexpression) of GATA3.[5] The relationship between the (logarithms of) expression levels is remarkably close to linear. The corresponding correlation coefficient is $r = -0.5413$. In fact, the six conditions which do not seem to correlate well (upper left corner in the figure) correspond to cell lines, more specially those whose GSTP1 expression level is extremely small. When cell lines are not taken into account, the correlation increases significantly to $r = -0.6744$ (see Fig. 4, top plots). Similarly, the middle plot of Fig. 4 shows that WNT6 also has significant anticorrelation ($r = -0.6328$) with GATA3. Another illustration of such anti-correlation is provided by the pair WNT6-XBP1; the differential expression plots are exhibited in the bottom row of Fig. 4. The corresponding correlation coefficients read $r = -0.6185$ when the complete dataset is used, and $r = -0.6309$ when cell lines are not taken into account. All this seems to indicate the presence of two antiregulated groups of genes, including GATA3, XBP1, ESR1 on one hand, and GSTP1, WNT6, CDH3 on the other hand. The gene–gene correlations are summarized in Table 1.

The comparison with clinical data is also quite interesting. Remarkably enough, the mixing matrix coefficients A_9^c correlate extremely well with ER (estrogen receptor) clinical data. Leaving aside the extreme left part of the pdf plot (which corresponds only to cell lines), the two modes correspond to ER positive (left mode) and ER negative (right mode) tumor cells, the middle part being a bit more mixed up. This again confirms results reported by Bertucci *et al.* (2000, 2002).

Further comments. To confirm our results, we have also checked the dependence with respect to the non-Gaussianity criterion. The three criteria shown in Fig. 1 have been tested extensively, and the results we obtain appear to depend only weakly on the chosen criterion.

Also, since the initial number of sources has to be fixed in advance, we have tested for possible dependence of the results in this parameter. Again, the significant results, such as the four sources described above, appear to remain stable within a reasonable range of values of the number of sources (say, between 12 and 25).

4.1.2. The case of the reduced dataset As stressed before, the cell lines seem to play a significant role in several independent sources estimated from the complete dataset (this is particularly the case for the immune response source and the stromal source, which clearly discriminate between cell lines and

[5]Similar results had been previously reported by Gruvberger *et al.* (2001).

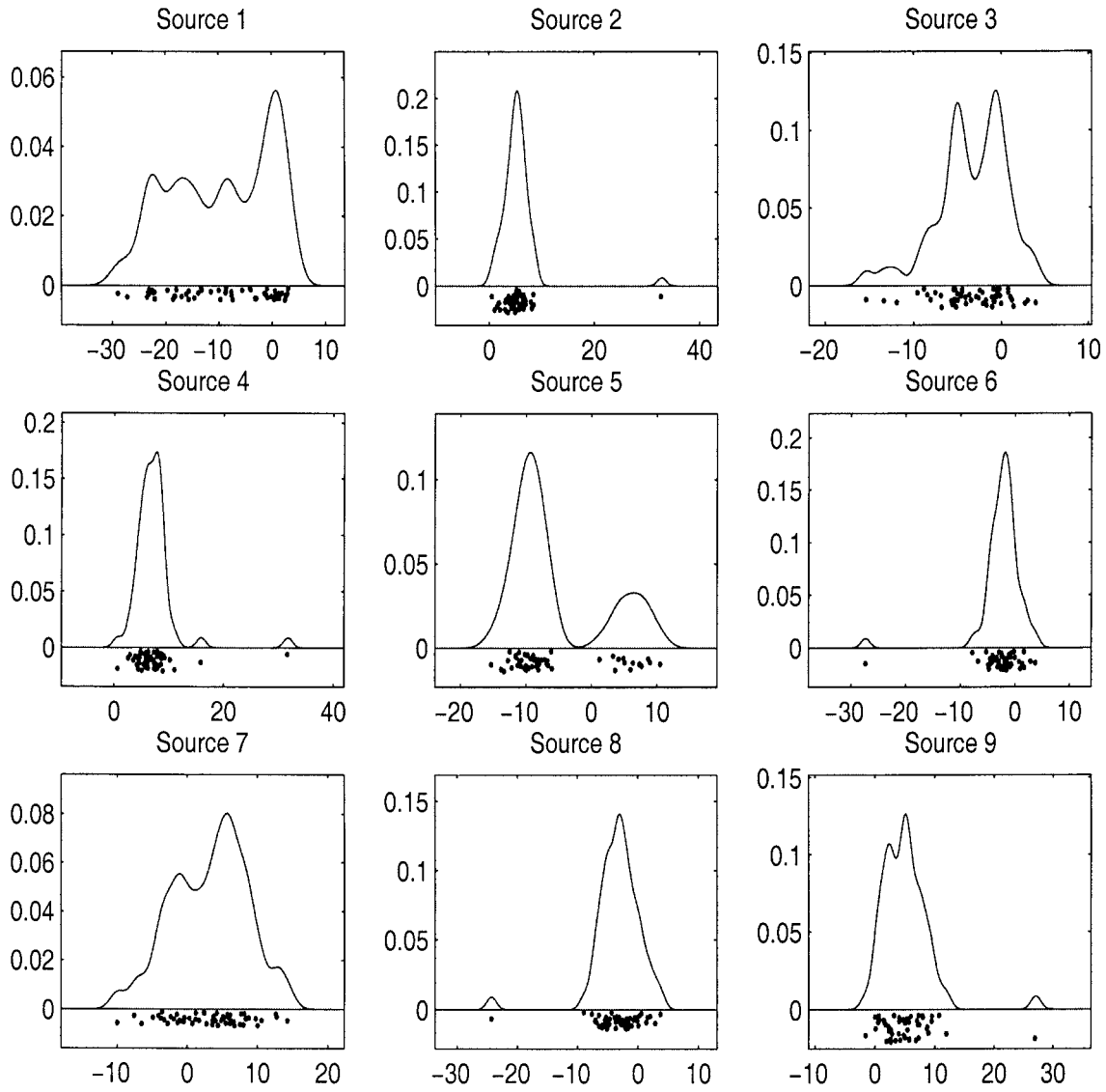

FIG. 5. Breast cancer data: pdf of estimated mixing matrix coefficients for the nine more credible consensus sources (small dataset).

tumors). When cell lines are not taken into account, the main features of the other interesting sources turn out to remain quite stable. The blind source separation has been performed on the reduced dataset, using equivalent parameters (same number of sources, dimension reduction to 25, same non-Gaussianity criterion). Figure 5 shows the pdf of mixing matrix coefficients for nine consensus sources with maximal credibility (more than 70%). As a result, the most significant components discussed above are still present in this new situation. In particular, the stablest source (source **1**, credibility index: 100%) is still the one governed by the immune response genes. We recall that in the case of the complete dataset, that source yielded a clear separation between cell lines and tumor cells. When cell lines are taken away, that source still shows up.

This is not the case of the source that was characterized by collagen genes (source **4**), which in the complete dataset also yielded a clear-cut separation betwen cell lines and tumors and does not appear any more in the reduced dataset.

The source involving CIDEA and PRLR (source **2** in the complete dataset) is still present and has the same spectacularly bimodal distribution (source **5**), with a somewhat smaller credibility index (85%). This

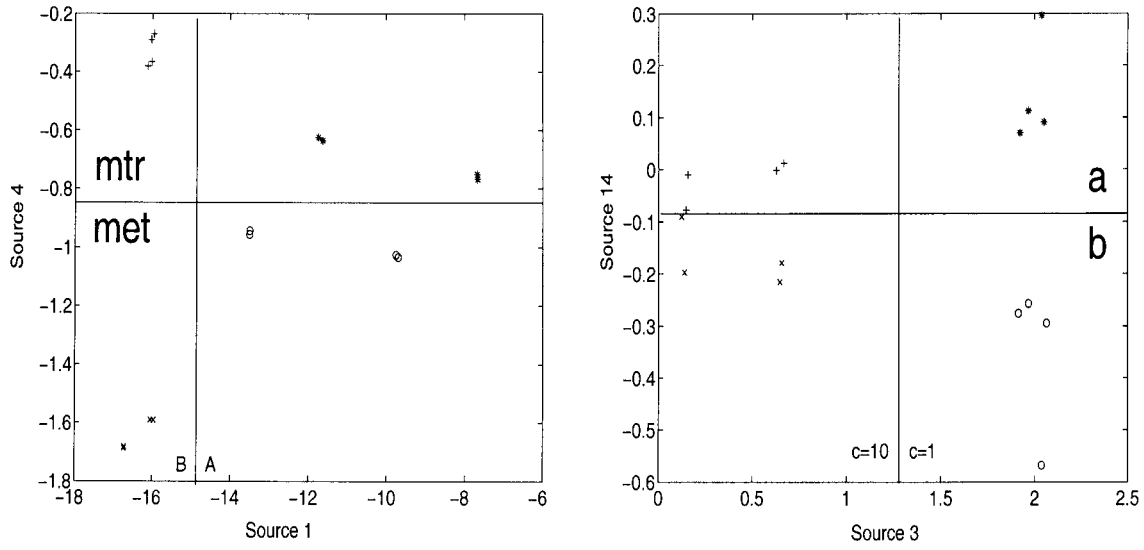

FIG. 6. *Bacillus subtilis* sulfur data mixing matrix coefficients: projections of the conditions on four distinct sources. **Left:** source 1 (date effect) and source 4 (sulfur source). **Right:** source 3 (concentration effect) and source 14 ("spot" effect).

is also the case of the source involving ESR1, GATA3, XBP1 (source **9** in the complete dataset and source **3** in the reduced one), with similar credibility. In both cases, the significantly under- and overexpressed genes are essentially the same in each source, and the discriminating power (in terms of ER status, for instance) is also stable.

4.2. Sulfur metabolism Bacillus subtilis *data*

The blind source separation method was applied to a dataset on the expression variations of 4,107 genes of *Bacillus subtilis* across two sulfur nutrients (methionine and methylthioribose), in different conditions: two different sulfurs (mtr or met), two different dates (days A and B), two different RNA concentrations (1 μg and 10 μg), and two spots per experiment (spots a and b), all together 16 different conditions. The relative effect of these factors has been studied using the analysis of variance (see Sekowska *et al.* [2000] and Didier *et al.* [2002]). The data under study are logarithms of expression levels (in the absence of zero values, no "random" correction was necessary). The independent component analysis turns out to be able to isolate such effects in a quite simple way. In the numerical experiment reported here, the same procedure as above was used, using the Φ_{tanh} non-Gaussianity criterion, without prior dimension reduction, and 100 runs were done seeking each time 16 sources. The obtained consensus sources turn out to be remarkably stable (seven consensus sources obtained credibility 100%, the fourteenth—out of 16—having credibility 94%). Out of these consensus sources, four particular ones (namely, sources **1**, **3**, **4** and **14**) turned out to have a simple and clear interpretation, and an additional one (source **2**) presented interesting characteristics. Since the number of conditions (16) was fairly small, the results are displayed differently in Fig. 6. The other consensus sources we obtained sometimes presented interesting patterns in terms of condition grouping (though not as clear as the four sources examined below), but the corresponding overexpressed and underexpressed genes were mainly genes with unknown function.

4.2.1. Methionine versus methylthioribose: Source **4**. The mixing matrix coefficients for consensus source **4** appear to form different groups and appear in the following order: the four metB conditions, then the four metA, the four mtr A, and finally the four mtr B. The two sulfur sources are then clearly separated by the horizontal line in the left of Fig. 6. That source does not make any significant difference between the two spots ("a" and "b").

Among the identified genes, this particular source shows a very significant antiexpression of two families of genes: a group arg* of 6 arginine biosynthesis genes, together with the 3 yqi* mentioned by Sekowska

et al. (2000) as arginine transporters, 6 ydc* and 6 ydd* genes, and the groups flg* (4 genes), flh* (4 genes), fli* (11 genes), and ytm* (4 genes) on the other hand. We therefore first recover the results published by Sekowska *et al.* (2000) stressing the role of arginin biosynthesis genes in sulfur metabolism, and almost all of the genes identified in Table 1 of Sekowska *et al.* (2000) appear in the list above. Interestingly enough, the blind source separation also points out other gene families (namely, the unknown ydc* and ydd* genes on one hand, and the flg*, flh*, fli* flagellar proteins and other genes related to mobility on the other hand). We do not have any simple biological explanation to such a very strong appearance of mobility-related genes in this group.

4.2.2. The date effect: Source **1**. The difficulty of controlling precisely the experimental conditions from one day to another was mentioned as an important source of variation. The "date" effect is present in the ICA analysis. The separation appears to be mainly governed by a "day" effect (A or B), and a concentration effect for day A. More precisely, conditions appear in the following order: first the B-day conditions, then the A-day conditions, with a clear cut separation (see the vertical line in the left plot of Fig. 6). This confirms the results obtained by Sekowska *et al.* (2000, Table 3): day A data is characterized by an overexpression of genes involved in competence (11 com* genes, nucA). More generally, most of the genes appearing in Table 3 of Sekowska *et al.* (2000) appear as significantly over- or underexpressed in this particular source.

4.2.3. Other unexplained effects. The blind source separation algorithm also provides sources which may be put into correspondence with well-defined factors, even though there does not seem to exist any simple biological explanation or any well-defined gene family specifically involved.

For instance, the effect of concentration was identified by Sekowska *et al.* (2000) as the most important source of variance, even though no simple explanation could be given. The consensus source **3** estimated by ICA provides a clear cut separation between the conditions corresponding to different RNA concentrations (see the vertical line in the right plot of Fig. 6). The left-hand part of the points corresponds to RNA amount of 10 μg, while the right-hand part corresponds to 1 μg of RNA. A closer examination of the main genes involved in this particular source did not lead us to any sensible interpretation, exhibiting neither well-identified gene families, nor coherent sets of genes of unknown function.

Similarly, and even though the "spot" effect was mentioned as weak and disregarded by Sekowska *et al.* (2000), it appears significantly in the consensus source **14** (credibility 94%), as may be seen on the right-hand plot of Fig. 6. More precisely, the conditions appear in the following order: the "b" spot first, then the "a" spot (even though the two conditions close to the horizontal line are too close to each other to yield a really significant order). Again, the main genes involved in this particular source are genes with unknown function, and they did not lead us to any sensible interpretation.

5. CONCLUSIONS

Blind source separation appears as a promising tool for exploratory analysis of gene expression data, as already remarked by Liebermeister (2002) and Hori *et al.* (2001). The additional "consensus sources" search techniques yields stable estimates for sources.

In the two examples we have studied in this article, this approach was able to identify consensus-independent sources which have a good biological coherence and put them into correspondence with consistent classes of conditions. Moreover, it could do so without any a priori information (unlike ANOVA and related techniques) or prior gene filtering. As such, it may also be used as a "class discovery method," like the methods described by Ben-Dor *et al.* (2001) or von Heydebreck *et al.* (2001), even though this is not the main goal of the approach.

Some aspects of the analysis described here require further investigations and developments. Among them, the algorithmic part is one of the most important. Like many data-analysis and clustering techniques, ICA provides estimates which turn out to depend on parameters such as the required number of sources, the initialization of the algorithm, (even though the really significant results turn out to be remarkably stable). The consensus source search represents a significant improvement in this respect. However, this is

a point we plan to investigate further, for example, by studying variants (avoiding fixed-point algorithms) to the optimization algorithms we have used here.

Another important point is the a posteriori validation of the technique. In our work, the results are validated by taking into account prior knowledge about the problem, which is not taken into account in the estimation of the sources (namely, the adequacy of the sources with known facts about conditions, or the coherence of the sets of genes significantly involved in a given source).

APPENDIX: CONSENSUS SOURCES ESTIMATION

When running ICA $\mathcal{N}$ times, we obtain $\mathcal{N}$ times M candidate sources, denoted by $\mathbf{S}_m^n$, $n = 1, \ldots \mathcal{N}$, $m = 1, \ldots M$, which we normalize to unity. We recall that each of these is an N_g-dimensional vector (N_g being the number of genes). Out of these, the significant ones are expected to show up a large number of times, modulo small perturbations, and in uncontrolled order. We outline below the procedure we use for searching such significant sources.

We first compute the similarity matrix of (normalized) source scalar products and record the pairs of indices $[(n, m), (n', m')]$ such that $\mathbf{S}_m^n$ and $\mathbf{S}_{m'}^{n'}$ are *similar*, in the following sense

$$\mathbf{S}_m^n \sim \mathbf{S}_{m'}^{n'} \quad \text{if} \quad n \neq n' \quad \text{and} \quad |\langle \mathbf{S}_m^n, \mathbf{S}_{m'}^{n'} \rangle| \geq \tau, \tag{13}$$

where $\tau \in [0, 1)$ is a fixed threshold (typically, $\tau = 0.9$) (the sources being normalized, this amounts to considering the sources whose Pearson coefficient exceeds τ in absolute value). We then obtain an $\mathcal{N}M \times \mathcal{N}M$ matrix $\mathcal{M}$, whose nonzero entries correspond to similar sources and are set to 1. Equivalently, $\mathcal{M}$ is the adjacency matrix of a graph whose vertices are the estimated sources $\mathbf{S}_m^n$ and whose edges connect similar sources. Consensus sources are then defined from maximal connected subgraphs, by averaging the corresponding sources, and their credibility is obtained as their relative frequency in the $\mathcal{N}$ simulations.

More precisely, we used the following approximate scheme, which amounts to gradually "peeling off" the set of estimated sources: for a fixed value of τ, initialize the similarity matrix $\mathcal{M}^{(1)} = \mathcal{M}$ and consider the set of all $n(1) = \mathcal{N}M$ estimated sources. Then do the following iteration:

while $dim(\mathcal{M}^{(k)}) > 0$, do

- Among the $n(k)$ remaining sources, pick the source $\mathbf{S}^{(k)}$ with maximal number $|N(\mathbf{S}^{(k)})|$ of neighbors, and denote by $c(k) = |N(\mathbf{S}^{(k)})|/\mathcal{N}$ the corresponding *credibility index*.
- Calculate the corresponding average source

$$\overline{\mathbf{S}}^{(k)} = \frac{1}{|N(\mathbf{S}^{(k)})|} \sum_{\mathbf{S}' \sim \mathbf{S}^{(k)}} \text{sign}(\langle \mathbf{S}^{(k)}, \mathbf{S}' \rangle)\, \mathbf{S}'. \tag{14}$$

- Remove the sources in $\mathbf{S} \sim \mathbf{S}^{(k)}$ from the list of estimated sources (yielding $n(k)$ sources) and the corresponding entries from the similarity matrix (now of dimension $n(k) \times n(k)$).

We then obtain an *ordered* set of consensus sources $\overline{\mathbf{S}}^{(1)}, \overline{\mathbf{S}}^{(2)}, \ldots$ together with their credibility index $c(1), c(2), \ldots$. The consensus sources whose credibility exceeds a fixed value are finally retained. Even though such a procedure is not optimal in a general situation, it is fairly simple and turned out to perform very satisfactorily in the situation at hand. This procedure mainly depends upon two parameters: the threshold τ and the final number of consensus sources. The result turns out to depend weakly on the value of τ: this comes from the fact that the sources are either very close to each other, or significantly different. In the first case, their scalar product is quite close to 1, and setting τ to 0.9 or 0.7 does not make much difference. Otherwise, the scalar product is very small (we recall that the dimension of the sources is N_g, usually a large number).

ACKNOWLEDGEMENTS

We wish to thank R. Houlgatte from the Marseille TAGC group and A.S. Carpentier and A. Hénaut from the Laboratoire Génome et Informatique of the Génopole d'Evry for stimulating discussions and for providing us the datasets discussed in this paper. We are also grateful to the anonymous referees for their valuable suggestions and criticisms and for bringing several references, including Liebermeister (2002), Hori *et al.* (2001), and Durbin *et al.* (2002) to our attention.

REFERENCES

Alter, O., Brown, P., and Botstein, D. 2003. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl. Acad. Sci. USA* 100(6), 3351–3356.

Beirlant, J., Dudewicz, E., Gyorfi, L., and van der Meulen, E. 1997. Nonparametric entropy estimation: an overview. *Int. J. Math. Statist. Sci.* 6(1), 17–39.

Ben-Dor, A., Friedman, N., and Yakhini, Z. 2001. Class discovery in gene expression data. *RECOMB 2001*, 31–38.

Ben-Dor, A., Shamir, R., and Yakhini, Z. 1999. Clustering gene expression patterns. *J. Comp. Biol.* 6(3–4), 281–297.

Bertucci, F., Houlgatte, R., Benziane, A., Granjeaud, S., Adelaide, J., Tagett, R., Loriod, B., Jacquemier, J., Viens, P., Jordan, B., Birnbaum, D., and Nguyen, C. 2000. Gene expression profiling of primary breast carcinomas using arrays of candidate genes. *Human Mol. Genet.* 9(20), 2981–2991.

Bertucci, F., Nasser, V., Granjeaud, S., Eisinger, F., Adelaide, J., Tagett, R., Loriod, B., Giaconia, A., Benziane, A., Devilard, E., Jacquemier, J., Viens, P., Nguyen, C., Birnbaum, D., and Houlgatte, R. 2002. Gene expression profiles of poor-prognosis primary breast cancer correlate with survival. *Human Mol. Genet.* 9(8), 863–872.

Cardoso, J.F. 1998. Blind signal separation: Statistical principles. *Proc. IEEE* 9(10), 2009–2025.

Chiappetta, P., Roubaud, M.-C., and Torrésani, B. 2002a. Classification mixte pour l'analyse de données d'expression. *Proc. JOBIM'02*, 53–54.

Chiappetta, P., Roubaud, M.-C., and Torrésani, B. 2002b. Séparation de sources en aveugle pour l'analyse de données d'expression. *Proc. JOBIM'02*, 131–136.

Didier, G., Brézellec, P., Remy, E., and Hénaut, A. 2002. GeneANOVA—Gene expression analysis of variance. *Bioinformatics-Applications Notes* 18(3), 490–491.

Dudoit, S., Yang, Y., Callow, M., and Speed, T. 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12(1), 111–139.

Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. 2002. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics* 18(suppl. 1), S105–S110.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863–14868.

Ghosh, D. 2002. Singular value decomposition regression models for classification of tumors from microarray experiments. *Pacific Symposium on Biocomputing*, 18–29.

Gruvberger, S., Ringnér, M., Chen, Y., Panavally, S., Saal, L., Borg, A., Fernö, M., Peterson, C., and Meltzer, P. 2001. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.* 61, 5979–5984.

Hori, G., Inoue, M., Nishimura, S., and Nakahara, H. 2001. Blind gene classification based on ICA of microarray data. *ICA 2001*.

Hyvärinen, A. 1999. Fast and robust fixed point algorithms for independent component analysis. *IEEE Trans. on Neural Networks* 10(3), 626–634.

Hyvärinen, A., and Oja, E. 2000. Independent component analysis: Algorithms and applications. *Neural Networks* 13, 411–430. See also *Independent Component Analysis: A tutorial*, available at *www.cis.hut.fi/projects/ica*.

Liebermeister, W. 2002. Linear modes of gene expression determined by independent component analysis. *Bioinformatics* 18, 51–60.

Perou, C., Jeffrey, S., van de RijnDagger, M., Rees, C., Eisen, M., Ross, D., Pergamenschikov, A., Williams, C., Zhu, S., Lee, J., Lashkari, D., Shalon, D., Brown, P., and Botstein, D. 1999. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Nat. Acad. Sci. USA* 16, 9212–9217.

Peterson, L. 2002. CLUSFAVOR 5.0: Hierarchical cluster and principal-component analysis of microarray-based transcriptional profiles. *Genome Biol.* 3(7), software 0002.

Renyi, A. 1970. *Foundations of Probability*, Holden-Day, San Francisco.

Sekowska, A., Robin, S., Daudin, J., Hénaut, A., and Danchin, A. 2000. Extracting biological information from DNA arrays: An unexpected link between arginine and methionine metabolism in *Bacillus subtilis*. *Genome Biology* 2(6), 19.1–19.12.

Shannon, C. 1949. *A Mathematical Theory of Communication*, University of Illinois Press.

von Heydebreck, A., Huber, W., Poustka, A., and Vingron, M. 2001. Identifying splits with clear separation: A new class discovery method for gene expression data. *Bioinformatics* Suppl. 1, S107–S114.

Wall, M., Dyck, P., and Brettin, T. 2001. Svdman—Singular value decomposition analysis of microarray data. *Bioinformatics* 17(6), 566–568.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., John, A., Olson, J.J.R.M., and Nevins, J.R. 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Nat. Acad. Sci. USA* 98(20), 11462–11467.

Address correspondence to:
B. Torrésani
LATP, CMI
39 Rue Joliot-Curie
F-13453 Marseille Cedex 09
France

E-mail: torresan@cmi.univ-mrs.fr