BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 15 2008, pages 1688–1697
doi:10.1093/bioinformatics/btn245

Gene expression

Knowledge-based gene expression classification via matrix factorization

R. Schachtner[1], D. Lutter[1,2,3], P. Knollmüller[1], A. M. Tomé[4], F. J. Theis[1,2], G. Schmitz[3], M. Stetter[5], P. Gómez Vilda[6] and E. W. Lang[1,*]

[1]CIML/Biophysics, University of Regensburg, D-93040 Regensburg, [2]CMB/IBI, GSF Munich, [3]Clinical Chemistry, University Hospital Regensburg, D-93042 Regensburg, Germany, [4]IEETA/DETI, Universidade de Aveiro, 3810-193 Aveiro, Portugal, [5]Siemens Corporate Technology, Siemens AG, Munich, Germany and [6]DATSI/FI, Universidad Politécnica de Madrid, E-18500 Madrid, Spain

Received on September 24, 2007; revised on May 14, 2008; accepted on May 23, 2008

Advance Access publication June 5, 2008

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Modern machine learning methods based on matrix decomposition techniques, like independent component analysis (ICA) or non-negative matrix factorization (NMF), provide new and efficient analysis tools which are currently explored to analyze gene expression profiles. These exploratory feature extraction techniques yield *expression modes* (ICA) or *metagenes* (NMF). These extracted features are considered indicative of underlying regulatory processes. They can as well be applied to the classification of gene expression datasets by grouping samples into different categories for diagnostic purposes or group genes into functional categories for further investigation of related metabolic pathways and regulatory networks.

Results: In this study we focus on unsupervised matrix factorization techniques and apply ICA and sparse NMF to microarray datasets. The latter monitor the gene expression levels of human peripheral blood cells during differentiation from monocytes to macrophages. We show that these tools are able to identify relevant signatures in the deduced component matrices and extract informative sets of marker genes from these gene expression profiles. The methods rely on the joint discriminative power of a set of marker genes rather than on single marker genes. With these sets of marker genes, corroborated by leave-one-out or random forest cross-validation, the datasets could easily be classified into related diagnostic categories. The latter correspond to either monocytes versus macrophages or healthy vs Niemann Pick C disease patients.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: elmar.lang@biologie.uni-regensburg.de

1 INTRODUCTION

Modern signal processing and machine learning techniques provide appropriate tools to analyze high-throughput datasets like microarrays. Despite the fact that many problems still remain to be solved (Dougherty and Datta, 2005; Dougherty *et al.*, 2005; Quackenbush, 2001), some consensus is slowly reached as to how data should be analyzed properly (Allison *et al.*, 2006).

Raw gene expression level measurements need sophisticated preprocessing (Wu and Irizarry, 2007) encompassing background correction, summarization, normalization (Baldi and Hatfield, 2002; Hochreiter *et al.*, 2006) and missing value imputation (Troyanskaya *et al.*, 2001), which is often done using software available from the chip producer (Affymetrix, 2002).

After preprocessing, normalized gene expression levels can be analyzed using feature extraction (Guyon and Elisseeff, 2003) and classification (Dudoit *et al.*, 2002) methods. Any statistical analysis of gene expression probe level data, however, has to face the 'large N, small M' problem setting, where N denotes the number of genes (= features, variables, parameters) and M denotes the number of samples (= experiments, environments, tissues). Also overfitting has to be avoided to construct a classifier with a good generalization ability (Spang *et al.*, 2002). Any robust classifier needs a sample-per-feature (SpF) ratio of 5-to 10-fold, while with usual microarray probe level measurements the SpF amounts to $1/50-1/200$ roughly. Hence a substantial reduction of the feature space dimensionality via gene or feature selection is often the only way out of this SpF dilemma.

Traditionally two strategies exist to analyze such sets of gene expression signatures: *Supervised approaches* and *Unsupervised approaches*. *Supervised approaches* afford prior knowledge such as class labels, clinical outcomes, prior densities, etc. and a truly representative set of training data. They are generally used for classification of malignancies within a discriminant analysis. *Unsupervised approaches* explore correlations in the high-dimensional data space and find appropriate transformations to identify relevant subspaces and group observations accordingly. However, such approaches often need additional constraints to yield unique answers but they allow for the detection of new, yet unknown classes (Saidi *et al.*, 2004). For a detailed account of the relevant literature see the extended 'Introduction' in the accompanying Supplementary Material.

There is a recent interest in applying exploratory matrix factorization (MF) techniques, like principal component analysis (PCA), independent component analysis (ICA) or non-negative

[*]To whom correspondence should be addressed.

matrix factorization (NMF), to gene expression level measurements with microarrays (Liebermeister, 2002). In this study we propose to include diagnostic knowledge and explore the potential of matrix decomposition techniques to identify and extract marker genes from microarray data sets and classify these datasets according to the diagnostic classes they represent. Note that the feature extraction process via exploratory matrix decomposition techniques is unsupervised, but the identification of the most relevant features follows the supervision of diagnostic information available. Preliminary work along these lines has been presented recently at a conference (Schachtner *et al.*, 2007a). Corresponding supervised feature extraction and classification techniques like support vector machines (SVM) have been applied to the same dataset and are discussed in short as well. For a more detailed discussion of these supervised techniques, though applied to different datasets, see (Schachtner *et al.*, 2007b).

2 THE MONOCYTE–MACROPHAGE DATASET

For our analysis we combined the gene-chip results from three different experimental settings to the monocyte–macrophage (MoMa) dataset (Lutter *et al.*, 2008). In each experiment human peripheral blood monocytes were isolated from healthy donors (Experiment 1 and 2) and from donors with Niemann Pick type C disease (Experiment 3). Monocytes were differentiated to macrophages for 4 days in the presence of M-CSF (50 ng/ml, R&D Systems). Differentiation was confirmed by phase contrast microscopy. Gene-expression profiles were determined using Affymetrix HG-U133A (Experiment 1 and 2) and HG-U133plus2.0 (Experiment 3) Gene Chips covering 22 215 probe sets and about 18 400 transcripts (HG-U133A). Probe sets only covered by HG-U133plus2.0 array were excluded from further analysis. In Experiment 1 pooled RNA was used for hybridization, while in Experiment 2 and 3 RNA from single donors were used. The final dataset consisted of seven monocyte and seven macrophage expression profiles and contained 22 215 probe sets. After filtering out probe sets which had at least one absent call, 5969 probe sets remained for further analysis. Exp. 1–7 refer to monocytes and Exp. 8–14 to macrophages. Exp. 1–4 and 8–11 stem from healthy subjects, the rest from diseased subjects.

3 METHODS

The data are traditionally represented as a $N\times M$-dimensional data matrix $\mathbf{X}$ whose M columns represent *gene expression signatures* (GES) of N genes during M experiments or environmental conditions while the N rows represent *gene expression profiles* (GEP) of each of the N genes across all M experimental conditions. Column vectors are denoted as, for example, $\mathbf{x}_{*m}$, while row vectors are denoted as $\mathbf{x}_{n*}$ in the following. The index m is always signifying an environmental condition in the following whereas the index n always refers to a certain gene.

3.1 Gene selection schemes

A general problem with microarray datasets is the problem of overfitting which arises whenever the number N of parameters (genes) is large compared to the number M of samples (experimental conditions). One way around this problem is to preselect a reduced number of genes which are ranked according to some scoring scheme. The data matrix $\mathbf{X}$ contains in its columns the GESs observed during M experimental conditions. Out of these, K genes and their corresponding class label were preselected from the training dataset, using different selection methods and scoring schemes, to result in a $(K+1)\times M$-dimensional data matrix whose column vectors served as input to the Lagrangian SVM (LSVM) classifier which thus operated in a $(K+1)$-dimensional subspace of the N-dimensional gene space.

3.1.1 Random gene picking First and simplest, random gene picking was used to preselect K-tuples of genes whose gene expression profiles were used to train the LSVM classifier. If the resulting decision hyperplane correctly classified the training set, a leave-one-out (LOO) cross-validation was applied to estimate the classification error. With the current dataset single marker genes could always be identified which could classify the dataset according to the classification task considered, be it monocyte versus macrophage or healthy versus NPC disease. Thus the algorithm was used with $k=1$ only.

3.1.2 Score-based gene selection Thereby the following scoring criteria are considered:

1. *FCh*: A simple and often used score is the Fold Change (FCh).
2. *w-score*: In Golub *et al.* (1999), an empirical w-score was proposed. This score has been criticized, however, to yield incorrect units in the related discriminant function (Dudoit *et al.*, 2002).
3. *T-score*: In Liu *et al.* (2005) a score based on a likelihood ratio of in class and between class variances was suggested.
4. *c-score*: Galton (1888, 1889) proposed a statistical correlation score nowadays known as Pearson correlation (Pearson, 1901), which can be used to measure the similarity between the n-th GEP $\mathbf{x}_{n*}$ of the data matrix and a design vector $\mathbf{d}$ reflecting the diagnostic knowledge available from the experimental design.
5. *SAM*: The significance analysis of microarrays (SAM) represents a variance stabilized version of a t-test (Tusher *et al.*, 2001).

The gene selection can now proceed by choosing either a fixed number of genes with the highest scores or by defining a threshold and selecting all genes with scores above this threshold. Here a set of 50 genes is selected always.

3.1.3 SVM-based gene selection A gene selection method different from the selection schemes discussed above can be derived from the SVM directly (Barnhill *et al.*, 2002). A SVM estimates an optimal hyperplane, which is characterized by a vector $\mathbf{w}$ normal to the hyperplane, separating the dataset into appropriate subspaces. Gene expression signatures $\mathbf{x}_{*m}$, representing the expression levels of all genes in gene space, can be projected onto these normal vectors via dot products. Hence the components of $\mathbf{w}$ indicate the importance of a gene for the classification task. Genes which have small components only in $\mathbf{w}$ can be removed as their associated unit vectors almost lie parallel to the hyperplane, hence they are orthogonal to the direction of optimal class discrimination, represented by $\mathbf{w}$, and will not contribute to the classification.

Note that all these scores are based on the discriminative power of *single* gene statistics. Thus they ignore the joint discriminative power of a group of genes where each single gene might exhibit a low individual score. This is clearly a deficit of these gene selection schemes.

3.2 Feature generation and selection schemes

MF techniques like ICA or NMF seem promising in generating features suitable for diagnostic classification purposes. A key feature of ICA as well as NMF is the ability to identify patterns that together explain the observed GESs as a linear combination of *expression modes* (ICA) or *metagenes* (NMF), respectively. Hence they overcome the limitations inherent to single gene statistics as discussed earlier. In the following we will discuss these feature generation techniques in their application to microarray datasets. For feature selection we consider a modified Pearson score where we replace the observed GEPs $\mathbf{x}_{n*}$ of the data matrix by corresponding component profiles obtained from the feature generation step by applying matrix decomposition techniques to the data matrix. Note that while the feature generation techniques are completely unsupervised, the subsequent feature selection and classification affords the inclusion of diagnostic knowledge through a corresponding design vector $\mathbf{d}_i$ for class i and renders the methods semi-supervised instead.

3.2.1 Data representation and preprocessing The gene expression levels are represented by an $(N\times M)$ data matrix $\mathbf{X}=[\mathbf{x}_{*1}\cdots\mathbf{x}_{*M}]$ with each column $\mathbf{x}_{*m}$ representing the expression levels of all genes in one of the M experiments or environmental conditions. With NMF, a decomposition is then sought according to $\mathbf{X}=\mathbf{WH}$ which is not unique, of course, and needs further specification. The columns of $\mathbf{W}$ are called *metagenes*. They represent the GESs combined together according to the weights contained in the rows of $\mathbf{H}$ which are called *meta-experiments*. Note that the data matrix is non-square with $N\approx 10^3 M$ which renders a transposition of the data matrix necessary when techniques like ICA are applied. Hence ICA follows the data model: $\mathbf{X}^T=\mathbf{AS}$ where the columns of matrix $\mathbf{A}$ represent the basis vectors of a new representation which is obtained by grouping together the observed GEPs according to the weights contained in the rows of $\mathbf{S}$ which are called *expression modes*. Note that the latter are constrained to be as statistically independent as possible. Accordingly, the columns of matrix $\mathbf{A}$ may be named *feature profiles*. Further, each row $\mathbf{a}_{m*}$ contains the weights to combine the independent *expression modes* to the observed GESs.

The raw expression data have been normalized per chip. As normalization and summarization tool MAS 5.0 has been applied by the medical doctors who owned and produced the datasets. The raw datasets have never been available to us for alternative preprocessing. Hence, more advanced preprocessing tools like RMA (Bolstad *et al.*, 2003; Irrizarry *et al.*, 2003), GCRMA (Wu *et al.*, 2004), FARMS (Hochreiter *et al.*, 2006; Talloen *et al.*, 2007) or DFW (Chen *et al.*, 2007) could not be applied and, in part, have not been available yet at the time the data were produced and processed.

3.2.2 ICA—Analysis The $M\times N$ data matrix $\mathbf{X}^T$ was used as input to the JADE algorithm (Cardoso and Souloumiac, 1993, 1996) which is a nearly exact algebraic algorithm focusing on the 4th order cumulant of the distribution of expression levels within the *expression modes*. It was preferred over stochastic algorithms like the fastICA which yield slightly different results in every run and need several repetitions to identify the robustly estimated *expression modes*. Note that JADE encompasses centering and whitening procedures as preprocessing steps. The number of extracted *expression modes* $K\leq M$ need not correspond to the maximum possible and can be chosen deliberately. The output is a $K\times M$ demixing matrix $\mathbf{B}$, which allows the computation of the corresponding *expression modes*. It is defined through the relation $\mathbf{BX}^T=\mathbf{Y}^T=\mathbf{BAS}=\mathbf{DPS}$ with $\mathbf{D}$, a diagonal scaling matrix and $\mathbf{P}$, a permutation matrix. Typical results show *expression modes* which exhibit both positive and negative expression levels, though the raw expression profiles only have positive expression levels, of course. Note that either the *expression modes* or their related basis vectors can be normalized to unity because of the inherent scaling indeterminacy of ICA.

An ICA-based gene grouping scheme to analyze gene expression profiles was proposed by (Lee and Batzoglou, 2003). A thorough discussion of this rather classical ICA approach, applied to the current dataset, can be found in Lutter *et al.* (2008), where also a discussion of the biological and medical implications of the findings is discussed in greater detail. Hence we abstain here from any such discussion and put the emphasis on the new methodological aspect. In contrast to this classical analysis, here a different approach is taken which utilizes basic properties of the matrix decomposition model and incorporates diagnostic knowledge to evaluate the structure of the *feature profiles*, i.e. the basis vectors of the new representation, and deduce a corresponding *expression mode* to identify a set of marker genes with sufficient discriminative power concerning the classification task at hand.

ICA essentially seeks a new representation of the observed dataset with the columns of matrix $\mathbf{A}$ representing the new basis vectors. Whereas the original dataset $\mathbf{X}$ is interpreted as M data vectors (experiments) in N dimensions (genes), with ICA the transposed data $\mathbf{X}^T$ matrix is considered. Hence, the data are interpreted as N data vectors, each representing the *expression profile* of one of the N genes in an $M\ll N$-dimensional feature space. The latter is spanned by the M-dimensional column vectors of matrix $\mathbf{A}$ which represent the new coordinate system. Hence, individual *gene expression profiles* are mapped onto *feature profiles* and the *component profiles*, i.e. the columns of $\mathbf{S}$, associated with the *expression profiles* by the rules of matrix multiplication contain the weights with which every *feature profile* contributes to each observed *expression profile*. If the matrix decomposition is adequate, a concise analysis of the structure of these *feature profiles* comprising the columns of matrix $\mathbf{A}$ might hopefully provide insights into the structure of the dataset itself.

After the ICA decomposition $\mathbf{X}^T=\mathbf{A}\cdot\mathbf{S}$, the matrix $\mathbf{S}$ contains $K\leq M$ supposedly statistically independent *expression modes* $\mathbf{s}_{m*}\in\mathbf{R}^{1\times N}, 1\leq m\leq (K\leq M)$ forming the rows of $\mathbf{S}$. Now note that to every *gene expression signature* a corresponding row of matrix $\mathbf{A}$ is related which contains the weights with which each of the $K\leq M$ independent *expression modes* contributes to the *gene expression signature* under consideration. Hence each column of $\mathbf{A}$ can be associated with one specific *expression mode* $\mathbf{s}_{m*}$ to which it is related by the rules of matrix multiplication. This observation forms the basis of the proposed feature selection scheme. In a first step an informative *feature profile* reflecting the available diagnostic information of the set of experiments is identified and in a second step the genes of the associated independent *expression mode* are analyzed with respect to their diagnostic classification potential.

Each investigated microarray dataset represents at least two different classes of cells, such as cell lines taken either from healthy subjects (Class 1) or patients suffering from any disease (Class -1). If the *gene expression signatures* $\mathbf{x}_{1*},\ldots,\mathbf{x}_{M*}$ of the different experiments are arranged according to their diagnostic class label, i.e. the j experiments of Class 1 constitute the first j rows of the data matrix $\mathbf{X}^T$, whereas the members of Class -1, i.e. $\{\mathbf{x}_{m*}\}_{m=j+1}^{M}$, are collected in the remaining rows, this assignment is also valid for the rows of matrix $\mathbf{A}$. Suppose one of the independent *expression modes* $\mathbf{s}_{m*}$ is reflecting a putative cellular gene regulation process which is related to the diagnostic difference between the classes. Then to every *gene expression signature* of Class 1, this characteristic *expression mode* should contribute substantially—signalled by a large weight in the corresponding *feature profile*—whereas its contribution to Class -1 experiments should be less (or vice versa). Since the k-th column of $\mathbf{A}$ contains the weights with which the k-th *expression mode* $\mathbf{s}_{k*}$ contributes to all observed *gene expression signatures*, this column should show large/small feature components according to the class labels. Hence, in contrast to the method in Lee and Batzoglou (2003), the clinical diagnosis of the experiments is taken into account. The strategy concentrates on the identification of a column of $\mathbf{A}$, which shows a class specific signature according to a design vector $\mathbf{d}$ containing the class labels of each experiment. The *expression mode* related to that specific column is assumed to provide a good candidate for further class specific analysis concerning the identification of marker genes. Informative columns were identified using the correlation coefficient $|\text{corr}(\mathbf{a}_{*n},\mathbf{d})|$ of each column vector of $\mathbf{A}$ with a design vector $\mathbf{d}$ whose m-th entry is $d_m=\pm 1$, according to the class label of experiment m.

3.2.3 Local NMF—analysis NMF replaces the assumption of statistically independent expression modes by a positivity constraint concerning the entries of the matrices into which the given data matrix of observed expression levels is to be decomposed. As the experimentally observed data correspond to fluorescence intensity levels, such positivity constraints seem more natural than independent *expression modes* which contain numerous negative entries. Such negative entries appear as well in the matrix of corresponding mixing coefficients meaning that different *expression modes* may partially compensate their respective contributions to the observed expression levels. Hence the constrained NMF model has the potential to identify sets of functionally related genes more accurately. Applying NMF, the data matrix corresponds to the usual $N\times M$ matrix $\mathbf{X}=[\mathbf{x}_{*1}\cdots\mathbf{x}_{*M}]$. Each column of $\mathbf{X}$, called a gene expression signature, comprises the gene expression levels of all genes resulting from one experiment. After applying the LNMF algorithm (Li *et al.*, 2001), the data matrix is decomposed into two new matrices $\mathbf{W}$ and $\mathbf{H}$. The columns of $\mathbf{W} = [\mathbf{w}_{*1} \cdots \mathbf{w}_{*k}]$ are called *metagenes*. One of them is expected to be characteristic of a regulatory process, which is responsible for the class specific difference in the observed experiments. Its contribution to the observed *gene expression signatures* is contained in the rows of matrix $\mathbf{H}$ which are called *meta-experiments*. Once an informative *meta-experiment* is identified through its correlation to the design vector encompassing the diagnostic information available, further analysis can be focused only on the genes contained in the corresponding *metagene*.

The search strategy is similar to the one used in case of the ICA analysis discussed earlier. As before all experiments were classified according to available diagnostic information and labeled accordingly. The correlation coefficients between every *meta-experiment* $\mathbf{h}_{m*}$ and $\mathbf{d}$ are then computed. Empirically, a correlation coefficient $|\mathrm{corr}(\mathbf{h}_{m*}, \mathbf{d})| > 0.9$ signifies a sufficient similarity between $\mathbf{h}_{m*}$ and $\mathbf{d}$. Besides the maximum number of iterations which controls the precision of the decomposition, the number of extracted basis components K, i.e. the *metagenes*, is the only adjustable parameter affecting the structure of $\mathbf{W}$ and $\mathbf{H}$. For several decompositions $\mathbf{X} = \mathbf{WH}$ using different numbers K of *metagenes* $[\mathbf{w}_{*1} \cdots \mathbf{w}_{*K}]$, the matrices $\mathbf{H}$ are studied with respect to the appearance of correlation coefficients $|\mathrm{corr}(\mathbf{h}_{m*}, \mathbf{d}|)$ close to 1. A *metagene* is considered informative only if all entries of the corresponding *meta-experiment* which belong to Class 1 are smaller than all other entries of that *meta-experiment* (or vice versa). After a total number of 5000 iterations, the cost function of the LNMF algorithm did not show noticeable changes with any of the datasets investigated. For $K = 2, \ldots, 49$, ten separate simulations were carried out each time and only the simulation showing the smallest reconstruction error was retained. Further matrix decompositions with $K = 50, 60, \ldots, 400$ *metagenes* were examined. In the latter case, only three simulations were performed for each K. Note that NMF traditionally chooses $K \ll M$ to go for a more compact representation of the data matrix. However, deliberately choosing $K \gg M$ opens the way to a sparse and hopefully more informative and straightforwardly interpretable representation of the dataset with co-regulated genes combined into one *metagene*.

3.3 Classifier

3.3.1 SVM classifier SVM (Schölkopf and Smola, 2002) are appropriate tools whenever data classification is the goal. They are based on geometric considerations in a vector space. Given an optimal separating hyperplane, characterized through its vector $\mathbf{w}_{\mathrm{opt}}$ normal to the hyperplane, the corresponding decision rule reads

$$f(\mathbf{x}) = sgn\left(\langle \mathbf{x}, \mathbf{w}_{\mathrm{opt}} \rangle + b\right) \tag{1}$$

where $\mathbf{w}_{\mathrm{opt}} = \sum_{m \in SV} y_m \alpha_m \mathbf{x}^{SV}_{*m}$ and y_m represents the class label, α_m represents a hyperparameter and $\mathbf{x}^{SV}_{*m}$ indicates the support vectors closest to the separating hyperplane. The solution to this quadratic optimization problem is implemented in the LSVM algorithm (Mangasarian and Musicant, 2001), a MATLAB version of which is available at www.cs.wisc.edu/~musicant/lsvm/ and has been used in this study. For cross-validation to evaluate the performance of the LSVM classifier, a LOO procedure was applied. Note that in every case the test sample has been taken out before the classifier was trained to avoid any bias in the decision making (Simon, 2003).

3.3.2 MF classifier MF of the data matrix has been considered a feature generation technique sofar. The resulting M-dim *feature profiles* (ICA: columns of $\mathbf{A}$) or the M-dim *meta-experiments* (NMF: rows of $\mathbf{H}$) have then been used for feature selection purposes by correlating them with a design vector $\mathbf{d}$ the components of which represent class labels encoding the diagnostic information available about the respective experiments or environmental conditions. But the known class labels of the GESs (ICA: $\mathbf{x}^T_{m*}$ or NMF: $\mathbf{x}_{*m}$) also translate to corresponding labels for the rows $\mathbf{a}_{k*}$ in case of ICA or the columns $\mathbf{h}_{*k}$ in case of NMF. Similarities between these row or column vectors, respectively, can thus be used to classify the observations directly without having to identify appropriate sets of marker genes. The method will be explained in the following referring to the NMF notation but can be translated to the ICA notation immediately by recognizing the correspondence of *metagene—expression mode* and *feature profile—meta-experiment*, i.e. $\mathbf{W} \triangleq \mathbf{S}^T$ and $\mathbf{H} \triangleq \mathbf{A}^T$, respectively.

Note that from $\mathbf{X} = \mathbf{WH}$ it follows that $\mathbf{W}^{\#} \cdot \mathbf{x}_{*m} = \mathbf{h}_{*m}$ where $\mathbf{W}^{\#}$ denotes the pseudo-inverse of $\mathbf{W}$. Now the similarities between the columns of $\mathbf{H}$ can be used to classify the observations. Though any method based on similarity measures can be used, we simply estimate the correlation coefficient, i.e. $c_m \equiv corr(\mathbf{h}^{test}, \mathbf{h}_{*m})$. Now for each class a separate index set I_i of indices is created, where I_1 encompasses all indices m for which $\mathbf{x}_{*m} \in$ Class 1 while I_{-1} contains all remaining indices. This thus results in two sets of correlation coefficients corresponding to the two assignments $m \in I_1$ or $m \in I_{-1}$. Two rules for class assignment were tested:

- Average correlation:

$$\mathrm{label}(\mathbf{h}^{test}_{*m}) = \begin{cases} 1 & \text{if } \langle c_m(1) \rangle > \langle c_m(-1) \rangle \\ -1 & \text{otherwise} \end{cases} \tag{2}$$

  where $\langle .... \rangle$ denotes an average of the correlation coefficients over the respective index set.

- Maximal correlation:

$$\mathrm{label}(\mathbf{h}^{test}_{*m}) = \begin{cases} 1 & \text{if } \max\{c_m(1)\} > \max\{c_m(-1)\} \\ -1 & \text{otherwise} \end{cases} \tag{3}$$

  where $\max\{c_m(\pm 1)\}$ denotes the maximal value of all correlation coefficients within either the set I_1 or I_{-1}.

Rule 1 thus assigns the class label according to an average correlation of the test vector with all vectors belonging to one or the other index set. Rule 2 assigns the class label according to the maximal correlation occurring between the test vector and the members of each index set.

3.3.3 Random forest classifier Last but not least a random forest (RF) classifier (Breiman, 2001; Diaz-Uriarte, 2007; Diaz-Uriarte and de Andrés, 2006) was applied to a set of 50 genes which were selected by either a) the highest scores, b) the highest expression values in the most informative *expression mode* in case of ICA feature selection or c) the highest expression levels in the most informative *metagene* in case of NMF feature selection. RF is a classification algorithm which uses an ensemble of classification trees. Each tree is built using a bootstrap sample of the data, and at each node of the decision tree the candidate set of variables is a random subset of the variables. Hence RF uses both bagging and random variable selection which results in largely uncorrelated decision trees. RF shows improved accuracy in comparison to other supervised learning methods. Apart from this it provides a stability measure of the list of genes selected according to some well-defined measure of variable importance. This is a definite advantage over other cross-validation schemes, hence it was applied to all features (= gene lists) generated with other techniques in this contribution.

4 RESULTS AND DISCUSSION

The following results discuss the application of the various gene or feature selection schemes and classifiers to the microarray datasets. The goal was to identify an as small as possible set of marker genes which allow for a diagnostic classification of the given microarray experiments devoted to the investigation of the related disease. These marker genes would allow the design of special purpose chips which could provide a less costly alternative to genome-wide microarray diagnostics.

The cases to be distinguished by the classifier in the MoMa datasets are the following:

- Case 1: monocyte versus macrophage (MoMa)
- Case 2: healthy versus Niemann Pick C disease (HeDi)

Table 1. The classification task was to classify monocytes versus macrophages (Case 1:MoMa)

No.	SGP/SVM-LOO	ICA-GEL-neg	ICA-RF-neg	ICA-GEL-pos	ICA-RF-pos	NMF-GEL	NMF-RF
1	PABPC4	*NFKBIA*	H3F3A	**GPNMB**	CD59	**GPNMB**	CPVL
2	*SNURF/SNRP*	S100A9	PPP1R15A	MMP9	**CTSB**	HLA-DRB1	CST3
3	*PEBP1*	*IL8*	HSPA5	**CTSB**	**ADFP**	CD74	**ADFP**
4	ITGAL	FCN1	H3F3A	FUCA1	**LIPA**	***SAT***	**LIPA**
5	ZNF331	S100A8	RGS2	**LIPA**	CTSL	*PSAP*	MS4A6A
6	CRYL1	CSPG2	CYP1B1	CD63	**K-ALPHA-1**	HLA-DRB1	HLA-DPA1
7		TNFAIP3	CD83	LAMP1	***GM2A***	HLA-DRB1	C12orf8
8		*DUSP1*	CYBB	TFRC	HEXB	*GRN*	***GM2A***
9		PRG1	HNRPH1	CSTB	PPIA	***GM2A***	CECR1
10		FPR1	CYP1B1	**K-ALPHA-1**	***SAT***	*GRN*	BASP1

SGP/SVM-LOO: genes of the dataset which lead to an error rate $\epsilon(\text{LOO}) \leq 2$ when genes were selected by either single gene picking or a SVM and LOO cross-validation was invoked. The genes are ranked according to their corresponding c-scores. ICA-GEL-neg/pos: the 10 most strongly expressed genes of *expression submodes* $\mathbf{s}_{6*}^{neg/pos}$, respectively, when a total of $k=8$ *expression modes* were extracted by ICA. NMF-GEL: the 10 most highly expressed genes in the most informative *metagene* $\mathbf{w}_{*28}$ selected by LNMF. NMF-RF: the most informative genes of *metagene* $\mathbf{w}_{*28}$ according to the RF classification of the 50 most highly expressed genes of the most informative *metagene* $\mathbf{w}_{*28}$ selected by NMF.

4.1 Gene selection and classification

4.1.1 Random gene picking and classification The simplest possibility tested was the identification of randomly picked single genes which were able to classify the datasets into the given classes according to both classification cases using the LSVM classifier and LOO cross-validation. Hence 14 LOO test have been performed. Results obtained with the MoMa dataset concerning the Case 1 classification are listed in Table 1, first column (SGP/SVM-LOO). Only genes with a classification error rate $\epsilon(LOO) < 2$ are listed. Concerning Case 2 classification, a total of 531 genes resulted from gene picking or SVM selection which showed a classification error rate $\epsilon(LOO) = 0$ when LOO cross-validation was applied. These genes were ranked according to their c-score and the 10 genes with the highest c-scores (see later) are listed in Table 2, first column (SGP/SVM-LOO). Please, note that different genes might have the same Affy-id, hence the same gene name (Loci-id) might appear multiply in the following lists. Entries in bold face in these tables represent genes which were also selected as marker genes with other methods, hence appear in several columns of the table with the same type of classification (Case 1 or Case 2). Entries in italics in these tables represent marker genes which also appear in lists related with the other type of classification (Case 1 and Case 2), hence appear in both tables.

4.1.2 SVM-based gene selection and classification The recursive gene elimination procedure discussed earlier has been tested with the MoMa dataset. As the LSVM algorithm is not able to handle datasets with many thousand genes, each dataset has been partitioned into subsets of roughly 1000 genes each. For each subset an optimal separating hyperplane has been estimated using LSVM. Two strategies were followed to filter out a set of diagnostic marker genes:

1. Remove the gene with the smallest squared contribution to **w** and run LSVM on the reduced dataset. When no solution can be obtained with the reduced set, then stop. Merge the 100 most important genes of each set and repeat the procedure.
2. Remove the 100 least important genes with the smallest components in **w** as long as more than 200 genes are in the set. For smaller sets remove genes step by step.

In each case the algorithm stopped at one of the genes listed in Tables 1 or 2 corroborating the findings already achieved with single gene picking and LOO cross-validation. The comparison of the scores of these genes with respect to the different scoring schemes show that the genes selected with either the single gene picking or the SVM-based selection scheme rank quite differently in the different scoring schemes. Choosing the 50 genes with the highest score of every scoring list, most of the genes in Table 1 would not have been selected at all by simple score ranking. The best correspondence is achieved with a ranking according to the c-score/SAM-score. Note that the SAM-score ranking is identical to the c-score ranking. Hence, we only will consider this latter scoring scheme further on.

4.1.3 Score-based gene selection and RF classification The score-based gene selection methods discussed earlier yield quite different collections of genes. The various scores of all genes of the dataset were calculated and the genes were sorted in descending order according to these scores. The 50 most highly ranked genes according to the different scores were then selected for further evaluation using a RF classifier. In the following we concentrate only on the more interesting Case 2 (HeDi) classification. Furthermore we only present results in case of the c-score which turned out to be the most reliable and, furthermore, yielded an identical ranking as the SAM - score. Table 2 summarizes the 19 genes with the highest importance according to the mean decrease accuracy and mean decrease Gini index (Breiman, 2001) estimated by the RF classifier (Fig. 1).

4.2 Feature generation, selection and classification

In the following we apply the matrix decomposition techniques discussed earlier for feature generation and selection which incorporates diagnostic knowledge available about the experiments. Sets of marker genes will result, which were subjected to a RF classifier to identify the most relevant genes for the classification task at hand.

4.2.1 Analysis of feature profiles and expression modes generated by ICA Using a decomposition into $K = M = 14$ independent

Table 2. The classification task was to classify healthy versus Nieman Pick C disease (Case 2: HeDi)

No.	SGP/SVM-LOO	C-Sc-RF	ICA-GEL-neg	ICA-GEL-pos	ICA-RF-neg	ICA-RF-pos	NMF-GEL	NMF-RF
1	RPS6KA4	*PEBP1*	**OAZ1**	3GM2A	RPL11	**SOD2**	**SAP18**	RY1
2	C14orf131	CD163	**C6orf62**	STAB1	**RPL7**	ACTB	**PNN**	RAD23B
3	**SAP18**	DIP	**ARPC2**	*GM2A*	ACTR2	NOTCH2NL	**RAB1A**	UBE2L3
4	COX7c	unknown	**RPL7**	HLA-A	**RAB1A**	***GRN***	**SUMO1**	MIF
5	**CHMP5**	***GRN***	**S100A4**	**SOD2**	**ALOX5AP**	*DUSP1*	**RAB31**	NUDT21
6	**SUMO1**	MMD	ITM2B	*NFKBIA*	HSPA8	HLA-C	NEDD8	PCMT1
7	**PNN**	STX7	ARHGDIB	*IL8*	SERPINA1	***GRN***	RPS25	**SAP18**
8	**OAZ**1	MPST	TMSB4X	HLA-B	**S100A4**	AHNAK	ATP6Y1C1	SNX3
9	TDE2	***GRN***	**ALOX5AP**	*SAT*	H3F3A	HLA-F	**CHMP5**	CASP1
10	**RAB1A**	PRKACB	HLA-DRA	PSAP	**ARPC2**	**CTSD**	TSPYL1	**RAB31**
11		GUSB			**C6orf62**			
12		GALC			HNRPA1			
13		NPC2			S100A10			
14		SPTBN1			**RAB31**			
15		RRAGD						
16		C5orf13						
17		*SNRP/SNURF*						
18		GRB10						
19		**CTSD**						

SGP/SVM-LOO: the first 10 marker genes out of a list of 531 genes of the dataset which lead to an error rate $\epsilon(\mathrm{LOO}) = 0$ when genes were selected by either single gene picking or a SVM and LOO cross-validation was invoked. The genes are ranked according to their corresponding c-scores. The first two genes had a positive c-score, all others a negative c-score. C-Sc-RF: list of genes of the dataset which, after ranking by their c-score, were selected as most informative by a RF classifier. Only the 50 highest ranked genes were input to the RF-classifier. ICA-GEL: 10 most strongly expressed genes of *expression mode* $\mathbf{s}_{3*}$ related with case 2 when a total of $k=8$ *expression modes* were extracted by ICA. ICA-RF: the most informative genes, according to a RF classifier, of the submodes $(\mathbf{s}_{3*}^{neg}, \mathbf{s}_{3*}^{pos})$ of *expression mode* $\mathbf{s}_{3*}$ related with case 2 when a total of $k=8$ *expression modes* were extracted by ICA. Only the 50 most highly expressed genes were input to the RF-classifier. NMF-GEL: the 10 most highly expressed genes in the most informative *metagene* $\mathbf{w}_{*13}$ selected by LNMF. NMF-RF: the most informative genes of *metagene* $\mathbf{w}_{*13}$ according to the RF classification of the 50 most highly expressed genes of the most informative *metagene* $\mathbf{w}_{*13}$ selected by NMF.

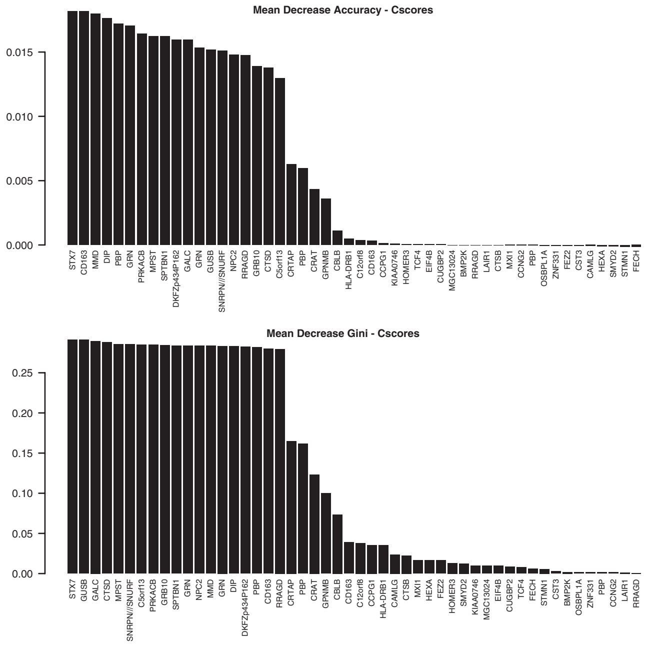

Fig. 1. Mean decrease accuracy and Gini index determined with a RF classifier of the 50 genes of the dataset with the highest c-score. Note that both indices label the same 19 genes as most informative though their ordering is slightly different.

expression modes, column $\mathbf{a}_{*7}$ of the resulting mixing matrix $\mathbf{A}$ showed a moderately strong correlation $|\mathrm{corr}(\mathbf{a}_{*7}, \mathbf{d}_{*1})| = 0.7$ with the design vector $\mathbf{d}_{*1} = (d_{1,1} = -1, \ldots, d_{7,1} = -1, d_{8,1} = 1, \ldots, d_{14,1} = 1)$ to discriminate GESs taken from monocytes from those taken from macrophages. Column $\mathbf{a}_{*1}$ showed a correlation coefficient $|\mathrm{corr}(\mathbf{a}_{*1}, \mathbf{d}_{*2})| = 0.95$ with design vector $\mathbf{d}_{*2} = (d_{1,2} = 1, \ldots, d_{4,2} = 1, d_{5,2} = -1, \ldots, d_{7,2} = -1, d_{8,2} = 1, \ldots d_{11,2} = 1, d_{12,2}, \ldots, d_{14,2})$. The signature of Column 7 is not very clear cut. Hence a systematic investigation of the structure of the mixing matrices was carried out while increasing the extracted number of *expression modes* from $K = 2, \ldots, 14$. The necessary dimension reduction step can be done during the whitening step of the JADE algorithm. The information loss is not critical in any case as the first three principal components cover 96.1% of the variance. Note that such an ordering principle does not hold in case of ICA. It is thus unclear whether such a dimension reduction removes informative components which posses large higher order correlations but accidentally only small second order correlations. It is because of this uncertainty that we performed a systematic investigation of the decomposition up to the full rank of the data matrix. The resulting maximal correlation coefficients $|\mathrm{corr}(\mathbf{a}_{*k}, \mathbf{d}_{*i})|$ showed little variation in both cases with average values $\langle|\mathrm{corr}(\mathbf{a}_{*k}, \mathbf{d}_{*1})|\rangle_K = 0.79$ and $\langle|\mathrm{corr}(\mathbf{a}_{*k}, \mathbf{d}_{*2})|\rangle_K = 0.94$. Shallow maxima occur at $k = 3$ in Case 1 and at $k = 8$ in Case 2. Figures 2 and 3 present the *feature profiles* $\mathbf{a}_{*6}$ and $\mathbf{a}_{*3}$ maximally correlated with design vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ and the related *expression modes* $\mathbf{s}_{6*}$ and $\mathbf{s}_{3*}$. A list of the 10 most strongly expressed genes in each of the extracted *expression modes* is given in Tables 1 and 2.

Table 3. List of marker genes of the MoMa dataset, Case 1, according to Table 1, first column and their corresponding rank according to the various scores listed above

Gene symb.	c	w	T^+	T^-	SAM	FCh
MoMa dataset: Case 1 classification						
PABPC4	>50	>50	>50	>50	>50	>50
SNURF	01	49	>50	>50	01	>50
PEBP1	09	>50	23	>50	09	09
PEBP1	02	>50	>50	>50	02	24
ITGAL	>50	>50	>50	>50	>50	>50
ZNF331	45	22	>50	>50	45	>50
CRYL1	>50	>50	>50	>50	>50	>50

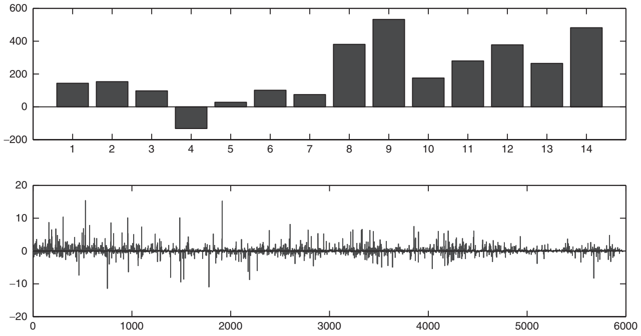

Fig. 2. ICA(JADE): *feature profile* $\mathbf{a}_{*6}$ for $k=8$ and the related *expression mode* $\mathbf{s}_{6*}$. *Feature profile* $\mathbf{a}_{*6}$ shows a strong correlation with the design vector $\mathbf{d}_1$ in Case 1 (monocyte versus macrophage).

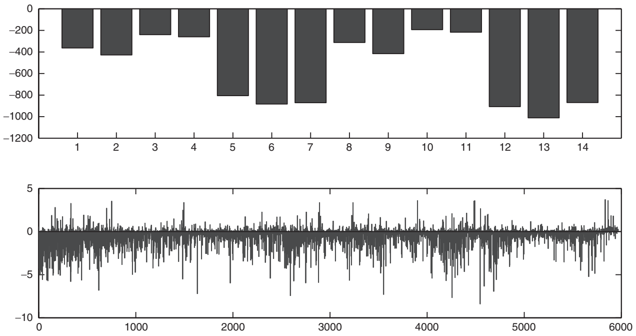

Fig. 3. ICA(JADE): *feature profile* $\mathbf{a}_{*3}$ for $k=8$ and the related *expression mode* $\mathbf{s}_{3*}$. The *feature profile* $\mathbf{a}_{*3}$ shows a strong correlation with the design vector $\mathbf{d}_2$ in Case 2 (healthy versus diseased).

These marker genes are involved in a gene regulation network and all genes in this network can be associated with the MeSH term *gene expression regulation*, except FUCA1 and STAB1. For a thorough discussion of the related pathways identified by applying BiblioSphere MeSH- and GeneOntology filter tools see (Lutter *et al.*, 2008).

Having selected the most informative *expression modes* $\mathbf{s}_{6*}$ (Case 1) and $\mathbf{s}_{3*}$ (Case 2), their 50 most strongly expressed genes have been selected from every submode and supplied to a RF classifier. In Table 1, the columns labeled ICA-GEL-neg/pos, ICA-RF-neg/pos contain either the 10 genes with the highest expression level (GEL) in the most informative *expression mode* or the most informative genes selected by a RF classifier. The columns labeled NMF-GEL/RF contain the corresponding genes from the most informative *metagene* selected with LNMF feature selection. The most highly expressed 50 genes of the *expression submodes* or of the corresponding *metagene* have been subjected to a RF classification. The importance of the variable decreases exponentially in all three examples, hence only the first 10 genes are listed as their importance was then already down to one- third of the maximal value. The corresponding results obtained in the more interesting classification Case 2 are listed in Table 2. According to the mean decrease accuracy estimated, 14 and 10 genes stand out as most informative as illustrated in Figure 4. They are listed in columns labeled ICA-RF-neg/pos and NMF-RF in Table 2, respectively.

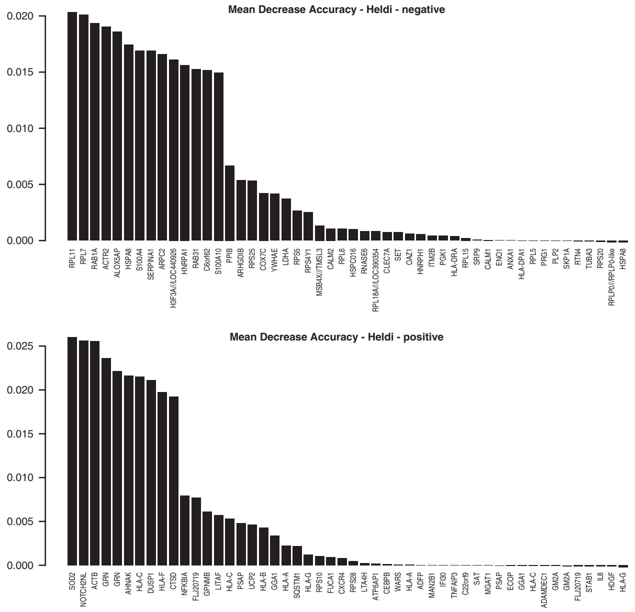

Fig. 4. Mean decrease accuracy according to a RF classification of the 50 most highly expressed genes in *expression submode* $\mathbf{s}_{3*}^{neg}$ and $\mathbf{s}_{3*}^{pos}$ selected by ICA with $k=8$ extracted *expression modes* in total. Only the 14 and 10 genes, respectively, with the highest mean decrease accuracy are considered informative.

4.2.2 Analysis of meta-experiments and metagenes generated by LNMF A LNMF analysis was also performed on the 14 experiments of the dataset. Again the number k of extracted *metagenes* was varied systematically to identify an optimal decomposition of the $N\times M$ data matrix $\mathbf{X}$. For every k, the correlation coefficients between the *meta-experiments* and the design vectors $\mathbf{d}_{i*}, i=1,2$ were computed. Again a RF classifier was used to select the most informative genes from the most informative *metagene*.

Monocyte versus macrophage: For $k>100$, several *meta-experiments* show small expression levels for all monocyte experiments compared with larger expression levels for the macrophage experiments, indicated by a large correlation coefficient

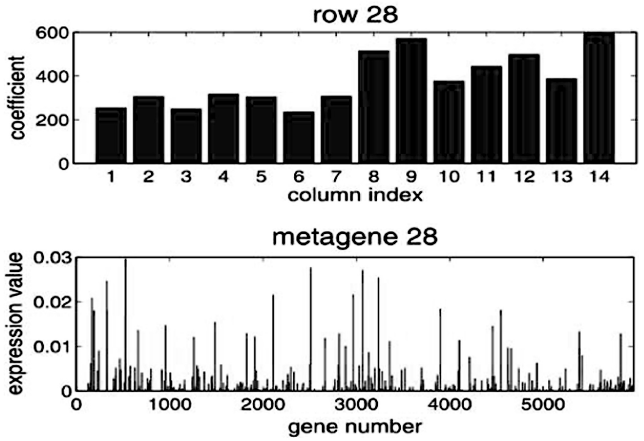

Fig. 5. Signature of *meta-experiment* $\mathbf{h}_{28*}$ of $\mathbf{H}_{k=29}$ and corresponding GES of *metagene* $\mathbf{w}_{*28}$ of $\mathbf{W}_{k=29}$.

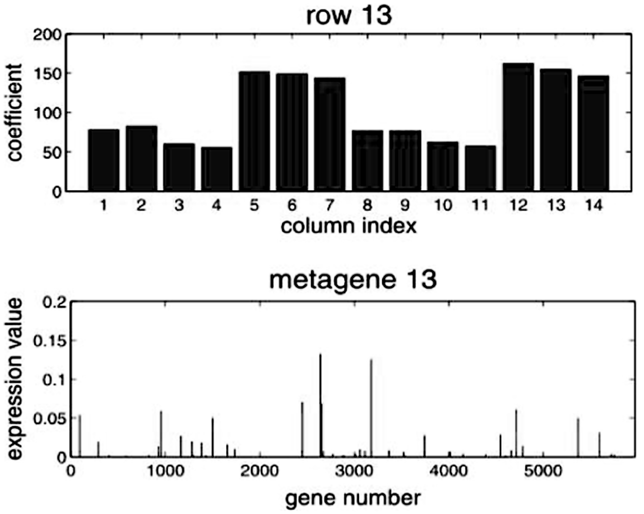

Fig. 6. Signature of *meta-experiment* $\mathbf{h}_{13*}$ and corresponding *metagene* $\mathbf{w}_{*13}$.

of $|\mathrm{corr}(\mathbf{h}_{k*}, \mathbf{d}_{i,*})| > 0.9$. Up to $K = 90$ mostly one significant meta-experiment was observed, for $K > 90$, except $K = 120, 170$ and 190, at least two significant meta-experiments were detected. Rows of $\mathbf{H}$ related to the reverse case $el(macrophage) < el(monocyte)$ do not appear at a comparable level of correlation to the design vector. Figure 5 exhibits the signature of row $\mathbf{h}_{28*}$ of $\mathbf{H}_{k=29}$ and the related *metagene*. The 50 most highly expressed genes in *metagene* $\mathbf{w}_{*28}$ have also been subjected to a RF classification to obtain a measure of importance of these genes for the decision in question (Case 1 MoMa). Only 4 genes stand out as informative (Fig. 7) but the mean decrease accuracy decreases exponentially, hence only the 10 most important genes are listed in Table 1 together with the 10 most highly expressed genes in *metagene* $\mathbf{w}_{*28}$.

Healthy versus diseased: In this case, the number of *meta-experiments* with a strong correlation with the design vector reflecting overexpressed genes in case of cell lines taken from Niemann Pick C patients increases nearly linearly with increasing k.

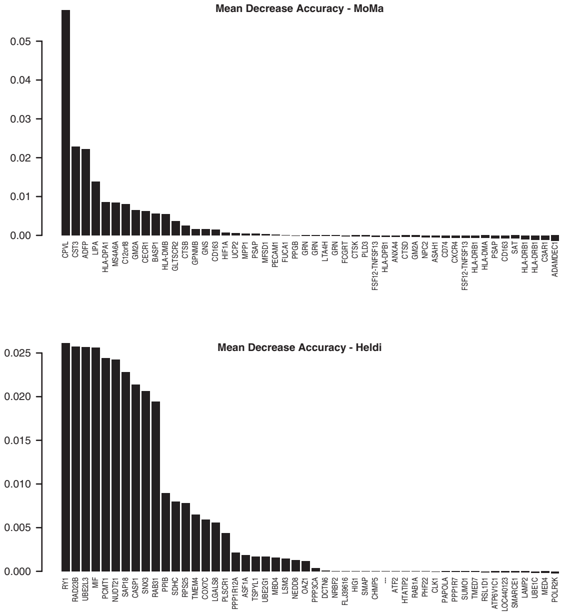

Fig. 7. Mean decrease accuracy according to a RF classification of the 50 most highly expressed genes in *metagene* $\mathbf{w}_{*28}$ and *metagene* $\mathbf{w}_{*13}$ selected by LNMF.

In case of underexpressed *metagenes* related to the disease, only a few significant *meta-experiments* appear for $K > 60$. As an example, a decomposition in $K = 370$ *metagenes* is considered. *Meta-experiment* $\mathbf{h}_{13*}$ yields $|\mathrm{corr}(\mathbf{h}_{13*}, \mathbf{d}_{2*})| = -0.98$ with respect to the separation between the classes 'healthy' and 'diseased' (Fig. 6). The 10 most strongly expressed genes in *metagene* $\mathbf{w}_{*13}$ which qualify as marker genes for the discrimination between healthy subjects and Niemann Pick C patients are listed in Table 2 in column labeled NMF-GEL. Again the 50 most highly expressed genes in *metagene* $\mathbf{w}_{*13}$ have been subjected to a RF classification, where 10 marker genes, listed in Table 2 in column labeled NMF-RF, stand out as most informative for Case 2 (HeDi) classification (Fig. 7).

4.3 MF classifier

Applying the MF classifier described eralier, the similarity between either the rows of matrix $\mathbf{A}$ (ICA) or the columns of matrix $\mathbf{H}$ (NMF) was studied using the dataset with LOO cross-validation. Note that the matrix decomposition step used the complete data matrix, though. As can be seen from Table 4, in most cases one false classification occurred in Case 1 classification leading to the suspicion of a falsely classified data sample which could be identified as Experiment 4. Compared to ICA-based feature selection and matrix decomposition, the corresponding LNMF-based feature selection and matrix decomposition lead to a more robust classification of all four diagnostic classes underlying the dataset with respect to a variation of the number k of extracted features. However, with features encompassing more than seven genes a close to perfect classification with a close to zero classification error resulted.

4.3.1 Comparison of expression modes and metagenes Though both the ICA and the LNMF algorithm lead to data matrix

Table 4. Number of false classifications of the MF classifier using ICA (left) or NMF (right) for matrix decomposition and LOO cross-validation

K	JADE MoMa		LNMF MoMa		JADE HeDi		LNMF VHeDi	
	avg	max	avg	max	avg	max	avg	max
2	11	11	4	9	11	12	3	5
3	3	5	1	1	5	5	0	0
4	3	4	1	1	5	3	0	0
5	2	2	1	1	4	2	0	0
6	3	2	1	1	4	1	0	0
7	1	1	1	1	1	0	0	0
8	1	1	1	1	0	0	0	0
9	1	2	1	1	0	0	0	0
10	2	2	2	1	0	0	0	0
11	1	2	1	1	0	0	0	0
12	1	2	1	2	0	0	0	0
13	1	2	2	2	0	0	0	0
14	1	2	2	2	0	0	0	0

Class labels: Mo - Monocyte, Ma - Macrophage, He - healthy, Di - Niemann Pick C disease.

decompositions which finally lead to robust and efficient diagnostic classification of the datasets, they nonetheless resulted with groups of strongly over- or under-expressed genes in the related metagenes and expression profiles which showed only partial overlap. It is interesting to compare the distribution of correlation coefficients of the individually observed expression profiles of the identified marker genes for both algorithms. It turned out that the LNMF algorithm results in a much narrower distribution of c-values meaning that it yields a much more consistent set of diagnostic marker genes when measured by the correlation to the diagnostic design vector.

5 SUMMARY AND CONCLUSION

5.1 Summary

In this study microarray datasets from a monocyte–macrophage differentiation study have been analyzed. Unsupervised feature generation strategies based on MF techniques were combined with knowledge-based feature selection and classification strategies referring to diagnostic information available about the class labels of the experiments. The importance of the generated features, i.e. either the *expression modes* or the *metagenes*, was assessed by measuring the correlation of the related *feature profiles* or *meta-experiments* with the diagnostic design vector of the experiment under consideration. Two strategies were then followed to achieve classification: either an MF classifier was applied directly or a set of marker genes was extracted from the most informative *expression mode* or *metagene* and classification was achieved applying an SVM classifier with LOO cross-validation. Marker genes were extracted based either on their expression level in the most informative feature vector or by applying an RF classifier to evaluate their level of importance for the classification decision at hand.

5.2 Conclusion

Concerning the identified marker genes, a remarkable number of marker genes could be corroborated by at least one other method, mostly of the MF venue. These genes are indicated in bold face in Tables 1 and 2. It is interesting to see that in case of ICA feature generation most of the genes with the highest expression level in the most informative *expression submode*, listed under ICA-GEL-neg, are corroborated by the RF classifier as being especially important for the classification decisions at hand. This lends credit to the ICA—feature generation and—selection procedure which indeed results in an informative set of marker genes. But in case of the NMF-GEL/RF methodology only little correspondence between both lists can be seen. Hence in the NMF methodology strong expression levels cannot be considered a valid selection criterion to identify marker genes though these genes belong to the feature vector which was identified as most relevant for the classification decision at hand according to the diagnostic knowledge available.

Also, there is still considerable divergence between the results obtained with the ICA and the NMF methodologies for both classification cases considered. One of the reasons might be the different constraints active in both feature generation methods (ICA versus NMF) investigated. The constraints operative with ICA assume statistical independence of the generated *expression modes*, thus only features which comply to this property can be detected. On the other hand, the sparseness and positivity constraints of the LNMF algorithm lead to *metagenes* which contain localized features (= few highly expressed values and many zeros) and only positive expression levels. The gene expression signatures of the underlying biological regulatory processes are expected to have non-Gaussian distributions, hence uncorrelated *metagenes* and independent *expression modes* are expected to be different. Further selection of marker genes solely based on their expression level in the feature vectors is not a statistically valid criterion, hence results might not be expected to conform to marker genes selected by the statistically well-founded RF method or the supervised techniques discussed as well. Still Tables 1 and 2 show a couple of genes which have been identified consistently with different methods for the classification tasks at hand. These genes can be considered well-justified marker genes.

Strong overlap is also observed between the lists resulting from the SGP/SVM-LOO and the NMF-GEL methodology in Case 2 classification meaning that each of these strongly expressed genes of the most informative *metagene* is able to also individually classify the cell line investigated. Note that only two of these genes, i.e. SAP1 and RAB1A, are also identified by the ICA-GEL/RF methodology, hence are identified by three of the four methodologies investigated. In Case 1 classification, no correspondence between the standard methodologies and the MF-based techniques can be observed, however. Further note that no single gene could be identified with these standard techniques which could classify monocytes versus macrophages perfectly. These observations might indicate that the joint discriminative power of a set of genes is necessary to successfully classify Case 1. If so then none of the gene which perform best individually belongs to this set.

All matrix decompositions strategies lead to classification results of the MF classifier with classification errors comparable to those obtained with supervised techniques. Since the quality of the results did not change when the number *K* of extracted basis vectors was increased, the classification performance did not favor one of the tested LNMF runs. Despite that one monocyte sample must definitely be considered to be an outlier, i.e. this probe set is expected to be falsely classified in the original dataset. No single tested decomposition, neither by JADE nor by LNMF, classified these samples correctly.

Concerning more standard supervised techniques, marker genes were selected by their ability to individually classify the cell lines under study applying an SVM classifier with LOO cross-validation. Preselection strategies based on various single gene scoring schemes were considered as well. It seems that the Pearson correlation score achieved the most useful results in what concerns reproducibility of the extracted marker genes by other methods. The resulting list of marker genes was for the HeDi classification problem also compared to an importance ranking resulting from the application of an RF classifier to the 50 most highly ranked genes according to all scores employed. No overlap is observed between both lists. However, 6 out of 10 genes of the SGP/SVM-LOO list could be confirmed by other methods but only one gene of the C-Sc-RF list is corroborated also by the ICA-RF-pos methodology.

Though the classification errors are comparable, the matrix decomposition-based approach has the advantage of not having to deal with each gene in isolation as is the case with the supervised techniques considered. The matrix decomposition technique instead provides a feature selection and classification tool which uses diagnostic knowledge available and the joint discriminative power of a group of most informative genes. If it only were to classify the type of cell lines investigated, the MF classifier does well and identification of marker genes is not necessary.

ACKNOWLEDGEMENTS

Funding: This work is supported by grants from Siemens AG, Munich, the DFG (Graduate College 638) and the DAAD (PPP Luso - Alem$\tilde{a}$ and PPP Hispano - Alemanas) which is gratefully acknowledged.

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2002) *Affymetrix Microarray Suite User Guide*. Affymetrix Santa Clara, CA.

Allison,D. *et al*. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.

Baldi,P. and Hatfield,W. (2002) *DNA Microarrays and Gene Expression*. Cambridge University Press.

Barnhill,S. *et al*. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**, 185–193.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Cardoso,J.-F. and Souloumiac,A. (1993) Blind beamformimg for non-gaussian signals. *IEEE Proc.*, **F140**, 362–370.

Cardoso,J.-F. and Souloumiac,A. (1996) Jacobi angles for simultaneous diagonalization. *SIAM J. Math. Anal. Appl.*, **17**, 161–164.

Chen,Z. *et al*. (2007) A distribution free summarization method for affymetrix genechip arrays. *Bioinformatics*, **23**, 321–327.

Diaz-Uriarte,R. (2007) Genesrf and varselrf: a web-based tool and r package for gene selection and classification using random forest. *BMC Bioinformatics*, **8**, 328–335.

Diaz-Uriarte,R. and de Andrés,S.A. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3–16.

Dougherty,E. and Datta,A. (2005) Genomic signal processing: diagnosis and therapy. *IEEE Signal Proc. Mag.*, **22**, 107–112.

Dougherty,E. *et al*. (2005) Research issues in genomic signal processing. *IEEE Signal Proc. Mag.*, **Nov**, 46–68.

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

Galton,F. (1888) Co-relations and their measurement, chiefly from anthropometric data. *Proc. R. Soc.*, **45**, 135–145.

Galton,F. (1889) Co-relations and their measurement, chiefly from anthropometric data. *Nature*, **39**, 238.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.

Hochreiter,J. *et al*. (2006) A new summarization method for affymetrix probe level data. *Bioinformatics*, **22**, 943–949.

Irrizarry,R.A. *et al*. (2003) Summaries of affymetrix genechip probe level data. *Nucleic Acids Res.*, **31**, 1–8.

Lee,S.-I. and Batzoglou,S. (2003) Application of independent component analysis to microarrays. *Genome Biol.*, **4**, R76.1–R76.21.

Li,S. *et al*. (2001) Learning spatially localized, parts-based representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1.

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

Liu,Z. *et al*. (2005) Gene expression data classification with kernel principal component analysis. *J. Biomed. Biotechnol.*, **2**, 155–159.

Lutter,D. *et al*. (2008) Analysing M-CSF dependent monocyte/macrophage differentiation and meta-clustering with independent component analysis derived expression modes. *BMC Bioinformatics*, in press.

Mangasarian,O. and Musicant,D. (2001) Lagrangian support vector machines. *J. Mach. Learn. Res.*, **1**, 161–177.

Pearson,K. (1901) On lines and planes of closest fit to points in space. *Phil. Mag.*, **2**, 559–572.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nature*, **2**, 418–427.

Saidi,S. A. *et al*. (2004) Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene*, **23**, 6677–6683.

Schachtner,R. *et al*. (2007a) Blind matrix decomposition techniques to identify marker genes from microarrays. In Davies *et al*. (eds), *Lecture Notes in Computer Science*, vol. 4666. Springer.

Schachtner,R. *et al*. (2007b) Routes to identify marker genes for microarray classification. In Dittmar,A. and Clark,J. (eds), In *Proceeding of the 29th International Conference on IEEE Engineering in Medicine and Biology Society. IEEE-EmBC*, Lyon, France, pp. 4617–4620.

Schölkopf,B. and Smola,A. (2002) *Learning with Kernels*. MIT Press.

Simon,R. (2003) Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). *SIGKDD Explor.*, **5**, 31–36.

Spang,R. *et al*. (2002) Prediction and uncertainty in the analysis of gene expression profiles. *In Silico Biol.*, **2**, 33–58.

Talloen,W. *et al*. (2007) I/NI-calls for the exclusion of non-informative genes: a highly effective filtering tool for microarray data. *Bioinformatics*, **23**, 2897–2902.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, **98**, 5116–5121.

Wu,Z. and Irizarry,R. (2007) A statistical framework for the analysis of microarray probe-level data. *Ann. Appl. Stat.*, **1**, 333–357.

Wu,Z. *et al*. (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.