Vol. 23 no. 1 2007, pages 44–49
doi:10.1093/bioinformatics/btl550

Gene expression

Inferential, robust non-negative matrix factorization analysis of microarray data

Paul Fogel[1], S. Stanley Young[2,*], Douglas M. Hawkins[3] and Nathalie Ledirac[4]

[1]Consultant, 4 rue Le Goff, F-75005, Paris, France, [2]National Institute of Statistical Sciences, PO Box 14006, Research Triangle Park, NC 27709-4006, USA, [3]School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street NE, Minneapolis, MN 55455, USA and [4]Laboratoire de Toxicologie Cellulaire et Moléculaire, Centre de Recherche INRA, 400 Route des Chappes, 06903 Sophia-Antipolis, France

Received on July 23, 2006; revised on October 20, 2006; accepted on October 22, 2006
Advance Access publication November 8, 2006
Associate Editor: David Rocke

ABSTRACT

Motivation: Modern methods such as microarrays, proteomics and metabolomics often produce datasets where there are many more predictor variables than observations. Research in these areas is often exploratory; even so, there is interest in statistical methods that accurately point to effects that are likely to replicate. Correlations among predictors are used to improve the statistical analysis. We exploit two ideas: non-negative matrix factorization methods that create ordered sets of predictors; and statistical testing within ordered sets which is done sequentially, removing the need for correction for multiple testing within the set.

Results: Simulations and theory point to increased statistical power. Computational algorithms are described in detail. The analysis and biological interpretation of a real dataset are given. In addition to the increased power, the benefit of our method is that the organized gene lists are likely to lead better understanding of the biology.

Availability: An SAS JMP executable script is available from http://www.niss.org/irMF

Contact: young@niss.org

Supplementary information: http://www.niss.org/irMF

INTRODUCTION

The 'omic' technologies, such as genomics, proteomics and metabolomics, aim to create a numeric profile of a biological sample that captures the state of the sample at that point of time. So whereas tens to hundreds of samples and two or just a few groups are typically under consideration, there can be several hundred to many thousands of measured attributes of each sample. In this paper, we will focus on microarray examples although the methods apply to any two-way data table where there are correlations among the columns. We will also focus on follow up studies where there are likely to be tens of rows and hundreds of columns in the two-way table under consideration.

The classic method for factoring a two-way table, $X \sim= LSR'$, is the singular value decomposition (SVD), where X is the two-way table with n rows and p columns, with rows representing the n samples and columns representing the p genes, L is a matrix of left eigenvectors and is $n \times k$, S is a $k \times k$ diagonal matrix of eigen values, and R' is a $k \times p$ matrix of right eigenvectors. The relationship, $X = LSR'$, is exact if X is of rank k. The study of L, S and R' often gives important insights into the nature of X. Indeed, SVD is the mathematical basis of most linear statistical methods (Good, 1969). There is an interpretative problem with SVD; multiple, distinct mechanisms can be subsumed within a single, right eigenvector so that great subject matter knowledge and analysis skill can be required to draw sound conclusions from the right eigenvectors. The elements of a right eigenvector can be viewed as regression coefficients of regressing columns of X on the corresponding left eigenvector. Also, the eigenvectors are orthogonal and their squared elements sum to 1. All these attributes make for good mathematical properties, but can make for problematic subject matter interpretation. For example, genes, proteins and metabolites could be in common over two interlocking biochemical pathways so we should want a math/stat procedure that would not obscure what is happening.

Recently interest has focused on a different matrix factorization method, non-negative matrix factorization (NMF), (Lee and Seung, 1999). Here X has only non-negative elements and the factorization is restricted to have non-negative elements as well. A most interesting claim is made: the factorization puts independent mechanisms into separate vectors. This claim is still unresolved. There is empirical evidence to support the claim and conditions necessary for its truth have been studied and given by Donoho and Stodden (2003). NMF is being successfully used on microarray data (Kim and Tidor, 2003; Brunet *et al*., 2004; Gao and Church, 2005). We will focus on NMF in this paper.

From the beginning of modern statistical sciences, e.g. Fisher (1925), statisticians and experimental scientists have been concerned with multiple testing. Research workers ignore multiple testing at their peril. A large factor in the recent, very expensive, failure of experiments to confirm expectations (Ioannidis, 2005) can be attributed to ignoring multiple testing, which is not surprising as there is active teaching against multiple testing adjustments (Rothman, 1990). Recent research on multiple testing has focused on two strategies. One strategy is to control the probability of making any false claim among all the questions under consideration, the family-wise error rate (Westfall and Young, 1993). The other strategy is to control the expected fraction of claims that are likely to be false, the false discovery rate (FDR), (Benjamini and Hochberg, 1995). Where experimental work is expensive, difficult to replicate

[*]To whom correspondence should be addressed.

and claims are disruptive, FWE is a logical choice. Where follow up work is relatively easy to do and false claims can be readily identified, FDR is the method of choice. For the omic sciences it is normal to follow up experiments with additional testing so we will focus on FDR. For both FWE and FDR, it is advantageous to be able to take into account correlation structures.

Our strategy will be to use NMF to capture correlation structures in the dataset and then use a sequential form of testing (Kropf and Lauter, 2002) to control the error rate of our procedure. If a rejection is in error with probability alpha, it does not matter what you do from then on since the maximum Type I error probability is alpha. If a rejection is correct, you can go to the next test at level alpha. There are two advantages of our method, increased statistical power and easier data interpretation. We first review non-negative matrix factorization. Next we introduce our inference-based method to link gene sets to class labels. We give simulation results indicating that our method is more powerful than the FDR of Benjamini–Hochberg. We apply our method to a well-studied dataset and note that the correlated gene sets found are supported by biological literature. Finally, we offer some discussion of our new method.

MATERIALS AND METHODS

Data

Golub *et al*. (1999) gave data for relating a gene expression from 7129 genes to disease status. There are 11 patients with acute myeloid leukemia (AML) and 27 with acute lymphoblastic leukemia (ALL), which can be further divided into T and B cell subtypes (19 and 8 patients, respectively). We consider only the 5000 genes used by Brunet *et al*. (2004).

Dataset: The leukemia data, containing 38 bone marrow samples hybridized on Affymetrix Hu6800 chips, is a reduced version of the original data used in Golub *et al*. (1999). The URLs for data, row and column labels are as follows: http://www.pnas.org/content/vol0/issue2004/images/data/0308531101/DC1/08531DataSet1.txt; http://www.pnas.org/content/vol0/issue2004/images/data/0308531101/DC1/08531DataSet2.txt; and http://www.pnas.org/content/vol0/issue2004/images/data/0308531101/DC1/08531DataSet3.txt.

NMF updating rules

Write the expression data as an $n \times p$ array X with rows representing the n samples and columns representing the p genes. The updating rules proposed by Lee and Seung (1999) can be viewed as a modification of the Gabriel and Zamir (1979) alternating least-squares algorithm (ALS). This approach uses all of the observed data and does not require imputation of missing data. If the dataset is complete, this alternating least-squares algorithm gives the first term of the conventional SVD. Note that the elements of the first term will be all non-negative provided that the matrix X has only non-negative cells. However, to obtain further terms of the factorization, the residual matrix, which elements may be negative, is used so non-negativity is no longer guaranteed. Lee and Seung modified the ALS algorithm to handle all k terms (we give methods to select k later) of the factorization at a time so they end up with non-negative elements as for the first term of conventional SVD. We want to approximate X with a bilinear form $x_{ij} = \sum_{1 \leq v \leq k} r_{vi} c_{vj} + e_{ij}$. We begin with a tentative estimate of the column factors c_{vj} and of the row factors r_{vi}, $1 \leq v \leq k$, scaled by the sum of their elements to eliminate the degeneracy associated with the invariance of the matrix factorization under the transformation $r_{vi} \rightarrow \lambda_v r_{vi}$ and $c_{vj} \rightarrow \lambda_v^{-1} c_{vj}$. For any u, $1 \leq u \leq k$, we consider the original array corrected for factors other than u: $\tilde{x}_{ij} = x_{ij}(r_{ui} c_{uj})/\hat{x}_{ij} = r_{ui} z_{ij} c_{uj}^{1/2}$ where $z_{ij} = (x_{ij}/\hat{x}_{ij})\, c_{uj}^{1/2}$ (note that using × and ÷ operators guarantees the non-negativity of $\tilde{x}_{ij}$). Regarding $z_{ij} = \tilde{r}_{ui} c_{uj}^{1/2} + \tilde{e}_{ij}$ as a regression of the i-th row of Z on the square root of the column factors $c_{uj}^{1/2}$ identifies $\tilde{r}_{ui}$ as the coefficient of a no-intercept regression: $\tilde{r}_{ui} = \sum_j (x_{ij}/\hat{x}_{ij})\, c_{uj}$ leading to the updated row factor: $r_{ui} \leftarrow r_{ui} \tilde{r}_{ui}$ after proper scaling. Then switching roles, we take the updated row factors $r_{vi}, 1 \leq v \leq k$ as given and use regression of all non-empty cells in exactly the same way to calculate fresh estimates of the column factors c_{vj}.

Lee and Seung show that their algorithm leads to the minimization of a divergence criterion. Here we show that their algorithm is essentially a modified form of the ALS algorithm. The most striking difference between Lee–Seung and Gabriel and Zamir algorithm is in the multiplicative updating rule, which guarantees that row and column factor elements will remain non-negative throughout the iterative process.

Clustering

For each observation (row of X), the elements of the left eigenvectors put weights on the right eigenvectors. These define class characteristics so each observation can be assigned to the class for which it has the highest weight. Likewise, the columns of X can be assigned the number of the right component with the highest element.

Sparse NMF

A sparse matrix is one among those elements are zero or near zero. There are at least two reasons to try to form sparse eigenvectors: it is very unlikely that all genes are involved in a specific mechanism, and sparse vectors are easier to interpret. Hoyer (2004) suggested a sparseness constraint on eigenvectors within the updating rules of NMF. We use a simpler algorithm, which takes advantage of the local nature of found solutions. First, we let the process converge for a set number of iterations, we then fix the smallest elements of the right (or left) eigenvectors to zero for a small number of iterations, and then we let the process continue.

Sequential procedure

The original approach of NMF, as described above, is stochastic: Choose a number of components k, guess a set of k left and right eigenvectors, and apply updating rules. If k is large, the numerous guesses increase the risk of converging to a local minimum. Brunet *et al*. (2004) propose to repeat the whole process many times and build a consensus matrix to see whether results are consistent across different trials. In contrast, our implementation of NMF is more directed. The idea is to build intermediate matrix factorizations starting from robust estimates of the column effects of the residual matrix. We start with one component and guess a set of column markers (trial right eigenvector) using the column medians. We also need to guess row markers (trial left eigenvector) before updating rules start, so we set them all to 1. We use the same strategy each time we add a new component into the model, except that we use the column medians of $\hat{x}_{ij}/x_{ij}$ where the approximation $\hat{x}_{ij}$ is from the preceding model. The sequential procedure facilitates our implementation of robust NMF.

Robust NMF

To make the method robust to outliers in the original table, we use a Least Trimmed Square approach as for robust SVD (Liu *et al*., 2003). We identify the most discordant observations and remove them from the fitting process. The outlier list is updated as the factorization proceeds. To start the process, elements are selected at random for the outlier list so our robust LTS implementation is stochastic. This is actually the reason why we apply a sequential procedure. We could not reasonably nest one stochastic process within another, as computation time would be excessive.

Optimal number of components

In developing an approximation to the matrix X, the number of right and left eigenvectors, k, needs to be specified or determined. Our method for finding the optimal number of components, k, is adapted from Zhu and Ghodsi (2006). One assumes that eigen values follow a mixture distribution of two normal distributed populations, the first one corresponding to the components that should be selected. We calculate the profile likelihood for any hypothesis of k significant components under the assumption that both populations have same standard deviation. We select the hypothesis number that has the highest profile likelihood. In a complementary test, we take advantage of the non-orthogonality of eigenvectors, which is specific to NMF. For a number of components k, we calculate the determinant of the k vectors $\widehat{X}_u, 1 \leq u \leq k$ after proper normalization, where $\widehat{X}_u$ is the approximation to X obtained with u components, reshaped into a column vector. As long as each component carries specific signal, this determinant is smoothly decreasing. An abrupt decrease in the determinant reflects a high level of correlation between the last introduced component and the preceding ones, suggesting that most of the signal is already explained.

Inference

Assume that rows of the matrix correspond to samples of different types or classes (e.g. control, disease or drug groups) and columns correspond to response variables. As it is likely that only a limited number of response variables are linked to each class label, we are interested in finding such variables or 'predictors' that would allow predicting the class of any sample which type is unknown. Non-negative matrix factorization is used to create ordered sets of response variables. Here each set corresponds to a particular right eigenvector, which has been ordered by decreasing values of its elements. It is important to note that the ordering is totally unsupervised (the information on sample group is not used). In order to identify predictors, we take advantage of the ordering of variables within each set and test each variable sequentially so there needs be no correction for multiple testing (since the ordering is not supervised) (Kropf and Lauter, 2002). This sequential form of testing controls the error rate of our procedure and leads to an increase in statistical power as shown by our simulation results (see next section). Finally, we link each set of predictors to the class label of the top element of the corresponding ordered left-hand eigenvector.

Gene expression levels typically vary on different scales, and the impact on matrix factorization and subsequent ordering of variables should be considered. Just as the prime factors of 225, 15×15, are larger than those of 25, 5×5, genes with larger variance will have larger elements in the right and/or left eigenvectors. If there is a control group, the variances of the genes can be normalized using this group. If there is no such control group, our sequential testing procedure will likely be corrupted by genes with high variance. Two things to note: first, it is important to remove outliers or genes with outliers before applying our method and second, the profile likelihood method can be used to find the real elements within an eigenvector.

RESULTS

Simulations

As we can test genes within an eigenvector without any adjustment for multiple testing we expect greater statistical power.

Two groups: We considered one normal and one treated group with following settings for the numbers of regulated genes (Table 1). Upregulated and downregulated genes were simulated in equal proportion.

Three groups: We considered one normal, N and two treated groups, T1, T2, with following settings for the number of regulated genes (Table 2). Upregulated and downregulated genes were simulated in equal proportion. *Note*: Some genes are induced by both treatments.

Baseline expression:

(i) Normal genes (Normal Baseline): 100

(ii) Upregulated genes: 1.5 × Normal Baseline

(iii) Downregulated genes: 0.67 × Normal Baseline

Table 1. Two groups: we considered one normal and one treated group with following settings for the numbers of regulated genes

No. of regulated genes	No. of unregulated genes	% Regulated genes
20	80	20
40	160	20
40	360	10
40	760	5

Upregulated and downregulated genes were simulated in equal proportion.

Table 2. Three groups: We considered one normal, N, and two treated groups, T1 and T2, with following settings for the numbers of regulated genes

No. of genes regulated by T1	No. of genes regulated by T2	No. of genes regulated by T1 and T2	No. of unregulated genes	% Regulated genes
6	6	8	80	20
10	10	20	160	20
10	10	20	360	10
10	10	20	760	5

Upregulated and downregulated genes were simulated in equal proportion. *Note*: Some genes are induced by both treatments.

Distribution: We used a log-normal distribution:

$$\text{Gene expression} = \text{Baseline} \times \exp(\mathbf{N}(0, 0.25)).$$

For support on the choice of this distribution, see Durbin *et al.* (2002). *Note*: since $\exp(1.5 \times 0.25) = 1.45$, nominal modulation levels 1.5 and 0.67 ensure strong overlapping between distribution of unregulated and regulated genes.

Correlation structure: We added a correlation structure between regulated genes in the following way:

(i) For each mechanism, set up a common profile:

$$\text{Profile(T1)} = \text{Baseline} \times 1.5 \times \exp(\mathbf{N}(0, 0.25))$$
$$\text{Profile(Other)} = \text{Baseline} \times \exp(\mathbf{N}(0, 0.25))$$

(ii) For each gene that belongs to the same mechanism, add Poisson noise to the profile:

$$\text{Measured Expression} = \lambda \times \text{Random Poisson}(\text{Profile}/\lambda).$$

The parameter λ allows controlling the correlation level through adjusting the standard deviation of the Poisson distribution (since $\sigma(\text{Poisson}[\mu/\lambda] \times \lambda) = \sqrt{\mu\lambda}$). In the simulation, we used $\lambda = 2$, which ensures a correlation level ranging between 0.7 and 0.9.

Methodology: We compared two methods for selecting genes:

(i) ANOVA with multiplicity correction using a linear step-up procedure, Benjamini and Hochberg (2000), to control FDR at nominal level 0.05 and 0.025 [where the number of regulated genes is estimated through Storey (2002)].

(ii) Sequential ANOVA, where the order of hypothesis is defined by the order in which genes appear in NMF components.

Table 3. Two groups, levels 0.05 and 0.025

	BH ANOVA FDR (%)	Power (%)	Sequential ANOVA FDR (%)	Power (%)	FWE (%)
2 Groups, level 0.05					
20% Regulated genes, 100 genes	6.9	47.8	1.7	61.6	20.0
20% Regulated genes, 200 genes	5.3	49.0	1.0	58.8	22.0
10% Regulated genes, 400 genes	5.7	34.2	1.7	59.3	31.0
5% Regulated genes, 800 genes	2.7	19.4	1.3	59.5	30.0
2 Groups, level 0.025					
20% Regulated genes, 100 genes	4.5	38.1	0.6	55.0	8.0
20% Regulated genes, 200 genes	1.9	40.3	0.5	50.2	13.0
10% Regulated genes, 400 genes	4.0	24.5	0.6	47.2	13.0
5% Regulated genes, 800 genes	1.8	17.3	0.6	49.8	13.0

Table 4. Three groups, levels 0.05 and 0.025

	BH ANOVA FDR (%)	Power (%)	Sequential ANOVA FDR (%)	Power (%)	FWE (%)
3 Groups, level 0.05					
20% Regulated genes, 100 genes	4.8	62.2	2.4	73.1	33.0
20% Regulated genes, 200 genes	4.9	64.5	1.2	70.3	31.0
10% Regulated genes, 400 genes	4.5	50.1	2.2	66.4	44.0
5% Regulated genes, 800 genes	3.9	31.1	2.6	61.3	39.0
3 Groups, level 0.025					
20% Regulated genes, 100 genes	2.5	46.4	0.9	64.8	13.0
20% Regulated genes, 200 genes	2.4	49.9	0.8	60.0	21.0
10% Regulated genes, 400 genes	1.6	30.6	0.9	52.0	18.0
5% Regulated genes, 800 genes	1.2	20.5	1.4	50.7	23.0

Alpha-level is set at 0.05 and 0.025. We test only the components associated with any of the two treatments. To identify those components, we look at the top elements of the left-hand eigenvectors. If the corresponding samples belong to any of the treatment groups, then we decide that the corresponding right eigenvector is linked with this treatment group and we run sequential ANOVA on this particular component. Note that NMF is run actually twice: on the original raw data and on the inverse of raw data to detect separately upregulated and downregulated genes. To ensure maximal power of our procedure, no Bonferroni correction is applied over the sequences.

In both cases, ANOVA was run on log data, given the distribution model used. NMF was run with two and three components for the 2 groups and 3 groups design, respectively.

Simulation results: The following results are based on 100 runs for each setting of the experimental design:

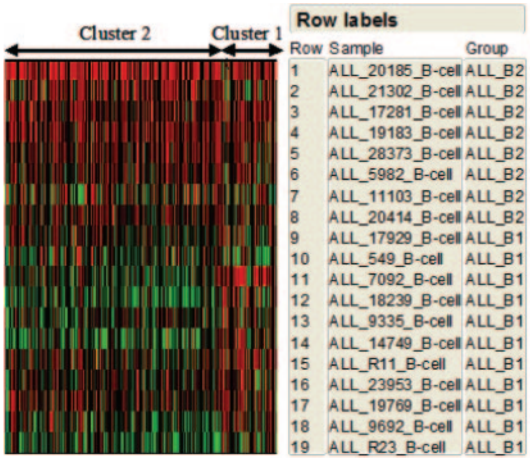
(i) *False discovery rate and power*: FDR is lower and power is higher. Note that at alpha level = 0.05, sequential ANOVA ensures FDR < 0.025. We can therefore compare the power of sequential ANOVA at alpha level 0.05, Table 3 (2 groups, level 0.05) and Table 4 (3 groups, level 0.05), with the power of BH ANOVA at FDR level 0.025, Table 3 (2 groups, level 0.025) and Table 4 (3 groups, level 0.025). If we do so, the difference in power is even more substantial (e.g. in the 3 groups simulation >20%).

(ii) *Family wise error*: Since we use uncorrected alpha in our ANOVA test, for the 2 groups simulation we apply sequential testing on 2 components twice (for up and downregulated genes) so we expect FWE = 2 × 2 × alpha = 0.20 and 0.10 at the 0.05 and 0.025 level, respectively. In the same way, for the 3 groups simulation we expect FWE = 2 × 3 × alpha = 0.30 and 0.15 at the 0.05 and 0.025 levels, respectively. The observed percentages are consistent with theory.

Biological example

Brunet *et al*. (2004) note that the leukemia dataset has become a benchmark in the cancer classification community. To compare their results with the output of our robust NMF, we ran the analysis on the same 5000 genes and found essentially the same results when we asked for two and three components: the biological distinction among AML, ALL_B and ALL_T subtypes was precisely recovered. Here we focus on a four-component model, which further divides the ALL_B samples into two subtypes ALL_B1 and ALL_B2. Scree plots and related tests, Zhu and Ghodsi (2006), suggest that this model should be optimal.

Brunet *et al*. (2004) note that the separation of ALL_B into two classes ALL_B1 and ALL_B2 is clear by their model selection, but its biological significance is unclear. We used the profile likelihood method described in the Materials and methods section to identify the genes that are responsible for the separation into two



Row labels

Row	Sample	Group
1	ALL_20185_B-cell	ALL_B2
2	ALL_21302_B-cell	ALL_B2
3	ALL_17281_B-cell	ALL_B2
4	ALL_19183_B-cell	ALL_B2
5	ALL_28373_B-cell	ALL_B2
6	ALL_5982_B-cell	ALL_B2
7	ALL_11103_B-cell	ALL_B2
8	ALL_20414_B-cell	ALL_B2
9	ALL_17929_B-cell	ALL_B1
10	ALL_549_B-cell	ALL_B1
11	ALL_7092_B-cell	ALL_B1
12	ALL_18239_B-cell	ALL_B1
13	ALL_9335_B-cell	ALL_B1
14	ALL_14749_B-cell	ALL_B1
15	ALL_R11_B-cell	ALL_B1
16	ALL_23953_B-cell	ALL_B1
17	ALL_19769_B-cell	ALL_B1
18	ALL_9692_B-cell	ALL_B1
19	ALL_R23_B-cell	ALL_B1

Fig. 1. Schematic view of genes distinguishing ALL_B1 from ALL_B2.

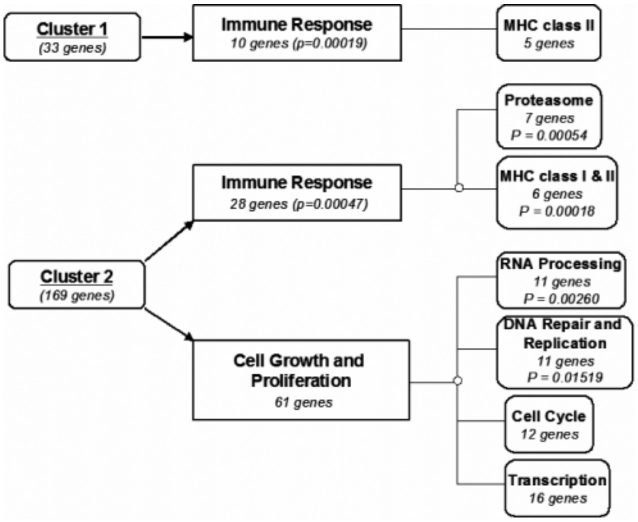


Fig. 2. Biological functions of the most discriminating genes.

subtypes and obtained two clusters of genes as each subtype was linked to one right eigenvector (Fig. 1).

We further examined the biological relevance of these clusters. Following the exclusion of unmapped gene IDs and gene repeat, the classification of genes eligible for generating classes linked to ALL_B1 and ALL_B2 was finally performed with ~80% of the initial gene lists, namely cluster1 and cluster 2 lists.

The Cluster 1 list of 33 genes (Supplemental information) includes genes coding for proteins known to be involved in the immune response and lymphatic system development (16 genes) (Fig. 2). Among them, several genes are related to the major histocompatibility complex (MHC) class II, which are involved in the presentation of antigenic peptides to T cells, such as HLA-DMA, HLA-DPA1, HLA-DPB1, HLA-DRA, HLA-DRB1 and CD74. As expected, these genes are mostly expressed in ALL-B, to a lesser extent in AML, but not in ALL-T.

Several genes, VPREB1, TCL1A, IGHM, CD79A or TCF4, are selectively expressed in ALL-B, and are related to genes expressed at the early stage of B-cell development, which characterize ALL-B. In addition, CXCR4, expressed in all samples, is strongly upregulated in ALL-B, up to 4-fold over AML and ALL-T. This gene encodes a chemokine receptor involved in pre-B cell growth stimulating that has been recently described to be antagonized by CD24 selectively expressed in ALL-B (Schabath *et al*., 2006). The co-expression of CD24 and CXCR4 might contribute to altered hematopoiesis in acute leukaemia.

The cluster 2 list of 169 genes (Supplemental information) includes many of cluster 1 genes (16 genes) required to differentiate ALL-B from ALL-T and AML. Cluster 2 contains additional genes necessary for the identification of the two ALL-B subtypes (Fig. 2).

We first identify a set of genes reflecting the immune response mechanism involving both antigen presentation and immunoproteasome pathway. In addition to the cluster 1 genes already described, there are genes related to MHC class I (HLA-E and HLA-F), which have been related to intracellular peptide presentation, with a 2-fold increased expression level of HLA-E in ALL-B2 compared with ALL-B1, ALL-T and AML. These peptides are generated during the degradation of intracellular proteins by the proteasome. Also in this set of genes are several genes related to 20S immunoproteasome that are predominantly expressed in ALL-B2 subtype, 2-fold higher than in ALL-B1, such as PSMA4, PSMA6, PSMB1, PSMB10 (MECL1), PSME2, in addition to PSMB8 (LMP7) and PSMB9 (LMP2) more specifically expressed in ALL-B2 compared with other leukemia. PSMB8 (LMP7), PSMB9 (LMP2) and PSMB10 (MECL1) are INF-γ-inducible subunits that have been shown to be strongly expressed in lymphoid tissues. In addition, concentration of proteasome have been reported to be abnormally high in leukemia cells (Kumatori *et al*., 1990) and support the recent interest for specific inhibitors of proteasome as therapeutic target in cancer therapies (Spano *et al*., 2005; Schabath *et al*., 2006).

Continuing the discussion of cluster 2, we identify a second set of genes reflecting the proliferative status of ALL-B2 compared with ALL-B1. This set of genes includes several genes related to DNA repair, transcription and replication, which seem to be differentially regulated in the two classes of ALL-B. Some of these genes are overexpressed in ALL-B2, such as HMGN1, HMGN2, CHD4, H3F3A, SMARCA4, TOP2B that play important roles in the regulation of transcription. Genes related to DNA replication are also overexpressed in ALL-B2 such as HMGB2, NAP1L1 as well as SSBP1 involved more specifically in mitochondrial replication. Numerous genes related to RNA splicing are highly expressed in ALL-B compared with other leukemia, with at least a 2-fold increase expression level in ALL-B2 compared with ALL-B1 (HNRPA1, HNRPA2B1, HNRPF, SFRS3, SFRS5, SFRS10, SFRS11, SNRPE, SNRPN, U2AF1). As might be expected, this set of genes also includes cell cycle-related genes, such as CCND3 and CCND2, that are checkpoint regulators of the progression from G_1 to S phase of the cell cycle, and show upregulation in ALL-B2. The gene CCNG1 is associated with the progression from G_2 to M phase and is selectively expressed in ALL-B2; it has been recently defined as a host risk factor for treatment-related myeloid leukemia (t-ML) (Bogni *et al*., 2006). Increased expression of these cyclins strongly supports the more proliferative status of ALL-B2. We can also include *PCNA* gene, frequently used as a proliferation marker that is predominantly expressed in ALL-B2 and ALL-T (2-fold above the expression level observed in ALL-B1 and AML samples). Moreover, the gene *CD*81, which plays an important role in positive regulation of B-cell proliferation, is upregulated in ALL-B1 (2-fold over AML and ALL-T) and ALL-B2 (3-fold over AML and ALL-T).

In addition, the cluster 2 list includes genes related to electron transport chain (COX5A, COX5B, COX7B, UQCRB, UQCRFS1, UCP2 and NDUFV2) involved in energy production. Only 5 out of 33 genes (cluster 1) demonstrate downregulated expression patterns in ALL-B2 compared with All-B1: HBA2, HBB, HSPB1, SOX18 and SRP68.

Taken together, upregulation of the expression of ALL-B2 genes may mainly reflect a more proliferative nature of ALL-B2 compared with ALL-B1, with higher rate of transcription and replication processes, more proteasomal activity and more energy production. This ALL-B2 subtype is also characterized by specific expression of recently identified surface antigen CD164 and is for the first time associated with the expression of KLRK1 receptor, normally expressed by natural killer cells and CD8(+) T cells, and involved in cell cytotoxic response.

Note that for gene clusters given by WebGestalt (Zhang *et al*., 2005), *P*-values are given. *P*-values are based on a hypergeometric test; all of them are quite significant.

DISCUSSION

When and why will this analysis strategy, inferential robust non-negative matrix factorization, irNMF, work? The strategy is conceptually simple. First, non-negative matrix factorization is used to create groups of genes that are moving together in the dataset. The error rate to be controlled is allocated over these groups. Within each group, genes are tested sequentially. The strategy should be effective if there are sets of genes moving together so that group formation reflects biological reality. For example, if cancer cells are compared with non-cancer cells, there are likely to be large blocks of correlated genes that differ between the two cell types.

As we do no multiple testing adjustments within the sequence of genes, we should have higher power and simulations support higher power for irNMF. Sets of genes are identified so the biological interpretation should be more straightforward and we think that it is for the Golub dataset.

On the leukemia dataset, our robust implementation of NMF gives four, almost perfect, clusters with only one AML misclassified among ALL_B1 samples; the standard implementation resulted in two errors. Also, on the computational side, convergence was obtained in one-third the number of iterations of the standard NMF updating process.

SVD attempts to separate mechanisms in an orthogonal way, although nature is all but orthogonal. As a consequence, SVD components are unlikely to match with real mechanisms and so are not easily interpreted. On the contrary, NMF appears to match each real mechanism with a particular component.

Conflict of Interest: none declared.

REFERENCES

Bogni,A. *et al*. (2006) Genome-wide approach to identify risk factors for therapy-related myeloid leukemia. *Leukemia*, **20**, 239–246.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.

Benjamini,Y. and Hochberg,Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Behav. Educ. Statist*., **25**, 60–83.

Brunet,J.P. *et al*. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, **101**, 4164–4169.

Donoho,D. and Stodden,V. (2004) When does non-negative matrix factorization give a correct decomposition into parts? In Thrun,S., Saul,L. and Scholkopf,B. (eds), *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Durbin,B.P. *et al*. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**(Suppl. 1), S105–S110.

Fisher,R.A. (1925) *Statistical Methods for Research Workers*. Hafner, New York.

Gabriel,K.R. and Zamir,S. (1979) Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, **21**, 489–498.

Gao,Y. and Church,G. (2005) Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics*, **21**, 3970–3975.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Good,I.J. (1969) Some applications of the singular decomposition of a matrix. *Technometrics*, **11**, 823–831.

Hoyer,P.O. (2004) Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res*., **5**, 1457–1469.

Ioannidis,J.P.A. (2005) Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, **294**, 218–228.

Kim,P.M. and Tidor,B. (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res*., **13**, 1706–1718.

Kropf,S. and Lauter,J. (2002) Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data. *Biometric. J*., **44**, 789–800.

Kumatori,A. *et al*. (1990) Abnormally high expression of proteasomes in human leukemic cells. *Proc. Natl Acad. Sci. USA*, **87**, 7071–7075.

Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Liu,L. *et al*. (2003) Robust singular value decomposition analysis of microarray data. *Proc. Natl Acad. Sci. USA*, **100**, 13167–13172.

Rothman,K.J. (1990) No adjustments are needed for multiple comparisons. *Epidemiology*, **1**, 43–46.

Schabath,H. *et al*. (2006) CD24 affects CXCR4 function in pre-B lymphocytes and breast carcinoma cells. *J. Cell Sci*., **119**, 314–325.

Spano,J.P. *et al*. (2005) Proteasome inhibition: a new approach for the treatment of malignancies. *Bull Cancer*, **92**, 945–952.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc. B*, **64**, 479–498.

Westfall,P.H. and Young,S.S. (1993) *Resampling-based Multiple Testing*. Wiley, New York.

Zhang,B., Kirov,S. and Snoddy,J. (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*., **33**(Web Server issue), W741–W748.

Zhu,M. and Ghodsi,A. (2006) Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Comput. Stat. Data Anal*., **51**, 918–930.