BIOINFORMATICS

Vol. 18 no. 1 2002
Pages 51–60

Linear modes of gene expression determined by independent component analysis

Wolfram Liebermeister

Theoretische Biophysik, Institut für Biologie, Humboldt-Universität zu Berlin, Invalidenstraße 42, 10115 Berlin, Germany and Max-Planck-Institut für molekulare Genetik, Ihnestraße 73, 14195 Berlin, Germany

Received on March 22, 2001; revised on May 29 and July 20, 2001; accepted on August 3, 2001

ABSTRACT

Motivation: The expression of genes is controlled by specific combinations of cellular variables. We applied Independent Component Analysis (ICA) to gene expression data, deriving a linear model based on hidden variables, which we term 'expression modes'. The expression of each gene is a linear function of the expression modes, where, according to the ICA model, the linear influences of different modes show a minimal statistical dependence, and their distributions deviate sharply from the normal distribution.

Results: Studying cell cycle-related gene expression in yeast, we found that the dominant expression modes could be related to distinct biological functions, such as phases of the cell cycle or the mating response. Analysis of human lymphocytes revealed modes that were related to characteristic differences between cell types. With both data sets, the linear influences of the dominant modes showed distributions with large tails, indicating the existence of specifically up- and downregulated target genes. The expression modes and their influences can be used to visualize the samples and genes in low-dimensional spaces. A projection to expression modes helps to highlight particular biological functions, to reduce noise, and to compress the data in a biologically sensible way.

Availability: The FastICA algorithm (Hyvärinen, *IEEE Trans. Neural Netw.*, **10**, 626–634, 1999) is freely available at http://www.cis.hut.fi/projects/ica/fastica/. Additional matlab scripts and detailed results can be downloaded from http://www.molgen.mpg.de/research/lehrach/projects/genica/

Contact: wolfram.liebermeister@rz.hu-berlin.de

INTRODUCTION

Cells react to external stimuli and to their internal needs by the induction or repression of genes, among other things by up- or downregulating the amounts of corresponding mRNA molecules. Samples of cells or tissues that represent different biological situations or experimental treatments show characteristic expression patterns. Gene expression is controlled by a combination of mechanisms including networks of signaling substances, transcription factors and their binding sites in the promoter regions of genes, as well as modifications of the chromatin structure and different types of posttranscriptional regulation. Thus, the expression of each gene relies on the specific processing of a number of regulatory inputs, which are still unknown in most cases. The gene Endo16 in the sea urchin, whose regulatory function has been presented in the style of a computer program (Yuh *et al.*, 1998), remains a rare exception.

Genomic-scale gene expression data, which are provided by high-throughput methods like the microarray technology, give new insights about the regulatory machinery behind gene expression. Different types of genetic network models (see for example Liang *et al.*, 1998; Wahde and Hertz, 2000; Weaver *et al.*, 1999; D'haeseleer *et al.*, 1999; Friedman *et al.*, 2000) have been used to represent coregulation or feedback relations between genes. However, constructing detailed genomic-scale networks in an unsupervised fashion, i.e. from expression data alone, suffers from two drawbacks: the measurement errors are still large, and a realistic network model would involve many quantities besides the observed mRNA concentrations, like the amounts of the gene products, metabolites, signaling molecules, or transcription factors. However, the coregulation of genes may be described by a small number of 'effective' regulators, each acting on a large set of genes and varying between distinct biological situations.

Clustering is a widely used tool in the analysis of gene expression: with two-way hierarchical clustering (Eisen *et al.*, 1998), the genes and samples are organized in tree structures. After a rearrangement, clusters become visible in the data matrix. Self-organized maps (Tamayo *et al.*, 1999) determine gene clusters of similar size, which are joined by a predefined topology. Ben-Dor *et al.* (1999) proposed an algorithm to recover clusters from noisy data. Gene shaving (Hastie *et al.*, 2000) determines optimal small clusters of marker genes that show a large variance either across all samples or between predefined sample groups.

In contrast to clustering, linear models rely on the idea of a combinatorial control, describing the expression levels of genes as linear functions of common hidden variables. Ideally, these variables may be related to distinct biological causes of variation, like regulators of gene expression, cellular functions, or responses to experimental treatments. The 'reduce' model (Bussemaker *et al.*, 2001) is based on the occurrence of common motifs in the genes' promoter sequences, assuming that each regulator acts on the genes via a particular promoter element. Principal Component Analysis (PCA; Raychaudhuri *et al.*, 2000; Wen *et al.*, 1998), and singular value decomposition (Alter *et al.*, 2000) decompose the gene profiles into linear combinations of 'eigengenes', the eigenvectors of the covariance matrix. Holter *et al.* (2001) fitted the time-behaviour of eigengenes by a linear dynamical model. The 'plaid model' (Lazzeroni and Owen, 2000) decomposes the expression matrix into a sum of submatrices, each related to specific subsets of genes and samples.

In this paper, we study the application of Independent Component Analysis (ICA) (see Hyvärinen *et al.*, 2001) to gene expression data: the variables (corresponding to the samples) are linearly transformed to so-called 'independent components' with minimal statistical dependencies between them. ICA has been used for blind source separation, denoising, and sparse coding. In the context of gene expression, we propose to regard the independent components as linear influences of unobserved variables, which we term 'expression modes'. Each component defines corresponding groups of induced and repressed genes. Samples and genes can be visualized by projecting them to particular modes or to their influences, respectively. In the two data sets studied, we found that the dominant modes could be related to particular biological or experimental effects. We projected the data to selected modes in order to highlight these factors and to filter out other sources of variation. Reducing the number of data dimensions may be useful to simplify further analysis, while maintaining the most relevant biological information.

METHODS

Independent component analysis

We consider a data matrix $X = (X_{il})$ whose rows correspond to individuals (genes) and whose columns correspond to the variables (cell samples)[†]. The column means have been shifted to zero. The ICA model

$$X_{il} = \sum_k S_{ik} A_{kl}$$

splits the data matrix into a matrix product $X = SA$ (see Figure 1), subject to the condition that the statistical dependence between the columns of S be minimized. The new variables, contained in the columns of S, are called 'independent components'. The statistical dependence between variables can be quantified by the mutual information $I = \sum_k H_k - H$, where H_k and H denote the entropy of the kth variable and the total entropy, respectively (see, for instance, Cover and Thomas, 1991). As the total entropy H remains constant under linear transformations, minimizing the mutual information I is equivalent to minimizing the marginal entropies H_k. Among the distributions with unit variance, the normal distribution has the maximal entropy value H_N, so ICA determines directions where the distribution of the data is as non-normal, and thus as informative, as possible. As a side-effect, ICA can identify components that are 'approximately sparse', showing an increased fraction of values around zero.

We used the FastICA algorithm, which has been published by Hyvärinen (1999). As illustrated in Figure 2, the matrix A is split into the product $A = R\, C^{1/2}$, where the 'dewhitening matrix' $C^{1/2}$ representing the linear correlations is calculated from the data covariance matrix C. The remaining rotation R is chosen such that the statistical dependence between the independent components is minimized. In order to avoid the time-consuming calculation of the H_k, FastICA substitutes the difference $H_N - H_k$ by a 'contrast function'

$$J_G(k) = |\langle G(S_{ik})\rangle_i - \langle G(\nu)\rangle|.$$

J_G applies some even, non-quadratic function $G(\cdot)$ (we chose the Gaussian function because of its robustness properties) to each variable $S_{\cdot k}$ and to a normally distributed variable ν, returning the absolute difference of the mean values. R is initialized with random values and then iteratively adjusted to maximize the J_G until a convergence criterion is met.

Like PCA, ICA removes all linear correlations. By introducing a non-orthogonal basis, it also takes into account higher-order dependencies in the data. If the data lack such higher order structure, for instance if they are normally distributed, the solution is not unique. The ICA model leaves some freedom of scaling and sorting: by convention, the independent components are scaled to unit variance, while their signs and their order can be chosen arbitrarily. The number of independent components equals the number of variables, but it may be reduced, for instance by removing weak principal components before applying the ICA, which considerably decreases the computational costs.

[†] In the ICA literature, the problem is usually formulated using the transposed matrix X^{T}.

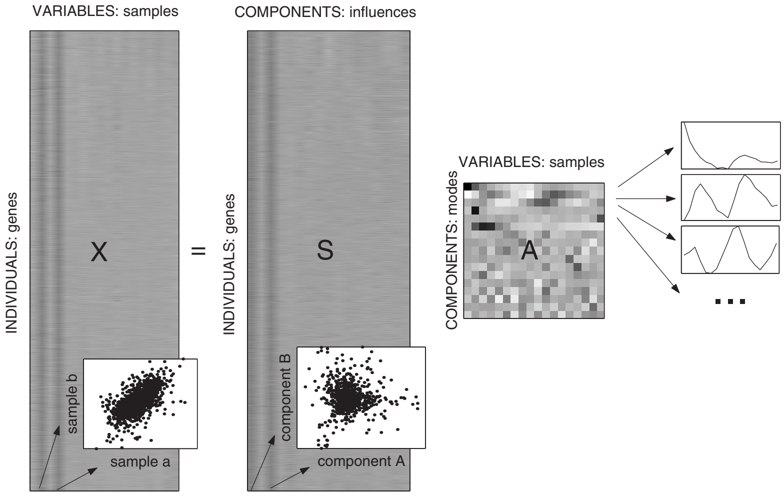

Fig. 1. Independent Component Analysis. ICA splits the gene expression matrix X (coded by shades of grey) into a matrix product $X = SA$, introducing new variables ('independent components', contained in the columns of S) with minimal statistical dependencies between them. The two lower panels show scatterplots between two variables (shaded columns of X) and between two independent components (shaded columns of S). The independent components represent the data with respect to a new basis formed by the rows of the 'mixing matrix' A. The first three basis profiles are shown in the panels on the right. The data describe the α factor time-course of the yeast cell cycle experiment (see below).

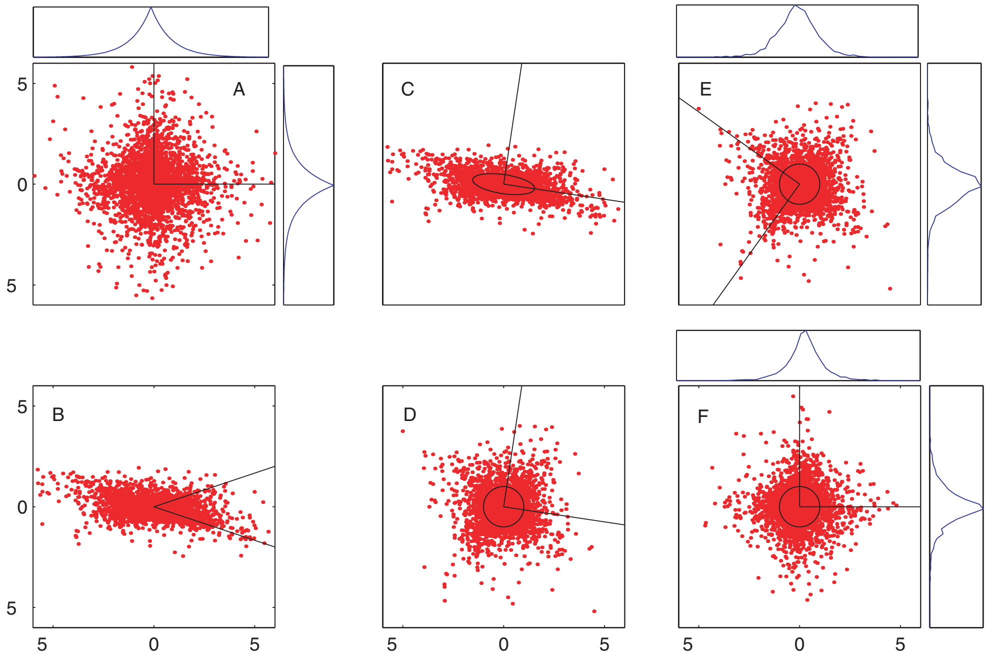

Fig. 2. Reconstructing artificial data using ICA. Left: N data points were produced by (a) choosing independent coordinates (S_1, S_2) from the two-sided exponential distribution and (b) shearing the data cloud by a linear transformation A. The centered data are contained in a $N \times 2$ matrix X. ICA reconstructs the unsheared data up to scaling, permutation, and reflection of the axes, based on the knowledge that the coordinates were independent. Middle: (c) linear correlations between the two variables are represented by the covariance matrix C: its eigenvectors point along the axes of an ellipse defined by $\mathbf{x}\, C^{-1}\, \mathbf{x}^{\mathrm{T}} = 1$. ICA 'whitens' the data (d) by stretching them to unit variance along these directions, thereby removing the linear correlations. Right: the whitened data (e) are rotated to independent components (f) maximizing the 'contrast function' J_G, a dissimilarity between their marginal distribution and the normal distribution.

Interpreting linear models of gene expression

The gene expression profiles (rows of X) can be regarded as points in a multidimensional space with dimensions corresponding to the different samples. A linear model $X = SA$ represents the data by new variables (the rows of S) or, alternatively stated, with respect to a new set of basis vectors (the rows of A). To interpret the linear decomposition, we propose a model of gene expression based on the following assumptions:

(1) The sample expression profiles are determined by a combination of hidden regulatory variables. We call these variables ‘expression modes’.

(2) The genes’ responses to these variables can be approximated by linear functions.

Expression mode k is characterized by its profile over the samples (kth row of A) and by its linear influences on the genes (kth column of S). If logarithmic data are used, the linear combination of inputs corresponds to a multiplicative rather than to an additive processing.

It would be useful to detect modes related to distinct biological processes involved in gene regulation: direct or indirect regulators of mRNA synthesis, like transcription factors or external stimuli, or cellular tasks that require the activation or repression of cooperating genes, for instance shock responses or the regulation of metabolic pathways. Modes might also describe general differences between individuals or tissue types, or different compositions of tissue samples. In order to determine biologically meaningful modes, the statistical assumptions underlying a linear model should reflect plausible properties of effective biological regulators.

PCA assumes its first components to capture a maximal amount of data variance. This constrains the modes, as well as their influences, to be orthogonal. Although the biological interpretation of individual principal components is not obvious, PCA can be expected to separate a subspace of biological effects from the subspace of weaker noise components. The plaid model (Lazzeroni and Owen, 2000) determines modes that are active only in subsets of the cell samples (sparse A), acting on distinct (but overlapping) groups of genes (sparse S). The ICA model states that different modes exert independent influences on the genes. As a consequence, ICA is sensitive to modes whose influences on the genes follow a ‘supergaussian’ distribution with large tails and a pronounced peak in the middle. These modes correspond to regulators which specifically act on (possibly overlapping) sets of target genes and have little effect on the others. In cases where the data cloud contains a few pronounced gene clusters on the surface and less structure in the central part, the ICA modes will be attracted by those clusters. Relationships between clustering and ICA have been described by Hyvärinen (1998).

If data sets contain much more samples than genes, ICA may be applied to the cell sample profiles using the transposed data matrix X^{T}. In this case, the ICA model states that the modes themselves assume their values independently, while the influences might be correlated. Such expression modes may be interpreted as signals with optimal coding properties.

RESULTS

We applied ICA to publicly available sets of microarray data. In the experiments studied, cDNA (reverse-transcribed mRNA) populations from the sample being studied and from a reference sample were stained with different fluorescent dyes and both hybridized to the same chip. The gene expression matrix X contains the log-ratios $X_{ik} = \log_2(R_{ik}/G_{ik})$ between the red (experiment) and green (reference) intensities. As the mean values for genes and samples mainly reflect the experimental procedure, we shifted them to zero and then replaced the missing values by zeros. For the independent components, we adopted the following conventions:

(1) Assuming that some of the components were of biological significance while others represented noise, we sorted them in order to discriminate roughly between those groups. When compared to noise, the biological components should be more informative, showing a large contrast J_G, and they should also capture a higher amount J_A of the data variance. With centered data and components scaled to unit variance, the variance explained by component k is proportional to $J_A(k) = \sum_l A_{kl}^2$. To take both properties into account, and without considering a biological meaning behind the exact order, we sorted the components according to a linear combination

$$s^{(k)} = cJ_G^{(k)}/\langle J_G\rangle + (1-c)J_A^{(k)}/\langle J_A\rangle$$

of both quantities, scaled by their mean values, with some arbitrary $c \in [0, 1]$.

(2) For each component, the sign was chosen such that the mean influence was higher than the median. Accordingly, a mode will rather induce than repress genes, which is of course not more than a convention: when a mode is downregulated, the genes repressed by it are upregulated.

(3) By setting the gene mean values to zero, we implicitly shifted the mode mean values to zero as well, although one could also have shifted the modes to zero for a chosen reference sample.

For the whole data sets, performing an ICA took about 4 min on a 900 MHz PC, and about 8 s when projecting the data to the first 10 principal components. The dominant modes were quite reproducible, whereas their order varied in some cases.

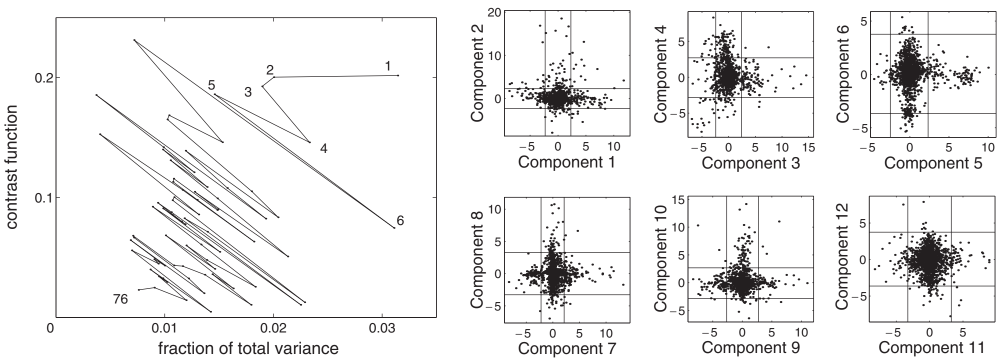

Fig. 3. ICA of cell cycle data (Spellman *et al.*, 1998). Left: sorting the 76 independent components. Each component is characterized by the fraction of the data variance it captures (abscissa) and by the contrast J_G (ordinate) measuring the non-normality of its distribution. J_G indicates, among other things, the presence of outliers. The components are connected by lines to indicate their order according to a linear combination of both quantities. Components with small values on both axes are likely to represent noise. Right: the first 12 independent components (columns of S). Each panel shows the values of two subsequent components plotted against each other, with the genes represented by dots. In our interpretation, the components are related to unobserved variables called 'expression modes', describing the modes' linear influences on the genes. For each component, outliers from the normal distribution (thresholds shown as lines) are regarded as highly induced or repressed genes.

Yeast cell cycle data

We applied ICA to data from Spellman *et al.* (1998) who studied the expression of 6178 Open Reading Frames (ORFs) during the cell replication cycle in the budding yeast *Saccharomyces cerevisiae*. Within separate experiments, cell cultures were synchronized with different methods: addition of the α mating pheromone, which arrests cells in G1 phase, blocking of the cell cycle regulators Cdc15 and Cdc28 (Cho *et al.*, 1998), and selection of small G1 cells. Besides, the effects of two cyclins were investigated: Cln3 induces the 'start' transition from G1 phase to S phase, when budding and DNA synthesis take place, and Clb2 induces progress through mitosis (M phase), involving separation of the chromosomes and cell division.

The data set (http://cellcycle-www.stanford.edu/) contains 77 samples in total, but shifting the gene mean values to zero confined the data to a 76-dimensional subspace. We ordered the independent components as described above (see Figure 3, left panel) with $c = 0.5$, putting similar weight on variance and contrast. The first 12 components are shown in the right panel of Figure 3.

For each component k, sets of induced and repressed genes were determined by the following iterative procedure: the gene with the largest absolute influence value $\max_i(|S_{ik}|)$ was considered as an outlier and excluded until all remaining values were situated within n_σ standard deviations from their median. Thus, each mode defined two groups of genes that showed a strong positive or negative response. Thus, we used the components to determine differentially expressed genes, in analogy to using the log-ratios between two samples. Setting $n_\sigma = 4$, we found 2546 genes to be strongly influenced by some mode, while a number of about 40 would be expected from normally distributed data. Due to their high contrast J_G, the dominant modes defined large sets of target genes, which often contained subgroups related to particular biological functions, mostly consistent with the mode's profile over the samples (see Table 1 for a selection of modes). However, all sets contained also considerable numbers of genes with unrelated functions. The genes corresponding to lower-scoring modes generally did not share any obvious biological roles.

Cell-cycle behaviour is mainly manifested by the modes 1, 2, and 4, which show a periodic behaviour with a slow decay in amplitude, possibly due to desynchronization (see Figure 4). Mode 1, which oscillates between M and S phase, is induced by Clb2 and repressed by Cln3, while mode 4 peaks in early G1 and does not respond in the cyclin experiments. Mode 2 is also active in G1, but remains rather weak during the first cell cycle round in the α and *cdc28* experiment, and it appears shifted to M-phase during the elutriation time-course. In contrast to mode 4, it has a larger influence on metabolic genes than on essential cell-cycle processes. Mode 3, which reflects the response to the mating α factor, decays during the G1 phase after α release. Many modes are activated specifically in some of the experiments, or even in single samples. For instance, mode 5 seems to represent an induced protein production in one of the *cdc28* samples and might be filtered out as an experimental artefact.

Table 1. Selected modes from the cell cycle data. For each mode, target genes were selected as shown in Figure 2. The modes were characterized according to functionally related groups among these genes. Some of the mode profiles are described in brackets, with cell-cycle oscillations indicated by a dot. Detailed information is available on the web

	Description	Induced functions	Repressed functions
1	Mitosis versus replication ●	M cyclins, mitosis, MCM complex, cytoskeleton, cell wall, stress, mating cascade, H^+-transport, galactose, secreted acid phosphatases	S phase cyclins, DNA replication, histones, spindle pole duplication, bud emergence, cell wall
2	G1 ●	G1/S cyclins, stress, mating, cell wall, lipid production	Energy and amino acid metabolism
3	Mating response	Mating, cell wall, metabolism	G2/M and S cyclins, histones, stress, metabolism
4	Replication/budding versus separation ●	G1/S cyclins, MCM, DNA replication/repair, chromatin, subtelomerically encoded genes	G2/M cyclins, histones, cell wall
5	Translation	Ribosomal, proteins, sugar metabolism	Ribosomal
6	Growth	Cell wall, sugar	RNA processing
7	Sporulation ●	Sporulation, proteins, metabolism	Meiosis-specific
10	(Single *cdc15* sample)	Meiosis, proteins	
11	(Decrease in elutriation experiment)	Stress, metabolism, Cu/Fe transport	Cyclins
14	Galactose ●	Galactose metabolism	Hexose transport, sugar
15	(● During *cdc15, cdc28*)	Galactose, protein targeting	Stress
16	(Rising during *cdc15*)	Mating α type, stress	
18	(Single *cdc15* sample)	Meiosis, proteins	
19	Late mating response	Mating, meisosis, proteins, metabolism	
21	Ribosomes	Ribosomes	
22	Oxidative/osmotic stress		Oxidative/osmotic stress, sugar
23	Ribosomes (falling in elutriation)		Ribosomes, translation
24	Stress		Stress
25	Methionin ●	Methionin metabolism	Sugar

Alter *et al.* (2000) reanalyzed the data by applying singular value decomposition to the separate experiments. To compare the PCA and ICA approaches, we generated PCA modes from the data set as a whole (shown in the web supplement). With both methods, most of the cell-cycle behaviour is captured by a small number of modes. However, we found that the separation into oscillatory, spiky, and noise-like patterns was more distinct with ICA. Besides, the dominant PCA modes varied during all time-courses, while various ICA modes seemed to remain inactive within some of the experiments, which was not forced by the method itself. In contrast to ICA, PCA could identify weak components that remain constant within the four experiments, varying only between them. They may be related to different experimental conditions, reference samples, and normalization schemes.

Dimension reduction

Dimension reduction may be useful to compress data sets before further calculation-intensive study. Assuming that cell cycle behaviour is sufficiently captured by the modes 1, 2, and 4, one may omit the remaining modes, thereby compressing the data from 76 to 3 dimensions. Such a projection to biologically relevant directions should improve predictions of cell-cycle regulated genes from the expression data. In order to test this, we studied a list of 551 genes that are controlled by known cell-cycle promoter elements (taken from the web supplement of Spellman *et al.*, 1998). Scoring all ORFs by the variance of their expression levels over the samples, we predicted the n_{pred} highest-scoring genes to be contained in the list. Figure 5 shows the number of successful predictions as a function of n_{pred}, based on the original data as well as on different kinds of filtered data: projecting the

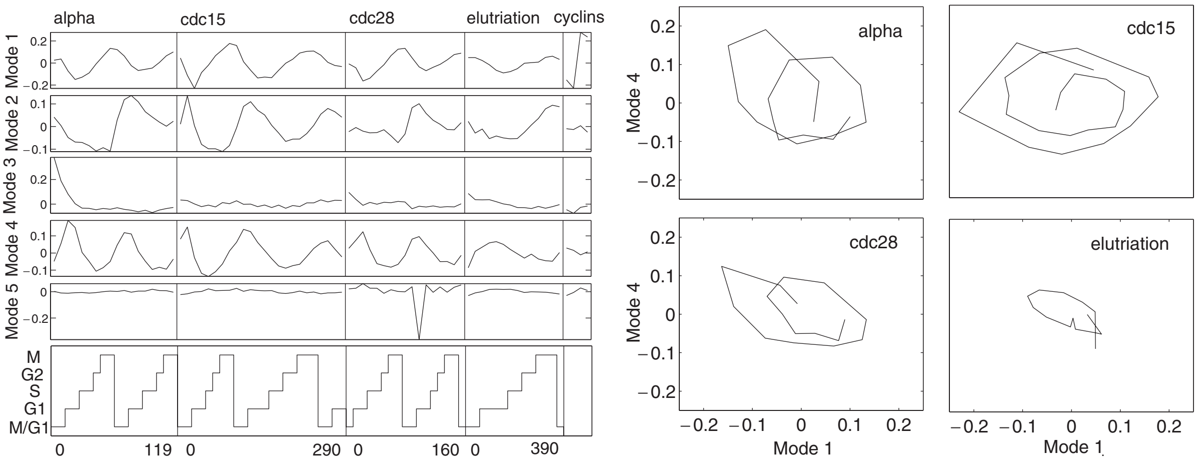

Fig. 4. Expression mode profiles (rows of *A*) calculated from the cell cycle experiments. Left: levels of the first 5 modes. The samples shown on the abscissa represent time-courses following different methods of cell synchronization (mating α factor, *cdc15*, *cdc28*, sorting by elutriation), as well as the activation of cyclins Cln3 and Clb2 (two samples each). The numbers indicate duration in minutes. Cell cycle phases are indicated in the lower panel. The corresponding sets of target genes (see Figure 2) confirm that modes 1, 2, and 4 are related to the cell cycle, while the modes 3 and 5 correspond to the mating response and to protein translation, respectively. Right: expression modes 1 and 4, plotted against each other. The four experiments are shown in separate panels. Samples are joined to indicate their time order.

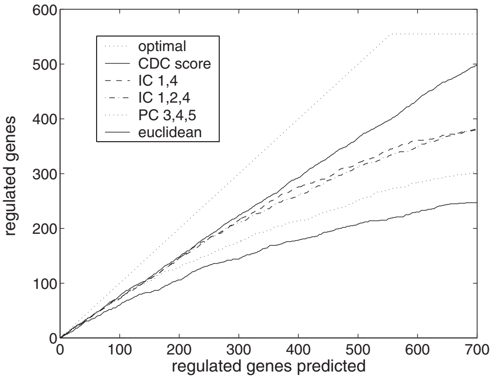

Fig. 5. Filtering the expression data using ICA improves a prediction of cell-cycle regulated genes. We scored the yeast ORFs by different methods to predict 551 genes controlled by a number of known cell-cycle promoter elements (see text). The reliability of each prediction method was assessed by plotting the numbers of successful predictions versus the number of genes predicted. Upper dotted line: perfect prediction. Lower solid curve: genes were scored by the variance of the gene expression profiles. Replacing the gene profiles by the influence values of cell-cycle-related ICA modes 1, 2, and 4 (dashed–dotted curve) or 1 and 4 (dashed curve) improved the prediction. Projecting the data to the most cell-cycle related principal components 3–5 (lower dotted curve) performed less well. The ‘aggregate cdc score’ (Spellman *et al.*, 1998) yielded the best results (upper solid curve).

profiles to the cell-cycle related principal components 3–5 improved the prediction considerably, and replacing the gene expression profiles by the influence values of the cell-cycle related ICA modes had an even larger effect. The best prediction was achieved using the ‘aggregate cdc score’ (Spellman *et al.*, 1998), which compares the gene expression profiles to sine and cosine waves and to the profiles of known cell-cycle regulated genes.

B-cell lymphoma data

We applied ICA to a second data set related to different cell types rather than to time-courses. Alizadeh *et al.* (2000) investigated the expression of 4026 human genes in 96 samples of normal and malignant lymphocytes. The ‘lymphochip’ used in this study contains clones from lymphoid cDNA libraries as well as genes related to immune-response and oncogenesis. The samples included T-cells, activated blood B-cells, B-cells from the Germinal Centre (GC), six leukemia cell lines (WSU1, Jurkat, U937, OCI Ly12, OCI Ly13.2, SUDHL5), and cells from three types of lymphomas: Follicular Lymphoma (FL), Chronic Lymphocytic Leukemia (CLL), and Diffuse Large B-cell Lymphoma (DLBCL). We downloaded the data from http://llmpp.nih.gov/lymphoma/ and analyzed them as described above. The cell samples are visualized in Figure 6 by scatter-plotting the first 12 expression modes (see also Table 2). We compared the modes to the gene clusters that had been determined in the original work using hierarchical clustering. Although the clusters and modes are not equivalent, as the modes describe additive effects, we found some relations between them: modes 2 and 5, which show the highest variance among the modes, point towards the ‘proliferation’ and ‘lymph node’ gene clusters , while mode 8 and 12 are related to

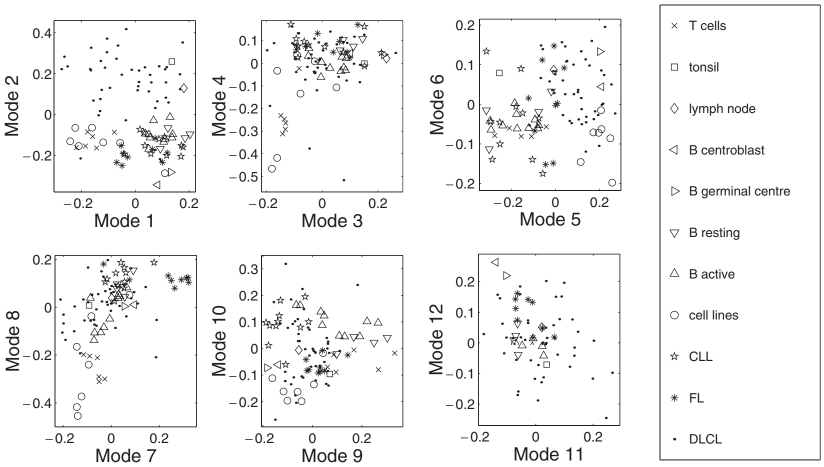

Fig. 6. Cell samples from the lymphoma data set (Alizadeh *et al.*, 2000). The axes represent levels of the first 12 expression modes, each panel showing a projection to two subsequent modes (rows of A). In some of the projections, clusters of cell types (indicated by different symbols) become visible. A description of the modes in terms of related cellular functions is given in Table 2.

Table 2. The first 12 expression modes inferred from the lymphoma data. Modes are characterized by the cell types in which they are most up- /downregulated (compare Figure 6) and by functions of their target genes. More information is available on the web

	Mode	Upregulated in	Downregulated in	Functions induced	Functions repressed
1	B-cell activation	Lymph node, tonsil, blood B, CLL, SUDHL6	T-cells, Jurkat, U937, OCI	Immunoglobulins, differentiation	
2	Lymph node	DLBCL, lymph node, tonsil		Interferon-induced genes, activation, defense	
3		Lymph node, tonsil, GC	T-cells, Jurkat, U937, OCI	Immunoglobulins	
4	MHC		T-cells, Jurkat, U937	MHC	
5	Proliferation	DLCL, cell lines, GC	T-cells, active B CLL, tonsil	Cell cycle	Interferon-inducible
6		DLCL, GC tonsil, lymph node		Immunoglobulins	
7	FL	FL		Anti-proliferative	
8	B versus T-cells	CLL, FL	Jurkat, OCI, T-cells	B receptors	T receptors
9		Blood B, T-cells SUDHL6, Jurkat, U937	GC, CLL	Adhesion, proliferation, shock, signaling	
10		Blood B, CLL	Cell lines	B receptors	
11	T activation	T-cells	Active B, FL, CLL	T activation, chemokines, T receptors (CD3), interferon-inducible genes	Adhesion
12	GC	GC, FL	OCI	B activation	Homing

the ‘pan B-cell’ and the ‘germinal center B-cell’ cluster, respectively. The authors of Alizadeh *et al.* (2000) stated that genes from ‘T-cell signature’ appearing in DLBCL samples indicated the presence of T-cells in the biopsies. Mode 11, which is related to this cluster, may be expected to describe the contamination with T-cells and might be filtered out to correct the DLBCL expression patterns for this particular effect.

DISCUSSION

As increasing amounts of gene expression data become available, there is a growing need for visualization tools that reduce the data to their most relevant aspects. At the moment, the most widely used method for organizing and visualizing expression data is clustering, which determines a (possibly large) number of gene groups, each showing a particular behaviour across the whole set of samples. In contrast to that, linear models explicitly describe a superposition of a smaller number of regulatory effects: the genes respond to different combinations of common input variables, and the regulatory functions are approximated by linear responses. While clustering may identify groups of genes that respond to particular sets of variables, it does not represent the combinatorial structure itself. By projecting the data to smaller subspaces spanned by ‘interesting’ modes, special aspects of the coregulation structure become highlighted, while partial information about all genes and samples is maintained. Such projections may be useful for visualization, defining problem-relevant metrics, and systematic denoising and dimension reduction. Reducing the complexity of gene regulation to a small number of key variables may also be a first step towards simple dynamic models of gene regulation.

How could such effective variables be related to detailed genetic network models? In a general setting, a cell may be described as a dynamic system with a large number N of (mostly unobserved) variables, among them the measured mRNA levels. The system is constrained to a submanifold $\mathcal{M} \subset \mathbb{R}^N$ of cell states that can occur in the experiment studied. In the neighbourhood of some reference state, this manifold $\mathcal{M}$ may be approximated by a hyperplane M_T of low dimensionality q, which can be parametrized by some set of coordinates. With p samples, M_T needs to have $q = p - 1$ dimensions at most, but one may reduce the number of dimensions even further and try to explain the remaining variation by experimental noise. The remaining coordinates (‘expression modes’) describe common variations of the cell variables, for instance a whole pathway of interacting signaling molecules rather than a single particular substance. As long as the complete set of cell variables has not been defined, we cannot determine how the hyperplane is embedded in the cell-model space $\mathbb{R}^N$. However, linear models may be used to define expression modes by their values across the samples, based on their linear effects on the mRNA levels.

ACKNOWLEDGEMENTS

We thank A.Hyvärinen for providing the FastICA algorithm and D.Holste for stimulating discussions. This work was supported by a grant of the Deutsche Forschungsgemeinschaft (graduate programme 286).

REFERENCES

Alizadeh,A.A., Eisen,M.B. *et al.* (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10 101–10 106.

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

Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nature Genet.*, **27**, 167–171.

Cho,R. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Cover,T.M. and Thomas,J.A. (1991) *Elements of Information*. Wiley, New York.

D’haeseleer,P. *et al.* (1999) Linear modeling of mRNA expression levels during CNS development and injury. *Pac. Symp. Biocomput.*, **4**, 41–52.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Friedman,N., Linial,M., Nachman,I. and Pe’er,D. (2000) Using bayesian networks to analyze expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology on RECOMB 2000*. pp. 127–135.

Hastie,T., Tibshirani,R. *et al.* (2000) ‘Gene shaving’ as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**, research 0003.1–0003.21.

Holter,N.S., Maritan,A., Cieplak,M., Fedoroff,N. and Banavar,J.R. (2001) Dynamic modeling of gene expression data. *Proc. Natl Acad. Sci. USA*, **98**, 1693–1698.

Hyvärinen,A. (1998) Independent component analysis in the presence of Gaussian noise by maximizing joint likelihood. *Neurocomputing*, **22**, 49–67.

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

Hyvärinen,A., Karhunen,J. and Oja,E. (2001) *Independent Component Analysis*. Wiley, New York.

Lazzeroni,L. and Owen,A. (2000) Plaid models for gene expression data, *Technical Report*, Stanford Biostatistics Series.

Liang,S., Fuhrman,S. and Somogyi,R. (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, **3**, 18–29.

Raychaudhuri,S., Stuart,J.M. and Altman,R.B. (2000) Principal component analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.*, **5**, 455–466.

Spellman,P.T., Sherlock,G. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridiation. *Mol. Biol. Cell*, **9**, 3273–3297.

Tamayo,P., Slonim,D. *et al.* (1999) Interpreting patterns of gene expression with self-organising maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Wahde,M. and Hertz,J. (2000) Coarse-grained reverse engineering of genetic regulatory networks. *BioSystems*, **55**, 129–136.

Weaver,D., Workman,C. and Stormo,G. (1999) Modeling regulatory networks with weight matrices. *Pac. Symp. Biocomput.*, **4**, 112–123.

Wen,X. *et al.* (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA*, **95**, 334–339.

Yuh,C.-H., Bolouri,H. and Davidson,E.H. (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279**, 1896–1902.