

Extracting information from cDNA arrays

Hanspeter Herzel,^{a)} Dieter Beule, Szymon Kielbasa, and Jan Korbel
Institute for Theoretical Biology, Humboldt-University, Invalidenstr. 43, D-10115 Berlin, Germany

Christine Sers
*Institute of Pathology, University Hospital Charité, Humboldt-University, Schumannstr. 20/21,
 D-10117 Berlin, Germany*

Arif Malik, Holger Eickhoff, and Hans Lehrach
Max-Planck-Institute for Molecular Genetics, Ihnestr. 73, D-14195 Berlin, Germany

Johannes Schuchhardt
Institute for Theoretical Biology, Humboldt-University, Invalidenstr. 43, D-10115 Berlin, Germany

(Received 17 July 2000; accepted for publication 7 November 2000)

High-density DNA arrays allow measurements of gene expression levels (messenger RNA abundance) for thousands of genes simultaneously. We analyze arrays with spotted cDNA used in monitoring of expression profiles. A dilution series of a mouse liver probe is deployed to quantify the reproducibility of expression measurements. Saturation effects limit the accessible signal range at high intensities. Additive noise and outshining from neighboring spots dominate at low intensities. For repeated measurements on the same filter and filter-to-filter comparisons correlation coefficients of 0.98 are found. Next we consider the clustering of gene expression time series from stimulated human fibroblasts which aims at finding co-regulated genes. We analyze how preprocessing, the distance measure, and the clustering algorithm affect the resulting clusters. Finally we discuss algorithms for the identification of transcription factor binding sites from clusters of co-regulated genes. © 2001 American Institute of Physics. [DOI: 10.1063/1.1336843]

In the last decade the DNA-chip (or microarray) technology has been developed (see *The chipping forecast* Nature Genetics Suppl. 21, 1999 for a review). The technology is based on the strong binding affinities of DNA or RNA molecules to its complementary strand. On a planar surface a library of DNA molecules is fixed as a target. (There is no general consensus on the usage of the terms probe and target. In this paper we will use target for the library on the array and probe for the complex mRNA mixture extracted from the cells and hybridized to the array.) Then a radioactive or fluorescent labeled probe from a given tissue or cell line is hybridized with the target molecules. Complementary strands of target and probe molecules bind to each other and after a washing procedure the amount of fixed probe molecules can be measured using fluorescence or radioactivity. The current technologies use different ways of target fixation or probe labeling. Oligonucleotide glass chips are covered with thousand of rectangular domains containing gene specific DNA-sequences of about 25 bases.^{1,2} The mRNA probes are labeled with fluorescence markers. Alternatively, amplified genomic cDNA can be spotted on nylon filters or glass slides. Complex mRNA probes from tissue or cell lines are reverse transcribed to cDNA and labeled with red or green fluorescent dyes³ or with radioactive markers.⁴ After hybridization with the immobilized cDNA library the fluorescence or radioactivity of each

spot quantifies the amount of mRNA present in the original probe. The technologies allow the simultaneous measurement of thousands of mRNA concentrations. Monitoring the time dependence of expression levels can reveal potentially the structure and dynamics of complex gene regulatory networks. Approaching this goal requires the following first steps: (1) Critical assessment of data reliability (image analysis, normalization, calibration, reproducibility), (2) identification of co-regulated genes by cluster analysis, and (3) detection of regulatory mechanisms (promoters, enhancers, silencers). The resulting information can then be incorporated into network models. In this paper we discuss the three topics mentioned above using data obtained from cDNA arrays and publicly available gene expression time series.

I. INTRODUCTION

A long-term goal of high-throughput measurements in molecular biology is the reconstruction of complex gene control networks. Models of transcriptional regulation and of intracellular biochemical networks will lead to a better understanding of signaling pathways, cell-cycle dynamics, carcinogenesis, the effect of drugs, or the role of single nucleotide polymorphisms (SNPs).⁵⁻⁹ Currently detailed models of specific processes are being developed such as prokaryotic genetic circuits,¹⁰ glycolysis,^{7,11} cAMP signaling,⁸ calcium oscillations,^{12,13} bacterial chemotaxis,^{14,15} signaling cascades,¹⁶⁻¹⁸ cell cycle dynamics,¹⁹ morphogenesis,²⁰⁻²³ or the circadian clock.²⁴ Even in these specific subsystems a

^{a)}Electronic mail: h.herzel@biologie.hu-berlin.de

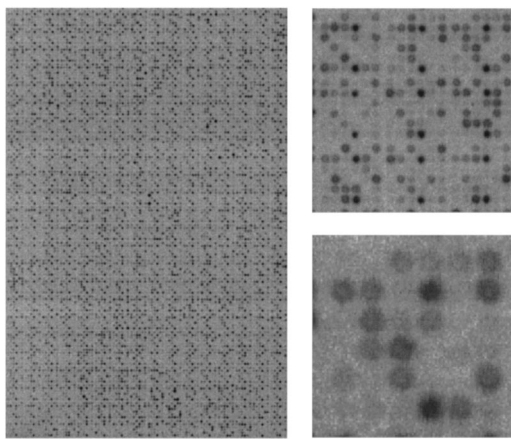


FIG. 1. Image of a cDNA microarray after hybridization with a radioactively labeled complex probe. The spotting pattern is organized in blocks of 6×6 spots. The upper enlargement shows an area of 3×3 blocks. A single block (below) contains double spots of the probe and control spots. In order to obtain a better resolution and a wider range the intensities are plotted on a logarithmic scale.

large number of interacting components are involved and the estimation of kinetic constants remains a serious problem. In the past decades information about regulatory networks has been obtained typically by single gene experiments such as knockouts, mutagenesis, or over-expression.

The situation has changed drastically in the last few years. Complete genomes are available and DNA-chips allow simultaneous measurement of thousands of mRNA concentrations. Furthermore, yeast-two-hybrid systems reveal protein-protein interactions²⁵ and protein abundances can be measured using two-dimensional gel electrophoresis.²⁶ These high-throughput technologies generate huge amounts of data and there is hope that the measurements can be used to reconstruct large regulatory networks. However, this *reverse engineering* approach requires a careful evaluation of the quality of the data.

In this paper we discuss the first steps of reverse engineering from large-scale mRNA measurements: Image analysis, normalization, reproducibility, identification of co-regulated gene clusters, and the detection of transcription factor binding sites. We focus on our data from cDNA arrays hybridized with radioactively labeled probes. Most of the methods, however, can be easily transferred to other technologies such as fluorescence labeling³ or oligo-nucleotide arrays.¹

II. RELIABILITY OF MICROARRAYS

A. Data acquisition

We present in this paper results of complex cDNA hybridizations using high-density arrays spotted on nylon filters by robots using a 16×24 pin matrix. The cDNA microarrays contain 16×24 blocks containing 6×6 spots each (compare Fig. 1). Every block contains 14 double spotted mouse clones and 8 control spots. In this way a mouse unigene library of 5376 clones is studied to determine tissue specific gene expression profiles.²⁷ Here we analyze in detail data from a mice liver probe ($1.5 \mu\text{g}$ mRNA) in order to assess

the reliability of the measurements. Arrays were scanned with a phosphor imager (Fuji BAS 5000, Raytest Germany).

The preparation of probes, targets and arrays lead to different statistical and systematic errors:

- (i) sample to sample fluctuations of the mRNA preparation;
- (ii) varying reverse transcription to cDNA and varying PCR amplification;
- (iii) fluctuations in labeling;
- (iv) spotting variations due to pin geometry and transported target;
- (v) varying target fixation, filter or hybridization inhomogeneities;
- (vi) dependence on hybridization parameters;
- (vii) unspecific hybridization or cross-hybridization within gene families;
- (viii) background noise and outshining from neighboring spots;
- (ix) image analysis (saturation, variations in spot shape).

Since in general the correct spot intensities are unknown, particular strategies have to be developed to quantify statistical and systematic errors in complex hybridization experiments. We apply the following methods (see Ref. 28 for details):

- (i) comparison of double spots (quantifying the variability for the same array and the same pin);
- (ii) analysis of control spots with *Arabidopsis thaliana* clones (variability from pin to pin, variations across the filter);
- (iii) reproducibility on different filters (eight different arrays have been compared²⁸);
- (iv) analysis of empty background spots (unspecific noise, outshining);
- (v) dilution series of the target (in Ref. 28 *Arabidopsis* controls were diluted by factors of 2, 4, 8, 16, and 32);
- (vi) dilution series of probes (see below).

Even though the correct mRNA abundance of each gene is unknown, a dilution of the probe should result in a well-defined scaling of the intensity by the corresponding dilution factor. In particular, dilution series allow reproducibility studies over a wide range of intensities. In this paper we analyze (the data are available on request) six filters with probe dilution factors 1, 2, 4, 8, 16, and 50.

B. Image analysis

The high density of spots, a wide range of spot intensities, varying spot shapes, inhomogeneities of filters or slides, and various artifacts such as fingerprints or scratches make an automatic image analysis difficult. The following steps are required for the calculation of expression levels:

- (i) grid finding and localization of spots;
- (ii) quantification of their intensities;
- (iii) correction of background and outshining;
- (iv) elimination of artifacts.

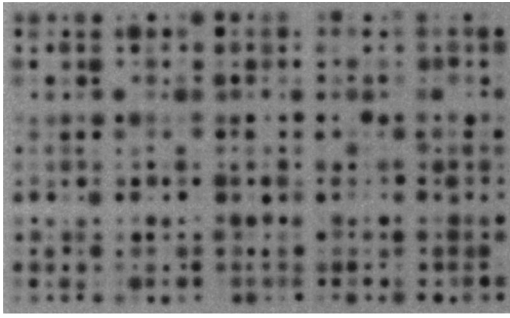


FIG. 2. Computer-generated surrogate image for the evaluation of image analysis software (see text for details). Again we plot intensities on a logarithmic scale.

The assessment of the quality of any image analysis is difficult since the true spot intensities are typically not known. Therefore, we developed a model to generate surrogate images for test purposes (Fig. 2). The Gaussian shape of the spots is blurred due to the finite range of the radiation (^{33}P). This process is modeled by a set of random walkers with constant energy deposition and exponentially distributed range. The overall spot intensities (number of walkers) are chosen from an exponential distribution, fluctuations of the spot locations and background noise are added. The scatter plots of Fig. 3 indicate that a reliable automatic extraction of spot intensities is possible even for relatively noisy images using proprietary software (GeneSpotter, MicroDiscovery Berlin).

C. Double spots on the same filter

Due to the uncertainties of the whole hybridization procedure as discussed above control experiments are necessary. Consequently hybridization is performed on several filters or slides in parallel.^{27,28} Moreover each clone is spotted twice on each filter. The two spot intensities, termed A and B in the following, allow the quantification of statistical errors within the blocks. Note that each block is spotted by the

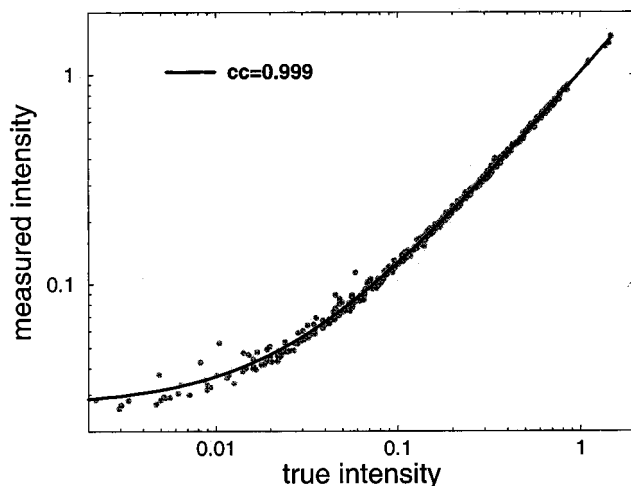


FIG. 3. Results of image analysis applied to Fig. 2. The scatter plot shows the successful extraction of the true spot intensities for high and medium intensities. For small intensities additive noise and outshining effects limit the reliability of the measured signals.

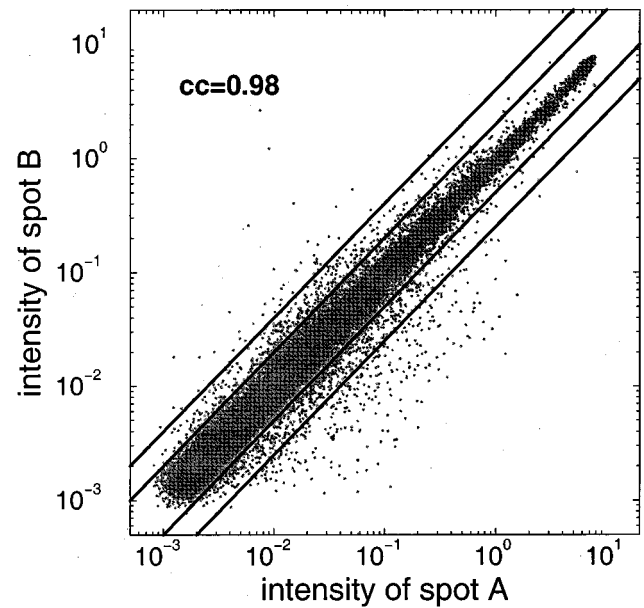


FIG. 4. Comparison of double spots on the six filters of the dilution series from a mouse liver probe (about 36 000 pairs). The correlation coefficient in the logarithmic presentation is 0.98. The lines parallel to the diagonal mark deviations by factors 2 and 4.

same pin. Consequently, pin-to-pin variability cannot be analyzed by such duplicates. A compensation of pin-to-pin variations has been achieved with *Arabidopsis thaliana* control spots as discussed elsewhere.²⁸

Figure 4 shows the reproducibility of duplicates for the six filters from the dilution series (cf. Sec. II A). Signal intensities range over three orders of magnitude. The nearly constant width of the scatter plot on the double-logarithmic scales indicates primarily multiplicative noise. In Fig. 5 we plot the relative fluctuations $|A - B| / (A + B)$ versus the mean intensity $(A + B) / 2$. A running median reveals that the fluctuations are below 10% over almost the whole range.

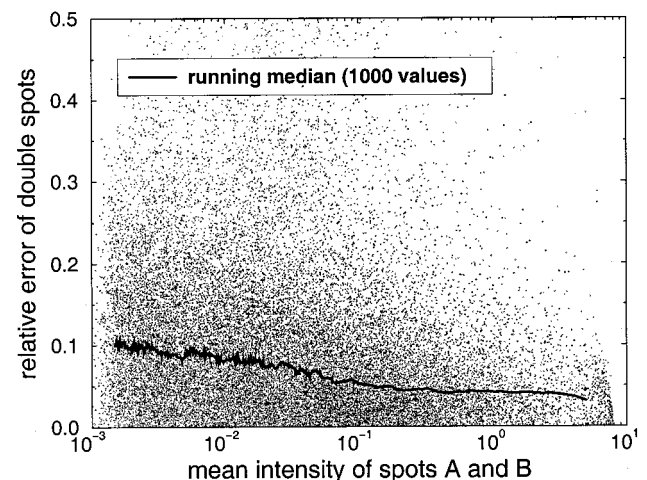


FIG. 5. Dependence of the relative errors $|A - B| / (A + B)$ of the double spots in Fig. 4 on the mean intensity $(A + B) / 2$. The running median indicates accuracies between 5% and 10%.

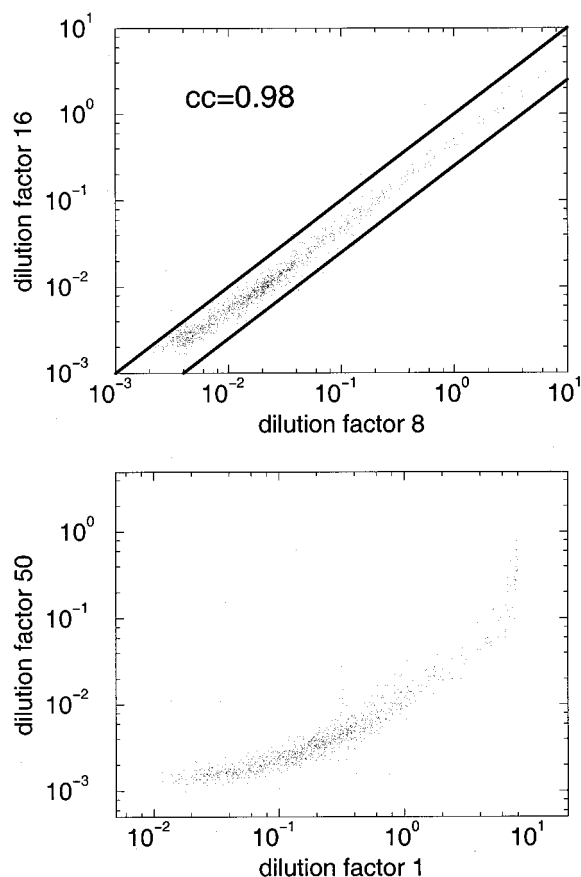


FIG. 6. Comparison of spot intensities from different filters of the dilution series. The upper graph shows that for intermediate intensities the expected ratio of 2 is found. The parallel lines mark deviations by a factor of 2 from the expected scaling. The lower graph shows distortions of the expected scaling due to saturation at large intensities and noise for small intensities.

D. Comparison of different filters

Even though the true spot intensities of our mouse clones are unknown the dilution series allows filter-to-filter comparisons since we can assume that the corresponding intensities scale by the chosen dilution factor. The scatter plots in Fig. 6 show the raw data, i.e., no normalization or background correction has been applied. It turns out that for intermediate dilution levels the signal from the eight times diluted probe is approximately twice the intensity of the 16 times diluted probe. The correlation coefficient of 0.98 indicates a fairly good filter-to-filter reproducibility. The median of the relative errors is below 7%. However, at small and large intensities systematic deviations from a straight line are visible. These effects are even more pronounced in the lower graph of Fig. 6 where the undiluted probe versus the maximally diluted probe (factor 50) is plotted. Here saturation effects (cutoff at 8.0) and additive background noise of the highly diluted probe are visible.

These results demonstrate that dilution series are an appropriate tool to evaluate the reliability of array data, to quantify statistical fluctuations, and to point to systematic artifacts. The calibration of the measurements using *Arabidopsis thaliana* clones and selected genes from the dilution series will be discussed elsewhere.

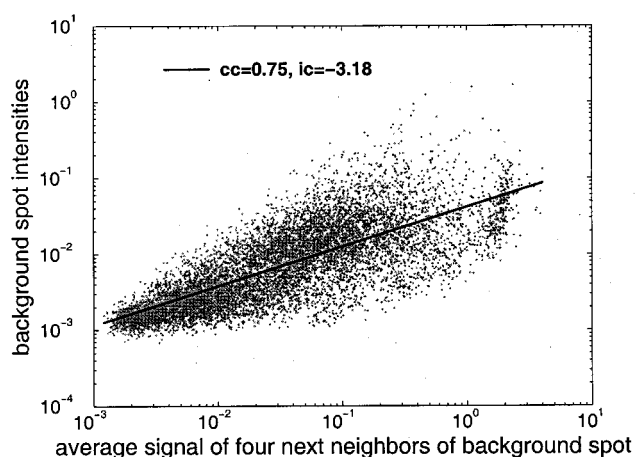


FIG. 7. Scatter plot of background spot intensities versus the average intensity of the neighboring spots. The correlation indicates outshining effect and the intercept points to an additive background noise below 10^{-3} .

E. Background and outshining

High density cDNA arrays with radioactively labeled probes achieve a remarkable sensitivity compared to other technologies.²⁹ However, scattering induces background noise and outshining. As shown in Fig. 6 these effects delimit the accuracy at low signal intensities. Fortunately, these artifacts can be easily quantified and partially corrected using empty control spots. Since each block contains eight of these controls (cf. setup in Fig. 1) we can quantify their statistical properties. In Fig. 7 the background versus the mean of the intensity of the neighboring spots is shown for the six dilution series. It turns out that there is a clear correlation as a result of outshining. Moreover, the axis intercept around 10^{-3} indicates additive background noise. The regression line provides an empirical model of background noise and outshining. Since this fit is reasonable over the whole range of intensities it can be assumed that it applies also to the mouse clones. Thus it can be exploited to reduce background noise and outshining for the clones of interest.

III. CLUSTERING GENE EXPRESSION TIME SERIES

As shown in the previous section array measurements are inherently noisy and even restricting ourselves to changes by at least a factor of 2, false positives will be obtained. These limitations apply to all current technologies.³⁰⁻³⁴ Nevertheless, cluster analysis is widely used to extract the relevant features of large-scale expression measurements.^{35,36} Clustering can be regarded as a powerful noise reduction procedure and leads to a compact representation of the data.

Since cluster algorithms analyze multiple measurements of many genes some general questions arise: How to preprocess the expression levels? What distance measure should be used? Which cluster algorithm should be applied? How many clusters are biologically meaningful?

A. The data

In the following we discuss publicly available gene expression time series (the data are available at [Downloaded 09 Mar 2001 to 130.149.160.103. Redistribution subject to AIP copyright, see <http://ojps.aip.org/chaos/chocpyrts.html>](http://genome-</p>
</div>
<div data-bbox=)

www.stanford.edu/serum/)—the reaction of human fibroblast cells to serum stimulation.³⁷ The fibroblasts were prepared in a nondividing state characterized by low metabolic activity and changed to a proliferating state through serum stimulation. The relative change in expression of 8613 human genes has been determined at 11 time points between 15 min and 24 h after stimulation using cDNA microarrays. An additional twelfth measurement was performed on exponentially growing fibroblasts. Typical expression ratios range from tenfold suppression to tenfold enhancement with respect to the nondividing reference state. As in Ref. 37 we analyze only those 517 genes that showed the largest changes in expression in response to serum stimulation.

Even though expression ratios are studied, the data span two orders of magnitude. If the raw data are processed the few largest ratios dominates the cluster analysis. In order to weight increase and decrease of expressions equally, the logarithm of the expression ratios were studied.³⁷

B. Clustering methods

There is no comprehensive classification of the multitude of clustering algorithms.^{38,39} The algorithms can be characterized by the partitioning result or the partitioning procedure. Even among nonoverlapping and complete partitionings one can distinguish between hierarchical and nonhierarchical partitionings that can be generated in an agglomerative or divisive way. The authors in Ref. 37 use average linkage clustering with the normalized scalar product [cf. Eq. (2)] as similarity score. They identified 10 clusters containing 484 genes of the total 517 genes. For the purpose of comparison we use the clustering results from the web supplement of Ref. 37. In this paper we restrict ourself to the non hierarchical divisive *k*-means clustering.⁴⁰ In this way we have control over the number of clusters and can compare different preprocessing procedures and distance measures easily.

C. Distance measures

Clustering requires distance or similarity measures between expression profiles **x** and **y** (in our case **x** and **y** refer to the *N*=12 measurements of two different genes). Wen et al.⁴¹ proposed the following distance *D*:

$$D(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^N (x_i - y_i)^2 + \sum_{i=1}^{N-1} (\Delta x_i - \Delta y_i)^2. \tag{1}$$

Here *i* labels the time points and $\Delta x_i = x_i - x_{i+1}$ and $\Delta y_i = y_i - y_{i+1}$ denote the changes of the expression levels.

In Ref. 37 the data were clustered using the similarity score

$$S(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^N \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right), \tag{2}$$

$$\sigma_x = \sqrt{\sum_{i=1}^N \frac{(x_i - \bar{x})^2}{N}}, \quad \sigma_y = \sqrt{\sum_{i=1}^N \frac{(y_i - \bar{y})^2}{N}}.$$

The authors of Ref. 37 set \bar{x} and \bar{y} to zero and thus *S*(**x**,**y**) becomes simply a normalized scalar product. Another sen-

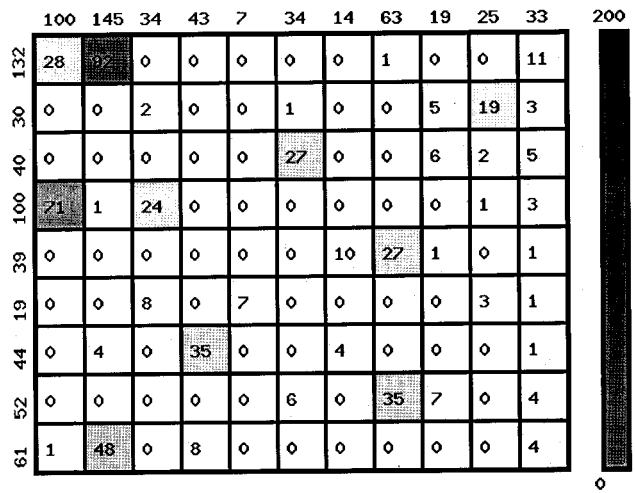


FIG. 8. Contingency table of the clusters identified by Iyer et al. (Ref. 37) by visual inspection of a dendrogram of a hierarchical clustering and the eight cluster identified in Ref. 43 by *k*-means clustering.

sible choice for \bar{x} and \bar{y} is the mean which leads to Pearson's correlation coefficient $\rho(\bar{\mathbf{x}}, \bar{\mathbf{y}})$. Both choices focus on the temporal dynamics of the expression ratios and pay little attention to the absolute size.

D. Comparison of algorithms and distance measures

First we compare the results of hierarchical clustering in Ref. 37 with *k*-means clustering. Figure 8 shows the contingency table of the partitionings. In both cases the logarithm base 2 and the distance $D(\mathbf{x}, \mathbf{y}) = 1 - S(\mathbf{x}, \mathbf{y})$ were used. It turns out that there is a reasonable agreement of both clustering results. Figure 9 shows the temporal development of eight clusters identified by *k*-means. The clusters represent rapid response to the serum stimulation (cluster 8), late up-regulation (clusters 4 and 6), peaks around six hours (clusters 5 and 7), or down-regulation at different times (clusters 1, 2, 3). Similar features are identified by the hierarchical clustering in Ref. 37 and, consequently, the agreement in Fig. 8 is not too surprising. In order to quantify the similarity of the clustering results we apply the Rand index.⁴² It gives the relative number of all possible $N(N-1)/2$ pairs of genes that are clustered in the same manner by both partitionings, i.e., both genes of a pair are assigned either to the same or to different clusters. The contingency table of Fig. 8 gives a Rand index of 0.85.

In order to interpret Rand indices we have to discuss their distribution for random contingency tables. Straightforward calculations show that for 517 genes and eight clusters random tables give Rand indices of 0.781 ± 0.001 .

Clustering of the raw data differs strongly from clusterings on the logarithmic scale. Thus the preprocessing by taking logarithms of the ratios has a strong effect on the clustering result. Contrarily, variations of chosen distance measures lead to similar clustering results. Table I shows Rand indexes from *k*-means clustering using different distance measures.

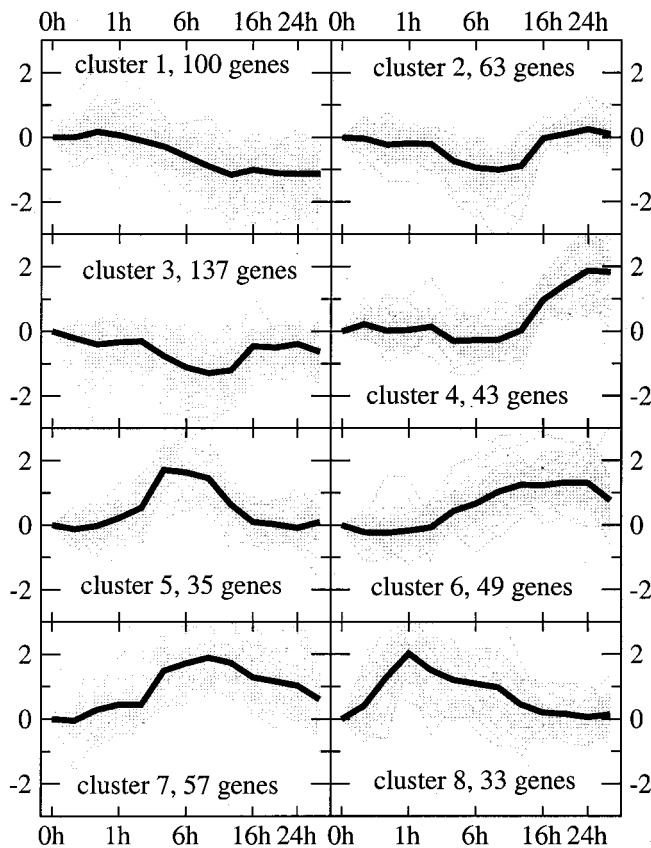


FIG. 9. Temporal development of eight different clusters identified in the fibroblast data set (\log_2 of the ratio with respect to the reference state over time). Each gene correspond to a gray line while the cluster average is shown by the thick black line.

E. How many clusters are significant?

As shown elsewhere^{35,43} random data also lead to partitionings that can be hardly distinguished from biologically meaningful clusters. Consequently, we have developed a criterion to estimate for k -means clustering the number of relevant clusters. Our approach (for details see Ref. 43) is based on the generation of random surrogate data sets. Since time series clustering exploits temporal correlations, the randomization of the time points for each single gene destroys the

TABLE I. Table of Rand indices comparing the different clusterings: (a) Hierarchical (average weighted linkage) clustering of \log_2 -ratio by Ref. 37 using normalized scalar product as similarity measure (10 clusters), (b) k -means clustering of \log_2 -ratio using normalized scalar product as similarity measure (eight clusters), (c) k -means clustering of \log_2 -ratio using Pearson correlation coefficient as similarity measure (eight clusters), (d) k -means clustering of \log_2 -ratio using Eq. (1) as distance measure (eight clusters), (e) k -means clustering of \log_2 -ratio using Euclidean distance (eight clusters).

Clustering	(a)	(b)	(c)	(d)	(e)
(a)	1	0.85	0.86	0.85	0.85
(b)		1	0.86	0.89	0.89
(c)			1	0.86	0.86
(d)				1	0.95
(e)					1

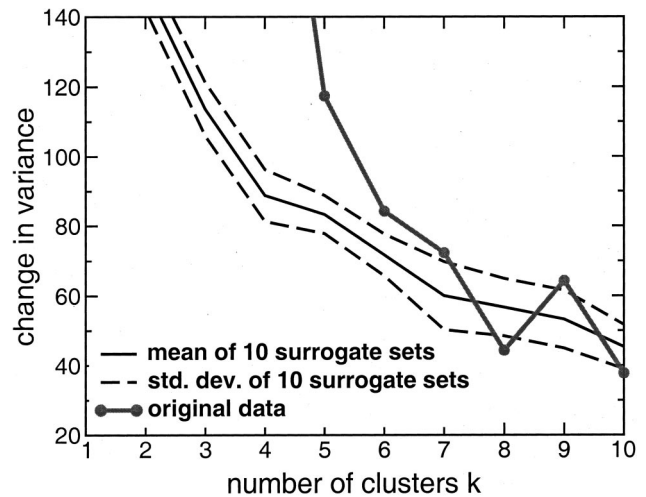


FIG. 10. Variance reduction for the fibroblast data set (solid line) and the surrogate data sets when the number of clusters is increased from k to $k + 1$. The mean variance of the surrogates and its standard deviation were estimated from 10 different surrogate sets.

relevant correlations. The resulting surrogate data share with the original data the distribution of expression levels but have to be considered random otherwise.

The k -means algorithm minimizes the internal variance V_k around k centers \mathbf{c}_j of the partitioning of N genes

$$V_k = \sum_{j=1}^k \sum_{i=1}^{N_j} D(\mathbf{x}_i, \mathbf{c}_j)^2, \quad \sum_{j=1}^k N_j = N. \quad (3)$$

The variance V_k can obviously be reduced simply by increasing the number of clusters. The variance decreases more strongly for the original data compared to the random surrogates. This fast decay of the variance can be attributed to the fact that the original data contain more structure due to temporal correlations reflecting gene regulation. The change in variance $V_k - V_{k+1}$ as plotted in Fig. 10 characterizes the gain due to the refinement of the clustering. For large k the gain for the original data exceeds no longer the gain of the surrogates. From Fig. 10 we deduce that for $k \geq 8$ the gain for the original data is not significantly different from the gain of the surrogates. Consequently, we can estimate that about eight clusters are meaningful in this case.

IV. FINDING TRANSCRIPTION FACTOR BINDING SITES IN CO-REGULATED GENES

Transcription of eukaryotic genes is mediated by a complex and extended regulatory system.⁴⁴ The transcriptional control of gene expression depends upon the specific interaction of proteins and short DNA motifs. Even though hundreds of binding sites are known⁴⁵ the detailed regulation of most genes is unknown. Under the assumption that co-regulated gene sets contain common motifs, clusters obtained from gene expression time series can be analyzed to predict transcription factor binding sites. Two different approaches are currently used to detect DNA regulatory elements: exhaustive analyses of oligo-nucleotide frequencies (see, e.g., Ref. 46) and optimization approaches using weight matrices (e.g., Gibbs sampling introduced by Ref. 47). The

TABLE II. Highly ranked hexanucleotides as computed by the program ITB are similar to the consensus sequences of previously characterized sites taken from TRANSFAC.

Family name (no. of genes)	Previously characterized element	Bound factors	ITB predict.	No. of motifs/seqs. match	Inf. cont. [bits]	Z-score	Rank
NIT(7)	GATAAG ^f	Gln3	GATAAG	26/6	7.2	13.9	1
MET ₁ (11)	TCACGTG ^c	Cbfl-Met4-Met28	CACGTG	13/11	12.8	13.6	1
MET ₂ (11)	AAAACGTGG ^c	Met31, Met32	ACYSKG	39/8	9.2	4.8	4
PHO(5)	CACGTKNG ^a	Pho4	ACGTGS	18/5	7.2	12.1	1
PDR(7)	TCCGCGGA ^b	Pdr1, Pdr3	CCGYGG	18/4	12.9	15.3	1
INO(10)	CATGTGAAWT ^j	Ino2/Opi1	CATGTG	15/9	10.5	7.9	2
TUP(25)	KANW ₄ ATSYG ₄ W ^e	Mig1	GYGGGG	33/18	8.3	11.7	1
YAP(16)	TACTAA ^a	Yap1	MTTASK	99/16	6.5	7.2	1
GAL(6)	CGGN ₅ WN ₅ CCG ^g	Gal4
GCN(38)	RTGACTCATNS ^a	Gcn4	TGACTC	44/26	7.1	12.5	1
MAT(10)	CRTGTNNW ^a	Mata2	CATGYA	21/7	6.3	5.7	3

^aReference 45.^bReference 76.^cReference 77.^dReference 78.^eReference 3.^fReference 79.^gReference 80.

latter approach is based on a multiple alignment procedure. In case of small co-regulated gene sets and weak signals this method is of limited use since often poly-A sequences or GC-rich regions are likely to be aligned by such algorithms.

A. The ITB algorithm

In this section we discuss the performance of a new algorithm ITB, an *Integrated Tool for Box-finding*. A detailed description of the algorithm can be found elsewhere.⁴⁸ The ITB algorithm exhaustively analyses regular expression-like patterns in upstream regions of genes, allowing gaps and the matching of more than one base at any position. The program looks for over-represented words from the alphabet ACGTWRKSYMN with the following meanings: W=A or T, R=A or G, K=G or T, S=C or G, Y=C or T, M=A or C, N=any of ACGT.

In order to score over-represented patterns \mathcal{W} a comparison with a training set is required. From this set the expected frequencies $n_{\text{exp}}(\mathcal{W})$ are calculated using Markov models of varying orders. A list of top-scoring motifs is created with the Z-score along the lines of⁴⁶

$$Z(\mathcal{W}) = \frac{n_{\text{obs}}(\mathcal{W}) - n_{\text{exp}}(\mathcal{W})}{\sigma_{\text{exp}}(\mathcal{W})}. \quad (4)$$

In order to calculate the standard deviations $\sigma_{\text{exp}}(\mathcal{W})$ we have to take into account self-overlaps of the motif. We adopted the expressions introduced in Ref. 49 to double-strand analysis and to the extended alphabet given above.

In the remainder of this section we consider the 10 highest Z-scores using a default word length of 6. In addition to the Z-score we give the information content to each detected motif according to Ref. 50.

B. Motifs in yeast clusters and RAS regulated genes

Since the whole yeast genome is known and many yeast expression data sets are available, the prediction of previously characterized yeast regulatory elements provides a good test case for any algorithm. We analyze 10 co-regulated

gene sets as introduced by Ref. 51. Moreover, we study another set of 10 MAT (mating-type) genes having the highest levels of change in transcription, when yeast mating types ‘‘a’’ and ‘‘α’’ are compared.⁵²

The 800 basepairs (bp) upstream regions of the yeast genes were retrieved from the MIPS database.⁵³ As a training set for the background model, a Markov chain of order 3, we use the complete set of 800 bp yeast upstream regions.

As discussed elsewhere,⁴⁸ other programs such as MEME,⁵⁴ ALIGN-ACE,⁵² or RSA-tools-oligo-analysis⁵¹ miss some of the known motifs. Table II shows that the ITB algorithm predicts 10 of 11 elements, most of them even as top scoring motif. Only the binding site of the transcription factor Gal4 regulating the GAL family was not detected. The corresponding motif CGGNNNNNWNNNNNCCG contains large gaps and cannot be found by the current version of the ITB algorithm which is biased towards compact motifs. Only ALIGN-ACE⁵² which is based on the Gibbs sampling algorithm⁴⁷ was able to detect this motif.

The identification of binding sites in mammalian genomes is much more difficult. First of all, the number of experimentally verified promoter regions is limited and the accuracy of promoter prediction *in silico* is still not satisfactory.^{55–57} Thus, our training had to rely on the 271 human promoter sequences from the eukaryotic promoter database.⁵⁸ Moreover, due to isochores,⁵⁹ poly-A sequences or CpG islands the regulatory regions are quite inhomogeneous and motifs appear also several kilobases away from the basal promoter. Consequently, the encouraging results from yeast clusters cannot be transferred easily to higher eukaryotes.

In the following we present a first attempt for a set of genes down-regulated via the H-RAS protein involving the RAF/MEK/ERK signaling cascade.

Recently, hundreds of genes have been identified by gene expression profiling^{60,61} that are up- or down-regulated by oncogenic RAS mutants. We examine promoter regions of the following five down-regulated genes: LOX, LOXL1, LOXL2, RIL/LIM, and TSP1. The ITB algorithm found the

significantly over-represented pattern CGARCG (Z-score 10.1) in the upstream regions of these genes. The pattern appears nine times in four of the promoter sequences and does not match any previously characterized site listed in the TRANSFAC database.⁴⁵ The motif was confirmed by runs with other background models (Markov chains of different order) and a search for 8-words.

In summary, we emphasize that the detection of transcription factor binding sites in higher eukaryotes is a rather complicated task. Lists of over-represented words are easily generated but the prediction of true sites requires adequate background models and scores that can suppress repeats or self-overlapping words such as poly-A or CGCGCG.

V. SUMMARY AND DISCUSSION

A. Summary

Even though microarray experiments have matured considerably in recent years, they are still inherently noisy. Since thousands of expression levels are measured in parallel, the appearance of false positives cannot be avoided. Nevertheless, microarray data provide a valuable tool for screening of many genes.

The dynamics of cluster averages is more reliable than single gene trajectories. However, clustering typically requires a preprocessing of the data and appropriate distance measures. These steps seem to be even more important than the choice of the clustering algorithm. The number of significant clusters can be estimated by analyzing randomized data.

Clusters of co-regulated genes can support the tremendously difficult analysis of eukaryotic gene regulation. Over-represented words in upstream regions of co-regulated gene sets are candidates for transcription factor binding sites. However, a careful choice of the background statistics and corrections for self-overlapping words are required to derive appropriate scores. Since microarrays are based on cDNA sequences, the corresponding upstream regions have to be extracted from genomic data. Hence the promoter analysis profits from genome sequencing projects as well as from large-scale expression analysis.

B. Outlook

The ultimate goal of microarray analysis is the detailed understanding of the entire gene regulatory network. The concept of *reverse engineering* was first discussed in connection with Boolean networks^{62,63} or continuous generalizations.⁶⁴ For such idealized models some promising results were obtained.⁶⁵ For example, networks of 160 genes have been reconstructed from less than 100 (simulated) measurements.⁶⁶ Recently, a series of 28 experiments has also been used to fit a linear model to the dynamics of 65 genes.⁶⁷

However, as discussed in this paper there are still numerous problems with microarray data. The accuracy is quite limited, different experiments are difficult to compare, and the number of measurements is rather limited even in large scale studies.^{35,68,69} Consequently, it is currently more appropriate to extract the dynamics of clusters instead of single

genes. In this way the noise present in single expression levels can be reduced and the resulting empirical models provide a compact representation of the data. For example, Wahde and Hertz⁷⁰ derived a regulatory network for four clusters derived from expression data of 112 genes from neuronal tissue.⁴¹

Even though currently regulatory networks cannot be derived solely from microarray data, existing models can be refined using expression profiles. For example hundreds of cell cycle-regulated genes have been identified using oligo-chips⁷¹ or cDNA arrays.⁷² DNA chips have been applied to classify subtypes of cancer^{73,74} and to study *Drosophila* development.⁷⁵ Moreover, signaling pathways can be explored. Recently, targets of the transcription factor MYC have been found using DNA chips.³³ In another study the MAPK pathways of yeast have been analyzed systematically with microarrays.⁶⁸ It turned out that expression profiles from 46 diverse experimental conditions provided valuable information about the architecture of the signaling network. This example illustrates that expression profiles with microarrays are particularly useful if they supplement existing models of regulatory networks.

ACKNOWLEDGMENTS

We thank I. Grosse, D. Holste, A. O. Schmitt, H. Seidel, and E. Wolski for stimulating discussions and K. Winkelhoefer for help with preparing the manuscript. We acknowledge support by the Deutsche Forschungsgemeinschaft and the German Ministry for Research and Education.

¹ D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, and E. L. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature Biotechnol.* **14**, 1675–1680 (1996).

² D. J. Lockhart and E. A. Winzeler, "Genomics, gene expression and DNA arrays," *Nature (London)* **405**, 827–836 (2000).

³ J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science* **278**, 680–686 (1997).

⁴ G. G. Lennon and H. Lehrach, "Hybridization analysis of arrayed cDNA libraries," *Trends Genet.* **7**, 314–317 (1991).

⁵ A. Arkin and J. Ross, "Computational functions in biochemical reaction networks," *Biophys. J.* **67**, 560–578 (1994).

⁶ D. Bray, "Protein molecules as computational elements in living cells," *Nature (London)* **376**, 307–312 (1995).

⁷ R. Heinrich and S. Schuster, *The Regulation of Cellular Systems* (Chapman & Hall, New York, 1996).

⁸ A. Goldbeter, *Biochemical Oscillations and Cellular Rhythms* (Cambridge University Press, Cambridge, 1996).

⁹ P. Smolen, D. A. Baxter, and J. H. Byrne, "Modelling transcriptional control of gene networks—Methods recent results, and future directions," *Bull. Math. Biol.* **62**, 247–292 (2000).

¹⁰ H. H. McAdams and A. Arkin, "Simulation of prokaryotic genetic circuits," *Annu. Rev. Biophys. Struct.* **27**, 199–224 (1998).

¹¹ P. J. Mulquoney and P. W. Kuchel, "Model of 2,3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: Equations and parameter refinement," *Biochem. J.* **342**, 581–596 (1999).

¹² G. Dupont, M. J. Berridge, and A. Goldbeter, "Signal-induced Ca^{2+} oscillations: Properties of a model based on Ca^{2+} -induced Ca^{2+} release," *Cell Calcium* **12**, 73–85 (1991).

¹³ T. Hofer, "Model of intercellular calcium oscillations in hepatocytes: Synchronization of heterogeneous cells," *Biophys. J.* **77**, 1244–1256 (1999).

¹⁴ D. Bray, R. B. Buorret, and M. Simon, "Computer simulation of the

- phosphorylation cascade controlling bacterial chemotaxis," *Mol. Biol. Cell.* **4**, 469–482 (1993).
- ¹⁵ N. Barkai and S. Leibler, "Robustness in simple biochemical networks," *Nature (London)* **387**, 913–917 (1997).
- ¹⁶ J. E. Ferrell, Jr. and R. R. Bhatt, "Mechanistic studies of the dual phosphorylation of mitogen-activated protein kinase," *J. Biol. Chem.* **272**, 19008–19016 (1997).
- ¹⁷ B. D. Aguda, "Instabilities in phosphorylation-dephosphorylation cascades and cell cycle checkpoints," *Oncogene* **18**, 2846–2851 (1999).
- ¹⁸ U. S. Bhalla and R. Iyengar, "Emergent properties of networks of biological signaling pathways," *Science* **283**, 381–387 (1999).
- ¹⁹ K. C. Chen, A. Csikasz-Nagy, B. Gyorfy, J. Val, B. Novak, and J. J. Tyson, "Kinetic analysis of a molecular model of the budding yeast cell cycle," *Mol. Biol. Cell* **11**, 369–391 (2000).
- ²⁰ H. Meinhardt, "Models for positional signalling with application to the dorsoventral patterning of insects and segregation into different cell types," *Development (Cambridge, U.K.)* **107**, 160–180 (1989).
- ²¹ J. Reintz, E. Mjolsness, and D. H. Sharp, "Model for cooperative control of positional information in *Drosophila* by bicoid and maternal hunchback," *J. Exp. Zool.* **271**, 47–56 (1995).
- ²² P. B. Singh and D. Brown, "Modelling the activity of the ultrabithorax parasegment-specific regulatory domains around their anterior boundaries," *J. Theor. Biol.* **186**, 397–413 (1997).
- ²³ L. Mendoza, D. Thieffry, and E. R. Alvarez-Buylla, "Genetic control of flower morphogenesis in *Arabidopsis thaliana*: A logical analysis," *Bioinformatics* **15**, 593–606 (1999).
- ²⁴ J.-C. Leloup and A. Goldbeter, "Modeling the molecular regulatory mechanism of circadian rhythms in *drosophila*," *BioEssays* **22**, 84–93 (2000).
- ²⁵ P. Uetz, G. Loic, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Quereshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadmodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, "A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*," *Nature (London)* **403**, 623–627 (2000).
- ²⁶ A. Pandey and M. Mann, "Proteomics to study genes and genomes," *Nature (London)* **405**, 837–846 (2000).
- ²⁷ H. Eickhoff, J. Schuchhardt, I. Ivanov, S. Meier-Ewert, J. O'Brien, A. Malik, N. Tandon, E. Wolski, E. Rohlf, R. Reinhard, W. Nietfeld, and H. Lehrach, "Tissue gene expression analysis using arrayed normalized cDNA libraries," *Genome Res.* **10**, 1230–1240 (2000).
- ²⁸ J. Schuchhardt, D. Beule, A. Malik, H. Wolski, E. Eickhoff, and H. Lehrach, "Normalization strategies for cDNA-microarrays," *Nucleic Acids Res.* **28**, e47 (2000).
- ²⁹ F. Bertucci, K. Bernard, B. Liorid, Y.-C. Chang, S. Granjeaud, D. Birnbaum, C. Nguyen, K. Peck, and B. R. Jordan, "Sensitivity issues in DNA array-based expression measurements and performance of nylon microarrays for small samples," *Hum. Mol. Genet.* **8**, 1715–1722 (1999).
- ³⁰ G. Piéto, O. Alibert, V. Guichard, B. Lamy, F. Bois, E. Leroy, R. Mariage-Samson, R. Houlgatte, P. Soularue, and C. Auffray, "Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of high density cDNA array," *Genome Res.* **6**, 492–503 (1996).
- ³¹ Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *J. Biomed. Opt.* **2**, 364–374 (1997).
- ³² S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz, "The transcriptional program of sporulation in budding yeast," *Science* **282**, 699–705 (1998).
- ³³ H. A. Collier, C. Grandori, P. Tamayo, T. Colbert, E. S. Lander, R. N. Eisenman, and T. R. Golub, "Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion," *Proc. Natl. Acad. Sci. U.S.A.* **97**, 3260–3265 (2000).
- ³⁴ D. Beule, J. Schuchhardt, A. Malik, H. Eickhoff, H. Lehrach, and H. Herzel, "Reliability of microarray data and clustering," in *Proceedings of the German Conference on Bioinformatics, Heidelberg, 2000*, pp. 167–174.
- ³⁵ M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863–14868 (1998).
- ³⁶ P. D'haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: From co-expression clustering to reverse engineering," *Bioinformatics* **16**, 707–726 (2000).
- ³⁷ V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. F. Lee, J. M. Trent, L. M. Staudt, J. Hudson, Jr., M. S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P. O. Brown, "The transcriptional program in the response of human fibroblasts to serum," *Science* **283**, 83–87 (1999).
- ³⁸ R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).
- ³⁹ B. Mirkin, *Mathematical Classification and Clustering* (Kluwer Academic, Dordrecht, 1996).
- ⁴⁰ D. Steinhausen and K. Langer, *Clusteranalyse* (Walter de Gruyter, Berlin, 1977).
- ⁴¹ X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi, "Large-scale temporal gene expression mapping of central nervous system development," *Proc. Natl. Acad. Sci. U.S.A.* **95**, 334–399 (1998).
- ⁴² W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Am. Stat. Assoc.* **66**, 846–850 (1971).
- ⁴³ D. Beule, J. Schuchhardt, and H. Herzel, "Clustering gene expression time series" (to be published).
- ⁴⁴ C.-H. Yuh, H. Bolouri, and E. H. Davidson, "Genomic cis-regulatory logic: Computational and experimental analysis of a sea urchin gene," *Science* **279**, 1896–1902 (1998).
- ⁴⁵ E. Wingender, P. Dietze, H. Karas, and R. Knuppel, "TRANSFAC: A database on transcription factors and their DNA binding sites," *Nucleic Acids Res.* **24**, 238–241 (1996).
- ⁴⁶ J. van Helden, M. del Olmo, and J. E. Perez-Ortin, "Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals," *Nucleic Acids Res.* **28**, 1000–1010 (2000).
- ⁴⁷ C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, "Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment," *Science* **262**, 208–214 (1993).
- ⁴⁸ Sz. M. Kielbasa, J. O. Korb, D. Beule, J. Schuchhardt, and H. Herzel, "Finding transcription factor binding sites in coregulated genes by exhaustive sequence search" (to be published).
- ⁴⁹ P. A. Pevzner, M. Yu. Borodovsky, and A. A. Mironov, "Linguistics of nucleotide sequences I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words," *J. Biomol. Struct. Dyn.* **6**, 1013–1026 (1989).
- ⁵⁰ T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht, "Information content of binding sites on nucleotide sequences," *J. Mol. Biol.* **188**, 415–431 (1986).
- ⁵¹ J. van Helden, B. André, and J. Collado-Vides, "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies," *J. Mol. Biol.* **281**, 827–842 (1998).
- ⁵² F. R. Roth, J. D. Hughes, P. E. Estep, and G. M. Church, "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation," *Nature Biotechnol.* **16**, 939–945 (1998).
- ⁵³ H. W. Mewes, K. Albermann, K. Heumann, S. Liebl, and F. Pfeiffer, "MIPS: A database for protein sequences, homology data and yeast genome information," *Nucleic Acids Res.* **25**, 28–30 (1997).
- ⁵⁴ T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology (AAAI, Menlo Park, California, 1994)*, pp. 28–36.
- ⁵⁵ J. W. Fickett and A. G. Hatzigeorgiou, "Eukaryotic promoter recognition," *Genome Res.* **9**, 861–878 (1997).
- ⁵⁶ M. Scherf, A. Klingenhoff, and T. Werner, "Highly specific localization of promoter regions in large genomic sequences by promoterinspector: A novel context analysis approach," *J. Mol. Biol.* **297**, 599–606 (2000).
- ⁵⁷ M. Reese, G. Hartzell, N. L. Harris, U. Ohler, J. F. Abril, and S. E. Lewis, "Genome annotation assessment in *Drosophila melanogaster*," *Genome Res.* **10**, 483–501 (2000).
- ⁵⁸ R. C. Prier, T. Junier, and P. Bucher, "The eukaryotic promoter database EPD," *Nucleic Acids Res.* **26**, 353–357 (1998).
- ⁵⁹ G. Bernardi, "The isochore organization of the human genome," *Annu. Rev. Genet.* **23**, 637–661 (1989).
- ⁶⁰ O. I. Tchernitsa, J. Zuber, C. Sers, R. Brinckmann, S. K. Britsch, V. Adams, and R. Schäfer, "Gene expression profiling of fibroblasts resistant toward oncogene-mediated transformation reveals transcription of negative growth regulators," *Oncogene* **18**, 5448–5454 (1999).
- ⁶¹ J. Zuber, O. I. Tchernitsa, B. Hinzmann, A. C. Schmitz, M. Grips, M. Hellriegel, C. Sers, A. Rosenthal, and R. Schäfer, "A genome-wide survey of RAS transformation targets," *Nature Genetics* **24**, 144–152 (2000).

- ⁶²S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *J. Theor. Biol.* **22**, 437–467 (1969).
- ⁶³S. A. Kauffman, *The Origin of Order* (Oxford University Press, Oxford, 1993).
- ⁶⁴L. Glass, "The logical analysis of continuous non-linear biochemical control networks," *J. Theor. Biol.* **39**, 103–129 (1973).
- ⁶⁵S. Liang, S. Fuhrman, and R. Somogyi, "Reveal, a general reverse engineering algorithm for inference of genetic network architectures," *Pacific Symposium on Biocomputing* **3**, 18–29 (1998).
- ⁶⁶T. Akutsu, S. Miyano, and S. Kuhara, "Algorithms for inferring qualitative models of biological networks," *Pacific Symposium on Biocomputing* **5**, 293–304 (2000).
- ⁶⁷P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, "Linear modeling of mRNA expression levels during CNS development and injury," *Pacific Symposium on Biocomputing* **4**, 41–52 (1999).
- ⁶⁸C. J. Roberts, B. Nelson, M. J. Marton, R. Stoughton, M. R. Meyer, H. A. Bennett, Y. D. He, H. Dai, W. L. Walker, T. R. Hughes, M. Tyers, C. Boone, and S. H. Friend, "Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles," *Science* **287**, 873–879 (2000).
- ⁶⁹T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburty, J. Simon, S. H. Friend, and M. Bard, "Functional discovery via a compendium of expression profiles," *Cell* **102**, 109–126 (2000).
- ⁷⁰M. Wahde and J. Hertz, "Coarse-grained reverse engineering of genetic regulatory networks," *BioSystems* **55**, 129–136 (2000).
- ⁷¹R. J. Cho, M. J. Campbell, E. A. Winzler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis, "A genomewide transcriptional analysis of the mitotic cell cycle," *Mol. Cell* **2**, 65–73 (1998).
- ⁷²P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Mol. Biol. Cell* **9**, 3273–3297 (1998).
- ⁷³T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science* **286**, 531–537 (1999).
- ⁷⁴A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature (London)* **403**, 503–511 (2000).
- ⁷⁵K. P. White, S. A. Rifkin, P. Hurban, and D. S. Hogness, "Microarray analysis of *Drosophila* development during metamorphosis," *Science* **286**, 2179–2183 (1999).
- ⁷⁶D. J. Katzmann, T. C. Hallstrom, Y. Mah, and W. S. Moye-Rowley, "Multiple Pdr1p/Pdr3p binding sites are essential for normal expression of the ATP binding cassette transporter protein-encoding gene PDR5," *J. Biol. Chem.* **271**, 23049–23054 (1996).
- ⁷⁷L. Kuras, H. Chrest, Y. Surdin-Kerjan, and D. Thomas, "A heteromeric complex containing the centromere binding factor 1 and two basic leucine zipper factors, met4 and met28, mediates the transcription activation of yeast sulfur metabolism," *EMBO J.* **15**, 2519–2529 (1996).
- ⁷⁸F. Paltauf, S. D. Kohlwein, and S. Henry, "Regulation and compartmentalization of lipid synthesis in yeast," in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression* (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1992), pp. 415–500.
- ⁷⁹B. Magasanik, "Regulation of nitrogen utilisation," in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression* (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1992), pp. 283–318.
- ⁸⁰M. Johnston and M. Carlson, "Regulation of carbon and phosphate utilisation," in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression* (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1992), pp. 193–281.