3

Computational Biology and Toxicogenomics

KATHLEEN MARCHAL, FRANK DE SMET, KRISTOF ENGELEN and BART DE MOOR

ESAT SISTA-SCD, K.U.Leuven,
Leuven-Heverlee, Belgium

1. INTRODUCTION

AQ1

Unforeseen toxicity is one of the main reasons for the failure of drug candidates. A reliable screening of drug candidates on toxicological side effects in early stages of the lead component development can help in prioritizing candidates and avoiding the futile use of expensive clinical trials and animal tests. A better understanding of the underlying cause of toxicological and pharmacokinetic responses will be useful to develop such screening procedure (1).

Pioneering studies (such as Refs. 2–5) have demonstrated that observable/classical toxicological endpoints are reflected in systematic changes in expression level. The observed endpoint of a toxicological response can be expected to result from an underlying cellular adaptation at molecular biological level. Until a few years ago studying gene regulation during toxicological processes was limited to the detailed study of a small number of genes. Recently, high-throughput profiling techniques allow us to measure expression at mRNA or protein level of thousands of genes simultaneously in an organism/tissue challenged with a toxicological compound (6). Such global measurements facilitate the observation not only of the effect of a drug on intended targets (on-target), but also of side effects on untoward targets (off-target) (7). Toxicogenomics is the novel discipline that studies such large scale measurement of gene/protein expression changes that result from the exposure to xenobiotics or that are associated with the subsequent development of adverse health effects (8,9). Although toxicogenomics covers a larger field, in this chapter we will restrict ourselves to the use of DNA arrays for mechanistic and predictive toxicology (10).

1.1. Mechanistic Toxicology

The main objective of mechanistic toxicology is to obtain insight in the fundamental mechanisms of a toxicological response. In mechanistic toxicology, one tries to unravel the pathways that are triggered by a toxicity response. It is, however, important to distinguish background expression changes of genes from changes triggered by specific mechanistic or adaptive responses. Therefore, a sufficient number of repeats and a careful design of expression profiling measurements are essential. The comparison of a cell line that is challenged with a drug to a negative control (cell line treated with a nonactive analogue) allows discriminating general stress from drug specific responses (10). Because the triggered pathways can be dose- and condition-dependent, a large number of experiments in different conditions are typically needed. When an in vitro model system is used (e.g., tissue culture) to assess the influence of a drug on gene expression, it is of paramount importance that the model system accurately encapsulates the relevant biological in vivo processes.

With dynamic profiling experiments one can monitor adaptive changes in the expression level caused by administering the xenobiotic to the system under study. By sampling the dynamic system at regular time intervals, short-, mid- and long-term alterations (i.e., high and low frequency changes) in xenobiotic-induced gene expression can be measured. With static experiments, one can test the induced changes in expression in several conditions or in different genetic backgrounds (gene knock out experiments) (10).

Recent developments in analysis methods offer the possibility to derive low-level (sets of genes triggered by the toxicological response) as well as high-level information (unraveling the complete pathway) from the data. However, the feasibility of deriving high-level information depends on the quality of the data, the number of experiments, and the type of biological system studied (11). Therefore, drug triggered pathway discovery is not straightforward and in addition is expensive so that it cannot be applied routinely. Nevertheless, when successful it can completely describe the effects elicited by representative members of certain classes of compounds. Well-described agents or compounds, for which both the toxicological endpoints and the molecular mechanisms resulting in them are characterized, are optimal candidates for the construction of a reference database and for subsequent predictive toxicology (see Sec. 1.2). Mechanistic insights can also AQ2 help determining the relative health risk and guide the discovery program towards safer compounds. From statistical point of view, mechanistic toxicology does not require any prior knowledge on the molecular biological aspects of the system studied. The analysis is based on what is called unsupervised techniques. Because it is not known in advance which genes will be involved in the studied response, arrays used for mechanistic toxicology are exhaustive, they contain cDNAs representing as much coding sequences of the genome as possible. Such arrays are also referred to as diagnostic or investigative arrays (12).

1.2. Predictive Toxicology

Compounds with the same mechanism of toxicity are likely to be associated with the alteration of a similar set of elicited genes. When tissues or cell lines subjected to such compounds are tested on a DNA microarray, one typically observes characteristic expression profiles or fingerprints. Therefore, reference databases can be constructed that contain these characteristic expression profiles of reference compounds. Comparing the expression profile of a new compound with such a reference database allows for a classification of the novel compound (2,5,7,9,13,14). From the known properties of the class to which the novel substance was classified, the behavior of the novel compound (toxicological endpoint) can be predicted. The reference profiles will, however, depend to a large extent on the endpoints that were envisaged (used the cell lines, model organisms, etc.). By a careful statistical analysis (feature extraction) of the profiles in such a compendium database, markers for specific toxic endpoints can be identified. These markers consist of genes that are specifically induced by a class of compounds. They can then be used to construct dedicated arrays (toxblots (12,15), rat hepato chips (13)). Contrary to diagnostic arrays, the number of genes on a dedicated array is limited resulting in higher throughput screening of lead targets at a lower cost (12,15). Markers can also reflect diagnostic expression changes of adverse effects. Measuring such diagnostic markers in easily accessible human tissues (blood samples) makes it possible to monitor early onset of toxicological phenomena after drug administration for instance during clinical trials (5). Moreover, markers (features) can be used to construct predictive models. Measuring the levels of a selected set of markers on, for instance, a dedicated array can be used to predict with the aid of a predictive model (classifier) the class of compounds to which the novel xenobiotic belongs (predictive toxicology). The impact of predictive toxicology will grow with the size of the reference databases. In this respect, the efforts made by several organizations (such as e.g., the International Life Science Institute (ILSI) http://www.ilsi.org/) to make public repositories of microarray data that are compliant with certain standards (MIAMI) are extremely useful (10,16).

1.3. Other Applications

There are plenty of other topics where the use of expression profiling can be helpful for toxicological research, including e.g., the identification of interspecies or in vitro-in vivo discrepancies. Indeed, results on the determination of dose responses and on the predicted risk of a xenobiotic for humans are often extrapolated from studies on surrogate animals. Measuring the differences in effect of administering well-studied compounds to either model animals or cultured human cells, could certainly help in the development of more systematic extrapolation methods (10).

Expression profiling can also be useful in the study of structure activity relationships (SAR). Differences in pharmacological or toxicological activity between structural related compounds might be associated with corresponding differences in expression profiles. The expression profiles can thus help distinguish active from inactive analogues in SAR (7).

Some drugs need to be metabolized for detoxification. Some drugs are only metabolized by enzymes that are encoded by a single pleiothropic gene. They involve the risk of drug accumulation to toxic concentrations in individuals carrying specific polymorphisms of that gene (17). With mechanistic toxicology, one can try to identify the crucial enzyme that is involved in the mechanism of detoxification. Subsequent genetic analysis can then lead to an a priori prediction to determine whether a xenobiotic should be avoided in populations with particular genetic susceptibilities.

2. MICROARRAYS

2.1. Technical Details

Microarray technology allows simultaneous measurement of the expression levels of thousands of genes in a single hybridization assay (7). An array consists of a reproducible pattern of different DNAs (primarily PCR products or oligonucleotides—also called probes) attached to a solid support. Each spot on an array represents a distinct coding sequence of the genome of interest. There are several microarray platforms that can be distinguished from each other in the way that the DNA is attached to the support.

Spotted arrays (18) are small glass slides on which presynthesized single stranded DNA or double-stranded DNA is spotted. These DNA fragments can differ in length depending on the platform used (cDNA microarrays vs. spotted oligoarrays). Usually the probes contain several hundred of base pairs and are derived from expressed sequence tags (ESTs) or from known coding sequences from the organism under study. Usually each spot represents one single ORF or gene. A cDNA array can contain up to 25,000 different spots.

GeneChip oligonucleotide arrays (Affymetrix, Inc., Santa Clara (19)) are high-density arrays of oligonucleotides synthesized in situ using light-directed chemistry. Each gene is represented by 15–20 different oligonucleotides (25-mers), that serve as unique sequence-specific detectors. In addition, mismatch control oligonucleotides (identical to the perfect match probes except for a single base-pair mismatch) are added. These control probes allow the estimation of cross-hybridization. An Affymetrix array represents over 40,000 genes.

Besides these customarily used platforms, other methodologies are being developed (e.g., fiber optic arrays (20) as well).

In every cDNA-microarray experiment, mRNA of a reference and agent-exposed sample is isolated, converted into cDNA by an RT-reaction and labeled with distinct fluorescent dyes (Cy3 and Cy5, respectively the "green" and "red" dye). Subsequently, both labeled samples are hybridized simultaneously to the array. Fluorescent signals of both channels (i.e., red and green) are measured and used for further analysis (for more extensive reviews on microarrays we refer to (7,21–23)). An overview of this procedure is given in Fig. 1. F1

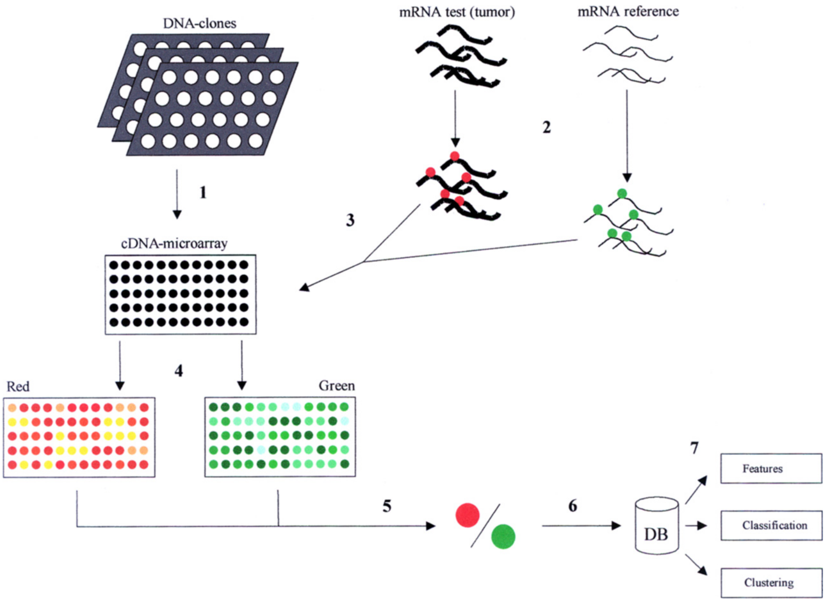


Figure 1 Schematic overview of an experiment with a cDNA microarray. (1) Spotting of the presynthesized DNA-probes (derived from the genes to be studied) on the glass slide. These probes are the purified products from PCR-amplification of the associated DNA-clones. (2) Labeling (via reverse transcriptase) of the total mRNA of the test sample (red = Cy5) and reference sample (green = Cy3). (3) Mixing of the two samples and hybridization. (4) Read-out of the red and green intensities separately (measure for the hybridization by the test and reference sample) of each probe. (5) Calculation of the relative expression levels (intensity in the red channel/intensity in the green channel). (6) Storage of results in a database. (7) Data mining.

2.2. Sources of Variation

In a microarray experiment, changes in gene expression level are being monitored. One is interested in knowing how much the expression of a particular gene is affected by the applied condition. However, besides this effect of interest, other experimental factors or sources of variation contribute to the measured change in expression level. These sources of variation prohibit direct comparison between measurements.

That is why preprocessing is needed to remove these additional sources of variation, so that for each gene, the corrected "preprocessed" value reflects the expression level caused by the condition tested (effect of interest). Consistent sources of variation in the experimental procedure can be attributed to gene, condition/dye, and array effects (24–26).

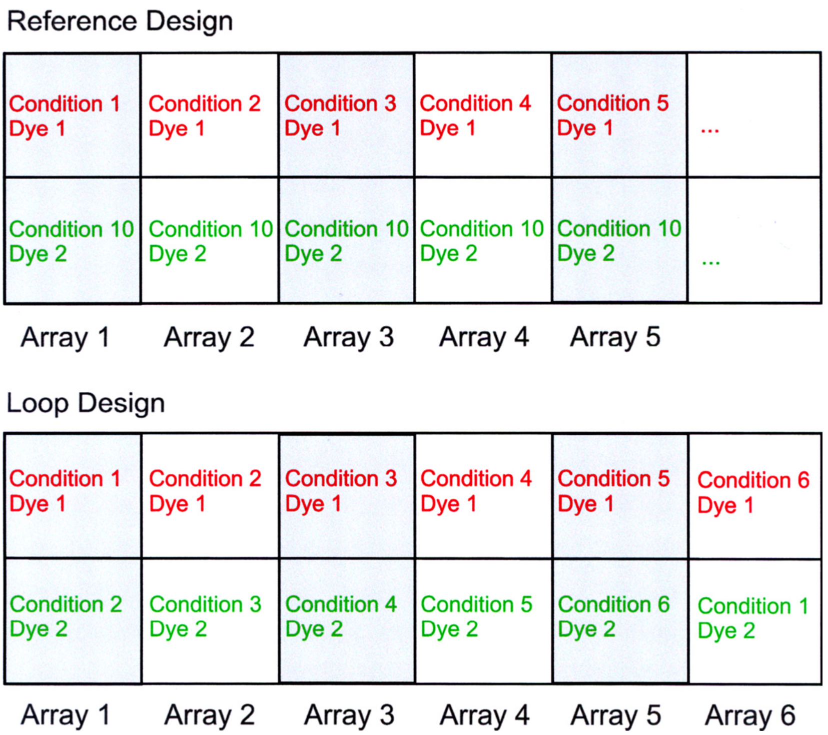
Condition and dye effects reflect differences in mRNA isolation and labeling efficiencies between samples. These effects result in a higher measured intensity for certain conditions or for either one of both channels. When performing multiple experiments (i.e., by using more arrays), arrays are not necessarily being treated identically. Differences in hybridization efficiency result in global differences in intensities between arrays, making measurements derived from different arrays incomparable. This effect is generally called the array effect.

The gene effect explains that some genes emit a higher or lower signal than others. This can be related to differences in basal expression level, or to sequence-specific hybridization or labeling efficiencies. A last source of variation is a combined effect, the array–gene effect. This effect is related to spot-dependent variations in the amount of cDNA present on the array. Since the observed signal intensity is not only influenced by differences in the mRNA population present in the sample, but also by the amount of spotted cDNA, direct comparison of the absolute expression levels is unreliable.

The factor of interest, which is the condition-affected change in expression of a single gene, can be considered to be a combined gene–condition (GC) effect.

2.3. Microarray Design

The choice of an appropriate design is not trivial (27–29). In Fig. 2 distinct designs are represented. The simplest microarray experiments compare expression in two distinct conditions. A test condition (e.g., cell line triggered with a lead compound) is compared to a reference condition (e.g., cell line triggered with a placebo). Usually the test is labeled with Cy5 (red dye), while the reference is labeled with Cy3 (green dye). F2



Reference Design

Array 1	Array 2	Array 3	Array 4	Array 5	
Condition 1 Dye 1	Condition 2 Dye 1	Condition 3 Dye 1	Condition 4 Dye 1	Condition 5 Dye 1	...
Condition 10 Dye 2	Condition 10 Dye 2	Condition 10 Dye 2	Condition 10 Dye 2	Condition 10 Dye 2	...

Loop Design

Array 1	Array 2	Array 3	Array 4	Array 5	Array 6
Condition 1 Dye 1	Condition 2 Dye 1	Condition 3 Dye 1	Condition 4 Dye 1	Condition 5 Dye 1	Condition 6 Dye 1
Condition 2 Dye 2	Condition 3 Dye 2	Condition 4 Dye 2	Condition 5 Dye 2	Condition 6 Dye 2	Condition 1 Dye 2

Figure 2 Overview of two commonly used microarray designs. (A) Reference design; (B) loop design. Dyel = Cy5; Dye2 = Cy3; two conditions are measured on a single array.

Performing replicate experiments is mandatory to infer relevant information on a statistically sound basis. However, instead of just repeating the experiments exactly in the way described above, a more reliable approach here would be to perform dye reversal experiments (dye swap). As a repeat on a second array: the same test and reference conditions are measured once more but the dyes are swapped, i.e., on this second array, the test condition is labeled with Cy3 (green dye), while the corresponding reference condition is labeled with Cy5 (red dye). This allows intrinsically compensating for dye-specific differences. When the behavior of distinct compounds is compared or when the behavior triggered by a compound is profiled during the course of a dynamic process, more complex designs are required. Customarily used, and still preferred by molecular biologists, is the reference design: different test conditions (e.g., distinct compounds) are compared to a similar reference condition. The reference condition can be artificial and does not need to be biologically significant. Its main purpose is to have a common baseline to facilitate mutual comparison between me samples. Every reference design results in a relatively higher number of replicate measurements of the condition (reference) in which one is not primarily interested, than of the condition of interest (test condition). A loop design can be considered as an extended dye reversal experiment. Each condition is measured twice, each time on a different array and labeled with a different dye (Fig. 2). For the same number of experiments, a loop design offers more balanced replicate measurements of each condition than a reference design, while the dye-specific effects can also be compensated for.

Irrespective of the design used, the expression levels of thousands of genes are monitored simultaneously. For each gene, these measurements are usually arranged into a data matrix. The rows of the matrix represent the genes while the columns are the tested conditions (toxicological compounds, timepoints). As such one obtains gene expression profiles (row vectors) and experiment profiles (column vectors) (Fig. 3). F3

3. ANALYSIS OF MICROARRAY EXPERIMENTS

Some of the major challenges for mechanistic and predictive toxicogenomics are in data management and analysis (5,10). In the following chapter, we give an overview of the state of the art methodologies for the analysis of high-throughput expression profiling experiments. The review is not comprehensive as the field of microarray analysis is rapidly evolving. Although there will be a special focus on the analysis of cDNA arrays, most of the described methodologies are generic and applicable to data derived from other high-throughput platforms.

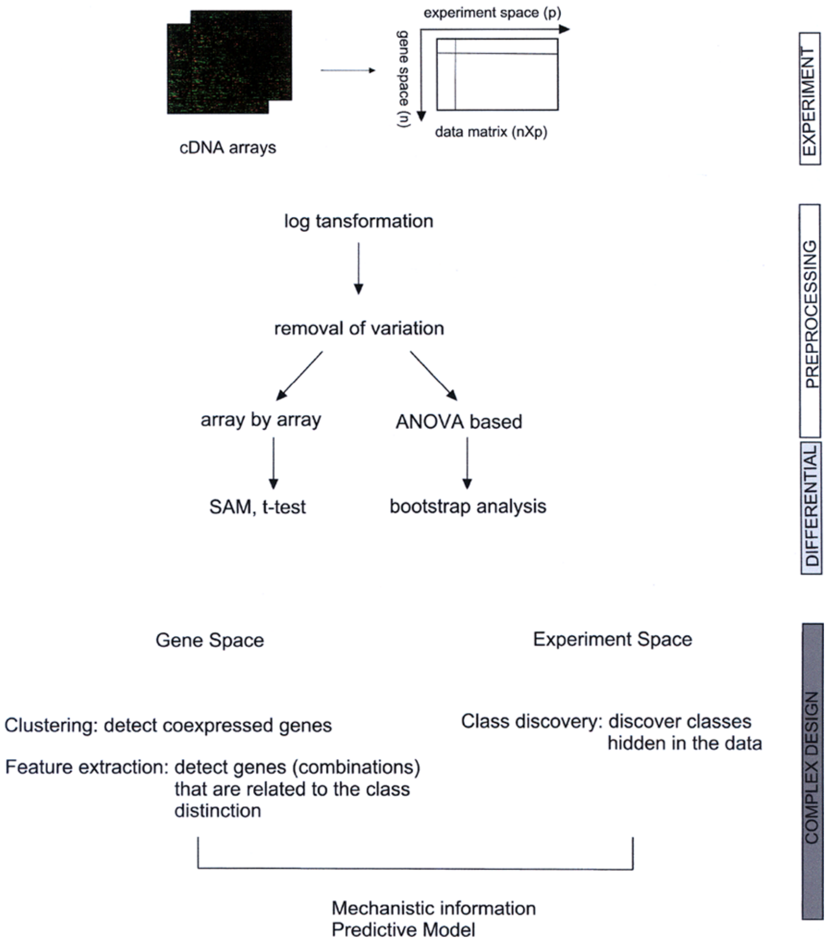


Figure 3 Schematic overview of the analysis flow of cDNA-microarray data.

3.1. Preprocessing: Removal of Consistent Sources of Variation

As mentioned before, preprocessing of the raw data is needed to remove consistent and/or the systematic sources of variation from the measured expression values. As such, the preprocessing has a large influence on the final result of the analysis. In the following, we will give an overview of the commonly used approaches for preprocessing: the array by array approach and the procedure based on analysis of variance (ANOVA) (Fig. 3). The array by array approach is a multistep procedure comprising log transformation, normalization, and identification of differentially expressed genes by using a test statistic. The ANOVA-based approach consists of a log transformation, linearization, and identification of differentially expressed genes based on bootstrap analysis.

3.1.1. Mathematical Transformation of the Raw Data: Need for a Log Transformation

The effect of the log transformation as an initial preprocessing step is illustrated in Fig. 4. In Fig. 4A, the expression levels of all genes measured in the test sample were plotted against the corresponding measurements in the reference sample. Assuming that the expression of only a restricted

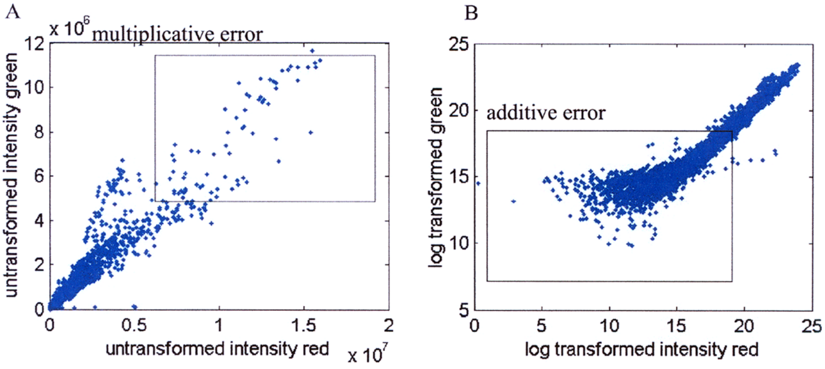


Figure 4 Illustration of the influence of log transformation on the multiplicative and additive errors. Panel A: representation of untransformed raw data. X-axis: intensity measured in the red channel, Y-axis: intensity measured in the green channel. Panel B: representation of $\log_2$ transformed raw data. X-axis: intensity measured in the red channel ($\log_2$ value), Y-axis: intensity measured in the green channel ($\log_2$ value). Assuming that only a small number of the genes will alter their expression level under the different conditions tested, for most genes the measurement in the green channel can be considered as a replica of the measurement in the red channel.

number of genes is altered (global normalization assumption, see below), measurements of the reference and the test condition can be considered to be comparable for most of the genes on the array. Therefore, the residual scattering as observed in Fig. 4A reflects the measurement error. As often observed, the error in microarray data is a superposition of a multiplicative error and an additive one. Multiplicative errors cause signal-dependent variance of residual scattering, which deteriorates the reliability of most statistical tests. Log transforming the data alleviates this multiplicative error, but usually at the expense of an increased error at low expression levels (Fig. 4B). Such an increase of the measurement error with decreasing signal intensities, as present in the log-transformed data, is however considered to be intuitively plausible: low expression levels are generally assumed to be less reliable than high levels (24,30).

An additional advantage of log transforming the data is that, differential expression levels between the two channels are represented by log(test) − log(reference) (see below statistical testing). This allows bringing levels of under- and over-expression to the same scale, i.e., values of underexpression are no longer bound between 0 and 1. AQ3

3.1.2. Array by Array Approach

In the array by array approach, each array is compensated separately for dye/condition and spot effects. A log (test/reference) = log (test) − log(reference) is used as an estimate of the relative expression. Using ratios (relative expression levels) instead of absolute expression levels allows compensating intrinsically for spot effects. The major drawback of the ratio approach is that when the intensity measured in one of the channels is close to 0, the ratio attains extreme values that are unstable as the slightest change in the value close to 0 has a large influence on the ratio (30,31).

Normalization methods aim at removing consistent condition and dye effects (see above). Although the use of spikes (control spots, external control) and housekeeping genes (genes not altering their expression level under the conditions tested) for normalization have been described in the literature, global normalization is commonly used (32). The global normalization principle assumes that only of a small fraction of the total number of genes on the array, the expression level is altered. It also assumes that symmetry exists in the number of genes for which the expression is increased vs. decreased. Under this assumption, the average intensity of the genes in the test condition should be equal to the average intensities of the genes in the reference condition. Therefore, for the bulk of the genes, the log-ratios should equal 0. Regardless of the procedure used, after normalization, all log-ratios will be centered around 0. Notice that the assumption of global normalization applies only to microarrays that contain a random set of genes and not to dedicated arrays.

Linear normalization assumes a linear relationship between the measurements in both conditions (test and reference). A common choice for the constant transformation factor is the mean or median of the log intensity ratios for a given gene set. As shown in Fig. 5, most often, the assumption of F5 a linear relationship between the measurements in both conditions is an oversimplification, since, the relationship between dyes depends on the measured intensity. These observed nonlinearities are most pronounced at extreme intensities (either high or low). To cope with this problem, Yang et al. (32) described the use of a robust scatter plot smoother, called Lowess, that performs local linear fits. The results of this fit can be used to simultaneously linearize and normalize the data (Fig. 5).

The array by array procedure uses the global properties of all genes on the array to calculate the normalization factor. Other approaches have been described that subdivide an array into, for instance, individual print tip groups, which are normalized separately (32). Theoretically, these approaches perform better than the array by array approach in removing position-dependent “within array” variations. The drawback, however, is that the number of measurements to calculate the fit is reduced, a pitfall that can be overcome by the use of ANOVA (see Sec. 3.1.3). SNOMAD offers a free online implementation of the array by array normalization procedure (33).

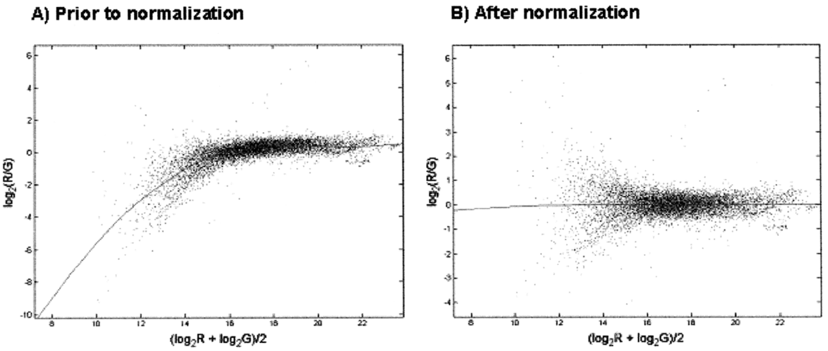


Figure 5 Illustration of the influence of an intensity-dependent normalization. Panel A: representation of the log-ratio $M = \log_2(R/G)$ vs. the mean log intensity $A = (\log_2(R) + \log_2(G))/2$. At low average intensities, the ratio becomes negative indicating that the green dye is consistently more intense as compared to the intensity of the red dye. This phenomena is referred to as the non-linear dye effect. Solid line represents the Lowess fit with *f* value of 0.02 ($R = $ red; $G = $ green). Panel B: Representation of the ratio $M = \log_2(R/G)$ vs. the mean log intensity $A = (\mathrm{logo}_2(\mathrm{R}) + \log_2(G))/2$ after performing a normalization and linearization based on the Lowess fit. Solid line represent the new Lowess fit with *f* value of 0.02 on the normalized data (R = red; G = green).

3.1.3. ANOVA-based preprocessing

ANOVA can be used as an alternative to the array by array approach (24,27). In this case, it can be viewed as a special case of multiple linear regression, where the explanatory variables are entirely qualitative. ANOVA models the measured expression level of each gene as a linear combination of the explanatory variables that reflect, in the context of microarray analysis, the major sources of variation. Several explanatory variables representing the condition, dye and array effects (see above) and combinations of these effects are taken into account in the models (see Fig. 6). One of the combined effects, the GC effect, reflects the expression of a gene solely depending on the tested condition (i.e., the condition-specific expression or the effect of interest). Similarly, the

$$I_{ijklmn} = \mu + B_m + D_l + A_{k(m)} + \left(AD\right)_{kl(m)} + P_{j(k(m))} + G_{i(j(o))} + E_{in(j(m))} + \varepsilon_{ijklmn}$$

Figure 6 Example of an ANOVA model. *I* is the measured intensity, *D* is the dye effect, *A* is the array effect, *G* is the gene effect, *B* is the batch effect (the number of separate arrays needed to cover the complete genome if the cDNAs of the genome do not fit on a single array), *P* is the pin effect, *E* is the expression effect (factor of interest). *AD* is the combined array–dye effect, ε is the residual error, *m* is the number of batches, *l* the number of dyes, *j* the number of spots on an array spotted by the same pin, and *i* the number of genes. The measured intensity is modeled as a linear combination of consistent sources of variation and the effect of interest Remark that in this model condition effect *C* has been replaced by the combined *AD* effect. AQ4

difference between the GC effects of two conditions reflects the differential expression. Of the other combined effects, only those having a physical meaning in the process to be modeled are retained. Reliable use of an ANOVA model requires a good insight into the experimental process. Several ANOVA models have been described for microarray preprocessing (24,34,35).

The ANOVA approach can be used if the data are adequately described by a linear ANOVA model and if the residuals are approximately normally distributed. ANOVA obviates the need for using ratios. It offers as an additional advantage that all measurements are used simultaneously for statistical inference and that the experimental error is implicitly estimated (36). Several web applications that offer an ANOVA-based preprocessing procedure have been published (e.g., MARAN (34), GeneANOVA (37)).

3.2. Microarray Analysis for Mechanistic Toxicology

The purpose of mechanistic toxicology consists of unraveling the genomic responses of organisms exposed to xenobiotics. Distinct experimental setups can deliver the required information. The most appropriate data analysis method depends both on the biological question to be answered and the experimental design. For the purpose of clarity, we make a distinction between three types of design. This subdivision is somewhat artificial and the distinction is not always clearcut. The simplest design compares two conditions to identify differentially expressed genes. Techniques developed for this purpose will be reviewed in Sec. 3.2.1. Using more complex designs, one can try to reconstruct the regulation network that generates a certain behavior. Dynamic changes in expression can be monitored as function of time. For such a dynamic experiment, the main purpose is to find genes that behave similarly during the time course, where often an appropriate definition of similarity is one of the problems. Such coexpressed genes are identified by cluster analysis (Sec. 3.2.2). On the other hand, the expression behavior can be tested under distinct experimental conditions (e.g., the effect induced by distinct xenobiotics). One is interested, not only in finding coexpressed genes but also in knowing the experimental conditions that group together based on their experiment profiles. This means that clustering is performed both in the space of the gene variables (row vectors) and in the space of the condition variables (column vectors). Although such designs can also be useful for mechanistic toxicology, they are usually performed in the context of class discovery and predictive toxicology and will be further elaborated in Sec. 3.3. The objective of clustering is to detect low-level information. We describe this information as low-level because the correlations in expression patterns between genes are identified, but all causal relationships (i.e., the high-level information) remains undiscovered. Genetic network inference (Sec. 3.2.3) on the other hand tries to infer this high-level information from the data.

3.2.1. Identification of Differentially Expressed Genes

When preprocessed properly, consistent sources of variation have been removed, and the replicate estimates of the (differential) expression of a particular gene can be combined. To search for differentially expressed genes, statistical methods are used that test whether two variables are significantly different. The exact identity of these variables depends on the question to be answered. When expression in the test condition is compared to expression in the reference condition, it is generally assumed that for most of the genes no differential expression occurs (global normalization assumption). Thus, the zero hypothesis implies that expression of both test and reference sample is equal (or that the log of the relative expression equals 0). Because in a cDNA experiment the measurement of the expression of the test condition and reference condition is paired (measurement of both expression levels on a single spot), the paired variant of the statistical test is used.

When using a reference design, one is not interested in knowing whether the expression of a gene in the test condition is significantly different from its expression in the reference condition since the reference condition is artificial. Rather, one wants to know the relative differences between the two compounds tested on different arrays using a single reference. Assuming that the ratio is used to estimate the relative expression between each condition and a common reference, the zero hypothesis now will be equality of the average ratio in both conditions tested. In this case, the data are no longer paired. This application is related to feature extraction and will be further elaborated in Sec. 3.3.1.

In this paragraph, a major emphasis will be on the description of selection procedures to identify genes that are differentially expressed in the test vs. reference condition.

The fold test is a nonstatistical selection procedure that makes use of an arbitrary chosen threshold. For each gene, an average ratio is calculated based on the different ratio estimates of the replicate experiments (log-ratio = log(test) − log(reference)). Average ratios of which the expression ratio exceeds a threshold (usually twofold) are retained. The fold test is based on the assumption that a larger observed fold change can be more confidently interpreted as a stronger response to the environmental signal than smaller observed changes. A fold test, however, discards all information obtained from replicates (30). Indeed, when either one of the measured channels obtains a value close to 0, the log-ratio estimate usually obtains a high but inconsistent value (large variance on the variables). Therefore, more sophisticated variants of the fold test have been developed. These methods simultaneously construct an error model of the raw measurements that incorporates multiplicative and additive variations (38–40).

A plethora of novel methods to calculate a test statistic and the corresponding significance level have recently been proposed, provided replicates are available. Each of these methods first calculates a test statistic and subsequently determines the significance of the observed test statistic. Distinct *t*-test like methods are available that differ from each other in the formula that describes the test statistic and in the assumptions regarding the distribution of the null hypothesis. *t*-Test methods are used for detecting significant changes between repeated measurements of a variable in two groups. In the standard *t*-test, it is assumed that data are sampled from a normal distribution with equal variances (zero hypothesis). For microarray data, the number of repeats is too low to assess the validity of this assumption of normality. To overcome this problem, methods have been developed that estimate the distribution of the zero hypothesis from the data itself by permutation or bootstrap analysis (36,41). Some methods avoid the necessity of estimating a distribution of the zero hypothesis by using order statistics (41). For an exhaustive comparison between the individual performances of each of these methods, we refer to Marchal et al. (31) and for the technical details, we refer to the individual references and Pan et al. (2002) (42). AQ5

When ANOVA is used to preprocess the data, significantly expressed genes are often identified by bootstrap analysis (Gaussian statistics are often inappropriate, since normality assumptions are rarely satisfied). Indeed, fitting the ANOVA model to the data allows the estimation of the residual error which can be considered as an estimate of the experimental error. By adding noise (randomly sampled from the residual error distribution) to the estimated intensities, thousands of novel bootstrapped datasets, mimicking wet lab experiments, can be generated. In each of the novel datasets, the difference in GC effect between two conditions is calculated, as a measure for the differential expression. Based on these thousands of estimates of the difference in GC effect, a bootstrap confidence interval is calculated (36).

An extensive comparison of these methods showed that a *t*-test is more reliable than a simple fold test. However, the *t*-test suffers from a low power due the restricted number of replicate measurements available. The method of Long et al. (43) tries to cope with this drawback by estimating the population variance as a posterior variance that consists of a contribution of the measured variance and a prior variance. Because they assume that the variance is intensity-dependent, this prior variance is estimated based on the measurements of other genes with similar expression levels as the gene of interest. ANOVA-based methods assume a constant error variance for the entire range of intensity measurements (homoscedasticity). Because the calculated confidence intervals are based on a linear model and microarray data suffer from nonlinear intensity-dependent effects and large additive effects at low expression levels (see also Sec. 3.1.1), the estimated confidence intervals are usually too restrictive for elevated expression levels and too small for measurements in the low intensity range. In our experience, methods that did not make an explicit assumption on the distribution of the zero hypotheses, such as Statistical Analysis of Microarrays (SAM) (41) clearly outperformed the other methods for large datasets.

Another important issue in selecting significantly differentially expressed genes is correction for multiple testing. Multiple testing is crucial since hypotheses are calculated for thousands of genes simultaneously. Standard Bonferroni correction seems overrestrictive (30,44). Therefore, other corrections for multiple testing have been proposed (45). Very promising for microarray analysis seems the application of the false discovery rate (FDR) (46). A permutation-based implementation of this method can be found in the SAM software (41).

3.2.2. Identification of Coexpressed Genes

3.2.2.1. Clustering of the Genes

As mentioned previously, normalized microarray data are collected in a data matrix. For each gene, the (row) vector leads to what is generally called an expression profile. These expression profiles or vectors can be regarded as (data) points in a high-dimensional space. Genes involved in a similar biological pathway or with a related function often exhibit a similar expression behavior over the coordinates of the expression profile/vector. Such similar expression behavior is reflected by a similar expression profile. Genes with similar expression profiles are called coexpressed. The objective of cluster analysis of gene expression profiles is to identify subgroups (= clusters) of such coexpressed genes (47,48). Clustering algorithms group together genes for which the expression vectors are "close" to each other in the high-dimensional space based on some distance measure. A first generation of algorithms originated in research domains other than biology (such as the areas of "pattern recognition" and "machine learning"). They have been applied successfully to microarray data. However, confronted with the typical characteristics of biological data, recently a novel generation of algorithms has emerged. Each of these algorithms can be used with one or more distance metrics (see Fig. 7). Prior to clustering, microarray data usually are filtered, missing values are replaced and the remaining values are rescaled. F7

3.2.2.2. Data Transformation Prior to Clustering

The "Euclidean distance" is frequently used to measure the similarity between two expression profiles. However, genes showing the same relative behavior but with diverging absolute behavior (e.g., gene expression profiles with a different baseline and/or a different amplitude but going up and down at the same time) will have a relatively high Euclidean distance. Because the purpose is to group expression profiles that have the same relative behavior, i.e., genes that are up- and downregulated together, cluster algorithms based on the Euclidean distance will therefore erroneously assign

Minkowski distance

$$d(x,y) = \sqrt[r]{\sum_{i=1}^{p} |x_i - y_i|^2} \qquad \begin{array}{l} r = 1 : \text{Manhattan distance} \\ r = 2 : \text{Euclidean distance} \end{array}$$

Pearson correlation distance

$$s(x,y) = \frac{\sum_{i=1}^{p} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p} (x_i - \bar{x})^2 \times \sum_{i=1}^{p} (y_i - \bar{y})^2}}$$

$$\bar{x} = \frac{1}{p} \sum_{i=1}^{p} x_i$$

$$\bar{y} = \frac{1}{p} \sum_{i=1}^{p} y_i$$

Figure 7 Overview of commonly used distance measures in cluster analysis. x and y are points or vectors in the p-dimensional space. x_i and y_i ($i = 1, \ldots, p$) are the coordinates of x and y. p is the number of experiments.

the genes with different absolute baselines to different clusters. To overcome this problem, expression profiles are standardized or rescaled prior to clustering. Consider a gene expression profile $g(g_1, g_2, \ldots, g_p)$ over p points (i.e., p time points or conditions) with average expression level μ and standard deviation σ. Microarray data are commonly rescaled by replacing every expression level g_i by

$$\frac{g_i - \mu}{\sigma}$$

This operation results in a collection of expression profiles all being 0 mean and with standard deviation 1 (i.e., the absolute differences in expression behavior have largely been removed). The Pearson correlation coefficient, a second customarily used distance measure, inherently performs this rescaling as it is basically equal to the cosine of the angle between two gene expression profile vectors.

As previously mentioned, a set of microarray experiments, in which gene expression profiles have been generated, frequently contains a considerable number of genes that do not contribute to the biological process that is being studied. The expression values of these profiles often show little variation over the different experiments (they are called constitutive with respect to the biological process studied). By applying the rescaling procedure, these profiles will be inflated and will contribute to the noise of the dataset. Most existing clustering algorithms attempt to assign each gene expression profile, even the ones of poor quality to at least one cluster. When also noisy and/or random profiles are assigned to certain clusters, they will corrupt these clusters and hence the average profile of the clusters. Therefore, filtering prior to the clustering is advisable. Filtering involves removing gene expression profiles from the dataset that do not satisfy one or possibly more very simple criteria (49). Commonly used criteria include a minimum threshold for the standard deviation of the expression values in a profile (removal of constitutive genes). Microarray datasets regularly contain a considerable number of missing values. Profiles containing too many missing values have to be omitted (filtering step). Sporadic missing values can be replaced by using specialized procedures (50,51).

3.2.2.3. Cluster Algorithms

The first generation of cluster algorithms includes standard techniques such as *K*-means (52), self-organizing maps (53,54) and hierarchical clustering (49). Although biologically meaningful results can be obtained with these algorithms, they often lack the fine-tuning that is necessary for biological problems. The family of hierarchical clustering algorithms was and is probably still the method preferred by biologists (49) (Fig. 8). According to a certain measure, the distance between every couple of clusters is calculated (this is called the pairwise distance matrix). Iteratively, the two closest clusters are merged giving rise to a tree structure, where the height of the branches is proportional to the pairwise distance between the clusters. Merging stops if only one cluster is left. However, the final number of clusters has to be determined by cutting the tree at a certain level or height. Often it F8

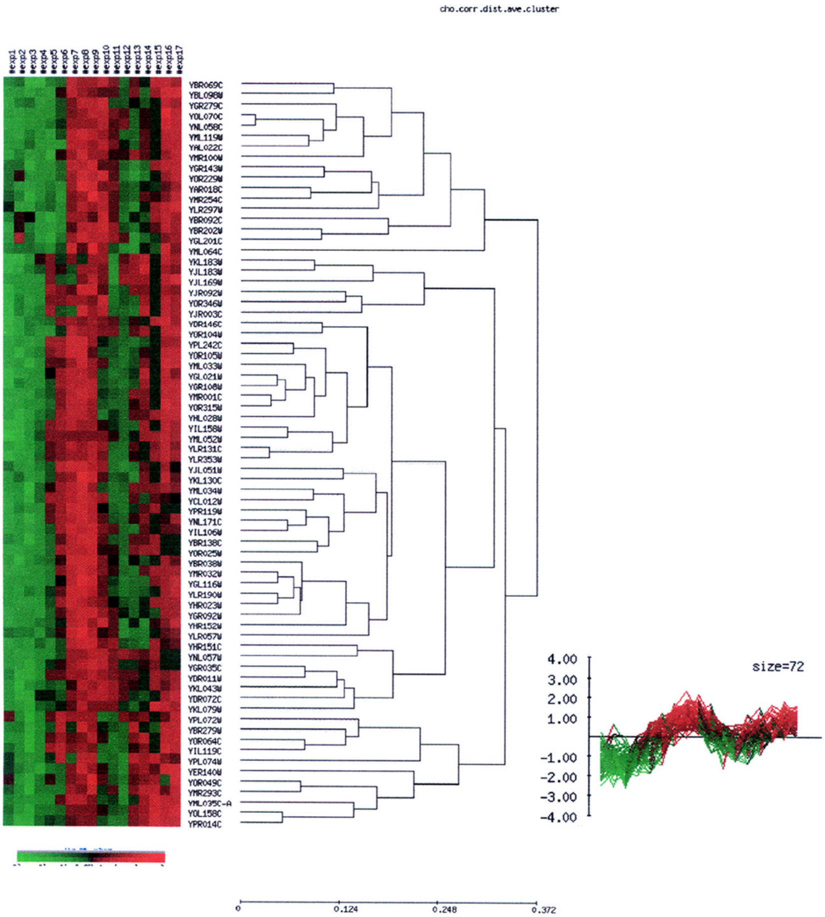


Figure 8 Hierarchical clustering. Hierarchical clustering of the dataset of Cho et al. (119) representing the mitotic yeast cell cycle. A selection of 3000 genes was made as described in Ref. 51. Hierarchical clustering was performed using the Pearson correlation coefficient and an average linkage distance (UPGMA) as implemented in EPCLUST (65). Only a subsection of the total tree is shown containing 72 genes. The columns represent the experiments, the rows the gene names. A green color indicates downregulation, while a red color represents upregulation, as compared to the reference condition. In the complete experimental setup, a single reference condition was used (reference design).

is not straightforward to decide where to cut the tree as it is typically rather difficult to predict which level will give the most valid biological results. Secondly, the computational complexity of hierarchical clustering is quadratic in the number of gene expression profiles, which can sometimes be limiting considering the current (and future) size of the datasets.

Centroid methods form another attractive class of algorithms. The *K*-means algorithm for instance starts by assigning at random all the gene expression profiles to one of the *N* clusters (where *N* is the user-defined number of clusters). Iteratively, the center (which is nothing more than the average expression vector) of each cluster is calculated, followed by a reassignment of the gene expression vectors to the cluster with the closest cluster center. Convergence is reached when the cluster centers remain stationary. Self-organizing maps can be considered as a variation on centroid methods that also allow samples to influence the location of neighboring clusters. These centroid algorithms suffer from similar drawbacks as hierarchical clustering: the number of clusters is a user-defined parameter with a large influence on the outcome of the algorithm. For a biological problem, it is hard to estimate in advance how many clusters can be expected. Both algorithms assign each gene of the dataset to a cluster. This is from a biological point of view counterintuitive, since only a restricted number of genes are expected to be involved in the process studied. The outcome of these algorithms appears to be very sensitive to the chosen parameter settings (number of clusters for *K*-means (Fig. 9)), the distance measure that is used and the metrics to determine the distance between clusters (average vs. complete linkage for hierarchical clustering). Finding the biological most relevant solution usually requires extensive parameter fine-tuning and is based on arbitrary criteria (e.g., clusters look more coherent) (55). F9

Besides the development of procedures that help to estimate some of the parameters needed for the first generation of algorithms (e.g., like the number of clusters present in the data (56–58)), a panoply of novel algorithms have been designed that cope with the problems mentioned above in different ways: self-organizing tree algorithm or SOTA (59)

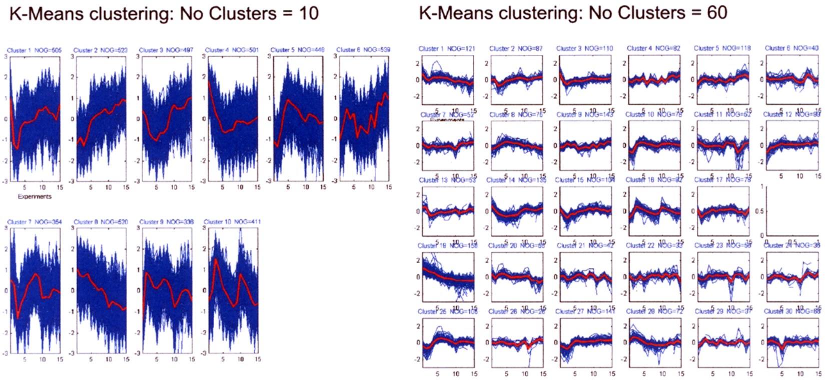


Figure 9 Illustration of the effect of using different parameter settings on the end result of a *K*-means clustering of microarray data. Data were derived from Ref. 119 and represent the dynamic profile of the cell cycle. The cluster number is the variable parameter of the *K*-means clustering. By underestimating the number of clusters, genes within a cluster will have a very heterogeneous profile. Since *K*-means assigns all genes to a cluster (no inherent quality criterion is imposed), genes with a noisy profile disturb the average profile of the clusters. When increasing the number of clusters, the profiles of genes that belong to the same cluster become more coherent and the influence of noisy genes is less exacerbating. However, when too high the cluster number, genes belonging biologically to the same cluster might be assigned to separate clusters with very similar average profiles.

combines self-organizing maps and divisive hierarchical clustering; quality-based clustering (60) only assigns genes to a cluster that meet a certain quality criterion; adaptive quality-based clustering (51) is based on a principle similar to quality-based clustering, but offers a strict statistical meaning to the quality criterion; gene shaving (61) is based on principal component analysis (PCA). Other examples include model-based clustering (56,58); clustering based on simulated annealing (57) and CAST (62). For a more extensive overview of these algorithms we refer to Moreau et al. (47).

Some of these algorithms determine the number of clusters based on the inherent data properties (51,58–60,63). Quality criteria have been developed to minimize the number of false positives. Only those genes are retained, in the clusters, that satisfy a quality criterion. This results in clusters that contain genes with tightly coherent profiles (51,60). Fuzzy clustering algorithms allow a gene to belong to more than one cluster (61). Distinct publicly available implementations of these novel algorithms are freely available for academic users (INCLUSive (64), EPCLUST (65), AMADA (66), Cluster (49), ...)

3.2.2.4. Cluster Validation

Depending on the algorithms and the distance measures used, clustering will give different results. Therefore validation, either statistically or biologically, of the cluster results is essential. Several methods have been developed to assess the statistical relevance of a cluster. Intuitively, a cluster can be considered reliable if the within cluster distance is small (i.e., all genes retained are tightly coexpressed) and the cluster has an average profile well delineated from the remainder of the dataset (maximal intercluster distance). This criterion is formalized by Dunn's validity index (67). Another desirable property is cluster stability: gene expression levels can be considered as a superposition of real biological signals and small experimental errors. If true biological signals are more pronounced than the experimental variation, repeating the experiments should not interfer with the identification of the biological true clusters. Following this reasoning, cluster stability is assessed by creating new in silico replicas (i.e., simulated replicas) of the dataset of interest by adding a small amount of artificial noise to the original data. The noise can be estimated from a reasonable noise model (68,69) or by sampling the noise distribution directly from the data (36). These newly generated datasets are preprocessed and clustered in the same way as the original dataset. If the biological signal is more pronounced than the noise signal in the measurements of one particular gene, adding small artificial variations (in the range of the experimental noise present in the dataset) to the expression profile of such gene will not influence its overall profile and cluster membership. The result (cluster membership) of that particular gene is robust towards what is called a sensitivity analysis and a reliable confidence can be assigned to the cluster result of that gene.

An alternative approach of validating clusters is by assessing the biological relevance of the cluster result. Genes exhibiting a similar behavior might belong to the same biological process. This is reflected by enrichment of functional categories within a cluster (51,55). Also, for some clusters, the observed coordinate behavior of the gene expression profiles might be caused by transcriptional coregulation. In such case, detection of regulatory motifs is useful as a biological validation of cluster results (55,70–72).

3.2.3. Genetic Network Inference

The final goal of mechanistic toxicology is the reconstruction of the regulatory networks that underlie the observed cell responses. A complete regulatory network consists of proteins interacting with each other, with DNA or with metabolites to constitute a complete signaling pathway (73). The action of regulatory networks determines how well cells can react or adapt to novel conditions. From this perspective, a cellular reaction against a xenobiotic compound can be considered as a stress response that triggers a number of specialized regulation pathways and induces the essential survival machinery. A regulatory network viewed at the level of transcriptional regulation is called a genetic network. This genetic network can be monitored by microarray experiments. In contrast to clustering that searches for correlation in the data, genetic network inference goes one step beyond and tries to reconstruct the causal relationships between the genes. Although methods for genetic network inference are being developed, the sizes of the currently available experimental datasets do not yet meet the extensive data requirements of most of these algorithms. In general, the number of experimental data is still much smaller than the number of parameters that is to be estimated (i.e., the problem is underdetermined). The low signal to noise level of microarray data and the inherent stochasticity of biological systems (74,75) aggravates the problem of underdetermination. Combining expression data with additional sources of information (prior information) can possibly offer a solution (76–79). Most of the current inference algorithms already make use of general knowledge on the characteristics of biological networks, such as the presence of hierarchical network structures (77,80), a powerlaw distribution of the number of connections (81), sparsness of a network (82,83), and a maximal indegree (maximal number of incoming and outgoing edges).

In order to unravel pathways, both dynamic and static experiments can be informative. However, most of the developed algorithms can only handle static data. Dynamic data can always be converted to static data by treating the transition from a previous time point to a consecutive time point as a single condition. However, this is at the expense of losing the specific information that can be derived from the dynamical characteristics of the data. Treating this biological time signals as responses of a dynamical system is one of the big challenges of the near future.

Networks are either represented graphically or by a matrix representation. In a matrix representation, each column and row represent a gene and the matrix elements represent causal relationships. In a graph, the nodes represent the genes and the edges between the nodes reflect the interactions between the genes. To each edge corresponds an interaction table (matrix representation) that expresses the type and strength of the interaction between the nodes it connects.

A first group of inference methods explicitly uses the graphical network representation. As such algorithms based on Boolean models have been proposed (84,85). Interactions are modeled by Boolean rules and expression levels are described by two discrete values. Although such discrete representations require relatively few data, the discretization leads to a considerable loss of information that was present in the original expression data. Most Boolean models cannot cope with the noise of the experimental data or with the stochasticity of the biological system although certain attempts have been made (86).

Bayesian networks (or belief networks) are from that perspective more appropriate (87). Because of their probabilistic nature, they cope with stochasticity automatically. Also, in this probabilistic framework, additional sources of information can easily be taken into account (76). With a few exceptions that can handle continuous data (88,89), most of the inference implementations based on Bayesian networks require data discretization. Bayesian networks can also cope with hidden variables (90). Hidden variables represent essential network components for which no changes in expression can be observed, either because of measurements error (then called missing variables), or because of biological reasons, e.g., the compound acts at posttranslational level. Inference algorithms based on Bayesian networks have been developed both for static data (76,88,89,91,92) and dynamic data (87,93,94).

The probabilistic nature of Bayesian networks certainly offers an advantage over the deterministic characteristics of Boolean networks. The downside, however, is the extensive data requirement that is much less explicit in the simpler Boolean models than in Bayesian networks. To combine the best of both methods, a hybrid model based on the use of Bayesian Boolean networks has been proposed. This method combines the rule-based reasoning of the Boolean models with probabilistic characteristics of Bayesian networks (95). A second group of methods uses the matrix, representation of a network. These methods are based on linear or nonlinear models, In linear models, each gene transcription level depends linearly on the expression level of its parents, for instance represented by linear differential equations (96,97). Nonlinear models make use of black box representations such as neural networks (98), nonlinear differential equations (99), or nonlinear differential equations based on empirical rate laws of enzyme kinetics (100). Nonlinear optimization methods are used to fit the model equations to the data and to estimate the model parameters. Estimating all of the parameters requires an unrealistic large amount of data. The matrix method of singular value decomposition (SVD) has been proposed to solve linear models more efficiently and to generate a family of possible candidate networks for the undetermined problem (101–104).

To this day, genetic network inference is, given the relatively small number of available experiments, an undetermined problem. The solution of any algorithm will therefore pinpoint a number of possible solutions, i.e., networks that are equally consistent with the data. To further reduce the number of possible networks, design methods have been developed (105). These methods predict, based on a first series of experiments, the consecutive set of experiments that will be most informative. Close collaboration between data-analysts and molecular biologists using experiment design procedures and consecutive series of experiments will be indispensable for biological relevant inference. Practical examples where genetic network inference has resulted in the reconstruction of at least part of a network are rare. Most of the successful studies use heuristic methods that are based on biological intuition and that combine expression data with additional prior knowledge (e.g., 77,106).

3.3. Microarray Analysis for Predictive Toxicology

Every toxicological compound affects the expression of genes in a specific way. Every gene, represented on the array, therefore, has a characteristic expression level triggered by the compound. All these characteristic gene expression levels contribute to a profile that is specifically associated with a certain compound (typical fingerprint or reference profile or experiment profile). Each reference profile thus consists of a vector with thousands of components (one, component for each probe present on the array) and corresponds to a certain column of the expression matrix (see Sec. 2.3). Assuming that compounds with a similar mechanism of toxicity are associated with the alteration of a similar set of genes, they should exhibit similar reference profiles, in our setup, a class or a group of compounds corresponds to the set of compounds that have a similar characteristic profile.

Based on this reasoning, reference databases are constructed. For each class of compounds, representatives, for which the toxicological response is well-characterized mechanistically are selected. For these representatives, reference profiles are assessed. The main goal of predictive toxicology is to determine the class to which a novel compound belongs by comparing its experiment profile to the reference profiles present in the database. However, due to its huge dimension (thousands of components), it is impossible to use the complete experiment profile at once in predictive toxicology. Prediction is based on a selected number of features (genes or combination of genes) that are most correlated with the class differences between the compounds (that are most discriminative). Identification of such features relies on feature extraction methods (Sec. 3.3.1). Sometimes the number of classes and the exact identity of classes present in the data are not known, i.e., it is not known in advance which of the tested compounds belong to the same class of compounds. Class discovery (or clustering of experiments) is an unsupervised technique that tries to detect these hidden classes and the features associated with them (Sec. 3.3.2). Eventually, once the classes and related features have been identified in the reference database, classifiers can be constructed that predict the class to which a novel compound belongs (class prediction or classification Sec. 3.3.3).

3.3.1. Feature Selection

Due to its high dimensionality, using the complete experiment profile to predict the class membership of a novel compound is infeasible. Dimensions need to be reduced, e.g., the profile consisting of the expression levels of 10,000 genes will be reduced to a profile that only consists of a restricted number of most discriminative features (e.g., 100). The problem of dimensionality reduction thus relates to the identification of the genes for which the expression profile is most correlated with the distinction between the different classes of compounds. Several approaches for feature selection exist, some of which will be elaborated below.

3.3.1.1. Selection of Individual Genes

The aim is to identify single genes the expression of which is correlated with the class distinction one is interested in. Features then correspond to these individual genes (i.e., single gene features). Because not all genes have an expression that contains information about a certain class distinction, some genes can be omitted when studying these classes. Contrary to class discovery, feature extraction as described here requires that the class distinction is known in advance (i.e., it is a supervised method). For this simple method of feature selection, standard statistical tests to identify two variables that are significantly different from each other are applicable (*t*-test, Wilcoxon rank-sum test,...—see Sec. 3.2.1). Other specialized methods have been developed such as the nonparameter rank based methods of Park et al. (107) or the measure of correlation described by Golub et al. (108).

Also here methods for multiple testing are required (see Sec. 3.2.1). Indeed, a statistical test has to be calculated for every single gene in the dataset (several thousands!). As a consequence, several genes will be selected coincidentally (they will have a high score or low *p*-value without having any true correlation with the class distinction, i.e., they are false positives).

Although frequently applied in predictive applications (109,110), using single gene features might not result in the best predictive performance. Indeed, in general, a class distinction is not determined by the activity of a single gene, but rather by the interaction of several genes. Therefore, using a combination of genes as a single feature is, a more realistic approach (see Sec. 3.3.1.2).

3.3.1.2. Selection of a Combination of Genes

In this section, methods for dimensionality reduction are described that are based on the selection of different combinations (linear or nonlinear) of gene expression levels as features.

Principal component analysis is one of the methods that can be used in this context (111). PCA finds linear combinations of the gene expression levels of a microarray experiment in such a way that these linear combinations have maximal spread (or standard deviation) for a certain collection of microarray experiments. In fact, PCA searches for the combinations of gene expression levels that are most informative. These (linear) combinations are called the principal components for a particular collection of experiments and they can be found by calculating the eigenvectors of Σ (co variance matrix of A—note that in this formula A has to be centralized, i.e., the mean column vector of A has to lie in the origin):

$$\Sigma = \frac{1}{p-1} A \cdot A'$$

where A is the expression matrix ($n \times p$ matrix—collection of p microarray experiments where n gene expression levels were measured). The eigenvectors or principal components with the largest eigenvalues also correspond to the linear combinations with the largest spread for the collection of microarray experiments represented by A. For a certain experiment, the linear combinations (or features) themselves can be calculated by projecting the expression vector (for that experiment) onto the principal components. In general, only the principal components with the largest eigenvalues will be used. So when (1) E ($n \times 1$) is the expression vector for a certain microarray experiment (where also n gene expression levels were measured), (2) the columns of P ($n \times m$ matrix) contain the m principal components corresponding to the m largest eigenvalues of A, and (3) F ($m \times 1$) is given by

$$F = P' \cdot E$$

then the m components of F contain the m features or linear combinations for the microarray experiment with expression vector E according to the first m principal components of the collection of microarray experiments represented by A.

As an unsupervised method, PCA can also be used in combination with, for example, class discovery or clustering. Also nonlinear versions of PCA (that use nonlinear combinations—kernel PCA—(112) and PCA-similar methods such as PLS (partial least squares) (113)) are available. AQ6

3.3.1.3. Feature Selection by Clustering Gene Expression Profiles

As discussed in Sec. 3.2.1, genes can be subdivided into groups (clusters) based on the similarity in their gene expression profile. These clusters might contain genes that contribute similarly to the distinction between the different classes of compounds. If the latter is the case, genes within a cluster of gene expression profiles can be considered as one single feature (mathematically represented by the mean expression in this cluster).

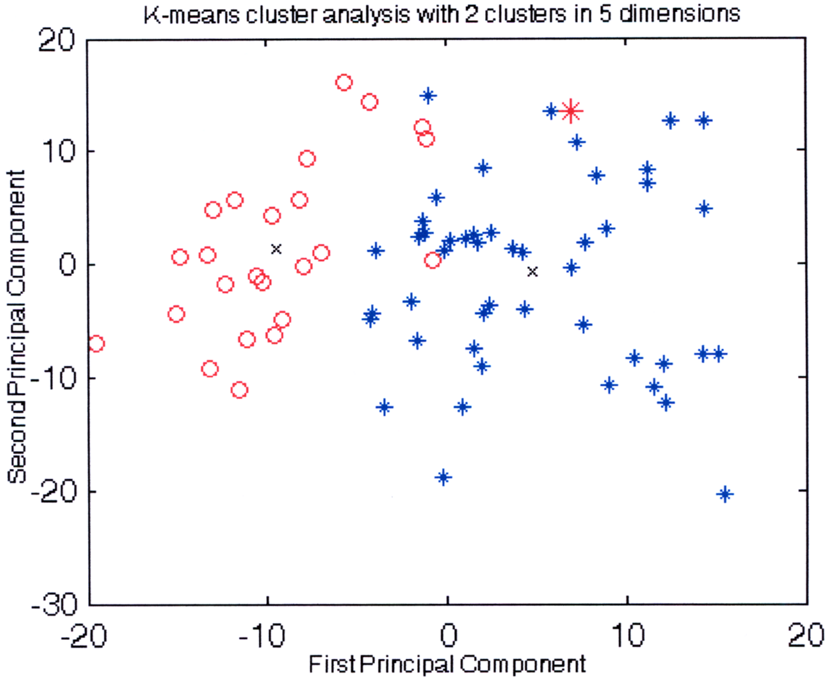
3.3.2. Class Discovery

Compounds or drugs can, according to their effects in living organisms, be subdivided in different classes. These effects are reflected in the characteristic expression profiles of cells exposed to a certain compound (fingerprints, reference profile). The knowledge of these different classes enables classification of new substances. However, the current knowledge of these different classes might still be imperfect. The current taxonomy may contain classes that include substances with a high variability in expression profile. Also current class borders might be suboptimal. All this suggests that a refinement of the classification system and a rearrangement of the classes might improve predicting the behavior of new compounds.

Unsupervised methods such as clustering allow automatically finding the different classes/clusters in a group of microarray experiments, without knowing the properties of these classes in advance (i.e., the classification system of the compounds to which the cells were exposed to is unknown). A cluster, in general, will group microarray experiments (or the associated xenobiotics) with a certain degree of similarity in their experiment expression profile or fingerprint. The distinct clusters identified by the clustering procedure will—at least partially—match with the existing classification used for grouping compounds. However, it is not excluded that novel, yet unknown entities or classes might originate from these analyses.

Several methods (e.g., hierarchical clustering (114), *K*-means clustering (115), self-organizing maps (108), . . .) discussed in Sec. 3.2.2.3 can also be used in this context (i.e., clustering of the experiment expression profiles or columns of the expression matrix instead of clustering the gene expression profiles or rows of the expression matrix). For some methods (e.g., *K*-means—is not able to cluster limited sets of high-dimensional data points), clustering of the experiment profiles must be preceded by unsupervised feature extraction or dimensionality reduction (Sec. 3.3.1) (Fig. 10). F10

When clustering gene expression profiles is performed concurrently with or in preparation of the cluster analysis of the experiment profiles, this is called biclustering. For instance, hierarchical clustering simultaneously calculates a



tree structure for both columns (experiments) and rows (genes) of the data matrix. One can also start with the cluster analysis of the gene expression profiles. Subsequently, one or a subset of these clusters (that seem biologically relevant) is selected. Cluster analysis of the experiments is based on this selection (114). Another technique is to find what is called "a bicluster" (106). A bicluster is defined as a subset of genes that shows a consistent expression profile over a subset of microarray experiments (and vice versa), i.e., one looks for a homogeneous submatrix of the expression matrix (116).

Figure 10 Illustration of class discovery by cluster analysis. The use of microarrays in toxicological gene expression is taking a lead from the work that has been carried out in the field of cancer research. From this field also the following example was taken because of its illustrative value. The dataset derived is from the study of Golub et al. (108) and describes a comparison between mRNA profiles of blood or bone marrow cells extracted from 72 patients suffering from two distinct types of acute leukemia (ALL or AML). Class labels (ALL or AML) were known in advance. In this example, it was demonstrated that the predefined classes could be rediscovered based on unsupervised learning techniques. Patients were clustered based on their experiment profiles (column vectors). Since each experiment profile consisted of the expression levels of thousands of genes (it represents a point in the n-dimensional space), its dimensionality was too high to use K-means clustering without prior dimensionality reduction. Dimensionality was reduced by PCA. The five principal components with the largest eigenvalues were retained and K-means clustering (two clusters) was performed in this five-dimensional space. Patients assigned to the first cluster are represented by circles, patients belonging to the second cluster by stars. Patients with ALL are in blue, and patients with AML are in red. Cluster averages are indicated by black crosses. For the ease of visualization, the experiments (patients) are plotted on the first two principal components. Note that all patients of the first cluster have AML and that almost all patients (with one exception) of the second cluster have ALL.

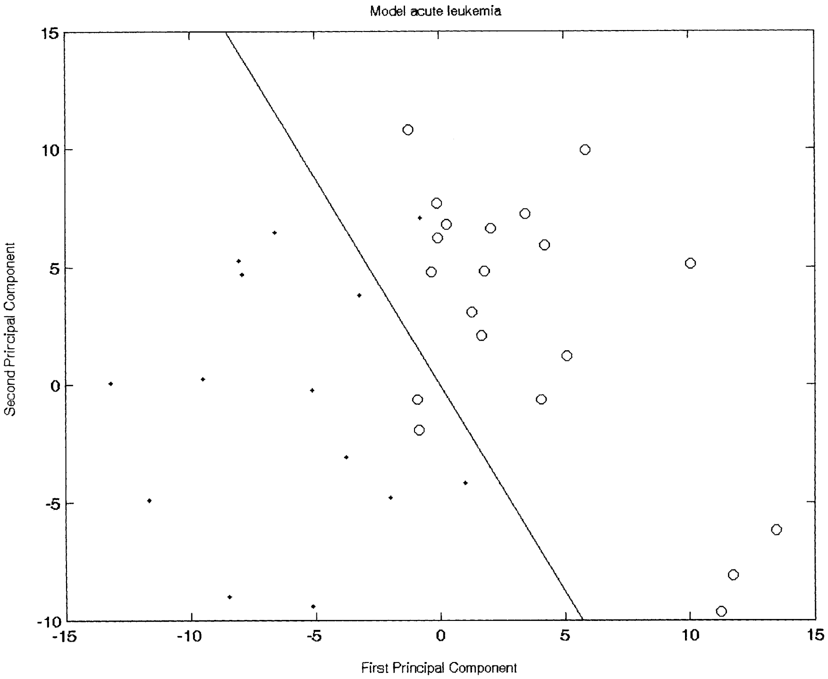


Figure 11 Example of a predictive method. This example resumes the example of Fig. 10 and illustrates the application of a classification model to predict the class membership of patients with acute leukemia based on their experiment profile. A linear classification model was built using Linear Discriminant Analysis based on the first two ($m=2$) principal components of the patients of a training set containing 38 patients. The line in this figure represents the linear classifier for which the parameters were derived using the patients of the training set. Only the patients of the test set (remaining 34 patients) are shown (after projection onto the principal components of the training set). The patients above the line are classified as ALL and below as AML. Note that this resulted in three misclassifications. Test set: ○ = ALL, · = AML.

3.3.3. Class Prediction

Predictive toxicogenomics tries to predict the toxicological endpoints of compounds, with unknown properties or side-effects, by using high-throughput measurements, such as microarrays. This implicates that first the class membership of the novel compound needs to be predicted. Subsequently, the properties of the unknown compound will be derived through extrapolation of the characteristics of the reference members of the class of compounds to which the unknown compound was predicted to belong.

To be able to predict the class membership of novel compounds, a classifier has to be built. Based on a set of features and a training set (reference database), a classifier model (like neural networks (111), support vector machines (112), linear discriminant analysis (111), Bayesian networks (117,118),...) will be trained. This means that the parameters of the model will be determined using the data in the training set (Fig. 11). This classifier is subsequently used to predict the class membership of a novel compound. F11

4. CONCLUSIONS AND PERSPECTIVES

Conclusively, the use high-throughput molecular biological data have much to offer the mechanistic and predictive toxicologist. The impact of these data on toxicological research will grow with the size of public datasets and reference databases. The combination and interpretation of all the data generated will be a major computational challenge for the future that can only be tackled by an integrated effort of both experts in toxicology and data analysis.

ACKNOWLEDGEMENTS

De Moor is a full professor at the K.U. Leuven. Kathleen Marchal is a postdoctoral researcher of the Belgian Fund for Scientific Research (FWO-VLaanderen); Frank De Smet is a research assistant of the K.U. Leuven. Kristof Engelen is research assistant of the IWT. B. This work is partially supported by: (1) IWT projects: STWW-00162, GBOU-SQUAD-20160; (2) Research Council KULeuven: GOA Mefisto-666, IDO genetic networks; (3) FWO projects: G.0115.01, G.0413.03, and G.0388.03; (4) IUAP V-22 (2002–2006).

REFERENCES

1. Ulrich R, Friend SH. Toxicogenomics aad drug discovery: will new technologies help us produce better drugs? Nat Rev Drug Discov 2002; 1:84–88.

2. Gerhold D, Lu M, Xu J, Austin C, Caskey CT, Rushmore T. Monitoring expression of genes involved in drug metabolism and toxicology using DNA microarrays. Physiol Genomics 2001; 5:161–170.

3. Waring JF, Ciurlionis R, Jolly RA, Heindel M, Uirich RG. Microanay analysis of hepatotoxins in vitro reveals a correlation between gene expression profiles and mechanisms of toxicity. Toxicol Lett 2001; 120:359–368.

4. Waring JF, Gum R, Morfitt D, Jolly RA, Ciurlionis R, Heindel M, Gallenberg L, Buratto B, Ulrich RG. Identifying toxic mechanisms using DNA microarrays: evidence that an experimental inhibitor of cell adhesion molecule expression signals through the aryl hydrocarbon nuclear receptor. Toxicology 2002; 181:537–550.

5. Amin RP, Hamadeh HK, Bushel PR, Bennett L, Afshari CA, Paules RS. Genomic interrogation of mechanism(s) underlying cellular responses to toxicants. Toxicology 2002; 181:555–563.

6. Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M. Interrelating different types of genomic data, from proteome to secretome: ‘oming in on function. Genome Res 2001; 11:1463–1468.

7. Clarke PA, te Poele R, Wooster R, Workman P. Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. Biochem Pharmacol 2001; 62:1311–1336.

8. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog 1999; 24:153–159.

9. Hamadeh HK, Amin RP, Paules RS, Afshari CA. An overview of toxicogenomics. Curr Issues Mol Biol 2002; 4:45–56.

10. Pennie WD, Kimber I. Toxicogenomics; transcript profiling and potential application to chemical allergy. Toxicol In Vitro 2002; 16:319–326.

11. Naudts B, Marchal K, De Moor B, Verschoren A. Is it realistic to infer a gene network from a small set of microarray experiments. Internal Report ESAT/SCD K.U.Leuven. http://www. esat.kuleuven. ac.be/~sistawww/cgi-bin/pub.pl.

12. Pennie WD. Use of cDNA microarrays to probe and understand the toxicological consequences of altered gene expression. Toxicol Lett 2000; 112:473–477.

13. de Longueville F, Surry D, Meneses-Lorente G, Bertholet V, Talbot V, Evrard S, Chandelier N, Pike A, Worboys P, Rasson JP, Le Bourdelles B, Remacle J. Gene expression profiling of drug metabolism and toxicology markers using a low-density DNA microarray. Biochem Pharmacol 2002; 64:137–149.

14. Gant TW. Classifying toxicity and pathology by gene-expression profile—taking a lead from studies in neoplasia. Trends Pharmacol Sci 2002; 23:388–393.

15. Pennie WD. Custom cDNA microarrays; technologies and applications. Toxicology 2002; 181–182:551–554.

16. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 2001; 29:365–371.

17. Gerhold DL, Jensen RV, Gullans SR. Better therapeutics through microarrays. Nat Genet 2002; 32(suppl):547–551.

18. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nat Genet 1999; 21:10–14.

19. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Genet 1999; 21:20–24.

20. Epstein JR, Leung AP, Lee KH, Walt DR. High-density, microsphere-based fiber optic DNA microarrays. Biosens Bioelectron 2003; 18:541–546.

21. Southern EM. DNA microarrays. History and overview. Methods Mol Biol 2001; 170:1–15.

22. Blohm DH, Guiseppi-Elie A. New developments in microarray technology. Curr Opin Biotechnol 2001; 12:41–47.

23. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nat Genet 1999; 21:33–37.

24. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. J Cornput Biol 2000; 7:819–837.

25. Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel H. Normalization strategies for cDNA microarrays. Nucleic Acids Res 2000; 28:E47.

26. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M, Johnston R. An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. Nucleic Acids Res 2001; 29:E41–E41.

27. Kerr MK, Churchill GA. Experimental design for gene expression microarrays. Biostatistics 2001; 2:183–201.

28. Yang YH, Speed T. Design issues for cDNA microarray experiments. Nat Rev Genet 2002; 3:579–588.

29. Churchill GA. Fundamentals of experimental design for cDNA microarrays. Nat Genet 2002; 32(suppl):490–495.

30. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. Bioinformatics 2001; 17:509–519.

31. Marchal K, Engelen K, De Brabanter J, Aerts S, De Moor B, Ayoubi T, Van Hummelen P. Comparison of different methodologies to identify differentially expressed genes in two-sample cDNA microarrays. J Biol Syst 2002; 10:409–430.

32. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust

composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002; 30:el5.

33. Colantuoni C, Henry G, Zeger S, Pevsner J. SNOMAD (Standardization and NOrmalization of MicroArray Data): web-accessible gene expression data analysis. Bioinformatics 2002; 18:1540–1541.

34. Engelen K, Coessens B, Marchal K, De Moor B. MARAN: a web-based application for normalizing micro-array data. Bioinformatics 2003; 19:893–894.

35. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 2001; 8:625–637.

36. Kerr MK, Churchill GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci USA 2001; 98:8961–8965.

37. Didier G, Brezellec P, Remy E, Henaut A. GeneANOVA—gene expression analysis of variance. Bioinformatics 2002; 18:490–491.

38. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol 2001; 8:37–52.

39. Ideker T, Thorsson V, Siegel AF, Hood LE. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. J Comput Biol 2000; 7:805–817.

40. Rocke DM, Durbin B. A model for measurement error for gene expression arrays. J Comput Biol 2001; 8:557–569.

41. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001; 98:5116–5121.

42. Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 2002; 18:546–554.

43. Long AD, Mangalam HJ, Chan BY, Tolleri L, Hatfield GW, Baldi P. Improved statistical inference from DNA microarray

data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in *Escherichia coli* K12. J Biol Chem 2001; 276:19937–19944.

44. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 2002; 18:1454–1461.

45. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report #578, Stanford University, 2000:1–38.

46. Storey JD, Tibshirani R. Statistical significance for genome-wide studies. Proc Natl Acad Sci USA 2003; 100:9440–9445.

47. Moreau Y, De Smet F, Thijs G, Marchal K, De Moor B. Functional bioinformatics of microarray data: from expression to regulation. IEEE Proc 2002; 30:1722–1743.

48. De Moor B, Marchal K, Mathys J, Moreau Y. Bioinformatics: organisms from Venus, technology from Jupiter, algorithms from Mars. Eur J Control 2003; 9:237–278.

49. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998; 95:14863–14868.

50. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altaian RB. Missing value estimation methods for DNA microarrays. Bioinformatics 2001; 17:520–525.

51. De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y. Adaptive quality-based clustering of gene expression profiles. Bioinformatics 2002; 18:735–746.

52. Tou JT, Gonzalez RC. Pattern classification by distance functions. Pattern Recognition Principles. Adison-Wesley, 1979:75–109. AQ7

53. Kohonen T. Self-Organizing Maps. Berlin, Germany: Springer-Verlag, 1997.

54. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of

gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999; 96:2907–2912.

55. Tavazoie S, Hughes JD, Campbell MJ, Cho PJ, Church GM. Systematic determination of genetic network architecture. Nat Genet 1999; 22:281–285.

56. Ghosh D, Chinnaiyan AM. Mixture modelling of gene expression data from microarray experiments. Bioinformatics 2002; 18:275–286.

57. Lukashin AV, Fuchs R. Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. Bioinformatics 2001; 17:405–414.

58. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. Bioinformatics 2001; 17:977–987.

59. Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics 2001; 17:126–136.

60. Heyer LJ, Kruglyak S, Yooseph S. Exploring expression data: identification and analysis of coexpressed genes. Genome Res 1999; 9:1106–1115.

61. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol 2000; 1:RESEARCH0003.

62. Ben Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. J Comput Biol 1999; 6:281–297.

63. Sharan R, Shamir R. CLICK: a clustering algorithm with applications to gene expression analysis. Proc Int Conf Intell Syst Mol Biol 2000; 8:307–316.

64. Thijs G, Moreau Y, De Smet F, Mathys J, Lescot M, Rombauts S, Rouze P, De Moor B, Marchal K. INCLUSive: INtegrated Clustering, Upstream sequence retrieval and motif Sampling. Bioinformatics 2002; 18:331–332. http://www.esat.kuleuven.ac.be/~dna/BioI/Software.html.

65. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2003; 31:68–71.

66. Xia X, Xie Z. AMADA: analysis of microarray data. Bioinformatics 2001; 17:569–570.

67. Azuaje F. A cluster validity framework for genome expression data. Bioinformatics 2002; 18:319–320.

68. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000; 406:536–540.

69. McShane LM, Radmacher MD, Freidlin B, Yu R, Li MC, Simon R. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. Bioinformatics 2002; 18:1462–1469.

70. Marchal K, Thijs G, De Keersmaecker S, Monsieurs P, De Moor B, Vanderleyden J. Genome-specific higher-order background models to improve motif detection. Trends Microbiol 2002; 11:61–66.

71. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol 2002; 9:447–464.

72. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 2001; 17:1113–1122.

73. Brazhnik P, de la Fuente A, Mendes P. Gene networks: how to put the function in genomics. Trends Biotechnol 2002; 20:467–472.

74. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science 2002; 297:1183–1186.

75. Rao CV, Wolf DM, Arkin AP. Control, exploitation and tolerance of intracellular noise. Nature 2002; 420:231–237.

76. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Combining location and expression data for principled discovery of genetic regulatory network models. Pac Symp Biocomput 2002; 437–449.

77. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 2002; 298:799–804.

78. Banerjee N, Zhang MQ. Functional genomics as applied to mapping transcription regulatory networks. Curr Opin Microbiol 2002; 5:313–317.

79. Zhang Z, Gerstein M. Reconstructing genetic networks in yeast. Nat Biotechnol 2003; 21:1295–1297.

80. Laub MT, McAdams HH, Feldblyum T, Fraser CM, Shapiro L. Global analysis of the genetic network controlling a bacterial cell cycle. Science 2000; 290:2144–2148.

81. Guelzim N, Bottani S, Bourgine P, Kepes F. Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet 2002; 31:60–63.

82. Rung J, Schlitt T, Brazma A, Freivalds K, Vilo J. Building and analysing genome-wide gene disruption networks. Bioinformatics 2002; 18(suppl 2):S202–S210.

83. Thieffry D, Salgado H, Huerta AM, Collado-Vides J. Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12. Bioinformatics 1998; 14:391–400.

84. Liang S, Fuhrman S, Somogyi R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pac Symp Biocomput 1998; 18–29.

85. Akutsu T, Miyano S, Kuhara S. Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. J Comput Biol 2000; 7:331–343.

86. Akutsu T, Miyano S, Kuhara S. Inferring qualitative relations in genetic networks and metabolic pathways. Bioinformatics 2000; 16:727–734.

87. Murphy K, Mian I. Modelling gene expression data using dynamic Bayesian networks. Technical Report 1999, Computer Science Division, University of California, Berkeley, CA. http://www.cs .berkeley.edu/∼murphyk/publ.html.

88. Yoo C, Thorsson V, Cooper GF. Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. Pac Symp Biocomput 2002; 498–509.

89. Imoto S, Goto T, Miyano S. Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. Pac Symp Biocomput 2002; 175–186.

90. Friedman N. Learning belief networks in the presence of missing values or hidden variables. Proceedings of the 14th International Conference on Machine Learning (ICML) 1997.

91. Pe'er D, Regev A, Elidan G, Friedman N. Inferring subnetworks from perturbed expression profiles. Bioinformatics 2001; 17(suppl 1):S215–S224.

92. Friedman N, Nachman I, Linial M, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol 2000; 7:601–620.

93. Ong IM, Glasner JD, Page D. Modelling regulatory pathways in *E. coli* from time series expression profiles. Bioinformatics 2002; 18(suppl 1):S241–S248.

94. Smith VA, Jarvis ED, Hartemink AJ. Evaluating functional network inference using simulations of complex biological systems. Bioinformatics 2002; 18(suppl 1):S216–S224.

95. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 2002; 18:261–274.

96. D'Haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 2000; 16:707–726.

97. Chen T, He HL, Church GM. Modeling gene expression with differential equations. Pac Symp Biocomput 1999; 29–40.

98. Wahde M, Hertz J. Coarse-grained reverse engineering of genetic regulatory networks. Biosystems 2000; 55:129–136.

99. Akutsu T, Miyano S, Kuhara S. Algorithms for inferring qualitative models of biological networks. Pac Symp Biocomput 2000; 293–304.

100. Kato M, Tsunoda T, Takagi T. Merring genetic networks from DNA microarray data by multiple regression analysis. Genome Inform Ser Workshop Genome Inform 2000; 11:118–128.

101. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 2000; 97:10101–10106.

102. Yeung MK, Tegner J, Collins JJ. Reverse engineering gene networks using singular value decomposition and robust regression. Proc Natl Acad Sci USA 2002; 99:6163–6168.

103. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV. Firadamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci USA 2000; 97:8409–8414.

104. Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 2000; 455–466.

105. Ideker TE, Thorsson V, Karp RM. Discovery of regulatory interactions through perturbation: inference and experimental design. Pac Symp Biocomput 2000; 305–316.

106. Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. Bioinformatics 2002; 18(suppl 1):S136–S144.

107. Park PJ, Pagano M, Bonetti M. A nonparametric scoring algorithm for identifying informative genes from microarray data. Pac Symp Biocomput 2001; 52–63.

108. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer:

class discovery and class prediction by gene expression monitoring. Science 1999; 286:531–537.

109. Xu J, Stolk JA, Zhang X, Silva SJ, Houghton RL, Matsumura M, Vedvick TS, Leslie KB, Badaro R, Reed SG. Identification of differentially expressed genes in human prostate cancer using subtraction and microarray. Cancer Res 2000; 60:1677–1682.

110. Wang K, Gan L, Jeffery E, Gayle M, Gown AM, Skelly M, Nelson PS, Ng WV, Schummer M, Hood L, Mulligan J. Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. Gene 1999; 229:101–108.

111. Bishop CM. Neural Networks for Pattern Recognition. New York: Oxford University Press, 1995.

112. Suykens J, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least Squares Support Vector Machines. Singapore: World Scientific, 2002.

113. Johansson D, Lindgren P, Berglund A. A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. Bioinformatics 2003; 19:467–473.

114. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Standt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000; 403:503–511.

115. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet 2001; 2:418–427.

116. Sheng Q, Moreau Y, De Moor B. Biclustering microarray data by Gibbs sampling. Bioinformatics 2003; 19(suppl 2):II196–II205.

117. Moreau Y, Antal P, Fannes G, De Moor B. Probabilistic graphical models for computational biomedicine. Methods Inf Med 2003; 42:161–168.

118. Jordan M. Learning in Graphical Models. Cambridge, MA, London: MIT Press, 1999.

119. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 1998; 2:65–73.

REFERENCES AS INTRODUCTORY TO TOXICOGENOMICS

1. Clarke PA, te Poele R, Wooster R, Workman P. Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. Biochem Pharmacol 2001; 62:1311–1336.
2. Amin RP, Hamadeh HK, Bushel PR, Bennett L, Afshari CA, Paules RS. Genomic interrogation of mechanism(s) underlying cellular responses to toxicants. Toxicology 2002; 181–182:555–563.
3. Ulrich R, Friend SH. Toxicogenomics and drug discovery: will new technologies help us produce better drugs? Nat Rev Drug Discov 2002; 1:84–88.
4. Vrana KE, Freeman WM, Aschner M. Use of microarray technologies in toxicology research. Neurotoxicology 2003; 24:321–332.

REFERENCES AS INTRODUCTORY TO METHODOLOGICAL REVIEWS

1. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet 2001; 2:418–427.

GLOSSARY

AQ8

Additive error: This represents the absolute error on a measurement that is independent of the measured expression level. Consequently, the relative error is inversely proportional to the measured intensity and is high for measurements with low magnitude.

Bayesian network: This represents a mathematical model that allows both a compact representation of the joint probability distribution over a large number of variables, and an efficient way of using this representation for statistical inference. It consists of a directed acyclic graph that models the interdependencies between the variables, and a conditional probability distribution for each node with incoming edges (see Chapter ???) AQ9

Class discovery: This represents the automatic identification of the hidden classes in a dataset without a priori knowledge on the class distinction. The data reduction or grouping is derived solely from the data. This can be obtained by using unsupervised learning techniques such as, e.g., clustering.

Classification/prediction: This represents determination for a certain experiment (microarray experiment of a certain compound) of its class membership based on a classifier or predictive model: objects are classified into known groups. Classification is based on supervised learning techniques.

Clustering: This represents unsupervised learning technique that organizes multivariate data into groups with roughly similar patterns, i.e., clustering algorithms group together genes (experiments) with a similar expression profile. Similarity is defined by the use of a specific distance measure.

Coexpressed genes: These are genes with a similar expression profile. Genes of which the behavior of the expression is similar in different conditions or at different timepoints.

Data matrix: This is a Mathematical representation of a complex microarray experiment. Each row represents the expression vector of a particular gene. Each column of the matrix represents an experimental condition. Each entry in the matrix represents the expression level of a gene in a certain condition.

Dedicated microarrays: These contain only a restricted number of genes, usually marker genes or genes characteristics for a certain toxicological endpoint. Using dedicated arrays offers the advantage of higher throughput screening of lead targets at a lower cost.

Diagnostic or investigative microarrays: These contain probes representing as much coding sequences of a genome as possible.

DNA Microarray: This is a High-throughput technology that enables the measurement of mRNA transcript levels at a genomic scale. DNA microarrays are produced by high density depositing thousands of individual spots (called probes) of synthetic unique oligonucleotides or cDNA gene sequences to a solid substrate such as a glass microscope slide or a membrane.

Dye reversal experiment: This is a specific type of experimental design used for cDNA arrays. On the first array the test condition is labeled with Cy5 (red dye), while the reference is labeled with Cy3 (green dye). On the second array, the dyes are swapped, i.e., reference condition is labeled with Cy5 (red dye), while the test is labeled with Cy3 (green dye)

Dynamic experiment: This is a complex microarray experiment that monitors adaptive changes in the expression level elicited by administering the xenobiotic to the system under study. By sampling the system at regular time intervals during the time course of the adaptation, short-, mid-, and long-term alterations in xenobiotic-induced gene expression are measured.

Expression profile of a gene: This is a vector that contains the expression levels of a certain gene measured in the different experimental conditions tested; corresponds to the row in the data matrix.

Expression profile of an experiment/compound: (also “fingerprint” or “reference pattern”) This is a vector that contains the expression levels of all genes measured in the specific experimental condition represented by the column; corresponds to the column in the data matrix.

FDR: The FDR (false discovery rate) is considered as a sensible measure of balance between the number of false positives and true positives. The FDR is the rate that the features called significant are truly null or the number of false positives among the features called significant.

Feature: This represents a gene (single feature) or combination of genes (complex feature) of which the expression levels are associated with a class distinction of interest (e.g., of which expression is switched on in one class and switched off in the other class).

Feature extraction: This represents mathematical or statistical methodology that identifies the features that are most correlated with a specific class distinction.

Filtering: This represents removal of genes from the dataset of which the expression does not change over the tested conditions, i.e., genes that are not involved in the process studied.

Global normalization assumption: This is a general assumption stating that, from one biological condition to the next, only of a small fraction of the total number of genes shows an altered expression level and that symmetry exists in the number of genes for which the expression is upregulated vs. downregulated.

Mechanistic toxicogenomics: This involves the use of high-throughput technologies to gain insight into the molecular biological mechanism of a toxicological response.

Missing values: These are gene expression values that could not be accurately measured and that were omitted form the data matrix.

Multiple testing: When considering a family of tests, the level of significance and power are not the same as those for an individual test. For instance, a significance of $\alpha = 0.01$ indicates a probability of 1% of falsely rejecting the null hypothesis (e.g., assuming differential expression while there is none). This means that for a family of 1000 tests, say every 1000 genes tested, 10 would be expected to pass the test although not being differentially expressed. To limit this number of false positives in a multiple test, a correction is needed (e.g., Bonferroni correction).

Multiplicative error: This represents the absolute error on the measurement increases with the measurement magnitude. The relative error is constant, but the variance between replicate measurements increases with the mean expression value. Multiplicative errors cause signal-dependent variance of the residuals.

Network inference: This represents reconstruction of the molecular biological structure of regulatory networks from high-throughput measurements, i.e., deriving the causality relationships between genetic entities (proteins, genes) from the data.

PCA: Principal component analysis (see other Chapter ?) AQ9

Predictive model or classifier: This represents a mathematical model (neural network, Bayesian model,...) of which the parameters are estimated by the use of a trainingsset (i.e., the reference database). The predictive model is subsequently used to predict the class membership of a novel compound, i.e., to assign a novel compound to a predefined class of compounds based on its expression profile.

Predictive toxicogenomics: This involves the prediction of the toxicological endpoints of compounds, with yet unknown properties or side-effects by the aid of high-throughput profiling experiments such as microarrays. A reference database of expression fingerprints of known compounds and a predictive model or classifier trained on this reference database are needed.

Preprocessing: This is a pretreatment process, that removes consistent and/or systematic sources of variation from the raw data.

Power: This represents the discriminant power of a statistical test (computed as $1-\beta$) and the probability of rejecting the null hypothesis when the alternative hypothesis is true (a decision known as a *Type II error*). It can be interpreted as the probability of correctly rejecting a false null hypothesis. Power is a very descriptive and concise measure of the sensitivity of a statistical test, i.e., the ability of the test to detect differences.

Probes: These resent the spots/oligonucleotides on the microarray that represent the different genes of the genome.

Reference databases: This is a compendium of characteristic expression profiles or fingerprints of well-described agents or compounds, for which both the toxicological endpoints and the molecular mechanisms resulting in them are characterized.

Rescaling microarray data: This represents transformation of the gene expression profiles by subtracting the mean expression level and by dividing by the standard deviation of the profile. This operation results in a collection of expression profiles all being 0 and with a standard deviation of 1.

Significance: This represents the significance level of a statistical test, referred to as α, and the maximum *probability* of accidentally rejecting a true *null hypothesis* (a decision known as a *Type I error*). The significance of a single result is also called its p-value, i.e., the lowest possible a that would lead to the acceptance of the null hypothesis for that result.

Static experiments: This is a complex microarray experiment that tests the induced changes in expression under several conditions or in different genetic backgrounds (gene knock out experiments). Samples are taken when the steady state expression levels are reached.

SVD: Singular value decomposition (see other Chapter ???) AQ9

Target: These are the labeled transcripts, present in the mRNA sample that is hybridized to the array.

Test statistic: This value is calculated from the data points (e.g., a mean) and used to evaluate a null hypothesis against an alternative hypothesis. In the framework of testing for differentially expressed genes, the null hypothesis states that the genes are not differentially expressed.

Toxicogenomics: This is a Subdiscipline of toxicology that combines large scale gene/protein expression measurements and the expanding knowledge of genomics to identify and evaluate genome-wide effects of xenobiotics.

Underdetermination: The number of parameters to be estimated exceeds the number of experimental data points. The mathematical problem has no single solution.