"Deep insight" into microarray technology

Claudia Schoch, Martin Dugas, Wolfgang Kern, Alexander Kohlmann, Susanne Schnittger, Torsten Haferlach

Laboratory for Leukemia Diagnostics, University Hospital Grosshadern, Department of Medicine III, Ludwig-Maximilians-University, 81366 Muenchen, Germany

So far the classification of tumors relies on the interpretation of clinical, histopathological, immunophenotypic, Cytogenetic and molecular genetic findings. Especially in hematological tumors a precise analysis of the malignant cells using classical methods such as cytomorphology and histology which both are supplemented by cytochemistry and multiparameter immunophenotyping are used in routine diagnostics for classification. Furthermore, insights into the genetic basis of the disease, i.e. disease-specific chromosomal aberrations and molecular alterations detected in the malignant cell clone, have substantially increased the importance of Cytogenetics, fluorescence in situ hybridization (FISH), and polymerase chain reaction (PCR) and their combination in establishing the diagnosis in each subentity. In the clinical setting this not only implies a better understanding of the course of distinct disease subtypes but also allows the selection of disease-specific therapeutic approaches, e.g. the use of all-trans retinoic acid (ATRA) in acute promyelocytic leukemia or of imatinib in chronic myeloid leukemia. This is also true for an early application of allogeneic transplantation strategies in AML with complex aberrant karyotypes. Given this genetic background the microarray technology may become an essential tool for the optimization of the classification of tumors and thus may be used as a routine method for diagnostic purposes in the near future.

Microarray technology

In general two different fields in microarray based techniques can be distinguished: those approaches dealing with copy number changes on the DNA level - so called Array- or Matrix-CGH (comparative genomic hybridization) - and those approaching gene expression measuring changes at the RNA level.

Array based CGH

Conventional comparative genomic hybridization (CGH) is a technique based on the use of genomic DNA as a probe. It provides an overview of DNA sequences copy number changes (losses, deletions, gains, amplifications) in a specimen and maps these changes on normal chromosomes (du Manoir et al., 1993; Kallioniemi et al., 1992). CGH is based on the in situ hybridization of differentially labeled total genomic test DNA and control reference DNA to normal human metaphase chromosomes. Copy number variations among the different sequences in the tumor DNA are detected by measuring the test/control fluorescence intensity ratio for each locus in the normal metaphase chromosomes. CGH only detects changes that are present in a substantial proportion of cells (>50%). It does not reveal translocations, inversions and other aberrations that do not change copy numbers. Therefore, CGH allows the comprehensive analysis of the entire genome in just one experiment providing information not only about the size of all chromosomal imbalances but also on their chromosomal band specific assignment. Using the conventional CGH approach hybridizing test and control DNA on metaphases the resolution is limited to the Cytogenetic band resolution, approximately 5-10 Mb for deletions and 2 Mb for amplifications. In 1997, Solinas-Toldo and co-workers established a matrix-based CGH array that replaces condensed metaphase chromosomes by defined cloned DNA probes immobilized on a glass surface as hybridization target, allowing automated analysis of genetic imbalances such as microdeletions and overrepresentations (Solinas-Toldo et al., 1997). This technique was called array CGH by others (Pinkel et al., 1998). More details on strategies for constructing
Gene expression analysis based on microarray technology

Gene expression analyses using microarrays have become an important part of the biomedical basic and clinical science. Common to all microarray approaches is the basic principle of complementary base pairing. Complementary nucleotide strands interact non-covalently and then can be detected (Southern et al., 1999). By the application of gene expression profiling the present gene expression status of a cell or a cell population is used to build a molecular fingerprint. To allow this the respective gene sequences are coated on a solid layer at high density. Thus, performing only one experiment the expression status of thousands of genes can be assessed simultaneously. Using specific software packages a judgement on the respective genes with regard to its expression is possible for distinct time points or disease states. Moreover, besides the qualitative assessment the results can be evaluated also quantitatively which may be highly relevant both in the pathogenesis of a malignant disease and for its clinical management as has been shown for the overexpression of human epidermal growth factor receptor-2 (HER2) in breast cancer. Gene expression profiles thus provide a molecular fingerprint of the transcriptome. If and when genes are expressed, however, is influenced by many developmental and tissue-specific factors.

Microarray platforms for gene expression analysis

Two basically different methods are available: desoxyribonucleic acid oligonucleotide (DNA) microarrays and arrays spotted with complementary DNA (cDNA) (figure 1). Gene-specific sequences are synthesized in situ at defined positions on the DNA-oligonucleotide microarrays (Lockhart et al., 1996). The sample to be analyzed (5 µg total ribonucleic acid (RNA), which is equivalent to 1 to 8 x 10⁶ cells) is translated into cDNA and subsequently amplified in an in vitro transcription. During this process biotinylated ribonucleotides are incorporated by the T7-RNA-polymerase into the growing cRNA strands. Per single experiment 10 µg fragmented and biotin-labeled cRNA are hybridized to the microarray. After staining of the DNA-cRNA hybrids using streptavidin-phycoerythrin their detection is accomplished by an argon laser (figure 2). On the present HG-U133A+B microarray (Affymetrix) for example about 33,000 human genes are represented each by 11 different oligonucleotides. In addition, sequence-specific oligonucleotides carrying a central point mutation are used as an internal mismatch control. The results obtained by this approach are highly reproducible. Optimizing the homogeneity of the cellular composition to which the microarray technology will be applied is essential for the selection of the respective samples. A significant mixture of malignant and non-malignant cells may lead to hybridization results not exactly reflecting the expression levels present in the tumor. Thus, microdissection of single tumor cells is applied in some analyses of solid tumors (Ramaswamy and Golub, 2002).

Figure 1 : The test (i.e. tumor) DNA and the control (i.e. normal) DNA are labeled with different fluorochromes and Cot-1 DANN is added. This mixture is hybridized onto the microarray. Appropriate software is used to calculate fluorescent intensity ratios that reflect genomic imbalances between test and control DNA. A ratio of 1 indicates equal copy numbers of genomic regions in the test and the control DNA. Deviations from this ratio are indicative of either gains or losses in the test genome.
Spotted cDNA-Arrays are applied mostly for specific questions defined by the user targeting on a selected number of gene sequences which are brought onto the respective surfaces as cDNA (e.g. glass slide, nylon membrane) (Duggan et al., 2004). The RNA of interest and a control RNA are then labeled by different fluorochromes and co-hybridized. This approach allows greater flexibility in single analyses as compared to the predefined whole genome DNA-oligonucleotide microarrays. However, more experience and skills are needed, with regard to the selection of sequences, the optimization of protocols for hybridization and washing procedures, the manufacturing of the array and the reproducibility in general.

For both methods the detection of the hybridization signals requires a specific microarray scanner connected to a database which is essential for the analysis of the large amount of data. In addition, various algorithms must be applied to optimize the evaluation of the data.

**Bioinformatics, data management, and functional annotation**

Microarray data is characterized by a large number (typically >20,000) of measurement values per patient sample. It is difficult to be interpreted, because the number of genes is much higher than the number of samples and the data correlation structure is not well understood. Therefore the challenge is to distinguish between random and significant patterns of gene expression. Clinical data is complicated, too. There are many different items, both quantitative and categorical, often with complex coding schemes. Clinical data is multi-dimensional, because each clinical data source (e.g. a specific lab technique) provides a different view on the same patient. For this reason a systematic approach in data management with detailed planning and documentation is strongly recommended to keep an overview with microarray analysis in a clinical setting.

Microarray studies generate vast quantities of data. Even small studies easily sum up to several gigabytes of raw expression data. Several steps have to be performed before the raw data are in shape for a biological or clinical evaluation. After quality control of the sample and the hybridization, image processing, translation of images into signal intensities and normalization data is prepared for data mining.

Quality control (QC) is a key issue in microarray analysis. It should be addressed both from a biological and a bioinformatic perspective. First, one should always look at the raw data, i.e. scanned chip images, to identify artifacts. QC parameters recommended by the microarray manufacturer should be assessed, for instance in Affymetrix arrays 3'/5' ratios for selected genes and percentage of "present calls". Reproducibility of microarray measurements should be assessed.
After image processing, the first analysis step is to produce a large number of quantified gene expression values. These values represent absolute fluorescence signal intensities as a direct result of hybridization events on the array surface. It is also possible to qualitatively rate gene expression as absent/present detection call calculations (Ramaswamy and Golub, 2002; Liu et al., 2002) Before analyzing the data it is a routine procedure to normalize the raw data (Quackenbush, 2002). This is a mandatory step in the data mining process in order to appropriately compare the measured gene expression levels. From the same experiments several data sets and lists of significant genes may be derived by different data normalization/calibration methods. Affymetrix chip data can be transformed by different techniques, for instance mas5 (http://www.affymetrix.com) or rma (http://www.bioconductor.org) (Irizarry et al., 2003). In addition, there are preprocessing techniques like global scaling, quantile normalization (dchip), (Li and Wong, 2001) variance stabilization (vsn) (Huber et al., 2002) and others. So far there is no generally accepted gold standard for quality control and data calibration.

Data mining, the discovery of non-obvious information, often uses mathematical techniques that have traditionally been used to identify patterns in complex data. Recently, they have been adapted for functional genomics needs. There are two different approaches to analyze the data, i.e. the unsupervised approach and the supervised approach. An unsupervised analysis does not use any a priori class definition, instead simply seeks to determine what structure is inherent in the data. A supervised analysis aims at uncovering putative associations. Therefore, it may bias the outcome by forcing a model onto the data.

Unsupervised analysis

Unsupervised analysis of microarray data aims to detect groups of samples or outliers without knowledge of the clinical features of each sample. For instance new subgroups of a disease with consistent "molecular signature" can be identified. Commonly used methods are principal component analysis (PCA) (Jolliffe, 1986) and hierarchical clustering (Eisen et al., 1998). PCA reduces dimensionality of the data set while retaining most of information contained therein via construction of a linear transformation matrix. Principal components are the projections of the data on the eigenvectors. In figure 3 each point corresponds to one patient; its coordinates are derived from the principal components. Hierarchical clustering is an unsupervised method for organizing expression data into groups with similar signatures (figure 4). There are several methods to calculate similarity (euclidean, manhattan, pearson) and several procedures to link similar sets of samples (single, average and complete linkage). Consequently, several different hierarchical clusterings can be generated from the same microarray data set, which may complicate interpretation. Hierarchical clustering can be used to reduce the complexity of the matrix-like data and to visualize it in a more understandable way.

Figure 3: Principal components are the projections of the data on the eigenvectors. In figure 3 a principal component analysis is shown based on gene expression signatures from n=800 genes which we identified to be differentially expressed when analysing AML samples with t(15;17) (n=40), t(8;21) (n=40), inv(16) (n=40), t(11q23)/MLL (n=40), or complex aberrant karyotypes (n=40). In the three-dimensional graph data points with similar characteristics will cluster together. Here, each patient’s expression pattern is represented by a single color-coded sphere. The feature space consisted of measured expression data from n=800 genes. The respective karyotype label, i.e. t(15;17), t(8;21), inv(16), t(11q23)/MLL, or complex aberrant was unknown to the algorithm. The labels and coloring of the classes were added after the analysis for means for better visualization. AML patients with t(15;17) are colored blue, t(8;21) red, inv(16) yellow, t(11q23)/MLL turquoise, and complex aberrant karyotype cases pink, respectively.
Figure 4: Hierarchical cluster analysis is a popular unsupervised method for arranging genes and patients according to underlying similarities in patterns of gene expression. In figure 4 a hierarchical clustering based on U133A microarray expression data of our adult AML samples (columns) is shown. This analysis is based on a subset of n=800 genes (rows) which we identified to be differentially expressed when analysing AML samples with t(15;17) (n=40), t(8;21) (n=40), inv(16) (n=40), t(11q23)/MLL (n=40), or complex aberrant karyotypes (n=40). The normalized expression value for each gene is coded by color (standard deviation from mean). Red cells indicate high expression and green cells indicate low expression. The respective karyotype label, i.e. t(15;17), t(8;21), inv(16), t(11q23)/MLL, or complex aberrant was unknown to the algorithm. The labels and coloring of the classes were added after the analysis for better visualization. AML patients with t(15;17) are colored blue, t(8;21) red, inv(16) yellow, t(11q23)/MLL turquoise, and complex aberrant karyotype cases pink, respectively.

Patterns in the data are discovered solely from the data itself as there is no previous knowledge or grouping of the data. Two-dimensional hierarchical clustering sorts both patients and genes according to similarities and leads to a tree-structured dendrogram which can easily be viewed and explored (Eisen et al., 1998).

It is clear that this hierarchical structure provides potentially useful information about the relationship between adjacent clusters. Common crossing points represent similar patient characteristics as well as similarities with regard to the co-expression of distinct genes (Eisen et al., 1998). It has to be kept in mind that both PCA and hierarchical clustering can be strongly influenced by selection of genes.

Supervised analysis

Typically, it is of great interest to correlate array data directly with e.g. clinical, cytomorphological, or Cytogenetic features. In supervised analysis the diagnosis or classification of each sample is known. The key issue of this evaluation is to identify significant differentially expressed genes. Usually classical biostatistical methods are applied, in particular t-test, ANOVA, correlation and regression methods. Given the large number of genes, results should be adjusted for multiple testing to exclude random patterns. A common approach is calculation of the false discovery rate (FDR), (Tusher et al., 2001) which is available in the SAM [http://www-stat.stanford.edu/~tibs/SAM/] and q-value software (Tusher et al., 2001; Storey and Tibshirani, 2003). There are different types of clinical response variables: Categorical (for example diagnosis A,B,C), quantitative (like creatinine level) and survival (e.g. disease-free survival time + survival-status). In each type of supervised analysis, a list is generated containing genes associated with the clinical response variable. The number of significant genes is determined by the choice of significance level.

Classification based on gene expression data

After detecting differential gene expression it is often necessary to accurately classify samples into known groups. There are quite a few methods to classify samples for instance support vector machines (SVM), (Vapnik, 1998) PAM, (Tibshirani et al., 2002) classification and regression trees (CART), k-Nearest-Neighbor (k-NN) and others. There is evidence that SVM-based prediction slightly outperforms other classification techniques (Furey et al., 2000; Brown et al., 2000). To estimate diagnostic accuracy, the prediction model is built using a training set of patient samples. Accuracy is determined based on predictions in an independent test set. Apparent accuracy usually is determined by 10-fold crossvalidation (10-fold CV):
The data set is divided into 10 approximately equally sized subsets, the prediction model is trained for 9 subsets and predictions are generated for the remaining subset. This training / prediction process is repeated 10 times to include predictions for each subset. Apparent accuracy is the overall rate of correct predictions. Random sets of training and test data can be generated iteratively to assess robustness of diagnostic accuracy. Since most software packages still rely on strong informatic/statistic knowledge and programming skills robust, intuitive and user-friendly software is needed. Bioconductor, for example, is an open source and open development software project to provide tools for the analysis and comprehension of genomic data (www.bioconductor.org/).

Functional Annotation

By use of detailed gene annotation it is possible to find functional groupings of genes based on their similarity among the gene expression profiles. This information can be used to get new insights into physiological pathways and may also help to characterize previously uncharacterized genes (Eisen et al., 1998). A structured and normalized annotation of the respective genes and gene products, essential for all evaluations, is provided by the Gene Ontology™ consortium (Ashburner et al., 2000). The three principles of organization and possibilities for the annotation are based on the description of the molecular function of the gene product (e.g. enzyme or transporter), of the biologic process in which one or more molecular functions are involved (e.g. cell growth or signal transduction), and on an assignment of the cellular localization (e.g. nucleus or integral membrane protein). In addition, detailed hierarchical models are provided, e.g. the metabolism of DNA is further separated into replication and repair of DNA. Public available databases like NetAffx provide regularly updated functional gene annotations (Liu et al., 2003). Furthermore, relevant gene informations are connected to other various databases like OMIM (www.ncbi.nlm.nih.gov/omim/), SWISSPROT (http://us.expasy.org/sprot/) and thereby substantiate the biologic knowledge about a gene and its gene product and facilitate the interpretation of microarray experiment results. Another way of accelerating the pace of data analysis is to approach the data from a higher level of organization instead from a gene-by-gene basis. MAPPfinder is such a useful application and integrates and links GO™ annotations to array expression data allowing to identify gene expression changes directly on particular pathways (Dahlquist et al., 2002; Doniger et al., 2003). The effective annotation of microarray experiments is a major task which is approached by the MGED group (Microarray Gene Expression Data Group, www.mged.org) (Brazma et al., 2001). This consortium defines standards for the annotation of microarray experiments (MIAME, Minimum Information About a Microarray Experiment) as well as a standard data-exchange format (MAGE-ML, Microarray Gene Expression Markup Language). Based on these standards global expression databases have been established with the aim to give access to, to share and to compare microarray data. In accordance with the MGED recommendations, ArrayExpress (www.ebi.ac.uk/arrayexpress) is such a public repository for microarray data. The NCBI has launched the Gene Expression Omnibus (GEO), a gene expression and hybridization array data repository, as well as an online resource for the retrieval of gene expression data from any organism or artificial source (www.ncbi.nlm.nih.gov/geo/).

Biological networks analysis

In order to evaluate the role of significantly deregulated genes in the pathogenesis of a certain disease the question arises whether some of these genes are involved in a common pathway. Such biological networks can be generated through the use of Ingenuity Pathways Analysis, a web-delivered application that enables scientists to discover, visualize and explore therapeutically relevant networks significant to their experimental results, such as gene expression array data sets. For a detailed description of Ingenuity Pathways Analysis, visit www.ingenuity.com.

First, genes have to be identified whose expression is significantly differentially regulated between two groups. For generating molecular networks that indicate how these genes may influence each other a cut-off of i.e. 5% FDR (q-value < 0.05) can be set. This data set containing gene identifiers and their corresponding expression signal intensities can be uploaded as a tab-delimited text file into the Ingenuity Pathways Knowledge Base. Then each probe set is automatically mapped to its corresponding data base gene object to designate so-called focus genes. Focus genes are genes from the analysis input data file that meet both of the following criteria: These genes have been designated as being of interest, i.e. a certain level of significance. They directly interact with other genes in the Ingenuity global molecular network, which consists of direct physical, enzymatic, and transcriptional interactions between mammalian orthologs from the published, peer-reviewed content in Ingenuity’s Pathways Knowledge Base (IPKB).

To start building the networks, the application queries the Ingenuity Pathways Knowledge Base for interactions between focus genes and all other gene objects stored in the knowledge base, and generates a set of networks with a network size of 35 genes/gene products. The application then computes a score for each network according to the fit of the user’s set of significant genes. The score is derived from a p-value and indicates the likelihood of the focus genes in a network being found together due to random chance. A
score of 2 indicates that there is a 1 in 100 chance that the focus genes are together in a network due to random chance. Therefore, scores of 2 or higher have at least a 99% confidence of not being generated by random chance alone. Biological functions are then calculated and assigned to each network.

The networks are displayed graphically as nodes (genes/gene products) and edges (the biological relationships between the nodes). The intensity of the node color indicates the degree of up- (green) or down-

(red) regulation. As described in the legend provided, nodes are displayed using various shapes that represent the functional class of the gene product. Edges are displayed with various labels that describe the nature of the relationship between the nodes (e.g., B for binding, T for transcription). The length of an edge reflects the evidence supporting that node-to-node relationship, in that edges supported by more articles from the literature are shorter. Two examples are shown in figures 5 and 6;
Results of microarray analyses

Results of genomic imbalances assessed by array based CGH

The sensitivity and quantitative capability of array based CGH for the measurement of gene dosage was analyzed by Pinkel et al. In the same study it was demonstrated that this technique was useful to detect DNA copy number aberrations in cancer. The first genome-wide array based CGH analysis of DNA gains and losses was performed in breast cancer (Pollack et al., 1999).

In chronic lymphocytic leukemia recurrent known as well as new genomic alterations were identified using array based CGH (Schwaenen et al., 2004). Several studies analysed DNA copy number changes in breast cancer. Array based CGH provided a higher resolution mapping of already known amplicons and revealed that amplicons are either simple or very complex, either showing a simple peak, i.e. the amplicon encompassing ERBB2 or in contrast the 11q amplicon showing a complex pattern with amplification of CCND1 usually accompanied by amplifications of several distinct adjacent copy number peaks as well as loss of copy number (Albertson, 2003). Also studies on classification of renal cell cancer, gastric cancer and liposarcoma based on copy number profiles assessed by array based CGH have been reported (Weisset al., 2004; Wilhelm et al., 2004).

Some studies have already been performed using genome wide array based CGH in combination with global gene expression profiling and found a good correspondence between copy number alterations and changes in gene expression (Hyman et al., 2002; Pollack et al., 2002). In a study of liposarcoma, copy number profiles had a greater power to discriminate between dedifferentiated and pleomorphic subtypes than expression profiling (Fritz et al., 2002).

Gene expression data

A magnitude of different questions was approached by microarray based gene expression analyses during the last years. Besides new insights into the pathophysiological and ontogenetic context (Enard et al., 2004) data on malignant diseases in particular are of high interest. Following the pivotal work of Golub et al. who provided data on the applicability of microarrays and new biostatistical methodologies (Golub et al., 1999) even more detailed analyses were performed. The published data can be separated in...
categories addressing different aspects: aim of
a) reproducibility of data on different array platforms or
different technologies,
b) “class prediction” (prediction of a tumor entity based
on specific gene expression profiles of selected
informative genes),
c) “class discovery” (discovery of new subentities
within groups formerly regarded as homogeneous),
d) prediction of prognosis,
e) prediction of response to therapy and f) new insights
into the pathogenesis.
During the recent months a large number of original
papers as well as reviews on gene expression analys
is in leukemia have been published. Therefore, only a few
examples are mentioned highlighting the different
aspects gene expression analysis based on microarrays
can address.
One of the first studies applying the microarray
technology to leukemia has been performed by Golub
et al. who demonstrated that both ALL and AML are
characterized by a distinct and specific gene expression
profile. In a pilot study bone marrow samples of 27
patients with ALL and 11 patients with AML have
been analyzed and discriminatory genes were identified
allowing the distinction of both of these large entities
of leukemias based on their gene expression profiles. A
total of 50 genes had been sufficient to classify acute
leukemias in this study: in 36 of 38 cases the molecular
diagnosis of leukemia was made correctly based on the
gene expression profile as analyzed on the microarray.
In a further set of 34 unknown samples which had not
been used to build up the classification model the
classification was also correct in the majority of cases,
i.e. 29 cases were assigned correctly. These analyses
represent the first and a major step towards molecular
diagnostics of acute leukemias (Golub et al., 1999).
Another report dealt with the question if AML with
trisomy 8 as the sole Cytogenetic aberration can be
separated from AML with a normal karyotype based on the
gene expression profile (Virtaneva et al., 2001). However,
in this analysis a totally correct separation has not been possible which may be due to trisomy 8
not being the primary lesion leading to AML but rather
a secondary aberration in addition to a molecular event.
Nonetheless, a gene dosage effect could be clearly
demonstrated since many genes which are coded on the
chromosome 8 showed an elevated gene expression in
general. This effect of gene dosage on gene expression
has been confirmed and was in addition also shown for
trisomy 11 and trisomy 13 as well as for monosomy 7
and deletion 5q in AML.
A different aspect with respect to the relationship
between genetic abnormalities on the genomic level and
gene expression was addressed in several studies. It
was demonstrated that Cytogenetically defined
subtypes in AML such as AML with t(8;21), AML with
t(15;17), and AML with inv(16) are characterized by
different and specific gene expression profiles (Schoch
et al., 2002). The basic genetic alterations lead to
patterns of gene expression that can be unequivocally
detected by microarrays. A minimum set of only 13
genes has been sufficient to accurately predict the
karyotypes in the respective AML samples (Schoch et
al., 2002). In a next step the analyzed cohort has been
extended by eight samples with normal bone marrow.
In this setting the expression profiles of 35 genes were
sufficient to predict with an accuracy of 100% if the
sample contained normal bone marrow, AML M2 with
t(8;21), APL with t(15;17), or AML M4eo with inv(16)
(Kohlmann et al., 2001). A further step consisted in the
addition of samples with AML carrying aberrations of
chromosome 11q23, i.e. MLL gene rearrangements,
representing an analysis of AML subtypes with recurring
chromosomal aberrations as defined by the
WHO (Kohlmann et al., 2002; Kohlmann et al., 2003).
In this analysis a minimum set of 39 genes has been
sufficient to classify based on the gene expression
profile samples as normal bone marrow or AML with
one of the aberrations t(8;21), t(15;17), inv(16), or
11q23-rearrangements. Thus, the differential
expression of these 39 candidate genes is sufficient to
classify Cytogenetically defined subtypes of AML and
to separate these from normal bone marrow. The
accuracy of this classification as determined by a leave-
one-out crossvalidation amounts to 100%.
Also in ALL data demonstrate that subtypes
characterized by specific genetic abnormalities show
distinct gene expression profiles. Armstrong et al. were
the first to demonstrate that childhood ALL with
chromosomal aberrations involving the MLL gene can
be regarded as a molecularly defined entity distinct
from other ALL (Armstrong et al., 2002). MLL-
positive ALL had a distinct gene expression profile
consistent with an early hematopoietic progenitor
expressing multilineage markers and specific HOX
genes. Also the comparison of these ALL with MLL
aberration with other ALL subtypes as well as with
AML samples resulted in a clear separation of all three
groups from each other.
Furthermore, the global gene expression profiles of
ALL (n=10) and AML (n=15) both t(11q23)/MLL-
positive was analyzed by Kohlmann et al. using U133A
arrays (Affymetrix). Based on 20 top-ranked genes
both leukemias were discriminated with a 100%
accuracy (permutation-based neighborhood analysis)
(Kohlmann et al., 2002).
A milestone in microarray analysis with respect to class
discovery, class prediction, prediction of outcome was
the report by Yeoh et al. on 327 childhood ALL (Yeoh
et al., 2002). Patients were discriminated according
to their Cytogenetic and immunological as well as to their
molecular subtype of ALL, i.e. T-ALL, E2A-PBX1,
BCR-ABL, TEL/AML1, MLL rearrangements, and
hyperdiploid ALL. Many of the relevant genes were
also verified in an analysis of corresponding adult ALL
patients (Kohlmann et al., 2004). As some patients
were not classified according to these subgroups Yeoh et al. postulated a novel subgroup of ALL that was characterized by high expression of genes including the receptor phosphatase PTPRM and LHFPL2. Ross et al. rehybridized 132 childhood ALL probes using the Affymetrix U133 set and identified almost 60% of new discriminating genes in comparison to their previous analysis. As a proportion of these new genes were highly ranked as class discriminators and led to an overall diagnostic accuracy of 97% in several analyses the authors suggested to assess these genes expression profiles in a prospective clinical setting. The main clinical focus should be at diagnosis of ALL with respect to accuracy, practicality, and cost effectiveness in comparison to standard diagnostic techniques.

A pivotal work with regard to “class discovery” by microarrays was presented by Alizadeh et al. who showed that cases with diffuse large-cell B-cell lymphoma can be separated into two molecularly different groups based on their gene expression profiles. This separation of the formerly homogeneously regarded group in addition resulted in two groups with highly differing prognoses. Interestingly, one of these two new subgroups was characterized by a gene expression pattern which was closely related to follicular B-cells. In contrast, the other group was related to activated B-cells as present in the peripheral blood. Thus, these contributions clearly demonstrated the possibility by gene expression profiling to define new subclasses within entities formerly regarded as homogeneous (Alizadeh et al., 2000). Further work also demonstrated the capability of this technique to separate a large variety of solid tumors (Ramaswamy et al., 2001).

In another paper Yagi et al. analysed 54 pediatric AML focused on the reproducibility of morphological subtypes of AML and especially on gene patterns to predict outcome (Yagi et al., 2003). After unsupervised clustering they were able to differentiate patients with t(8;21) from those with inv(16), and from those demonstrating an AML M4/5, or AML M7 phenotype or immunophenotype by specific gene expression signatures. Within this unsupervised analysis no specific profile was found that correlated to the prognosis of the patients. Since the inclusion of further cases with other FAB subtypes and Cytogenetic abnormalities (no karyotype available in 9 of 54 cases) resulted in an increased heterogeneity the authors restricted their further analyses to the genetically and morphologically better defined subentities. For further calculation data were analyzed supervised with respect to outcome and prognosis. A subset of 35 genes was selected that was independent from the morphology or karyotype of these patients, some of them are associated with regulation of the cell cycle or with apoptosis. By hierarchical cluster analysis patients could be classified into high-risk and low-risk groups with highly significant prognostics on EFS (p<0.001).

With respect to therapy response in ALL Hofmann et al. analyzed 25 bone marrow samples from 19 patients with Philadelphia-positive ALL all treated with the tyrosine kinase inhibitor imatinib using the HuGene FL arrays (Affymetrix) (Hofmann et al., 2002). The patients in this study were selected according to their Cytogenetic response to the drug and 95 genes were identified to predict the treatment outcome in their cohort. Another 56 genes were found to predict leukemia cells that had secondary resistance to the drug after remission had been achieved. These results point to further applications of microarray profiling leading to tailored therapy approaches and withhold therapies unlikely to induce remission. Especially in this field of gene expression profiling further studies with independent test cohorts are urgently needed.

In a recent study Cheok et al. examined gene expression profiles in 60 childhood ALL cases before and after in vivo treatment with methotrexate and mercaptopurine given alone or in combination (Cheok et al., 2003). A total of 124 differentially expressed gene before and after the corresponding treatment were identified capable of accurately discriminating the four possible treatment groups. Genes included those involved in apoptosis, mismatch repair, cell-cycle control, and stress response. These data indicate that leukemia cells in different patients react in a similar matter after specific treatments and therefore share common pathways of genomic responses to different drug schedules. It has further been published very recently that subdiscrimination in AML focussing on molecular markers such as FLT3-LM or CEPBA (Valk et al., 2004) and on prognostication in patients with normal karyotypes seems also possible by specific gene expression profiles and sophisticated biomathematical approaches (Bullinger et al., 2004; Grimwade et al., 2004).

**Future applications**

Microarrays provide a means for the comprehensive and simultaneous analysis of the expression status of thousands of genes. The resulting signature allows the identification of a distinct molecular phenotype. It is anticipated that the application of microarrays will significantly improve molecular diagnostics and will provide deep insights into the pathogenetic alterations of malignant and non-malignant diseases which will allow the development of novel treatment approaches. To realize these efforts it is important to cover well-defined questions in well-classified tumor samples. Furthermore, it is expected that the results of microarray experiments will allow the identification of disease-specific target structures and the design of novel and specific drugs.

The comparisons of different gene expression analyses will provide answers for both diagnostic and biological questions. Besides the above mentioned reports on
disease-specific gene expression profiles in a variety of malignancies even the potential of a primary tumor to progress into the metastatic status may be detected and predicted by microarray analyses (Phimister, 2002). Gene expression analyses are limited, however, to the transcriptional level and thus cannot be used to analyze alterations at the protein level, e.g. the phosphorylation and dephosphorylation of proteins. These aspects will be covered by novel techniques like proteomics.

With regard to leukemia the microarray analysis represents a novel promising method which may be used as a diagnostic tool in the near future. Today the diagnosis and subclassification of leukemia is based on the application of various techniques like cytomorphology, Cytogenetics, fluorescence in situ hybridization, multiparameter flow cytometry, and PCR-based methods which are time-consuming and cost-intensive and require expert knowledge in central reference laboratories. The microarray analysis serves the potential basis for the diagnosis of AML and other leukemias to be performed on a single platform at a high degree of validity and cost-effectiveness. Before this can be introduced into practice, however, extensive analyses will be necessary that compare the results of gene expression profiling with the results of methods considered standard for diagnostic purposes today.

References


This article should be referenced as such: