DNA MICROARRAY TECHNOLOGY IN MEDICAL DIAGNOSTICS

TEHNOLOGIJA DNA MIKROMREŽ V MEDICINSKI DIAGNOSTIKI

Damjana Rozman[1], Peter Juvan[2]

[1] Center za funkcijsko genomiko in bio-čipe, Medicinska fakulteta, Zaloška cesta 4, SI-1000 Ljubljana
[2] Fakulteta za elektrotehniko, Tržaška cesta 25, SI-1001 Ljubljana

Abstract

Background *DNA microarray technology opened a completely new venue in the biomedical research. By allowing to follow the expression of thousands of genes in a single experiment (ideally we can follow the expression of the entire genome), the microarray technology brings new perspectives for improvement of prognostics and diagnostics of complex human diseases. By SNP arrays that allow identification of individual patient's genotype for frequent genetic diseases, this technology contributes to the efforts towards personalized medicine.*

Methods *Different approaches to microarray technology are presented with some examples of its application in medicine and pharmacogenomics. Practical considerations for using microarrays are described, including selecting an appropriate platform, designing and performing an experiment, and managing data. Statistical issues in microarray data analysis are raised and the most common analysis techniques are listed. In conclusion the current state of microarray technology in Slovenia is addressed.*

Key words *microarray technology; human genome; gene expression; SNP*

Izvleček

Izhodišča *Tehnologija DNA mikromrež je prinesla nov vpogled v biomedicinske raziskave. Z zmožnostjo sledenja izražanja tisočerih genov hkrati (v limiti lahko sledimo izražanje celotnega genoma), tehnologija mikromrež prinaša nove obete za izboljšanje prognoze in diagnoze kompleksnih obolenj pri človeku. Pomaga pa tudi utirati pot osebni medicini, saj je s tematskimi SNP mikromrežami možno ugotavljati genotip bolnikov za bolj znana in pogostejša genetska obolenja.*

Metode *Predstavljeni so pristopi tehnologije mikromrež in primeri njene uporabe v medicini in farmakogenomiki. Podani so praktični napotki za uporabo mikromrež, kot so izbira primerne platforme, načrtovanje in izvedba poskusa, in upravljanje s podatki. Izpostavljen je statističen pogled na analizo podatkov mikromrež in naštete so najpogosteje uporabljene metode. Trenutno stanje tehnologije mikromrež v Sloveniji je opisano v zaključku.*

Ključne besede *tehnologija mikromrež; človeški genom; izražanje genov; SNP*

Introduction

Organisms are complex systems where thousands of genes and their products (RNA and proteins) interact with each other and make the mystery of life. Genes can be thought of as words in a dictionary of genome, and the procedures for exploring their expression on a large-scale as tools for reading and understanding the book of life.

The human genome consists of about 40.000 genes located on 23 pairs of chromosomes. One chromosome in each pair is inherited from the mother, the other from the father. Each chromosome contains a long molecule of DNA, the chemical of which genes

Corresponding author / Avtor za dopisovanje:
Damjana Rozman, Center za funkcijsko genomiko in bio-čipe, Medicinska fakulteta, Zaloška cesta 4, 1000 Ljubljana

are made. The DNA is a double-stranded molecule in which each strand is a linear array of units called nucleotides or bases. There are four different bases, called A, T, G and C. The bases on one DNA strand are precisely paired with the bases on the other strand, so that an A is always opposite to T and G opposite to C.

Until recently, genetic research has been limited to »single gene – single experiment«. That approach, used since the discovery of DNA in 1953, shaped lives of many researchers, and decades passed before their discoveries were fitted together. The cell is not just a bunch of isolated units; on contrary, signaling, metabolic, transporting and other processes intertwined with each other and in incessant contact with the environment make a life of a cell. Every perturbation, even if directed to a single cell's subunit, e.g. expression of a single gene, has a cascading effect on tens or even hundreds of other units in down-streaming pathways. In order to understand the principle of life we need to resort to global large-scale approaches which enable us to measure thousands of parameters in a single experiment. In the last decade, DNA microarray technology advanced into one of the most efficient techniques enabling us to study the expression of genes on a genome scale.

The microarray technology offers the promise of comprehensive study of complex diseases at a genomic level, potentially identifying novel molecular abnormalities, developing novel clinical biomarkers, and investigating drug efficacy. Possibilities to develop molecular profiles corresponding to therapeutic effects are in accordance with the concept of drug repositioning. Microarray technology has also the potential to contribute to the development of new biomarkers useful as predictors of disease etiology, outcome, and responsiveness to therapy, which is known as »personalized medicine«. We discuss herein the transcriptomic analysis as a novel post-genomic tool with great perspectives to aid in improved understanding, predicting and curing human diseases.[1]

Microarray technology and platforms

DNA microarrays consist of DNA molecules with known nucleotide sequences representing many genes organized into a matrix and bound onto a solid support, usually a microscopic slide. Each gene is represented by a single or more groups of identical DNA molecules called probes. Exposing a microarray to labeled sample extracts (single-stranded nucleic acids molecules) enables us to detect differences in expression of the targeted genes in different samples. In that process, referred to as hybridization, extracts from one or two samples of interest, each labeled with a unique fluorescent label, bind to specific microarray probes by matching complementary base pairs A-T and G-C. Reading the array with a dedicated laser scanner using different excitation wavelengths enables estimation of the strength of the hybridization signals, which are proportional to the amount of the genetic material in corresponding samples.

At first, microarrays were used to measure the expression of genes at the level of their transcription (amount of mRNA), also referred to as gene expression profiling. Nowadays they are also frequently used on a genome level. In addition to expression profiling, applications of this technology include:

- detection of DNA copy number aberrations using comparative genomic hybridization (CGH) arrays;
- identification of genetic variation that are thought to be the source of susceptibility to genetically caused diseases using single nucleotide polymorphism (SNP) arrays;
- sequencing portions of the genome in individuals, examination of novel transcripts, identification of specific transcription factor binding sites (chromatin immunoprecipitation-on-chip (ChIP-on-chip) studies) and methylation studies using arrays that cover an entire genomic region of interest by overlapping probes (genome tiling arrays);
- profiling of microRNAs, small non-coding RNA molecules functioning in post-transcriptional gene silencing, for better understanding of gene regulation, using new miRNA arrays.

Since its introduction in 1995, a variety of technologies for producing microarrays have been developed. In general we distinguish between three probe implementations:

- complementary DNA (cDNA) arrays where probes are synthesized from mRNA and subsequently deposited to a solid surface,
- oligonucleotide arrays where short sequences of nucleotides are synthesized *in situ* on the surface,
- oligonucleotide arrays where probes are synthesized in a tube and later spotted on the array surface.

The pioneers of microarray technology at Patrick Brown's laboratory at University of Stanford prepared 48 cDNA probes approximately a thousand base pairs (bp) long and printed them to a microscopic slide.[2] Nowadays usually 300–500 bp probes are used each representing a single gene, which are most often synthesized from mature (fully spliced) mRNA using the enzyme reverse transcriptase. Such classical approach is also used for production of genomic microarrays with probes more than a hundred thousand bp long. Oligonucleotides are either pre-synthesized and subsequently deposited to a solid surface or synthesized *in-situ*. Each target is represented by either one to three long (50–80 bp) or up to 40 short (25–30 bp) oligonucleotides.

The most common technique for depositing probes to a solid surface include a computer-controlled three dimensional motion robotic arm carrying pins to pick up small drops of solution from microtiter plates and contacting a flat solid surface carrying a relevant surface chemistry for attachment of nucleic acids (spotting). Non-contact printing is another technique, similar in terms of robotics, but instead of pins small dispensing systems are mounted to a robotic arm and droplets of samples are ejected onto a surface avoiding direct surface contact. Both systems are highly reliable, but cannot meet the density of *in-situ* synthesized microarrays pioneered by Affymetrix (http://

www.affymetrix.com). They developed the technology of light-directed DNA synthesis and patented the platform called GeneChip, indicating the correspondence of their production with computer chips. Short (25 bp) oligonucleotides are synthesized by directing light using lithographic masks to specific areas on the array surface, allowing chemical coupling to occur at specific sites. High density GeneChips allow for multiple probes for each expression or genotype measurement, achieving sufficient sensitivity even with short sequences. Each target is measured by 11 (expression) to 20 (genotype) probes matching the target sequence and an equal number of mismatch probes containing a single nucleic acid substitution located directly in the middle of the sequence. Agilent (Pao Alto, CA, USA) rely on long (60 bp) oligonucleotide arrays where probes are synthesized *in situ* using non-contact inkjet technology. Longer sequences are believed to be more sensitive due to the larger area available for hybridization and higher tolerance of sequence mismatches, though requiring only a single probe per target.

Many technical and analytical options are governed by selection of a microarray platform. Despite the fact that GeneChip platform, offering a costly closed-system solution, is currently accepted as the method for determining expression profiles with highest reproducibility, it is not an optimal choice for everyone, especially not for low-throughput research work. Conventional in-house spotted arrays are becoming increasingly popular, particularly for researchers interested in a specific subset of genes, allowing for their rapid customization. Although the price of commercial arrays is dropping rapidly, such boutique arrays allow for a large number of experimental conditions to be examined at a relatively low cost.

Gene expression analysis arrays

Today, both cDNA and oligonucleotide arrays are routinely applied to gene expression studies. In general we distinguish between whole-genome arrays, thematic arrays of commercial interest and custom in-house arrays. In case of whole-genome arrays, probes for as many as possible known genes are packed into an extremely dense matrix allowing for observation of expression of virtually every gene in a genome. Nowadays Affymetrix offers GeneChips covering more than 38.500 well-characterized human genes on a single slide (GeneChip Human Genome U133 Plus 2.0 Array) comprising of 1.3 million distinct probes, and new Agilent's 44K arrays include more than 41.000 unique human genes and transcripts printed in four replications on a single slide. Thematic arrays usually contain a subset of genes of their whole-genome counterparts, allowing cost-effective solutions to researchers interested in specific genes, especially in cancer biology (e.g. GeneChip Human Cancer G110 Array). Agilent recently launched eArray, an online array creation tool for designing custom arrays which can be printed and delivered worldwide within weeks. The service includes customization of probes by either choosing from optimized probes for given genes, uploading custom sequences, or having Agilent design probes, and selection of array format to accommodate the number of selected probes. In-house spotted arrays complement the use of commercially available arrays, offering unlimited possibilities for testing novel hypotheses and transferring discoveries into practice. Many of them are developing into novel diagnostic tools making them relevant for clinical diagnosis. Decision for in-house array production is based upon a balance between large amount of initial work for designing probes and printing process optimization on one hand and flexibility of their subsequent modification and low cost-production on the other.

DNA analysis arrays

DNA microarrays enable determination of nucleotide sequences on the chosen parts of DNA, which is a different approach as compared to expression analysis arrays. Even if the human genome has been mainly sequenced and the information available since 2001, there is still increasing need for (re)sequencing large parts of human DNA from individuals. The human genome is highly polymorphic which is frequently associated with disease. Thus, it is impossible to determine the »perfect« or »universal« sequence of the human genome, since we differ from each other for at least 0.1 % of the genome sequence. There are an increasing number of DNA microarray types as well as applications. We will focus on those that are closest to medical applications.

For the purpose of DNA sequencing tiling microarrays have been developed, where chromosomes are covered by 20 bases long overlapping oligonucleotides that cover the entire chromosome length. These arrays allow determination of individual's genotype and the nucleotide sequence of »healthy« and »disease« alleles. Tiling microarrays of some simple organisms, such as yeast *S. cerevisiae,* cover the entire genome of this microorganism,[3] while tiling human microarrays are currently available for individual human chromosomes (Agilent). In preparation of such microarrays one needs to take into account that every position in the genome can theoretically contain one of the four nucleotides.

A similar principle is also used in preparation of the single nucleotide polymorphism (SNP) arrays that are applied in medicine to characterize individual polymorphisms or mutations. The difference between tiling and SNP arrays is in the scale. While tiling arrays cover large portions of DNA, such as entire chromosomes or even the entire genome, the SNP arrays concentrate on, for example, a single polymorphic gene, taking into account all possible variances within it.

The comparative genomic hybridization (CGH) is another microarray application very relevant to studies of human diseases. It allows determination of quantitative changes on the genome level.[4] These include deletions (the absence of a particular genome part), insertions (inclusion of novel sequences in the genome), amplifications (multiplication of particular genome sequences), and chromosomal re-arrangements (aberrant shuffling of portions of different chro-

mosomes). Classical cytogenetic analyses that have been so far used to detect these chromosomal changes (karyotyping, chromosome banding, fluorescent *in situ* hybridization – FISH, etc) have a relatively low resolution and request analyses in dividing cells. CGH microarrays allow high resolution identification of aberrant sequences. The surface of a CGH array is covered by long (up to several hundred thousand bp) portions of DNA that is representing a defined part of a chosen chromosome. This allows precise determination and mapping of changes to the defined nucleotide sequence. Resolution is defined by the distance between particular DNA regions on the chromosome. CGH microarrays with overlapping large regions of DNA are being developed, allowing a single base resolution.

Applications in medicine and pharmacogenetics

Microarrays have been utilized to address in vitro pharmacology and toxicology issues and are being widely applied to improve the processes of disease diagnosis, pharmacogenomics, and toxicogenomics. The disease can be considered as a disturbed homeostasis, where moderation in one of the players has an effect on many other players. There are too many examples of application to be listed in this short review. Generally, any complex disease, such as cardiovascular, cancer, diabetes, metabolic syndrome, etc, can be approached by microarrays, studying either the modulated gene expression as a consequence of the disease, or the sequence of the gene that is associated with the disease (disease genotype). In the first case we compare the expression level of genes in the control population (healthy subjects, or in the case of cancer, healthy tissue from the same subject) with the expression level in the diseased state. In order to reach a general conclusion, i.e. modulated expression of genes particularly connected to the disease phenotype, samples of several individuals need to be analyzed together with several controls. As in any other biological research, the more samples we use the higher is the confidence in results. However, frequent daily limitations represent either the lack of proper patient samples or the cost of individual experiments. In the case that analyses lead to statistically significant modulated expression of a group of genes in a disease state, these genes represent novel biomarkers that can describe complex diseases. Based on such biomarkers novel classical or array-based diagnostic methods are being developed. In oncology, for example, several »onco-chips« are available[5] each including a collection of oncogenes and tumor suppressor genes whose expression is modulated in different types of cancer. Similarly, arrays are developed to monitor the progression of cardiovascular diseases, etc.

The medical application of SNP arrays includes identification of polymorphisms and mutations in patients. The microarray analysis from different individuals will show the presence or absence of a particular with the disease state linked SNP. For example, with the microarray that includes all so far known polymorphisms and mutations of the human cystic fibrosis (CF) gene, one can detect in a single experiment the CF genotype of an individual patient.[6] This is a tremendous benefit in the sense of time of analysis and well as the accuracy of detecting polymorphisms or mutations, since with classical techniques weeks or months were required and the polymorphisms or mutations were frequently missed. Another example of SNP-array application is in pharmacogenetics. CodeLink P450 SNP bioarray from Amersham Biosciences (New Jersey, USA) covers all known polymorphisms in the major human drug metabolizing enzymes of the cytochrome P450 family, such as the CYP1A1, 1A2, 1B1, 2C9, 2C19, 2D6, 2E, 3A4, and 3A5. This array is designed for screening clinical trial populations to determine their toxicogenetic profiles and for the discovery of novel associations between P450 genotypes and phenotypes.

The comparative human genome (CGH) microarrays have a high potential in oncology where clones of cells are prone to massive DNA rearrangements.[7] The technique allows, for example, a precise determination of the type of the cancer in an individual patient. Precise determination of the type contributes to better treatment which further improves the prognosis. The second common application of CGH arrays is determination of chromosomal aberrations for different genetic diseases that can be used in prenatal diagnosis of major chromosomal aberrations, such is the chromosome 21 trisomy known also as Down's syndrome. The analysis of DNA from different patients shows the presence or absence of the signal, that confirms the presence or absence of a disease-linked chromosomal abnormality.

Practical considerations for using microarrays

Microarray experiment will lead to biologically relevant conclusions only if the resulting data fits the analysis methods that will be used. Therefore designing a sound experiment is a more crucial step than it seems at the beginning, and there exist some general recommendations that should be followed.

To compare the abundance of different sample extracts the targets are first labeled with fluorescent label enabling the estimation of their amount being proportional to the strength of the signal emitted during the scanning. Hybridization, taking place in a hybridization buffer, can be performed manually or using an automatic hybridization station, the latter often producing more reproducible results.

Hybridized arrays are scanned and images analyzed to extract intensities of individual spots. Data is annotated and stored to a dedicated database. Data analysis starts with normalization, an adjustment made to accommodate for meaningful biological comparison. Individual arrays are examined for genes expressed differently between the observed samples. Due to unforeseen biological differences and inherent noise

in microarray replication of measurements is important, enabling distinction between the relevant biological differences and those attributed to differences between individuals. With increasing amount of microarray experiments the final challenge is to reverse engineer the regulatory mechanisms underlying the observed biological system and make an *in silico* model.

One versus two-color system

Historically, microarrays are employed to compare the transcript abundance between two different biological samples on the same array, each labeled with a unique color. In two-color experiments, where two different samples compete for the same probes, gene expression levels are expressed as ratios between the two samples, and a logarithmic transformation (usually base 2) is used to stabilize the variance in data and make its distribution symmetric. Many different strategies of pairing larger number of samples can be used, and it depends on the design of an experiment how many hybridizations will be needed in order to answer relevant biological questions with a certain confidence.

Recently increasing number of platforms are optimized to utilize one-color detection scheme (in contrast to two-color, or even both), allowing higher flexibility in designing experiments. An integral part of such systems is quality control, ensuring comparable conditions across non-competitive environments. Quality control probes, representing synthetic sequences not matching any of the target sequence, are usually scattered over the array in many repetitions. Corresponding spike-in control transcripts are added in increasing concentrations to both labeling reactions and hybridized together with biological samples. In case of equimolar concentrations their intensities should be similar for all samples.

The decision between a competitive or non-competitive hybridization largely depends on the selected platform that will be used. No matter whether we choose to compare samples on a single array using two different dyes or to hybridize them to individual arrays using a single dye, the measurements are subject to technical variation introduced by either differences between dyes or arrays, which decreases our ability of capturing relevant biological differences. For spotted arrays, competitive hybridization is the preferred choice as differences between individual arrays might introduce larger variation to data than that of using two different dyes. Nowadays industrial-scale *in situ* array production reached sufficient reproducibility of printing to allow for non-competitive hybridization thus bringing more flexibility to designing experiments and eliminating the need for different dyes.[8]

Designing an experiment

While there exist some general recommendations for how to design a microarray experiment, the details largely depend on biological question of interest, the complexity of the observed system, and the selected microarray platform. With emergence of one-color systems designing an experiment became less demanding, eliminating the problem of choosing which samples to pair.

In contrast to typical biomedical studies microarray experiments are due to their high costs often limited by the budget available, leading to statistical unsoundness and wrong conclusions. Such experiments merely generate new hypotheses which will require additional testing. The experiment should be grounded on a biological question of interest, which should be well-focused and limited enough to fit the budget.

In planning an experiment we should account for different sources of variation, which can be partitioned to biological variation inherent in individual samples, technical variation introduced during labeling and hybridization, and the error of a laser scanner. Individual components of variation can be estimated (and eliminated) by increasing the number of measured samples within individual groups, repeating the measurements of individual samples, and using arrays with replicated probes. Typical reproducibility of measurements of within-array replicates is more than 95 %, dropping to 60–80 % when using multiple arrays and further down to less than 30 % when multiple biologically similar samples are involved.[9]

A simple test of adequacy of a design involves counting the number of independent experimental units (e.g. samples from different patients) and subtracting the number of treatments (of which effects we wish to observe), which should yield at least 5. Sometimes a large number of samples are available, and samples receiving the same treatment are mixed together, forming pools. Pooling reduces biological component of variation which is important for controlling the number of false positive results of statistical hypothesis tests. Pooling should only be used if the amount of individual samples is insufficient for hybridizations

For two-color platform we need to decide which samples to hybridize together. An efficient design may be achieved following few simple rules.[10] We represent a design by a directed graph where nodes correspond to experimental units and arcs represent hybridizations. Two samples will be hybridized together as many times as there are arcs between them. Arcs also define how samples are labeled (e.g. Cy3 for their origins and Cy5 for the targets). In Figure 1 four different designs are represented involving two treatments (A and B). While the top two designs involve only technical replicates by swapping dyes, in the bottom designs two biological replicates of each treatment (denoted by 1 and 2) are planned, increasing the number of experimental units to four. Analytically we will be only able to compare sample pairs for which there exists a path in a design graph (e.g. using the design shown in Figure 1c we won't be able to compare A1 and B2). The longer the path, the less efficient the comparison will be. When designing an experiment we need to ensure that all the samples that are relevant for comparison are interconnected and that the paths between them are as short as possible. Cyclic designs such as shown in Figure 1d are very efficient, but only for small number of samples. For larger number of

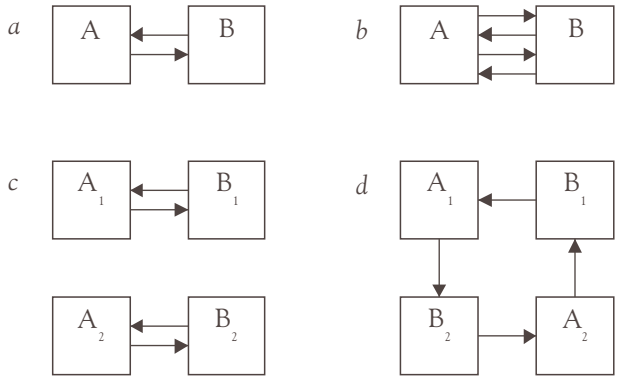

Figure 1. *Four direct experimental designs for a two-color platform involving two treatments (A and B). Nodes correspond to experimental units (e.g. samples from different patients) and arcs represent hybridizations together with the selection of dyes. Designs a and b consists of two and four technical replications of hybridizations, respectively, using dye-swap strategy. In designs c and d two biological replications of each treatment are planned (denoted by 1 and 2).*

Sl. 1. *Štirje neposredni načrti poskusa z dvobarvnimi DNA mikromrežami, ki vključujejo dva tretmaja (A in B). Vozlišča predstavljajo enote poskusa (npr. vzorce različnih bolnikov), povezave pa hibridizacije in izbiro barvil. Načrta a in b predvidevata dve oz. štiri tehnične ponovitve hibridizacij z zamenjavo barv; c in d sta osnovana na dveh neodvisnih bioloških ponovitvah tretmajev (ponazorjeno z 1 in 2).*

samples an indirect design such as shown in Figure 2 can be used where all samples of interest (A and B) are hybridized against a referential sample (Ref), which often has no biological relevancy. It is important that the reference sample is available in abundance (to make additional hybridizations possible) and that it lights up majority of spots. Universal references are commercially available or in-house reference sample can be prepared by mixing together all samples of interest.

Finally, a proper experimental design should account for randomization of all possible experimental units. Randomization ensures that data is not affected by systematic biases. If possible, treatments should be assigned to samples randomly, or samples large enough to represent the population differences should be used. As hybridizations are usually performed in batches random selection of samples and arrays is important.

Labeling, hybridization and scanning

Depending on the platform different strategies for labeling samples are used. A wide variety of different fluorescent dyes is available with cyanine (Cy3, Cy5) and Alexa Fluor family by Invitrogen being the most common. For spotted arrays three methods are employed in most laboratories: direct labeling where dyes are incorporated during the first cDNA strand synthesis from total RNA, postlabeling of cDNA where first-strand cDNA is initially labeled with amino-allyl

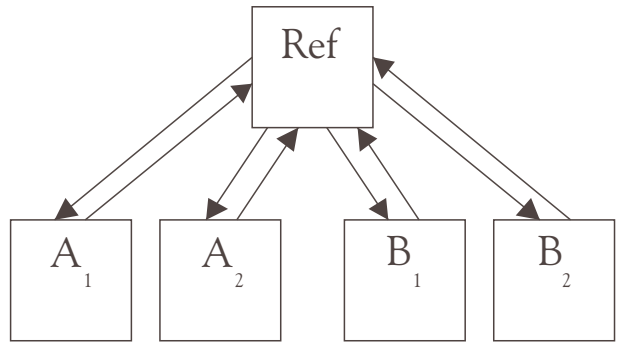

Figure 2. *An indirect experimental design for a two-color platform involving two treatments (A and B), four samples (A1,... B2) and a referential sample (Ref).*

Sl. 2. *Posreden načrt poskusa z dvobarvnimi DNA mikromrežami, ki vključuje dva tretmaja (A in B), štiri vzorce (A1,... B2) in referenčni vzorec (Ref).*

deoxyuridine triphosphate followed by chemical coupling of dyes, and dendrimer-based labeling where custom reverse transcriptase is used to enable subsequent attachment of fluorescent 3D dendrimers to hybridized cDNA (3DNA system by Genisphere). Other strategies were developed and optimized by commercial microarray suppliers, mostly solving the problem of low RNA input and signal amplification.

Labeled samples are spread over an array, covered and placed into hybridization chamber where left overnight at constant temperature. Afterwards slides are washed and dried. Alternatively, a hybridization station can be used to automate the above steps. Hybridized arrays are scanned with a dedicated laser scanner using excitation wavelengths and emission filters which are optimized for the dyes used. Scanned images are stored and analyzed using dedicated computer software to locate spots and extract their intensities. Raw data are stored, preprocessed and further analyzed using various computational methods.

Storing the data, standards and ontologies

Microarraray data largely depend on the protocols used and the experimental conditions, which need to be well-documented in order to fully comprehend the analysis results. The diversity of data describing microarray experiments and large amounts of measurements led to development computer standards to enable data management, storage and cross-platform comparison. Initiated by Microarray Gene Expression Data (MGED) Society (http://www.mged.org) the most important standards are:

- *MIAME* prescribing Minimum Information About a Microarray Experiment required to interpret and verify the results;
- *MAGE* utilizing expression data representation and exchange;
- *ontologies* regulating terminus for experiment description and biological material annotation;
- and *transformations* giving recommendations for microarray data transformations and normalization methods.

Standards do not prescribe the structure how microarray data should be stored, thus many different dedicated microarray data repositories were developed, most of them being compliant with the above standards. For their comparison refer to.[11] Standards and data storage are important, especially for publishing microarray results as nowadays majority of scientific journals require the data to be publicly available. Data is usually deposited to large public repositories such as ArrayExpress,[12] GEO[13] and CIBEX,[14] enabling their verification and cross-institutional comparison.

Data normalization

Microarray data normalization is computational procedure of adjusting the measured intensities and bringing them to a common level so that meaningful biological comparisons can be made. There are number of reasons for microarray data being imbalanced, including unequal quantities of sample extracts, differences in labeling reactions, variation in detection sensitivity of labels, etc.

In general we distinguish between two types of normalization strategies, depending on the number of probes and the way they were selected, and on the presence of spike-in control probes. If present, spike-in control transcripts added to samples during the labeling reaction may aid at estimating the differences between measurements and bringing them to a common level. Without spike-in controls other normalization strategies based on certain biological assumptions must be used. Most frequently employed total intensity normalization is based on assumption that the majority of transcripts are equal in both samples, therefore when taken as a whole their total intensity should be similar. Data is adjusted to meet the assumption. Such strategy can only be employed for genome-wide arrays with the number of probes large enough to accurately represent the population of the genome and with no bias at their selection. Another approach is to assume the existence of housekeeping genes which (if their number is large enough) can substitute for spike-in controls.

Independent of the strategy, various cofactors can be considered by the normalization algorithm. Most commonly used is the average intensity of spots (A) with the purpose of removing systematic bias introduced by a scanner due to nonlinear response to different intensities. Such bias can easily be observed by plotting log2 ratio of spot intensities (M) against A (referred to as MA-plot). Another cofactor, usually considered only for spotted array, is a physical location of spots on the array. Local normalization, where each part of an array corresponding to an individual printing pin is normalized separately, removes systematic bias introduced by inconsistencies among pins.

Statistical, data mining and network construction approaches

With microarray technology becoming less expensive and more widely available, increasingly complex biological questions are being addressed. The more complex the questions are the greater is the demand for statistical assessment of conclusions. Standard statistical and data mining approaches are inappropriate for microarray data analysis as larger number of experimental conditions than there are measured variables (genes) is required. Measuring many genes relative to few samples creates a high likelihood of finding false positive results. The main types of data analysis include selection of differentially expressed genes,[15] identification of markers for disease diagnosis, its outcome prediction[16] and identification of best treatment,[17] and finding new disease classes.[18]

The most common technique for microarray data analysis is clustering,[19] which is used to find groups of genes with similar expression profiles. Statistical hypothesis tests,[20] and their variants[21] are used for discovering differentially expressed genes. More advanced techniques such as probabilistic and information-based modeling are employed for reverse engineer genetic regulatory mechanisms by identifying influence interactions between genes and representing them as a gene network.[22] Mathematical models based on ordinary differential equations[23] are used for simulation purposes and prediction of system's response to various perturbations, with the ultimate goal to be able to simulate the functioning of a cell as a whole.[24]

Conclusion

Microarrays analysis is a powerful post-genomic technique with broad applications in understanding, diagnosing and treating human diseases. It is based either on detecting modulated gene expression in the diseased *versus* control (healthy state) or detects small (SNP) or large (CGH) nucleotide variations that are linked to the disease.

The microarray technology in Slovenia started in 2001 with the foundation of the Slovenian Consortium of Bio-Chips. The consortium members represent several Faculties from University of Ljubljana, clinical and research institutions and pharmaceutical industry. The research equipment for preparation and analysis of the low-density arrays as well as for analysis of high density arrays and Affymetrix GeneChips is located at the Centre for Functional Genomics and Bio-Chips at Faculty of Medicine, University of Ljubljana (CFGBC). The Centre was opened in June 2005, aimed to collect the microarray infrastructure in one place and to make it available to the broader scientific, clinical and industry environment. The Centre offers the equipment and the expertise in all types of microarray studies, especially to the members of the Slovenian Consortium of Bio-Chips. Currently, several projects are ongoing at CFGBC, including development of thematic microarrays for studies of cardiovascular disease and oncogenesis. For details about the Centre and its operation refer to http://cfgbc.mf.uni-lj.si.

References

1. Heidecker B, Hare JM. The use of transcriptomic biomarkers for personalized medicine. Heart Fail Rev 2007; 12: 1–11.

2. Schena M, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; 270: 467–70.
3. Gresham D, et al. Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. Science 2006; 311: 1932–6.
4. Oostlander AE, Meijer GA, Ylstra B. Microarray-based comparative genomic hybridization and its applications in human genetics. Clin Genet 2004; 66: 488–95.
5. Pusztai L, et al. Clinical Application of cDNA Microarrays in Oncology. Oncologist 2003; 8; 252–258.
6. Schrijver I, et al. Genotyping microarray for the detection of more than 200 CFTR mutations in ethnically diverse populations. J Mol Diagn 2005; 7: 375–87.
7. Yu W, et al. Development of a comparative genomic hybridization microarray and demonstration of its utility with 25 well-characterized 1p36 deletions. Hum Mol Genet 2003; 12: 2145–52.
8. Patterson TA, et al. Performance comparison of one-color and two-color platforms within the Microarray Quality Control (MAQC) project. Nat Biotech 2006; 24: 1140–1150.
9. Churchill GA. Fundamentals of experimental design for cDNA microarrays. Nat Genet 2002; 32 Suppl: 490–5.
10. Kerr MK, Churchill GA. Experimental design for gene expression microarrays. Biostatistics 2001; 2: 183–201.
11. Gardiner-Garden M, Littlejohn TG. A comparison of microarray databases. Brief Bioinform 2001; 2: 143–58.
12. Parkinson H, et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2005; 33: D553–5.
13. Barrett T, et al. NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Res 2005; 33: D562–6.
14. Ikeo K, et al. CIBEX: center for information biology gene expression database. C R Biol 2003; 326: 1079–82.
15. Claverie JM. Computational methods for the identification of differential and coordinated gene expression. Hum Mol Genet 1999; 8: 1821–32.
16. Van 't Veer LJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415: 530–6.
17. Golub TR, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286: 531–7.
18. Alizadeh AA, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000; 403: 503–11.
19. Eisen MB, et al. Cluster analysis and display of genome—wide expression patterns. Proc Natl Acad Sci USA 1998; 95: 14863–8.
20. Cui X, Churchill G. Statistical tests for differential expression in {cDNA} microarray experiments. Genome Biol 2003; 4: 210.
21. Cui X, et al. Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics 2005; 6: 59–75.
22. Bansal M, et al. How to infer gene networks from expression profiles. Mol Syst Biol 2007; 3: 78.
23. De Jong H, et al. Genetic Network Analyzer: qualitative simulation of genetic regulatory networks. Bioinformatics 2003; 19: 336–44.
24. Moraru II, et al. The virtual cell: an integrated modeling environment for experimental and computational cell biology. Ann NY Acad Sci 2002; 971: 595–6.