23 A Combinatorial Approach to Gene Expression Analysis: DNA Microarrays

*Concetta Ambrosino, Luigi Cicatiello, Claudio Scafoglio, Lucia Altucci and Alessandro Weisz**
Dipartimento di Patologia generale, Seconda Università degli Studi di Napoli, Napoli (Italy)

CONTENTS

*To whom correspondence should be addressed. Tel./fax (0039) 081 566-5702; alessandro.weisz@unina2.it

0-8493-2334-7/05/$0.00 + $1.50

I. INTRODUCTION

The microarray technology is based on analytical tools that parallelize the quantitative and qualitative analysis of nucleic acids, proteins and tissue sections one of its more recent evolutions-. By miniaturizing the size of the reaction and sensing area, microarrays allow to assess at the activity of thousands of genes in a given tissue or cell line at once in a rapid and quantitative way, and to carry out serial comparative tests in multiple samples. These tools, that stem from the innovations resulting from the technological improvements and knowledge arising from the genome sequencing projects, can be considered as a combinatorial technique that can rapidly provide significant information about complex cellular pathways and processes within one or few "mass scale" and comprehensive testing of a biological sample's composition. AQ5

DNA microarrays, the focus of this review, deals predominantly with their application to genome-wide gene expression analysis (gene expression profiling). They are formed by a planar support, usually a glass microscopy slide, allowing the binding of nucleic acids (cDNA or oligonucleotides) or, in other cases, proteins and oligopeptides. Are defined "probes" the detector molecules immobilized on the surface of the slide and "targets" the mixtures being interrogated [1]. The slides, each containing up to several thousands probes arranged in ordered arrays, are used to analyze labeled samples, generally prepared by fluorescent tagging of nucleic acids (DNA or RNA) extracted from a cellular or tissue sample under investigation. Specific binding of the unique components, of the tested sample, mix to its complementary probe, immobilized on the solid support, leading to the appearance of "spots" the glow of which is proportional to the activity of the expressed gene. AQ1

The microarray technology was developed at the Stanford University in the early 1990s [2]. From the beginning, it was clear that this technique could have the same impact in biomedical and biotechnological research as the "polymerase chain reaction" (PCR) had in the 1980s. PCR reactions are even now extensively used in microarray manufacturing.

The microarray technology is unique as no other analytical approach allows to explore to such an extent the biochemical complexity of biological samples and combines expertise from many different disciplines such as biology, chemistry, physics, engineering, mathematics, and computer science. The role of the recombinant DNA technology, developed in the 1970s, was important not only for the discovery of the enzymatic tools used in the microarray technology, such as RNA and DNA polymerases, but also for the tools and techniques it made available, in particular the cDNA libraries and nucleic acids hybridization protocols. The microarray technology required a modification of the hybridization techniques to make them suitable for a glass support. The first hybridization on glass was performed in the early 1990s. Rapid, efficient and cost effective chemical synthesis of natural and derivatized polynucleotides, is one more domain whose progress made possible the advent of the "microarray era." This also required the full development of the

fluorescent microscopy technology. Since 1970, fluorescent dyes are used for cell biology and microscopy studies, some of which are later being adapted for nucleic acid labeling. Microscopic analysis of chromosome structure by florescence *in situ* hybridization (FISH) is an example. The evolution of more sophisticated fluorescence microscopy devices, such as that of confocal microscopes in the 1990s, was necessary for the development of efficient microarray reading tools (see below). The above mentioned know-how, combined with the development of combinatorial oligonucleotides synthesis, the improvement of linker and surface synthesis technologies and detection methods led to the development of the first microarray assay in 1995. Furthermore, all this and the work that introduced robotization in microarrays manufacturing paved the way to the present success and diffusion of DNA microarray applications for gene expression profiling.

A. The DNA Microarray: Components and Characteristics

A DNA microarray, also known as DNA chip, gene chip or more generically biochip, is a microscopic slide on which multiple DNA samples are deposited (''spotted'') in predefined positions to constitute an ordered array of probe elements. The chemical nature of these probes in the arrays used for quantitative gene expression analysis can be different, e.g. DNA, PNA or RNA, although in most cases they are represented by cDNAs or chemically synthesized oligonucleotides. The amount of probes to be spotted (optimal probe concentration) as well as the number of spots for unit of area (optimal probe density) is first evaluated experimentally, as these parameters greatly depend upon the detection protocol and device to be implemented and the nature of the experimental test. The microarray surface plays an important role in determining the probe binding efficiency and specificity and the sensitivity of the detection step which greatly affects the quality of the data generated. To be used as analytical devices for genome-wide gene expression studies, arrays of probes which are planar, microscopic and specific are to be put in order. The array elements (''spots'') are put in orders in rows and columns that so the columns cross the rows in a perpendicular manner. The ordered elements have, as much as possible, the same size, spacing and a unique location on the array, to facilitates manufacture of the slide, as well as design and application of microarray reading devices and software for image and data analysis. Regularity of spot spacing is a prerequisite for correct data analysis, as it enables the use of standard analysis templates, while the uniformity of the spot size is required for quantitation and assay precision, as this ensures that the same amount of probe is spotted in each location. Quantitation templates are grids superimposed to the graphical image generated by the scanner, necessary to define the borders of each element and to calculate, for every one of them, its signal intensity and the relative statistics (''shape''), considering the pixels included within the area delimited by these borders. The presence of microscopic spots on the slide enables the examination of a large number of genes, up to an entire genome, with a single

test and is a necessary prerequisite for automation of the whole process. The slide surface has to be planar since a planar support allows an accurate scanning and imaging, due to the uniform detection distance between the optical element and the microarray surface. Furthermore, this surface needs to be impermeable to liquids, allowing the use of a small reaction (hybridization) volume. Specific recognition between the probes on the array and the target is the key criterion of microarray-based nucleic acids analysis, as it allows precise quantitation of the amount of molecules present in the sample. This aspect is also important for the statistical evaluation of the data (reproducibility, precision, confidence level, etc.).

B. The DNA Microarray Technology

Modern microarray technology interfaces biology, engineering and physics. Five steps are necessary to perform a microarray experiment: microarray manufacture, probe labeling, hybridization, detection and data analysis (Figure 23.1). In this section the general concepts of this technique are briefly discussed. An important prerequisite to what follows is the concept that the biological question addressed is a key determinant in the design of a microarray experiment, as its formulation dictates not only the technological platform to be selected but also other experimental parameters, such as positive and negative controls, reference and the strategies to apply for computational analysis of the resulting data.

Several criteria need to be kept into account for the choice of a microarray manufacture procedure, including the need to be able to use the technology in a given laboratory setting, the time required for its production or its analysis, the probe content (the total amount of DNA probes delivered on the slide) and density (number of target spots per unit area), the spot size, the purity and the reactivity of the element spotted and the overall costs. The microarray manufacturing technologies are in continuous evolution and they already ensure mass production of biochips as well as assay automation, leading to improved quality, dissemination and reproducibility of the experiments and, cost effectiveness in the near future. As more extensively reported in other sections that follow, DNA chip manufacturing technology follows two different approaches, known as ''delivery'' and ''synthesis.'' In the first case (Figure 23.1a) the probes, usually DNAs generated by PCR amplification of cloned cDNAs or synthetic oligonucleotides, are transferred to the slide surface by automatic robotic platforms. This was the way used for the preparation of the first microarray ever. The synthesis approach, currently applicable only to oligonucleotide arrays, foresees that all probes are synthesized *in situ* by light-driven chemical reactions, one base at a time, till the desired oligonucleotide is synthesized in each array position (Affymetrix® technology: Figure 23.1b).

The target for microarray experiments can be of different nature. Nucleic acids are more commonly used nowdays. Diverse techniques can be used for target labeling. Selection in each case dependsupon the molecular and biological nature of the target. Binding of the labeled target necleic acids to

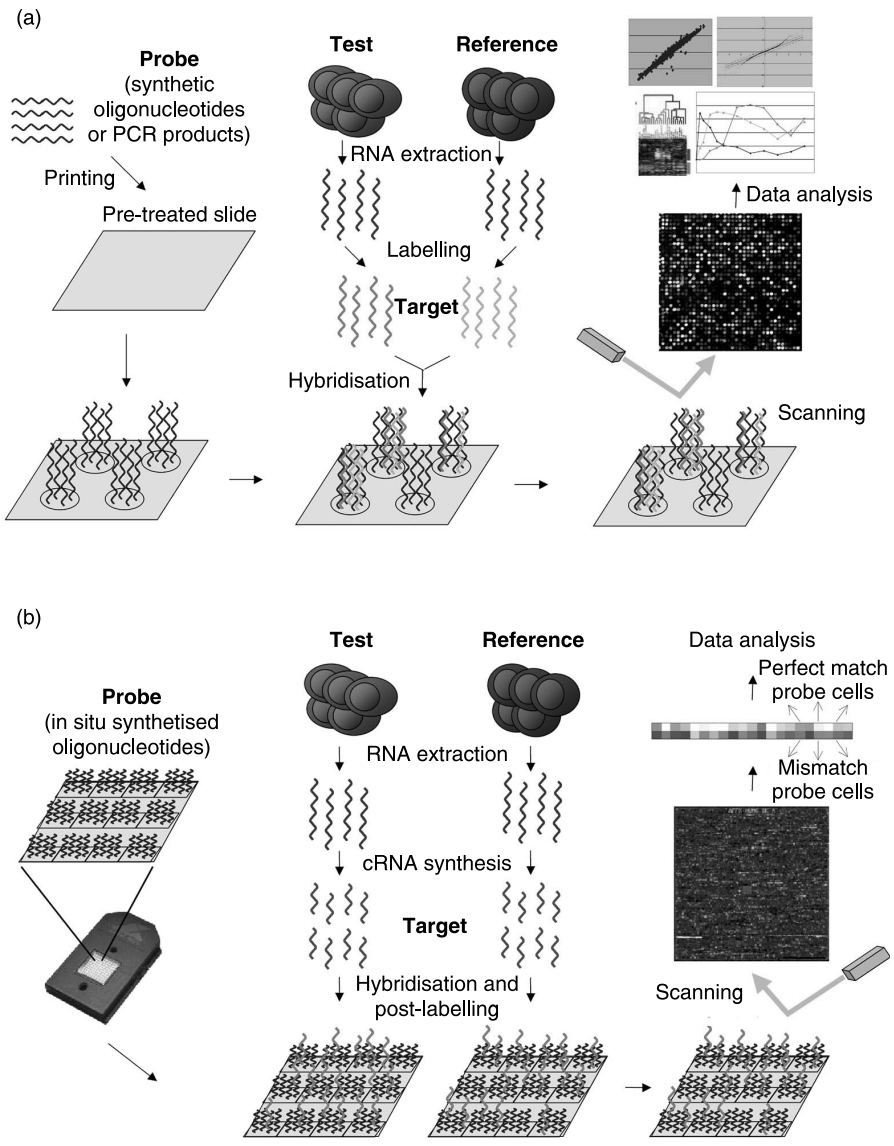

FIGURE 23.1 Schematic representation of genome-scale gene expression analysis with DNA microarrays. (a) DNA microarrays produced by probe deposition; (b) Oligonucleotide microarrays produced by *in situ* probe synthesis (Affymetrix® technology).

the probe onto the microarray occurs by molecular hybridization. Conditions for hybridization, and slide washing to remove targets unspecifically bound to the slide and/or to the probe, are experimentally determined, as they relate to the nature of the biochip and can be differentiated in order to achieve the degree of specificity desired. Once hybridized, the microarray is scanned with a

detection device in an automated manner. Fluorescent detection schemes, to detect very low concentrations of label, are generally used. The devices to detect the fluorescent signal on such slides are either (laserpowered) scanners or imagers. In both cases, the fluorescent dyes are excited and the emitted fluorescence is read by converting its stream of photons into an electrical current that, in turn, is transformed in to digital values that can be stored and analyzed by a computer. The processes by which numerical values are obtained from the data files in a format that can provide information helpful to determine each target's concentration in the original mixture, is known as quantitation. Bioinformatics and computational approaches are then used for data normalization, mining and modeling. In this respect, availability of comprehensive electronic libraries and database containing various functional and structural information on genes or proteins provide a great advantage in the interpretation of gene expression data. This is to be be discussed in detail later.

Each step of microarray technology is still evolving at a high pace and several modifications of the original, basic technology described here are being continuously introduced, while ever increasing new applications and improvements are further expected [1, 3–5].

II. A CLOSER LOOK TO THE DNA MICROARRAY TECHNOLOGY

A typical gene expression profiling experiment takes place in fue separate processes. They are (i) microarray fabrication, (ii) purification and labeling of the target material, (iii) hybridization, (iv) detection and (v) data analysis. The characteristics of each step was briefly discussed in the introduction. A closer look to each of these steps is the object of this section. Here we would mainly refer to biochips where the probe is constituted by nucleic acids (DNA microarrays).

A. Microarray Fabrication: Solid Support, Probe Synthesis and Immobilization Techniques

The microarray assay is based on hybridization reactions between labeled single stranded molecules in solution (target) and complementary molecules immobilized on the flat surface of the slide (probe). The fabrication of a microarray requires the synthesis of the target and its deposition on the slide surface (deposition technology). Alternatively, a different approach involving the synthesis of the target directly onto the surface can also be employed.

The slide surface has to be planar, uniform, inert and accessible. Several materials can provide slides with these characteristics, but the glass surface is the most commonly used for fluorescent labeling, since most plastics do not permit the use of fluorescent dyes. Different glasses are available that are suitable for slide preparation (borosilicate, fused silica, etc.) and all stable materials with low intrinsic fluorescence and reflectivity and an efficient transmission throughout the visible range. In order to allow the efficient

attachment of nucleic acids to the surface of the glass, various chemical pre-treatments of the slide are applied, including the most used: (i) poly-lysine covered glasses, with positive charged surface where unmodified nucleic acids can be bound and fixed by UV cross-linking; (ii) aldehyde covered slides, that attach DNAs carrying amino-modified end nucleotides and (iii) amine covered slides, that also provide a positive charged surface. The choice between these options will depend on the experimental procedures.

The first step of microarray fabrication process is the synthesis of the probe, commonly, constituted by PCR products or oligonucleotides.

1. Printing Microarrays with PCR Products

Usually primer pairs which are gene-specific or optimized to anneal in the vector sequence, are used to amplify any cDNA or express sequence tags (EST) from an available library. The amplicons are then deposited (printed) into the slide surface at pre-defined positions. Good quality PCR products are to be used for microarray construction and hence stringent QC procedures are applied, including analysis of each amplification product by agarose gel or capillary electrophoresis, to control amplification efficiency and quality of its products and to exclude production pipeline samples presenting a poor amplification or non specific bands and smears. The length of the PCR products used to generate microarrays varies generally between 300 and 800 nucleotides.

The main advantages of this approach is that it is not necessary to know the full sequence of the starting DNA clone (although it is preferable!) and the signals obtained are generally strong and hence the array is easier to implement and quality of the data generated is higher. The only disadvantage is denaturation step has to be introduced in to the procedure, due to the presence of a double stranded target on the microarray and other cross-hybridization problems are possible.

2. Printing with Oligonucleotides

For this strategy to be practiced 50 to 70mer oligonucleotides are generally used. The major benefits here with respect to cDNA probes are: (i) higher specificity, since the sequences are optimised to minimize cross-hybridization, (ii) the possibility to design several oligonucleotides for different parts of the same gene, and monitor the specificity of hybridization and detect alternative splicing products (different alleles) and (iii) the possibility to normalize the hybridization conditions by construction of oligonucleotide sets having similar melting-temperature (Tm).

In both the cases decribed above, the second step in microarray fabrication involves ordered deposition of the probes on the surface (spotting). The purified DNA probes (PCR fragment or oligonucleotides) are spotted on a modified glass slide in an ordered grid, by means of array spotters and robotic instruments that allow for precise deposition of few nanoliters of DNA

solution, with accuracy and reproducabiling leading to replicate uniformity. Two main spotting technologies are currently employed. They are ''the contact deposition'' and the noncontact, deposition. The ''contact deposition'' uses particular pins that aspirate by capillarity a small volume of pre-made target solutions and deposit a drop of it onto the slide surface on physical contact. The ''noncontact deposition'' distributes sub-nanoliters of the target solutions to specified locations. Two ''noncontact deposition'' technologies are mainly used: (i) the piezoelectric technology, which uses electricity in order to modify the morphology of a piezoelectric crystal that encircles a capillary containing nucleic acid solutions, resulting in a squeezing on the capillary and delivery jet onto the surface; (ii) the syringe-solenoid deposition, in which a positive pressure, produced from a syringe and regulated from a solenoid valve, allows the deposition of micro-volumes of the target solutions onto the slide, when the valve is opened. A third option, recently introduced, uses an adaptation of an ink jet printing technology to deliver sub-nanoliter droplets of DNA solution unto the glass surface.

All these approaches allow high-density spotting and are easy to implement at low costs.

3. Printing Microarrays by *in situ* Probe Synthesis

DNA targets are synthesized *in situ*, by a modified photo-lithography procedure, one base at a time for several cycles, until the desired sequence is obtained in each element of the array. The technology for semiconductor production are combined with photolithography in the Affymetrix® method, which uses ultraviolet light and solid-phase chemical synthesis for solid-state polynucleotide synthesis. In photo-lithography, the glass slide is modified with a surface providing the reactive amine groups modified with a photo-protecting group to control their reactivity. The amine group is activated by ultraviolet light. A predefined mask (photo mask) is applied to select the sites that have to be activated during each photo-activation step. In the de-protected regions, modified phosphoramidite nucleotides can then be covalently bound. The cycle of removing the photo-protecting group by UV-light and subsequently the coupling step facilitates oligonucleotide synthesis.

The advantages of this approach are similar to those involving delivered oligonucleotides. All the steps regarding sample production, handling and storage are eliminated and the oligonucleotides are produced directly using sequences from the databases. The use of ''perfect match'' vs ''mismatch'' probe pairs is a unique concept introduced in this case. For each probe, perfectly complementary to a target sequence (the ''perfect match'' probe; PM), an associate probe that carries a single base mismatch in its 13^{th} position is also synthesized in the same array (the ''mismatch'' probe; MM). This system allows the subtraction of the signals due to nonspecific cross-hybridization to the ''MM'' probe and provides a key information for signal specificity. Moreover, the chip-to-chip variations are significantly reduced. This kind of chips contain high number of targets (up to 400.000 oligonucleotide cells within 1.6 cm^2).

The main disadvantage of this approach is the short length of the *in situ* synthesized oligonucleotides (less then 30 nts) and the high manufacturing costs, although it might be the least expensive way to produce chips covering whole mammalian genome.

B. Target Preparation and Labeling Procedures

The method to label a target depends on its molecular nature, and the microarray technology implemented. Here we shall describe the labeling method used for a DNA chip with an RNA as the target molecule, which is used for gene expression profiling. Other target preparation protocols are found in Refs. 3–4.

The purity and integrity of the RNA isolated from tissues or cell lines are critical for microarray experiments. The most common method for total RNA isolation involves organic extraction of RNA from homogenized sampleslike guanidinium isothiocyanate or guanidinium hydrochloride. The RNA sample should be devoid of carbohydrates, DNA, lipids and proteins, as the presence of these contaminants may affect the performance of the sample in the downstream procedures. Several commercially available methods and buffers, are currently used. The RNA is significantly more labile than DNA because it is readily susceptible to degradation by endogenous and contaminating ribonucleases, which are stable and ubiquitous enzymes. To obtain a high quality RNA and to maintain its integrity during the subsequent procedures, several precautions are taken, which include: (i) processing the sample as soon as possible, and (ii) avoiding RNase contamination using disposable gloves and RNase-free glassware, plasticware, water and salt solutions, obtained by autoclaving or treatment with DEPC. The purified RNA is stored at –70°C.

Total RNA or (polyA+)RNA (mRNA) can be used for experiments. In the latter case, a purification step is necessary, as mRNA is isolated from total cellular RNA by affinity chromatography on oligo-dT immobilized to a solid support. The amount of the purified RNA determined by its dual wavelength absorbance at 260 nm and 280 nm and the quality checked by agarose gel or capillary electrophoresis.

In most recent microarray experiments, multicolor fluorescence labelling are used for simultaneous analysis of two or more samples in a single assay. For this, total RNA or mRNA are labelled with fluorescent nucleotides by a reverse transcription reaction. cyanines Cy3 and Cy5 are used for dual color analysis.

Various RNA labelling strategies involving direct or indirect labelling exist. Direct labelling is the most diffused method. The RNA template is converted to fluorochrome-labeled first-strand cDNA by a reverse transcription reaction (RT). Reverse transcriptase synthesizes cDNA using the RNA as a template, in the presence of oligo (dT) primers that hybridize with the poly-A tail of the mRNA, incorporating at the same time modified Cy3- or Cy5-conjugated deoxy-nucleotide triphosphates. This method is fast and simple. The main drawback is that cyanine-labeled nucleotides are not efficiently incorporated in

the polymerizing step, due to steric hindrance caused by the large fluorophores. After this step, the template is removed by RnaseH1 digestion or NaOH treatment and the free nucleotides are eliminated by gel filtration on dedicated columns. Generally, direct DNA labeling does not produce long cDNA molecules, and for this reason probes complementary to the nucleotide sequence near the poly-A plus tail of the target are recommended. In the indirect labelling, the 5-(3-amino-allyl)-2'-deoxynucleotide 5'-triphosphate (aa-dNTP) modified nucleotides are used. They are incorporated with the same efficiency of the unmodified nucleotides in the first-strand cDNA synthesis. After removal of the RNA template and purification of the amine-modified cDNA, the coupling reactions with N-hydroxysuccinimide-esters (NHS-esters) of Cy3 or Cy5 are performed to produce uniformly labelled probes. The labelled cDNA requires a re-purification step to remove the unincorporated, free Cy-Dye.

When the amount of the RNA sample is low (for example in tissues from biopsies), an RNA amplification step is needed, that involves labeling through a linear RNA amplification method. For this, the mRNA population is converted into a double-strand cDNA containing a strong promoter sequence from viral RNA polymerases, such as T3 or T7 phage promoters, by using an oligo(dT) primer including a 5' extension including the viral promoter for first strand cDNA synthesis, followed by complementary cDNA strand synthesis with DNA polymerases. Several RNA copies are synthesized form each template of double-strand cDNA in presence of Cy-Dye ribonucleotides and appropriate DNA-dependent RNA polymerases. If biotinylated ribonucleotides are used in the transcription amplifications to generate biotinylated cRNA, a post-labelling reaction is performed following target hybridization to the array and the washing steps, by staining with streptavidin-phycoerythrin conjugates (see http://www.affymetrix.com/technology/ge_analysis/index.affx for more details) or CyDye streptavidin conjugates (see: http://www4.amershambiosciences.com/aptrix/upp01077.nsf/Content/codelink_bioarray_system). Both technologies use short oligonucleotide probe microarrays (25mer- and 30mer-long, respectively). Fragmentation of cRNA before hybridization is necessary to avoid secondary structure of RNA interfering with the target-probe annealing.

A very recent labeling method instead of fluorescence, uses Resonance Light Scattering (RLS) technologies, based on the optical light scattering properties of nano-sized metal colloidal particles. In this technique, biotinylated nucleotides are used in the first-strand cDNA synthesis and the coupling steps are performed with anti-biotin- coated gold or silver particles. The main advantages of this technology are the high sensitivity and the absence of signal due to photochemical bleaching.

C. Hybridization and Washing

In the hybridization step, the ability of labeled targets to bind immobilized probes are tested. Referring again to the cases where DNA microarrays are

used, the labeled targets are first annealed with their complementary probes immobilized onto the slide surface. Subsequently, serial steps of post-hybridization washing are carried out in order to remove unbound materials, to improve the signal-to-noise ratio and to minimize cross-hybridization between labeled targets and probes. Hybridization and washing steps are performed in the dark, to avoid signal loss due to photo-bleaching of the fluorescent dyes. Several parameters are taken into account in order to obtain successful results, linked to the two main experimental variables: temperature of hybridization and composition of the hybridization solution. The base composition of the nucleic acids involved in the annealing reaction have a large effect on duplex yield of the solvents normally used for hybridization, due to the different stability of A : T vs G : C base pairs. Short oligonucleotides have extreme biases in composition and hence, correspondingly show large differences in melting temperature ($T_{m,}$ the temperature at which 50% of the target is denatured). As a rule-of-thumb, the addition of an A : T base pair increases the T_m by 2°C, compared with 4°C for a G : C pair. This difference is minimized by adding TMAC (tetra-methyl-ammonium chloride) to the hybridization mix. The optimum range of hybridization temperature in aqueous solutions is about 65–75°C. The degradation of the target due to high hybridization temperature is minimized by the use of 50% formamide, that allows the reduction of the hybridization temperature of about 25°C, thereby protecting the targets from degradation. Usually, in the presence of formamide, hybridization is carried out at 42°C. The other parameters to be considered in this process are the pH, (neutral value promotes hydrogen bonding between base-pairs), and salt concentration (improves the hybridization efficiency by shielding negative phosphate groups of the nucleic acid and minimizing electrostatic repulsion). The post-hybridization washings eliminate unbound labeled target. Usually these washes are performed in SSC/SDS solutions, progressively diminishing the concentration of salt. To improve the reproducibility of the microarray analysis by reducing the variability due to hybridization and washing steps, automatic hybridization-washing stations are developed, but their high cost is still a deterrent for their widespread diffusion. After the slides are washed, they are dried by blowing nitrogen gas or by low speed centrifugation at and quickly passed to the subsequent processing steps.

D. Target Detection

In microarray analysis detection is the step in which the signal for each spot is revealed and quantified, giving an image that is like photography of the microarray. In the subsement paragraphs we shall provide basic information on the currently used fluorescence detection systems for microarrays and the characteristics of relative detection devices.

1. Fluorescent Dyes

Fluorescence is the light emission process in which a fluorophore, a molecule able to adsorb light, after reaching an excited state releases light (photons) with

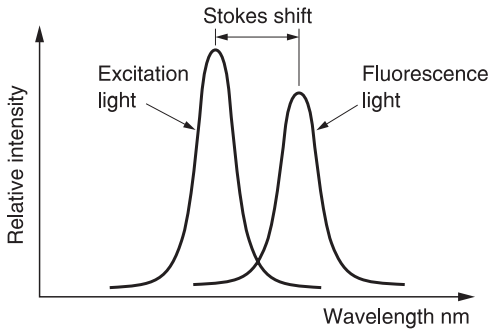

FIGURE 23.2 Stokes shift diagram depicting optimal relationships between excitation and fluorescence light of a fluorophore suitable for microarray applications.

less energy and, consequently, a longer wavelength than the exciting light (Stokes shift; Figure 23.2). in fluorophores the distance between excitation and emission wavelengths are greater and more suitable for microarray application. Part of the excitation energy is emitted in processes different than fluorescence (nonproductive fates), whose values have to be evaluated to choose the optimal fluorescent label and detection device. When a moderate excitation light is applied, a single fluorescent molecule can be cyclically excited and detected, increasing consistently the sensitivity of the assay. On the other hand, the signal emitted is reduced when the exciting light applied is too intense, a phenomenon causing permanent loss of fluorescence signal (photo-bleaching).

Several fluorescent labeling dyes are commercially available (Alexa series, Oregon green, Rhodamine and Cyanine dyes, etc.). They are all characterized by the presence of double bonds on every other carbon atom of a cyclic structure, containing the electron that once on excitation emitted fluorescent light. The cyanine dyes are the most widely used at the moment. They are bright, easily added to the nucleotides, stable to photo-bleaching and with a Stokes shift value of about 20 nm. The cyanine dyes used in microarray analysis are Cy3 (absorption at 550 nm and emission at 570 nm) and Cy5 (absorption at 649 nm and emission at 670 nm) that are already available as phosphoramidite derivatives.

The use of fluorescence in microarray technology has several advantages. Indeed, fluorescence significantly increases the sensitivity and the speed of the assay and enables the collection of very large amounts of data in automated fashion. Moreover, the fluorescent dyes are not too toxic stable and much safer than radioactivity. Their major advantage, however, is the spatial resolution they provide, as this allows correct assessment of both weak and strong signals even when they are emitted by elements located adjacent to one another on the array grid, as any signal spreading effect is avoided. This property allows the construction of high-density arrays. Two or more fluorescent dyes are often used in conjunction in microarray experiments, to increase the reliability of the comparative analyses and to permit the analysis of several differently labelled samples on the same microarray.

2. The Microarray Reading Devices

The two main detection systems currently used for microarray reading are the scanners and the imagers. The microarray scanners acquire the entire image by moving the array or the optic system back and forth in small increments (10 μm). The imagers are able to collect data from larger areas (1 cm^2) without any movement. Both devices have to include functions necessary to detect a microarray image, in particular: (i) a source of excitation light with a precisely defined wavelength, provided by a laser or a lamp, requires the addition of filters to provide beams of selected wavelengths; (ii) a system of optic lens collecting the fluorescent light in three dimensions, the numerical aperture, and a parameter defining the collection angle, e.g. collection efficiency of 50%; a lens with a numerical aperture of 1 collects the light over an entire hemisphere, (iii) the spatial addressing, that refers to fluorescence detection from a defined area of the glass slide, usually divided in pixels, smaller than the element size to reveal artifacts due to spotting problems or to dust on the slide; (iv) the ability to discriminate between excitation and emission light, the last being of much lower intensity than the first one and (v) the ability to convert the low level light into electrical signals.

The microarray detection devices, scanners and imagers, are sophisticated instruments that require components specifically designed to achieve the specifications listed above. Confocal scanners are widely used as microarray detection devices. Differing from a normal scanner, these systems have two focal points configured to limit the field of view in three dimensions, driving data acquisition one pixel at a time. In the basic design, a laser light is directed to an excitation filter allowing passage only to light of the wavelength of interest, usually corresponding to the excitation peak of the dye. The laser light is monochromatic (single color), coherent (the photons in the beam have the same phase) and collimated (highly parallel), implying that no other optic systems are required to get a beam that is really intense and precisely reflected. The device has to include multiple gas or solid-phase lasers, one for each wavelength required. The emitted light is reflected through the microscope objective by a beam-splitter, an optic filter that separates out the returning excitation beam. The main function of the splitter is to reject most of the reflected laser light (4% of the input), while allowing passage to most of the fluorescent light. The excitation of the fluorescent dye molecules results in the emission of fluorescent light in several directions, collected by the objective. The return beam goes to the beam-splitter, which in turns transmits only the fluorescence beam to a mirror directing it to the detector. This emission filter allows the passage of a selected narrow band of fluorescence, while rejecting the reflected laser light. At this point, the fluorescence beam is directed to the detector lens, that in turns focuses it on a detector device able to convert the beam into electrical signals. The most common detectors are photo-multiplier tube (PTM), which transforms the photons to an amplified electrical signal. The fluorescence beam is directed onto a light-sensitive surface of the PTM, the photo-cathode, that then releases electrons. These will jump onto a charged

electrode, inducing the release of a higher number of electrons (up to about 1 million-fold amplification) and are finally collected on the anode there by sending out of the PTM the electrical current that is easily recorded and measured. The recorded signal is converted to digital data, and stored as image (TIFF) files that are further computed.

The microarray imagers, on the other hand, are detection instruments able to capture images from a larger portion of the microarray in a single detection step. Conceptually, an imager is organized as a scanner but several technical details are different. The imagers have a white light source (polychromatic) that is directed to the optic filters to obtain the monochromatic beam, necessary to excite the fluorescent dye. The fluorescent beam is directed to a beam-splitter to remove the reflected light. A light sensor also known as the detector, located in the camera, captures the fluorescence light. The detector consists of a checkerboard matrix of light sensitive pixels, like for example in the charged-coupled devices (CCDs) in which a pixel is coupled to each photo-sensor allowing the charge accumulation of the photo-sensor to be transferred and amplified across the matrix. The sensing region, defining the CCD-chip, contains around 1,000 pixels in each direction. The smaller the pixels the higher is the image resolution. Only smaller images are captured each time as less charge is stored in the device. The amplified electrical signal is recorded, measured and transformed to digital data, as is the case with scanners.

E. Analysis of DNA Microarray Data

The raw data, generally fluorescence measurements extracted from the TIFF-format images generated by the scanning devices, require first scaling and normalization, to eliminate the systematic sources of variation between samples as well as the different intrinsic fluorescence labeling or hybridization efficiency among the two dyes analyzed in parallel, the unequal spreading of the hybridization mix on the array surface, and the variations in image analysis (laser power fluctuations, photo-multiplier gain adjustments, etc.). Indeed, raw analysis of data relative to replicate experiments reveal a high variability present, when comparing the same RNA labeled with two different dyes on a single array (self–self differential hybridization). As a consequence, at least three technical replicates (i.e. three different hybridization reactions for each sample-control pair) are required to allow an efficient analysis of variance and correction of systematic errors (biases) and variability due to stochastic events (variation).

The data analysis process, starting from normalization to eliminate casual sources of variability within and among arrays, proceeds through statistical analysis aiming at identifying those genes whose expression is significantly different in the two (or more) investigated samples. The so identified significant genes are then subjected to further bio-informatic analysis, to group them according to their expression patterns, functional role, etc., or to test their predictive value with respect to biological hypotheses [6].

AQ6

We shall describe here some of the approaches used for computation of data from ''two-dye'' comparative gene expression profiling analyses carried out with cDNA arrays.

The initial step proceeds through ''within-array normalization,'' aimed at correcting biases within the data sets due to intrinsic dye fluorescence, intensity-dependent or local (sub-array) variation. The more immediate and simplest way to address the problem is the total intensity normalization, which is based on two postulates: (i) the slide containing a large number of gene probes, is likely to show the same activity in both test and reference samples, (ii) the total mass is same for the two samples that are hybridized competitively to the array, so that the total fluorescence is same among them, even if some RNA species are over-represented in one sample and vice versa in the other. As a result, the data are adjusted so that either the sum, the mean or the median of the measured intensities are equal for the two fluorescence channel readings. A variation of this normalization step is to scale all data according to reference genes whose expression levels are assumed to be constant, based on biological and functional considerations (such as the so called ''housekeeping'' genes, for example). Probes for these reference genes are spotted on various regions of the array to correct sub-array variation and used as an alternative means to normalize data from replicate experiments. In any case, this type of scaling does not correct intensity-dependent variation, as standard deviation data often varies with the signal intensity because casual fluctuation affects signal detection more incisively at the lower end of the fluorescence scale than at its higher end.

To adjust this source of variation, a locally weighted linear regression (''lowess'': locally weighted scatter-plot smoothing, [7], that computes an intensity-dependent normalization factor for each gene, should be carried out [8]). In this way, however, local differences within the array (sub-array variation) are not addressed. The differences due to spatial location of the spots are in regard to those slides on which different arrays are spotted through different pins (the so-called ''print tip groups,'' ''pen groups'' or ''sub-grids''). The variance is due to slight differences in the geometry of pins, deformation of the pins after a long activity, or to unequal distribution of the hybridization efficiency over the slide surface. The same variability is observed among replicate slides. A possible solution is the execution of different ''lowess'' analyses for the various sub-grids, and a scaling of the obtained data similar to across-array normalization (see below).

A particular type of ''within-array'' analysis is the so called ''self–self'' hybridization [9], in which two dyes are used to label the same RNA species, so that the fluorescence values acquired by the scanner for each gene is supposed to be the same for the two channels. This approach allows the identification of the variability which depends only on systematic biases or on stochastic processes. Some Authors suggest the performance of some ''self–self hybridization'' for each experiment, to establish an error model used to correct data derived from experimental measurements.

The second step is the ''across-array normalization.'' The simplest way to compare replicate arrays is to scale their intensities according to a total fluorescence method, computing the standard deviation of replicates measurement for each gene, and excluding from the analysis those genes whose variance is too high. A complex approach to correct stochastic sources of variability is the ''variance regularization'' (10], which implies the adjustment of data relative to the different slides (or sub-grids) in order to center the fold-change distribution around zero (normalization step), and then multiplication of each element for the respective scaling factor, computed for each array by dividing its variance for the geometric mean of all the variances.

A different type of across-arrays data comparison is the ''flip-dye'' analysis, based on the inversion of the dyes used for labeling the test and reference samples in at least one replicate [dye swap, 11]. The comparison of at least two dye-swapped hybridization reactions reveal the presence of differences in fluorescence values which are not due to effective changes in RNA levels but instead to casual fluctuations; the genes for which the ratio between the fold-changes of the swapped arrays is far from one and should be excluded from further analyses. Therefore, in order to eliminate as many variability sources as possible, correct planning of a gene expression profiling test with microarrays should include not only replicate hybridizations, but also dye swapping. It has been even suggested that each sample should be subjected to balanced hybridization, carrying out as many labeling with Cy3 or Cy5.

The last step of this initial data analysis phase is the selection of significant genes whose expressions are different in the two compared samples. The first and simplest method used is based on computing fold-change differences for each gene, by averaging replicate results and choosing the first the cut-off value that defines differentially expressed genes. Generally, a gene is considered differentially expressed in two samples when the differences in mRNA detection among them by microarray hybridization are at least two-fold. If the data relative to many replicates are consistent, a lower cut-off (down to $\pm$1.5-fold change) is acceptable. However, a cut-off value has to be selected based on statistical significance, and for this reason a great number of computational approaches are introduced to compute the level of confidence associated with the selection of truly differentially expressed genes. The most used, among those methods, go from the standard t-test [12], to a Significance Analysis of Microarrays method [SAM, 13], to analysis of variance [ANOVA, 14] or application of Bayesian mathematics [15], to the ''maximum likelihood'' method [16].

The t-test analysis computes for each gene the probability that the difference between the mean fluorescence intensities of the test and reference samples is falsely called significant (p-value), by theoretical t-distribution or permutation test.

SAM involves a modified t-test and computes a ''False Discovery Rate'' (FDR, representing the expected incidence of false positives) for each chosen differential expression (significance) cut-off.

ANOVA takes in to account the different sources of variation, dependent on the arrays themselves. The different slides are spotted and hybridized under slightly different conditions. In the case of dyes, -one dye is often brighter than AQ5
the other, with regard to samples, their concentrations can be slightly different, in the case of genes, individual probes can show different efficiency of hybridization and with regard to the microarray elements a complete control over the amount of DNA immobilized on the slide is not possible. Numerous other mathematical methods are proposed, but their complete listing is beyond the scope of this review.

A different algorithm for data normalization and selection of significantly expressed genes is used for the Affymetrix®-type oligonucleotide arrays, in which each gene is represented by a probe set of 16–20 perfect match (PM) oligonucleotides, each of them paired with a single-based mutant (MM) to allow computing the quote of non specific annealing reaction (see above). The Affymetrix® algorithm first discriminates the genes effectively expressed from those whose levels are similar to MM, by executing a t-test for each probe set. Then the fluorescence value relative to the probe set is computed by averaging the intensities of the perfect matches subtracted of mismatch, and finally, a t-test is performed to compare the test and reference samples, hybridized to distinct biochips [http://www.affymetrix.com, 17].

The list of positive, differentially expressed genes obtained by either one of the above mentioned procedures are then subjected to other investigations to gather further insights on its biological meaning. A first analysis can be based on gene ''clustering,'' where in grouping of the genes according to similarities of their expression patterns in each sample is done. A basic principle of genome-wide expression analysis is that genes linked by similar expression profiles respond in a similar fashion to the environmental and internal signals reflecting the functional state of the cell, while the products they encode are likely to act in concert toward achieving a cellular phenotype. According to this view, data clustering is the first step for interpretation of microarray data toward identification of the biologically relevant processes they underscore. Different clustering algorithms have been proposed, among which the most used are: hierarchical clustering, K-means clustering, self-organizing maps, supervised clustering and Best Score Clustering [BSC, 18].

Hierarchical clustering [19] consists in computing the distances between each couple of genes of the studied list, thus constructing a distance matrix in which the distances of each gene from all the others is reported; the smallest distances computed will form the first cluster (composed by gene pairs). Another distance matrix is then constructed, considering now groups of genes (or ''objects'') instead of single genes, and the process of classification is repeated until a single group or cluster remains. The similarity hierarchy so computed can be represented likewise a phylogenetic tree, whose branch length is proportional to the correlation between the elements connected (expressed as Pearson's score). Besides clustering genes according to their expression pattern (gene clustering), samples can also be grouped according to their gene expression profile (array/sample clustering). This clustering is

effective in identifying the main groups of similar expression, but the hierarchic tree may be too rigid to represent the combinatorial complexity of gene expression data. Moreover, this method yields a large number of clusters in the tree-like structure, making it difficult to link the expression patterns to biological processes.

K-means clustering [19] is based on the assumption that a certain number of classes must be identified in the data set. The genes are first randomly assigned to these classes, and then rearranged in the clusters through successive steps, involving computation of the distance of each gene from the mean of each of the selected groups and shuffling it to the nearest class. This approach is used when *a Priori* hypothesis concerning the number of expected clusters is formulated in advance.

Self-organizing maps [20] are also based on the establishment of a certain number of nodes in a k-dimensional gene expression diagram followed by iterative mapping of the nodes in the space according to their distance from points corresponding to the gene expression values determined experimentally. The advantages of this method are the flexible structure of the clusters and their easy visualization and interpretation. The spatial representation of clusters better reflects the multiple distinct ways in which gene expression patterns can relate to each other.

Despite the many approaches proposed, one is still far from linking clusters to biologically relevant groups, as sufficient information about the biological role of the genes and the classes they group in are still missing. To address this problem, supervised clustering methods [21] are proposed, in which genes and other notions of interest are associated with labels that provide information about a pre-existing classification. The information used to drive the analysis may include knowledge of gene function or regulation, disease subtype or tissue origin of a cell type. The methods comprise a training phase (supervised learning), in which the expression profiles associated with each class are defined by using a set of informative genes, and a test phase, in which new genes are classified according to their similarity to the pre-defined classes.

More complex approaches to this problem involve the use of artificial neural networks [22], Bayesian networks [23] and support vector machines [24], which in turn are based on the same principle of supervised learning [25].

In parallel with the cluster analysis, a functional classification has to be carried out in order to identify groups of co-expressed or co-regulated genes that play a common or complementary role in the cellular homeostasis or in the response to external stimuli. An international effort, the Gene Ontology (GO) consortium [http://www.geneontology.org, 26], is currently under way to establish precise and univocal definitions of the involvement of all genes from various species in biological processes, including the molecular functions and cellular localizations to their products. The GO dictionary is organized in a hierarchical structure, in the form of Directed Acyclic Graphs (DAGs), in which each term belongs to a parent class and has in turn a certain number of child terms, going from the broader to the narrowest category [27]. The resource is public and available online through different browsers: AmiGo

(http://www.godatabase.org/cgi-bin/go.cgi), MGI (http://www.informatics.jax.org/userdocs/GO_help.shtml), QuickGo (http://www.ebi.ac.uk/ego/). Some data bases of GO annotations (i.e. matching single genes to the GO terms) have been compiled and are periodically updated (http://www.geneontology.org), while instruments are introduced to evaluate the statistical significance of the number of genes belonging to a given GO functional class identified in a microarray experiment [OntoExpress, http://vortex.cs.wayne.edu/Projects.html, 28]. Although the simple number of genes can be irrelevant, for activation of a biological process, the aggregation of many elements in a functional or spatial group may be an useful tool to drive further research about the biological meaning of the observed gene expression patterns.

In conclusion, microarray data analysis is a complex process, due to either the very large amount of information yielded by each single experiment or the high frequency of systematic and stochastic errors. Any of the different normalization and transformation methods described here can substantially modify a DNA microarray data set, so that a strenuous work of optimization and standardization of all steps of data gathering and analysis are required to gain the highest reproducibility and to allow direct comparison of data generated from multiple microarray analyses, especially when it is from different laboratories. Furthermore, for functional data analysis harmonization and integration of available databases [29] are required. These are the challenges that allow a capillary diffusion of this technology and the fulfillment of the expectations raised by its potentials.

III. APPLICATIONS OF THE MICROARRAY TECHNOLOGY FOR ASSESSMENT OF GENOME ACTIVITY IN NORMAL AND PATHOLOGIC CELLS AND TISSUES

AQ7

The DNA microarray technology has several applications. In the beginning it was applied for gene expression monitoring and then for mutation detection, mapping and evolutionary studies. Some of these aspects are discussed in this section.

Gene expression analysis through DNA microarrays is a powerful means to study the global profile of gene activity of any cell type or tissue, allowing many applications, to include molecular profiling of different tissues or stages of embryonic development, diseases classification according to gene expression signatures of pathologic specimens, identification of transcriptional modifications induced by drugs (pharmaco-genomics) and dynamic description of gene expression changes triggered by a particular stimulus in the cell, through single determinations or time-course analyses.

Sequencing of the entire genome of various species (Yeast, *C. elegans*, *A. thaliana*, *D. Melanogaster*, the house mouse and *H. Sapiens*) paved the way for a dynamic analysis of the genetic material of each cell type in the various differentiation and functional states (post-genomic or post-sequencing studies). Once all the genes present in the cells are identified the scientific community

needs to verify which of them is effectively activated in each given cell type. This leads to a functional genomic classification of all the tissues and organisms in all stages of differentiation and functional activation; each of these conditions are univocally defined by a specific combination of activated and/or repressed genes. DNA microarrays nowadays represent the analytical technology that best fulfill this need, and for this reason is currently applied at an increasing rate. This involves the concept that massive accumulation of gene expression data would occur quite rapidly and all this needs to be made rapidly and effectively available to all laboratories throughout the world. To this aim, publicly available gene expression data banks are organized, including Gene Expression Omnibus [www.ncbi.nlm.nih.gov/geo, 30], the Stanford University Microarray Database [genome-www5.stanford.edu/MicroArray/SMD/, 31] and the EMBL database [www.ebi.ac.uk/arrayexpress/, 32]. Furthermore, to make data from different laboratories directly and effectively comparable, a common scheme to standardize microarray data presentation is being studied (Minimum Information About Microarrays Experiments, [33–34].

These databases are constructed to include gene expression profiles not only of normal tissues but also of pathologic ones, as virtually all diseases are studied through DNA microarrays. The identification of pathologic gene expression patterns are useful for different aims, including: (i) definition of pathogenic alterations that underlie the disease, through the reconstruction of cellular pathways hyper-activated, or impaired, in the pathologic tissues; (ii) identification of gene expression patterns associated with the pathology, to be used for diagnostic applications; (iii) extraction of expression profiles useful for the prognostic evaluation and (iv) identification of new therapeutic targets through reconstruction of cellular pathways implicated in disease pathogenesis. Particular attention is given to molecular classification of cancer, as early diagnostic tools and more effective prognostic factors are actually required for most forms of malignant neoplasia. The complexity and wideness of microarray analysis provide an useful tool to investigate the heterogeneity of neoplastic diseases. In fact, large scale gene expression analyses show that each tumor has its own pattern of gene expression, which is different from that of other tumors derived from the same histological type [35]. The observation of specific expression profiles in many different tumors have suggested that the gene activation pattern (the so-called molecular signature, or portrait) is the result of a complex network of factors, including the genetic background of the patient, the tissue of origin, the grade of de-differentiation of the tumor, the clonal genetic alterations characterizing the neoplasia, the proliferation rate and the different cellular types that form the tumor mass. Therefore, different subsets of genes are identified in each tumor, reflecting the various components of its genetic background. Some genes are common to virtually all tumors or characterize the pathologic tissues from the normal counterparts. These genes are generally involved in cell-cycle control, adhesion and motility, apoptosis and angiogenesis [36–37]. On the other hand, supervised clustering methods show that other genes identified by microarray analysis allow to distinguish tumors according to their tissue of origin [38] or to discriminate the grade of

differentiation of neoplasia derived from the same tissue. Some studies show that gene expression analysis can even permit the distinction of functional subclasses among histologically homogeneous tumor samples, with different grade of malignancy and, therefore, quite different clinical outcome [39–41]. This is possible since the gene expression patterns reflect some biological properties of the tumor, that influence its ability to infiltrate the surrounding tissues, to give rise to metastasis [42–43] and to respond to therapy [44].

IV. APPLICATION OF MICROARRAY-BASED GENE EXPRESSION PROFILING ANALYSIS TO THE CHARACTERIZATION OF THE HORMONE-RESPONSIVE PHENOTYPE OF BREAST CANCER

Breast cancer is the most frequent malignant neoplasm in women, and is the best example of hormone-dependent cancer, defining in this way a tumor that needs a hormonal stimulus to grow and expand. Estrogens are the female sexual hormones and represent the main endogenous factor promoting breast cancer cells proliferation [45]. To increase our knowledge in the biological pathways involved in estrogen-dependent growth of breast cancer cells, gene expression analysis with cDNA microarrays was carried out on a hormone-dependent breast cancer cell line (ZR-75.1) before and after stimulation with a mitogenic dose of the natural estrogen 17β-estradiol. Time-course analysis was carried out in hormone-stimulated cells to provide a kinetic view of the gene responses to the hormone throughout a whole mitotic cycle (32 hrs in these cells). mRNA extracted from treated cells was used to synthesize cDNA labeled with the fluorescent dye Cy5, that was then mixed with an equal amount of a common reference target consisting of Cy3-labelled cDNA extracted from hormone-deprived cultures and hybridized to a glass array including 9,182 cDNA elements, representing 8,372 randomly selected unique gene/ESTs clusters. Three independent hybridization assays were performed for each sample pair and dye swapping (see above) was included in the protocol. 6,080 genes were selected as informative, or expressed in this cell line at detectable levels, by SAM statistics [13]; among these genes, 344 showed significant changes in activity in estrogen-treated cells [46–47]. We grouped estrogen-responsive genes through an unsupervised hierarchical clustering algorithm according to similarities in their inhibition or activation profiles in hormone stimulated vs control cells. Eight main clusters summarize the main patterns of gene expression changes detectable in estrogen growth-stimulated breast cancer cells (Figure 23.3). The first three clusters (1–3) group all down-regulated genes, with different kinetics of decrease in mRNA expression: significant down regulation occurring already 1 to 4 hrs into estrogen stimulation for the genes belonging to cluster 1, after 6 to 8 hrs for those of cluster 2 or after $\geq$12 hrs for genes of cluster 3. Clusters 4 to 6, instead, comprise activated genes whose transient expression patterns appear to be linked to cell cycle phasing, while the last two (clusters 7 and 8) include genes showing persistent activation by the hormone for up to 32 hrs, with RNAs starting to increase within the first 1 to 6 hrs of stimulation (cluster 7) or only after 8hrs (cluster 8).

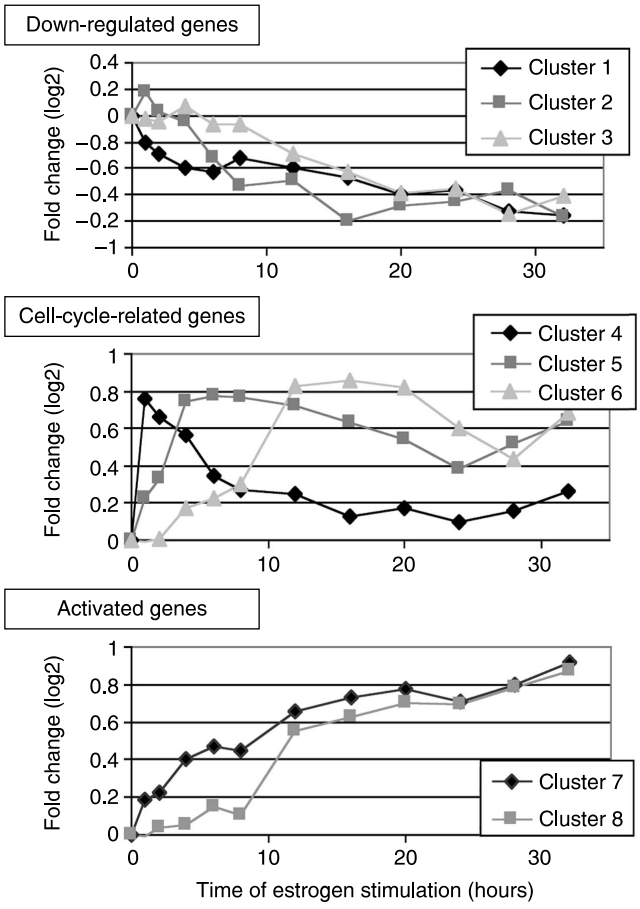

FIGURE 23.3 Example of the gene regulation patterns induced by a biological stimulus in responsive cells, in this case human breast cancer cells stimulated with a mitogenic dose of estrogens.

A functional classification of these genes according to Gene Ontology reveals that the cell cycle gene clusters (clusters 4, 5 and 6) comprise some important genes involved in cell cycle regulation (as c-*fos*, c-*jun*, c-*myc*, cyclin D1), while clusters 7 and 8 encompass a larger number of genes, with a general activation of some metabolic pathways, including glycolysis, nucleotide and cholesterol biosynthesis, revealing a clear activation of anabolic processes in these time windows.

This set of estrogen-regulated genes were then used to classify both breast cancer cell lines and specimens, whose expression data were publicly available in on-line databases (the NCI60 gene expression database for the molecular pharmacology of cancer [http://genome-www.stanford.edu/nci60/] and the Stanford University "Molecular Portraits of Human Breast Tumours" [48]

web site [http://genome-www.stanford.edu/breast_cancer/molecularportraits/], and the possibility to identify different subclasses with prognostic significance in an apparently homogeneous population of tumors was tested. A subset of 49 genes was able to distinguish between estrogen receptor (ER) positive and negative breast cancer cell lines or tumor specimens, thus confirming the predictive value of these genes [Figure 23.4 and Refs. 49–50].

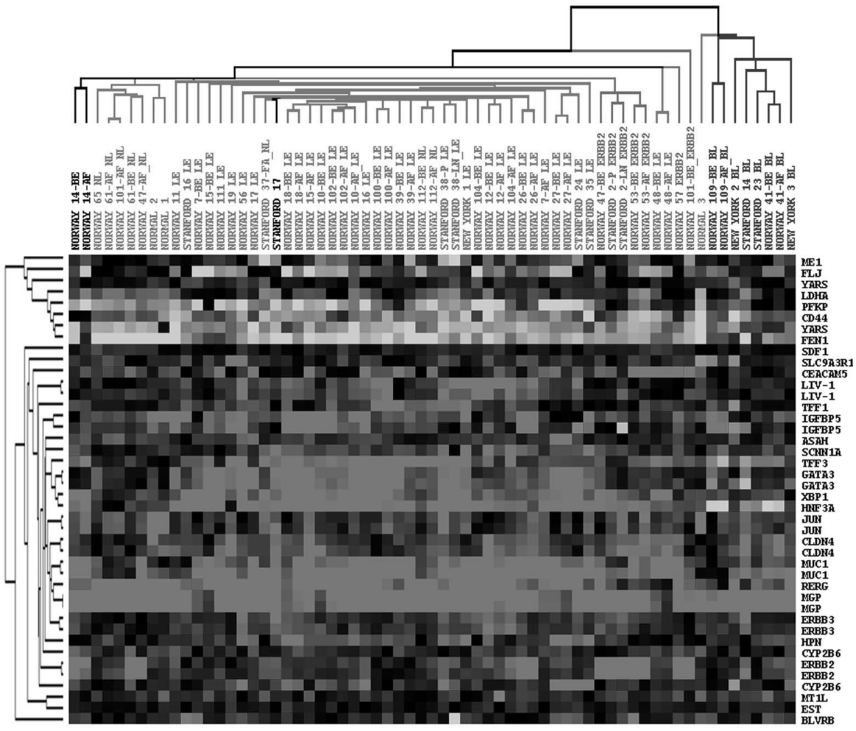

FIGURE 23.4 Clustering of breast cancer specimen according to gene expression profiling of a defined set of estrogen-responsive genes. Cluster analysis of 62 breast tumor surgical specimens and 3 normal mammary gland biopsies based on expression of a subset of 27 estrogen responsive genes identified in BC cell lines and 4 molecular markers of ESR1 (ER□) positive breast tumors. Expression data were from Perou *et al.* [48]; sample denomination has been maintained the same as in the original study. Colors highlights the molecular typing of the breast cancer samples, e.g. luminal epithelial/ER+ (LE, red), basal-like (BL, dark blue), *cErb-B2*-overexpression (ERBB2, green), normal-like (NL, light brown) and undetermined (black). The three normal breast tissue samples are also marked in light brown. Each column of the expression matrix represents the tissue sample indicated at the top and each row refer to a gene, colors of the matrix elements represent mRNA expression levels relative to a common reference sample (green for sample/reference ratios < 1, red for ratios > 1, black for ratios near 1 and gray for missing data). The top dendrogram represents hierarchical relationships between samples; the terminal branches are colored to reflect the known molecular nature of each tumor or normal tissue sample.

AQ8

This is fundamental for prognostic evaluation of these tumors, as ER expression is among the main prognostic factors for breast cancer, and reflects the disease responsiveness to hormonal therapy. These results confirm the possibility to discriminate between biologically different forms of cancers through the study of gene expression signatures related to relevant physiological or pathological stimuli.

ACKNOWLEDGMENTS

Preparation of this review and of the experimental data reported therein were supported by research grants from: Associazione Italiana per la Ricerca sul Cancro, Seconda Università degli Studi di Napoli (Fondi per la Ricerca di Ateneo e per Assegni e Dottorati di Ricerca), Ministero dell'Istruzione, dell'Università e della Ricerca (PRIN 2002067514_002 and FIRB RBNE0157EH) and the European Union (Contracts BMH4-CT98-3433, QLG1-CT-2000-01935 and QLK3-CT-2002-02029).

REFERENCES

AQ2

1. *Nature Genetics*. Vol. 32 (Supplement). Dec. 2002. AQ3
2. Schena M, Shalon D, Davis RW, Brown PO, Quantitative monitoring of gene expression pattern with a complementary DNA microarray, *Science*, 270:467–470, 1995.
3. Schena M, ed., *Microarray analysis*, Wiley-Liss, *Hoboken New Jersey*, USA, 2003.
4. Bowtell D, Sambrook J, eds., *DNA mycroarrays—A molecular cloning manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2003.
5. Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW, Microarray: Biotechnology's discovery platformfor functional genomics, *Trends Biotechnol*., 16:301–306, 1998.
6. Cleveland WS, Robust locally weighted regression and smoothing scatterplots, *J. Amer. Stat. Assoc*., 74:829–836, 1979.
7. Knudsen S, *A Biologists's Guide to Analysis of DNA Microarray Data*, New York, Wiley-Liss, 2002.
8. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, Saxild HH, Nielsen C, Brunak S, Knudsen S, A new nonlinear normalization method for reducing variability in DNA microarray experiments, *Genome Biol*., 3:research 0048, 2002. AQ9
9. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J, Within the fold: Assessing differential expression measures and reproducibility in microarray assays, *Genome Biol*., 3:research 0062, 2002. AQ9
10. 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*., 30: e15, 2002. AQ4
11. Churchill GA, Fundamentals of experimental design for cDNA microarrays, *Nat. Genet*., 32 Suppl:490–495, 2002.

12. Tsai CA, Chen YJ, Chen JJ, Testing for differentially expressed genes with microarray data, *Nucleic Acids Res.*, 31:E52, 2003
13. Tusher VG, Tibshirani R, Chu G, Significance analysis of microarrays applied to the ionizing radiation response, *Proc. Natl. Acad. Sci. USA.*, 98:5116–5121, 2001.
14. Kerr MK, Martin M, Churcill GA, Analysis of variance of gene expression microarray data, *J. Comput. Biol.*, 7:819–837, 2000.
15. Baldi P, Long AD, A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes, *Bioinformatics*, 17:509–519, 2001.
16. Ideker T, Thorsson V, Siegel AF, Hood LE, Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data, *J. Comput. Biol.*, 7:805–817, 2000.
17. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose MA, NetAffx: Affymetrix probe sets and annotations, *Nucleic Acids Res.*, 31:82–86, 2003.
18. Iazzetti G, Calabrò V, Saviozzi S, Weisz A, Lania L, Calogero R, BSC: A clustering program for DNA array expression data, *Proceedings of Biocomp, 2001*, Siena, 2001. http://obelix.bio.uniroma2.it/www/abstr_2001.html#iazzetti
19. Eisen MB, Spellman PT, Brown PO, Botstein D, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA.*, 95:14863–14868, 1998.
20. 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.*, 96:2907–2912, 1999.
21. Dettling M, Buhlmann P, Supervised clustering of genes, *Genome. Biol.*, 3: Research, 0069, 2002.
22. Mateos A, Herrero J, Tamames J, Dopazo J, Supervised neural networks for clustering conditions in DNA array data after reducing noise by clustering gene expression profiles, In: SM Lin, KF Johnson, eds., *Methods of Microarray Data Analysis II*, Boston: Kluwer Academic Publ, pp. 91–103, 2002.
23. Friedman N, Linial M, Nachman I, Pe'er D, Using Bayesian networks to analyze expression data, *FEBS Lett.*, 451:142–161, 1999.
24. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Jr Ares M, Haussler D, Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proc. Natl. Acad. Sci. USA.*, 97:262–267, 2000.
25. Tobler JB, Molla MN, Nuwaysir EF, Green RD, Shavlik JW, Evaluating machine learning approaches for aiding probe selection for gene-expression arrays, *Bioinformatics 18 Suppl.*, 1:S164–S171, 2002.
26. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, Gene ontology: Tool for the unification of biology, the gene ontology consortium, *Nat. Genet.*, 25:25–29, 2000.
27. Ashburner M, Ball CA, Blake JA, Butler H, Cherry JM, Corradi J, Dolinski K, Eppig JT, Harris MA, Hill DP, Lewis S, Marshall B, Mungall C, Reiser L, Rhee S, Richardson JE, Richter J, Ringwald M, Rubin GM, Sherlock G,

AQ9

Yoon J, Creating the gene ontology resource: Design and implementation, *Genome Res*., 11:1425–1433, 2001.

28. Facchiano A, A Weisz Internet tools for the analysis of gene expression by database integration, *Proceedings of the NETTAB 2001 Workshop: "CORBA and XML—Toward a bioinformatics integrated network environment"*, Italy, 2001, pp. 99–102.
29. Khatri P, Draghici S, Ostermeier GC, Krawetz SA, Profiling gene expression using onto-express, *Genomics*., 79:266–270, 2002.
30. Edgar R, Domrachev M, Lash AE, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Res*., 30:207–210, 2002.
31. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D, Cherry JM, The Stanford Microarray Database, *Nucleic Acids Res*., 29:152–155, 2001.
32. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca P -Serra, Sansone SA, ArrayExpress—a public repository for microarray gene expression data at the EBI, *Nucleic Acids Res*., 31:68–71, 2003.
33. 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*., 29:365–371, 2001.
34. Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Jr Stoeckert CJ, Brazma A, Design and implementation of microarray gene expression markup language (MAGE-ML), *Genome Biol*., 3:Research 0046, 2002.

AQ9

35. Chung CH, Bernard PS, Perou CM, Molecular portraits and the family tree of cancer, *Nat. Genet*., 32 Suppl:533–540, 2002.
36. Hanahan D, Weinberg RA, The hallmarks of cancer, *Cell*, 100:57–70, 2000.
37. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nat. Med*., 7:673–679, 2001.
38. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR, Multiclass cancer diagnosis using tumor gene expression signatures, *Proc. Natl. Acad. Sci. USA*., 98:15149–15154, 2001.
39. Weiss MM, Kuipers EJ, Postma C, Snijders AM, Siccama I, Pinkel D, Westerga J, Meuwissen SG, Albertson DG, Meijer GA, Genomic profiling of gastric cancer predicts lymph node status and survival, *Oncogene*, 22:1872–1879, 2003.
40. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL, Gene expression

patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *Proc. Natl. Acad. Sci. USA*., 98:10869–10874, 2001.

41. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S, Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nat. Med*., 8:816–824, 2002.
42. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH, Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, 415:530–536, 2002.
43. Kikuchi T, Daigo Y, Katagiri T, Tsunoda T, Okada K, Kakiuchi S, Zembutsu H, Furukawa Y, Kawamura M, Kobayashi K, Imai K, Nakamura Y, Expression profiles of nonsmall cell lung cancers on cDNA microarrays: Identification of genes for prediction of lymph-node metastasis and sensitivity to anti-cancer drugs, *Oncogene*, 22:2192–2205, 2003.
44. Okutsu J, Tsunoda T, Kaneta Y, Katagiri T, Kitahara O, Zembutsu H, Yanagawa R, Miyawaki S, Kuriyama K, Kubota N, Kimura Y, Kubo K, Yagasaki F, Higa T, Taguchi H, Tobita T, Akiyama H, Takeshita A, Wang YH, Motoji T, Ohno R, Nakamura Y, Prediction of chemosensitivity for patients with acute myeloid leukemia, according to expression levels of 28 genes selected by genome-wide complementary DNA microarray analysis, *Mol. Cancer Ther*., 1:1035–1042, 2002.
45. Weisz A, Estrogen regulated genes, In Oettel M, Schillinger E, eds., *Handbook of Experimental Pharmacology*, Vol. 135/I: Estrogens and Antiestrogens I. Berlin-Heidelberg-New York, Springer Verlag, 1999, pp. 127–151.
46. Cicatiello L, Facchiano A, Calogero R, De Bortoli M, Bresciani F, Weisz A, Gene expression monitoring in hormone-responsive human breast cancer cells during estrogen-induced cell cycle progression, *Proceedings of AACR/Nature Genetics Joint Conference: Oncogenomics: Dissecting Cancer through Genome Research*, Tucson, AZ, 2001. http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v27/n4s/full/ng0401supp_95a.html&_UserReference=C0A804EC4650B9B7E02E44E479153B0D3D72
47. Cicatiello L, Natoli G, Scafoglio C, Altucci L, Cancemi M, Facchiano A, Calogero R, Iazzetti G, De Bortoli M, Sfiligoi C, Sismondi P, Biglia N, Bresciani F, Weisz A, The gene expression program activated by estrogen in hormone-responsive human breast cancer cells, Submitted for publication.
48. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D, Molecular portraits of human breast tumours, *Nature*, 406:747–752, 2000.
49. Weisz A, Identification of the gene expression signature which characterizes human breast cancer cells response to estrogen, *Proceedings of Second International Cancer Congress: Translational Research in Cancer*, Rovigo, p. 83, 2001.
50. Weisz A, Basile W, Scafoglio C, Natoli G, Altucci L, Bresciani F, Facchiano A, Sismondi P, Cicatiello L, De Bortoli M, Molecular identification of ERalpha-positive breast cancer cells by the expression profile of an intrinsic set of hormone regulated genes, Submitted for publication.

AQ10 AQ11 AQ11