Historical Background and Anticipated Developments

BERTRAND JORDAN

Marseille-Génopole, Case 901, 13288 Marseille, France

ABSTRACT: Expression profiling using DNA arrays is often believed to have appeared during the second half of the 1990s, and to be based exclusively on nonisotopic methods. In fact, the first article describing the application of cDNA arrays to expression analysis was published in 1992, relied on radioactive labeling, and was a new development of "high-density" membranes used until then essentially for efficient screening of libraries. Several papers described the use of this technology for simultaneous expression measurement of thousands of genes at the time when the first glass microarrays were published. Simultaneously, oligonucleotide chips, originally developed for resequencing and mutation detection applications, were shown to be capable of expression measurement as well. The three approaches have developed over the years and still coexist, as each of them has specific advantages (and drawbacks); the major issues have become those of data quality, data analysis and storage (ideally in a common public database). Meanwhile, the technology continues to evolve. The most obvious trend is a shift towards using arrays of relatively long oligonucleotides that combine most of the advantages of very long (cDNA) and very short (25-mer) DNA segments. The search for better detection methods, ideally without labeling of the sample, is continuing, although it seems difficult to reach the required sensitivity. New materials for microarray manufacture and new implementations of existing methods have appeared. In addition, the field is progressively becoming segmented into high gene number, low volume (research) applications on the one hand, and low gene number, high throughput (diagnostic) uses on the other.

KEYWORDS: DNA arrays; expression; history; trends

HISTORICAL BACKGROUND

DNA arrays consist of a series of DNA segments regularly arranged on a support, and the expression measurement involves hybridizing the whole array with a labeled nucleic acid sample. The essential feature is parallel processing: in a single experiment, information is obtained on the expression level for each of the thousands of genes represented on the array. This parallelism has made the technology essential at a time when many megabases of genome sequence need to be understood in functional terms.

Address for correspondence: Bertrand Jordan, Marseille-Génopole, Case 901, 13288 Marseille, France. Voice: 33 (0)4 91 82 94 70; fax: 33 (0)4 91 82 94 71.
jordan@genopole.univ-mrs.fr

Ann. N.Y. Acad. Sci. **975: 24–32 (2002).**

Early DNA Arrays for Homology Studies and Library Access

DNA arrays already existed in the seventies as dot blots[1] and slot blots that allowed for homology determination or expression analysis on a series of samples, with radioactive labeling in almost all cases. A major change in the field came with the development in the late 1980s of robotic devices ("gridding robots") that made it possible to array bacterial colonies in compact and regular patterns. The resulting "high-density filters" contained typically 10,000 spots on a square $22 \times 22\,cm^2$ surface, corresponding to a "pitch" (center to center spacing) of approximately 2 millimeters (see FIGURE 1). These arrays were essentially used for library access, providing an efficient approach to genome analysis,[2,3] at a time when the path taken

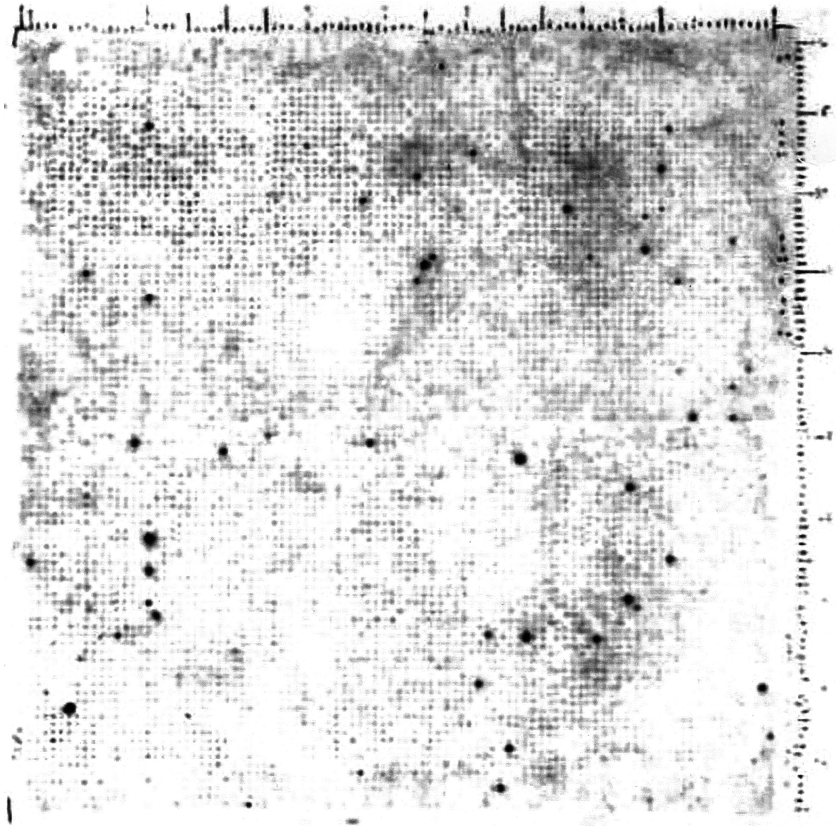

FIGURE 1. Early DNA array: 10,000 bacterial colonies containing DNA segments cloned in cosmids spotted on a 22 by $22\,cm^2$ nylon membrane. Hybridization with a radioactive probe prepared from the insert of a cDNA clone reveals positive spots indicating the corresponding genomic clones. The whole grid is visible because of light background hybridization. Successive hybridizations to search for various genomic clones are performed on the same membrane without stripping (to avoid loss of material), hence the large number of "positive" spots.

in the USA relied instead on PCR screening of cleverly arranged pools of clones.[4,5] Several resource centers accumulated various genomic and cDNA libraries and distributed them to laboratories as high-density membranes. When the user had identified a positive spot using a probe for a gene of interest, the center then provided the corresponding clone, an important advance at a time when library access was difficult and sequence information very meager.

Beginnings of Expression Analysis

The use, for expression analysis, of unordered or ordered colony filters containing cDNA clones began in the early 1980s with differential screening. Duplicate membranes, containing clones from conventional or subtracted cDNA libraries, were hybridized in parallel with complex labeled cDNA mixtures prepared from two different samples. The goal, of course, was to pin-point genes whose expression was different in the two conditions. Using radioactivity and X-ray film detection, the method was necessarily qualitative; it was also cumbersome and suffered from a number of technical problems, but it was nevertheless instrumental in isolating several important genes such as those coding for the T cell receptor[6] or the CTLA (Cytotoxic T Lymphocyte-Associated transcript) series of molecules.[7]

The concept of using imaging plate systems for quantitative acquisition of hybridization signals, allowing for more refined analysis of expression patterns, was discussed as early as 1990, with a first publication reporting actual data in 1992.[8] The full implementation of the technology took some time, however, as expression measurement with DNA arrays involves the quantification of very weak signals if data for genes expressed at very low levels is required. Artefacts leading to spurious data (especially with the unsequenced cDNA libraries of that period) had to be identified and eliminated,[9] and the fairly primitive image analysis programs of the period required serious improvement—a task made easier by increases in the computing power available to scientists. Automation of PCR and standardization of the plasmid vectors used for cDNA libraries led to a shift from colony filters to membranes on which amplified DNA was deposited.

High-Density Membranes, alias Macroarrays

During the first half of the 1990s, several groups worked out the methods, published proof-of-principle papers[9–11] and began accumulating expression data in different systems. Some of this work was performed in "gene discovery" mode, i.e., measuring the expression level in a number of conditions for a large set of genes, to highlight those that are the most relevant with regard to the biological question approached. Expression measurement was also done in an "expression profiling" mode, in which the set of genes (often more restricted) was chosen *a priori* and usually well known, and the objective was to obtain information on the samples: analysis of the expression profile for a series of tumors, for instance, in the hope of obtaining prognostic and therapeutic information (see, e.g., ref. 12). In this case the genes are used as tools to derive information on the samples. The high-density filter (under the more trendy name of "macroarray") is still widely used: for experiments of moderate scope, it performs quite adequately, requires only small samples, and blends well with existing laboratory equipment.

Miniaturization: cDNA Microarrays …

In the mid-1990s, miniaturization became a major issue in the further development of DNA arrays, with the aim of increasing the number of genes assayed in a single experiment, and also of reducing sample usage—although most current systems still require microgram amounts of messenger RNA (or a recourse to RNA amplification), a major limitation in practice. One avenue involved scaling down cDNA arrays, using optical detection methods (fluorescence) because of their superior resolution, and depositing the DNA spots on very planar supports (glass slides) to allow intensity measurement with confocal optics in order to achieve the required sensitivity. First published in 1995,[13] this approach has now been applied to many important studies. The use of fluorescence allows for dual labeling, simplifying comparisons and facilitating standardization of series of experiments; good sensitivity has been obtained, although sample requirements remain high. Microarrays can be constructed in the laboratory; the necessary equipment is commercially available, although the expense and logistics are not trivial. Ready-made arrays have appeared on the market, although this has been a relatively slow process owing to the time taken to build up the necessary logistics as well as to intellectual property issues.

Microarrays can also be produced on nylon membranes. Because of the intrinsic fluorescence of all nylon supports (so far), detection must be performed by enzymatic means that are convenient and affordable, but relatively insensitive,[14] or with (^{33}P) radioactive labeling, using high-resolution detectors that provide sufficient resolution to quantify arrays with 400 micron pitch. In this form the method makes possible expression profiling at reasonable sensitivity with very small biological samples.[15]

… and Oligonucleotide Chips

The other, competing approach is that of oligonucleotide chips, pioneered by the firm Affymetrix. These glass chips comprise hundreds of thousands of small (currently $18\times18\mu^2$) "features," each containing several million copies of a given oligonucleotide (20 to 25-mer). These arrays were originally developed for "quasi-sequencing" (mutation detection) applications.[16] They also allow for expression measurement, at the expense of assaying each gene with several (20 to 40) oligonucleotides and controls, in order to average out signal and background artefacts due to the vagaries of short oligonucleotide hybridization.[17] The manufacturing process, very similar to that of microelectronic devices, promises further miniaturization beyond the present chips that contain 500,000 features. Based solely on sequence knowledge, they do not require the cumbersome logistics of cDNA clone storage and PCR amplification, in contrast to cDNA arrays. However, this approach lacks flexibility; the arrays were initially very expensive and, accordingly, their use in the academic sector has developed slowly. Alternative approaches to oligonucleotide chips (notably different synthesis methods), and lower prices stemming from increasing competition, are increasing the popularity of this technology.

Data Acquisition, Storage, and Mining

The importance of software issues in expression measurement was not immediately recognized, but quickly became apparent as more and more data began to flow

from these experiments. It includes a number of aspects, from verification of the validity of measurements to sophisticated data mining through data representation and storage issues. Although available computing power has increased dramatically in the last decade, these issues are still far from being satisfactorily solved. An accessible and comprehensive account of this field has recently been published.[18]

Data Validation, Quality, and Statistical Issues

Many of the initial papers on expression analysis using DNA arrays relied on data whose validity was not proven. A number of issues have arisen, for example, errors in the cDNA collections used to produce the PCR products that are supposed to represent specific genes,[19] or in the sets of sequences used by manufacturers to derive oligonucleotide chips. The practice of systematically replicating measurements and of providing a statistical estimate of the variability in the data set has become general only recently.[20] Software suites dedicated to data acquisition and validation now exist, and publication criteria have become more strict; efforts are being made to define the information that should be provided with microarray results in order to make these usable by others, for example the MIAME (Minimum Information on A Microrarray Experiment) system developed by the European Bioinformatics Institute together with a wide group of users.[21]

LIKELY DEVELOPMENTS

Whole-Genome Chips?

With the completion of the human (and other) sequences, the race to higher density towards a "whole-genome" chip or microarray allowing for simultaneous measurement of the expression of complete sets of genes for a given organism will continue. Current limitations to minimum spot size and spacing for spotted arrays mean that the highest attainable densities remain below 5,000 spots per cm^2; thus the human or murine complement of genes is covered, at this time, by a set of several microarrays rather than a single array. Changes in spotting mechanisms and surface chemistry may allow for closer spacing, but in any case the future of very complex cDNA microarrays is probably limited because of the difficulty and expense involved in producing such large numbers of PCR products.

Concerning oligonucleotide chips, devices currently marketed by Affymetrix contain more than 500,000 short (20-mer) oligonucleotides on a single chip. However, because more than 20 oligonucleotides (including mismatched controls) are used for each gene, the latest "Human Genome U133" set from this firm requires two arrays to assay approximately 39,000 transcripts. The relatively poor yield of the photochemical on-chip synthesis process used does not allow (so far) for the manufacture of long oligonucleotides that would provide the same specificity with fewer "features." Increasing the chip density and placing one or a few million "features" on the surface of a microscope slide is probably feasible, although the resolution and sensitivity of reading devices would then have to be improved. In summary, assessing all human genes with a single Affymetrix chip will eventually be possible but may

represent the limit of this technology. New on-chip synthesis approaches allowing for the production of long oligonucleotides (see below) could remove this limitation.

In any case "full human genome" chips will definitely be expensive, and problems in data acquisition, storage and analysis, because of the sheer volume of data, will be a deterrent to their use. Smaller, specialized arrays are likely to remain attractive for many purposes.

From "Clone-Based" to "Sequence-Based Arrays"

DNA arrays based on the use of cDNA clones suffer from the difficulty, expense and tedium involved in assembling collections of thousands of verified cDNA clones and in producing sufficient amounts of purified DNA of each of them by PCR. Oligonucleotide chips do not suffer from the same problem. They completely eliminate the recourse to clones since they are based solely on sequence information—which is already vast for many organisms and is increasing at an explosive rate. In addition, economies of scale can be considerable, and will be reflected in the prices if competition increases—this is already the case to a certain extent.

A number of laboratories and firms are developing "on-chip" oligonucleotide synthesis techniques that rely, for example, on the fast dispensing of synthesis reagents to individual sites on the chip by print head-like devices; some of them have already begun to market their processes or products. Such procedures allow for the use of classical synthesis chemistry (rather than the less efficient photochemical method), making possible the manufacture of much longer oligonucleotides (40–60-mers) that in turn reduce the need for redundancy in the chip because of their higher specificity. This makes it easier to represent many genes on a chip since only one or a few (long) oligonucleotides are needed to assay each of them. In addition, these approaches are inherently more flexible that the Affymetrix photochemical method. The fabrication of a different chip simply involves reprogramming the dispensing of reagents, rather than the manufacture of a complete new series of precision masks for the photochemical procedure. Other firms offer arrays made with pre-synthesized (long) oligonucleotides. The development of these technologies will not depend solely on scientific and engineering advances, as intellectual property in this field is already a hotly contested issue. Hopefully, these conflicts will be resolved in a fashion that opens up competition.

Altogether, it is certain that sequence-based DNA chips will be increasingly used in the future, certainly for "standard" sets and possibly, depending on methods development, for more specialized arrays.

From "Home-Brew" to Commercial Chips

A definite shift towards the purchase of commercially manufactured devices is apparent. It does not make economic sense for individual groups or even research institutes to invest large resources in the construction of standard microarrays, a task that can be handled more efficiently by industry or, in some cases, by public resource centers. This is not to say that the manufacture of microarrays will disappear from the research environment: custom arrays allowing for the assay of limited, specialised sets of genes will remain useful in many cases, and maximum flexibility can be achieved by making them "in house." Alternately, some manufacturers may undertake

to produce such custom arrays, while others provide sets of "ready-to-spot" PCR products or oligonucleotides. The end result is likely to be a mixed situation in which large or standard sets of genes are assessed with commercial oligonucleotide chips or microarrays, while custom arrays are made in various academic-corporate arrangements. Of course in this context it is very important to standardize detection systems so that each type of industry-produced DNA array does not require its own proprietary scanning device.

From "Stand Alone" Array to Integrated "Lab-on-a-Chip"

Biochip technology is not limited to DNA arrays. The integration of a number of functionalities within chips whose dimensions are measured in centimeters is well underway; such devices can perform filtration, fluid handling, and reagent mixing, PCR reactions and even capillary electrophoresis. Their development is strongly stimulated by the need of pharmaceutical companies to perform literally millions of tests in the course of screening compounds for activities ("high throughput screening"), and by the requirement to do these assays very quickly, in a highly parallel mode and with the smallest possible amount of reagents.[22] At least for industrial and clinical systems, expression measurement (probably assessing limited numbers of genes) is likely to be packaged into such systems. This is for example the form in which expression measurement will penetrate in the clinical oncology laboratories—if indeed the clinical utility of such data is confirmed.

Detection without Labeling?

Fluorescent labeling is relatively cumbersome, interferes by steric hindrance with hybridization, and requires high-end, expensive detection systems; radioactive labelling is undesirable in many environments, and provides limited resolution even with high-performance (and costly) detectors. It would be very advantageous to achieve detection of the fact that a given location in the array has hybridized, and to quantify the extent of hybridization, by some other method. This should preferably involve the measurement of an electrical signal, and would, ideally, require no modification of the sample before hybridization. Much effort is devoted by many groups towards achieving this.[23,24] The approaches explored range from the detection of some subtle change of electrical properties upon hybridization to very exotic methods: microbalances "weighing" the extra mass of the hybridized material, or determination of the number of double-stranded (thus hybridized) molecules by atomic force microscopy. Proof of principle has been obtained for some of these approaches; it remains to be seen whether they can achieve the required sensitivity and throughput. If successful, they are likely to have an impact first in applications of DNA arrays such as bacterial identification or mutation detection, where a "yes/no" answer is often sufficient, rather that in expression measurement where accurate quantification is required.

More Sophisticated Data Interpretation and (Hopefully) Public Expression Databases

Software and bioinformatics development is a very important aspect that was not sufficiently taken into account at the beginning of the "DNA array revolution." Even today, the type of analysis performed on expression results remains relatively

unsophisticated. In addition, much of the actual data is still unavailable outside of the originator's laboratory and the selected data sets provided by some groups on their websites lack a common format making them directly usable by others. Great efforts are being made to develop better analysis software including both extensive statistical, correlation and clustering analysis, and direct links to current, constantly updated information available on the Web. In addition, serious attempts are underway to define a standard data format that would make it possible to store expression data in the way in which DNA sequences have been archived, and to make it thus generally available and useful to the research community. A number of repositories already exist (for up-to-date lists see http://www.ncgr.org/genex/ and http://www.biologie .ens.fr/en/genetiqu/puces/bddeng.html), but so far there is no unified system comparable to the Genbank and EMBL sequence databases. Of course the problem of data format and standardization is much more complex for expression data than for sequence information . . . . The "MIAME" standard developed at the European Bioinformatics Institute[21] and already referred to shows how far we still have to go.

Expression Measurement Is Here to Stay

This is an easy prediction to make. Undoubtedly other methods able to add functional significance to gigabases of DNA sequence will be streamlined, made more efficient and more amenable to large-scale implementation: protein interaction studies, proteomics in general, gene inactivation experiments in various model systems are bound to become faster, easier, cheaper. However large-scale expression measurement, enhanced by general availability of sequence data and boosted by technical development of DNA arrays, will certainly remain a major approach in fundamental and applied biology for quite a long time.

REFERENCES

1. KAFATOS, F.C., C.W. JONES & A. EFSTRATIADIS. 1979. Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. Nucleic Acids Res. **7:** 1541–1552.
2. LENNON, G.G. & H. LEHRACH. 1991. Hybridization analyses of arrayed cDNA libraries. Trends Genet. **7:** 314–317.
3. HOHEISEL, J.D., G.G. LENNON, G. ZEHETNER & H. LEHRACH. 1991. Use of high coverage reference libraries of *Drosophila melanogaster* for relational data analysis. A step towards mapping and sequencing of the genome. J. Mol. Biol. **220:** 903–914.
4. OLSON, M., L. HOOD, C. CANTOR & D. BOTSTEIN. 1989. A common language for physical mapping of the human genome. Science **245:** 1434–1435.
5. GREEN, E.D. & M.V. OLSON. 1990. Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. Proc. Natl. Acad. Sci. USA **87:** 1213–1217.
6. HEDRICK, S.M., D.I. COHEN, E.A. NIELSEN & M.M. DAVIS. 1984. Isolation of cDNA clones encoding T cell-specific membrane-associated proteins. Nature **308:** 149–153.
7. BRUNET, J.F., F. DENIZOT & P. GOLSTEIN. 1988. A differential molecular biology search for genes preferentially expressed in functional T lymphocytes: the CTLA genes. Immunol. Rev. **103:** 21–36.
8. GRESS, T.M., J.D. HOHEISEL, G.G. LENNON, *et al*. 1992. Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. Mamm. Genome **3:** 609–619.

9. NGUYEN, C., D. ROCHA, S. GRANJEAUD, *et al*. 1995. Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. Genomics **29:** 207–215.
10. ZHAO, N., H. HASHIDA, N. TAKAHASHI, *et al*. 1995. High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression. Gene **156:** 207–213.
11. PIETU, G., O. ALIBERT, V. GUICHARD, *et al*. 1996. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res. **6:** 492–503.
12. BERTUCCI, F., S.VAN HULST, K. BERNARD, *et al*. 1999. Expression scanning of an array of growth control genes in human tumor cell lines. Oncogene **18:** 3905–3912.
13. SCHENA, M., D. SHALON, R.W. DAVIS & P.O. BROWN. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:** 467–470.
14. CHEN, J.J., R. WU, P.C. YANG, *et al*. 1998. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. Genomics **51:** 313–324.
15. BERTUCCI, F., K. BERNARD, B. LORIOD, *et al*. 1999. Sensitivity issues in DNA array-based expression measurements and performance of Nylon microarrays for small samples. Hum. Mol. Genet. **8:** 1715–1722.
16. FODOR, S.P., J.L. READ, M.C. PIRRUNG, *et al*. 1991. Light-directed, spatially addressable parallel chemical synthesis. Science **251:** 767–773.
17. WODICKA, L., H. DONG, M. MITTMANN, *et al*. 1997. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nat. Biotechnol. **15:** 1359–1367.
18. BRAZMA, A., A. ROBINSON & J. VILOO. 2001. Gene expression data mining and analysis *In* DNA Microarrays: Gene Expression Applications. B. Jordan, Ed.: 106–129. Springer-Verlag. Berlin.
19. HALGREN, R.G., M.R. FIELDEN, C.J. FONG & T.R. ZACHAREWSKI. 2001. Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. Nucleic Acids Res. **29:** 582–588.
20. LEE, M.L., F.C. KUO, G.A. WHITMORE & J. SKLAR. 2000. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc. Natl. Acad. Sci. USA **97:** 9834–9839.
21. www.ebi.ac.uk/microarray/
22. TALARY, M.S., J.P. BURT & R. PETHIG. 1998. Future trends in diagnosis using laboratory-on-a-chip technologies. Parasitology **117** (Suppl.) :S191–S203.
23. SOUTEYRAND, E., J.P. CLOAREC, J.R. MARTIN, *et al*. 1997. Direct detection of the hybridization of synthetic homo-oligomer DNA sequences by field effect. J. Phys. Chem. B **101:** 2980–2985.
24. WANG, J., A. JIANG & B. MUKHERJEE. 1999. New label-free DNA recognition based on doped nucleic-acid probes within conducting polymer films. Anal. Chim. Acta **402:** 7–12.