GENE CHIPS AND DNA MICROARRAYS

Many, but not all, changes in cellular physiology are accompanied by changes in the transcription of genes. These transcriptional changes can be followed by measuring the levels of various mRNAs using hybridization. In the last several years, there has been great interest in developing methods to determine the levels of very large numbers of mRNAs using solid state hybridization arrays. The hope is that by determining a complete mRNA profile for a cell, it will be possible to design new drug screens, characterize various pathological states and understand interactions among genes that act in a pathway. Data from genome-wide transciptional profiling has been available for over a year. In contrast, efficient methods for monitoring protein-protein association and protein modification are only now being developed. We will therefor examine DNA microarrays first.

The first pan-genomic solid state nucleotide arrays to have been produced on a large scale contained genes from the budding yeast *S. cerevisiae*. *S. cerevisiae*, in addition to being the most widely studied simple eucaryote, is the first eukaryotic organism whose complete sequence has been determined. It is currently thought that S. cerevesiaie contains about 6200 functional genes. The first results from pan-genomic gene analyses in yeast were published in 1997 and the amount of information in public databases is increasing rapidly.

The first DNA arrays to be commercialized are the Gene Chips from Affymetix. However, at least four other companies are selling various forms of DNA arrays and over a dozen companies have announced their intention to do the same.

THE TECHNOLOGY OF MRNA PROFILING

mRNA profiling, as currently conceived, has five steps:

i) Constructing a DNA array comprised of either gene fragments or oligonucleotides
ii) Preparing labeled cDNA from control and experimental cultures of cells.
iii) Hybridizing the labeled cDNA to the DNA microarray
iv) Scanning the array to determine the levels of hybridized message at each position in the array
v) Analyzing the mRNA profile to recover meaningful biological data

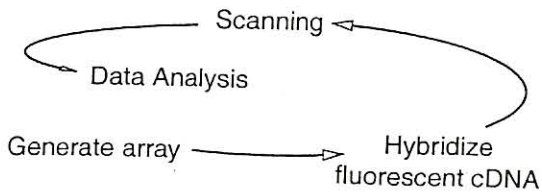


Figure 1. Overview of DNA microarray analysis. Fluorescent cDNA is prepared by reverse transcription of mRNA using red and green dyes (see below).

∞ - - ∞

The preparation of cDNAs and the hybridization of the cDNA to immobilized nucleotide targets make use of widely available and familiar technologies. We will not discuss them in any detail. However, a key aspect of the mRNA profiling is the simultaneous

measurement of signals for control and experimental cDNA samples. Typically, control cDNA is labeled green (with Cy3) and experimental cDNA is labeled red (with Cy5). The ratio of mRNA levels in experimental and control cells in then read directly on a single array as the ratio of red to green fluorescence (Figure 2).

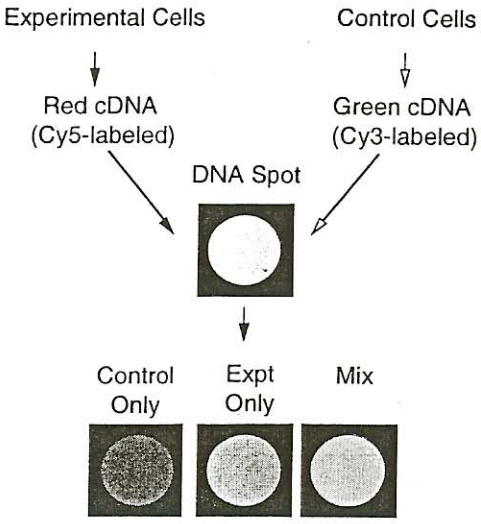


Figure 2. Schematic of an idealized DNA spot hybridized to control cDNA (green) and experimental cDNA (red). The red-green ratio is a measure of the change in mRNA levels between the control and experimental mRNA populations.

∞ - - ∞

Oligonucleotide Arrays (Affymetrix Gene Chips)

Two quite different methods have been developed for generating DNA arrays. The first, from Affymetrix, involves variations on methods developed in the semiconductor industry. In the Affymetrix approach, compact arrays of oligonucleotides are constructed using photolithography (Figure 3). The substrate for the Affymetrix arrays is derivatized silicon. Opaque masks are used to expose selected areas of the silicon chip to light. This releases blocking groups and exposes reactive moieties on the chip. By flooding the chip with a modified nucleotide, a base can be added selectively at the deprotected spots. This method allows the stepwise synthesis of oligonucleotides up to about 25 bases long.

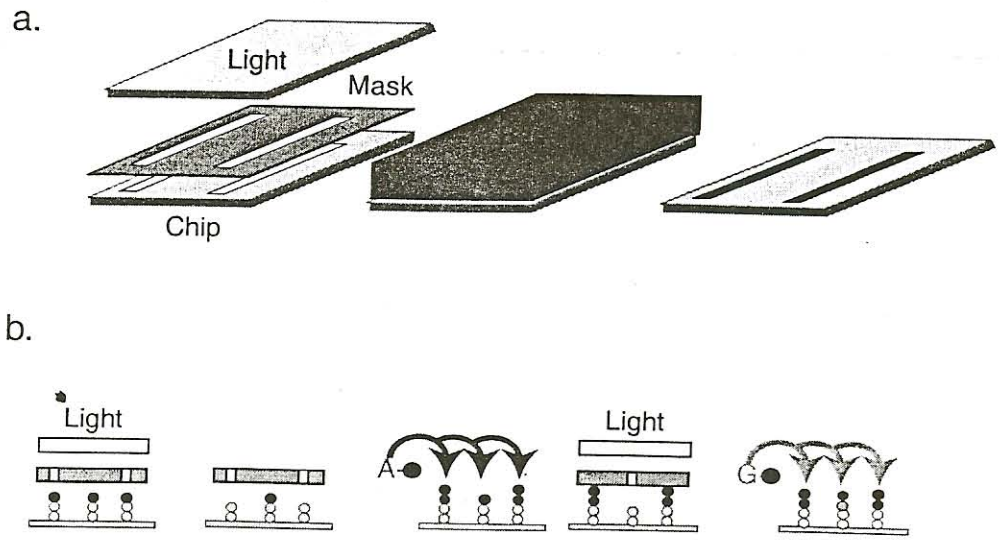


Figure 3. Schematic of the Affymetrix photolithography-based method of synthesizing oligonucleotide arrays. (a) A mask is used to selectively expose regions of the chip to light. In these regions, the light removes a protecting group. The chip is then flooded with a reactive nucleoside (red) resulting in the selective addition to the exposed region of the chip. (b) Cycles of synthesis and deprotection result in the removal of protecting groups (black circles) and the addition of bases (red and green)

∞ - - ∞

For every gene in an Affymetix Gene Chip, there are typically 20 perfect match oligonucleotides and 20 mismatch oligonucleotides. The use of multiple oligos increases the signal-to-noise ratio in the measurement and allows cross-hybridization among related genes to be detected and subsequently subtracted from the signal. A typical Gene Chip for yeast, (the yeast Ye6100 series) requires four arrays for the complete genome of 6100 open reading frames. With 40 oligos per gene, this means that there are ca. 60,000 olignucleotides on each array with a pitch (center to center distance) of 50 µm. The most recent generation of Gene Chips contains 400,000 oligonucleotides.

cDNA Microarrays

The second type of DNA array currently in use consists of spots of cDNA synthesized by PCR and arrayed on glass using a robotic spotter (Figure 2). Typically, 500 to 2000 bp fragments of each gene are amplified using PCR and then transferred to glass in small aliquots using a thin needle. A complete design for this type of spotting system has been published on the Web by Pat Brown at the Stanford Genome Center. Several organizations, including Stanford, Harvard University, Toronto University and Millennium Pharmaceuticals have built spotting robots from scratch, and six or more companies are marketing robots.

The published yeast arrays contain 6200 ORFS on a single slide using ca. 40µm spots arrayed with a pitch of 200 µm. Initial experiments suggest that the problem of cross-hybridization among related members of a gene family is not severe. The use of long fragments of DNA allows for high-stringency hybridization, apparently obviating the need for parallel (+) and (-) DNA controls as used in the Affymetrix Gene Chip system.

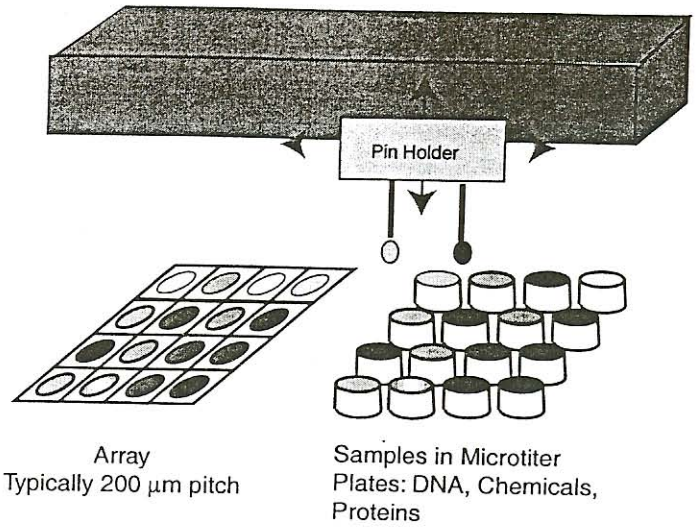


Figure 4. Schematic of a micro-arraying robot (spotter) transferring samples from a 96 or 384 well microtitier trays to a glass slide. The typical spot—to-spot distance (the pitch of the array) is 200 um and the feature size is usually about 50 um.

∞ - - ∞

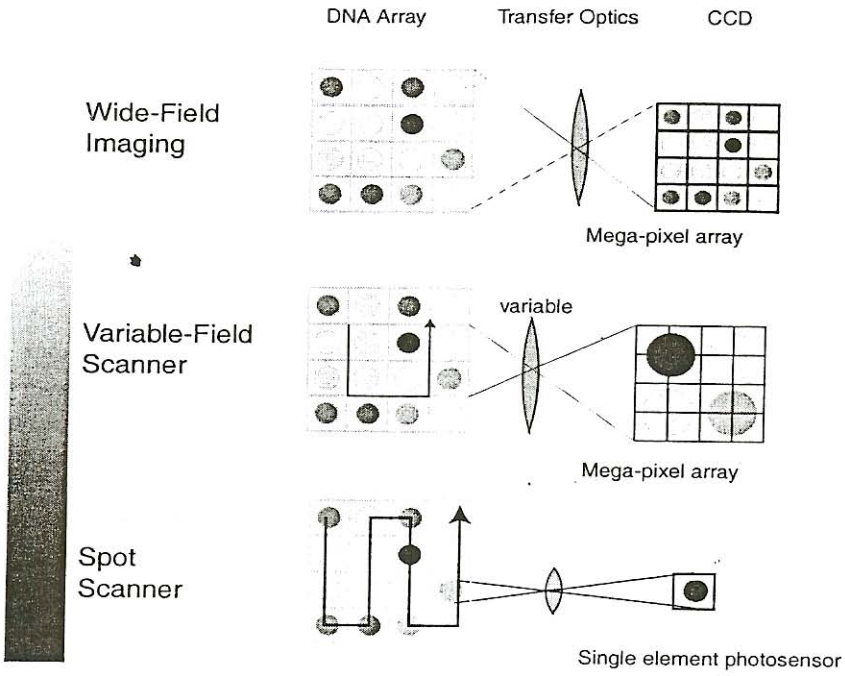
The obvious advantage of cDNA microarrays is that they are relatively inexpensive and can be customized without the need for complex and proprietary photolithography processes.

Scanning the Array

Following the hybridization of cDNA to arrays, the arrays are read using a fluorescence scanner. In the commonly used spot scanners (figure 5) Fluorescent probes are excited with a blue-green laser causing them to emit green-red light. The emitted light is collected by transfer optics and quantitated using a photomultiplier tube (PMT). Spot scanners apply a single reading element to the acquisition and measurement of photons arising from one element in an array and are basically fast, low resolution scanning confocal microscopes.

An alternative type of scanner, which we have been developing at MIT in collaboration with Applied Precision Inc. of Issaquah WA, uses wide-field imaging much like a microscope. In these scanners, the microarray is illuminated with white light from a metal halide bulb, the light is collected through a microscope objective and the array imaged on a CCD (charged coupled device) camera.

There are significant differences between the laser and wide field scanners. Our experiments suggest that scanning is a little appreciated but critical step in DNA array analysis. This is true because DNA micro-array data is very noisy and signal-limited



∞ - - ∞

Figure 5. Collecting data from arrays using various types of scanners. Existing instruments from HP and Molecular Dynamics are spot scanners that examine each element of an array in a serial fashion. Variable-field scanners use transfer optics to optimize the number of pixels used to acquire data from each element in an array. Wide-field imaging uses a wide-field lens to capture data from an entire array in one image.

∞ - - ∞

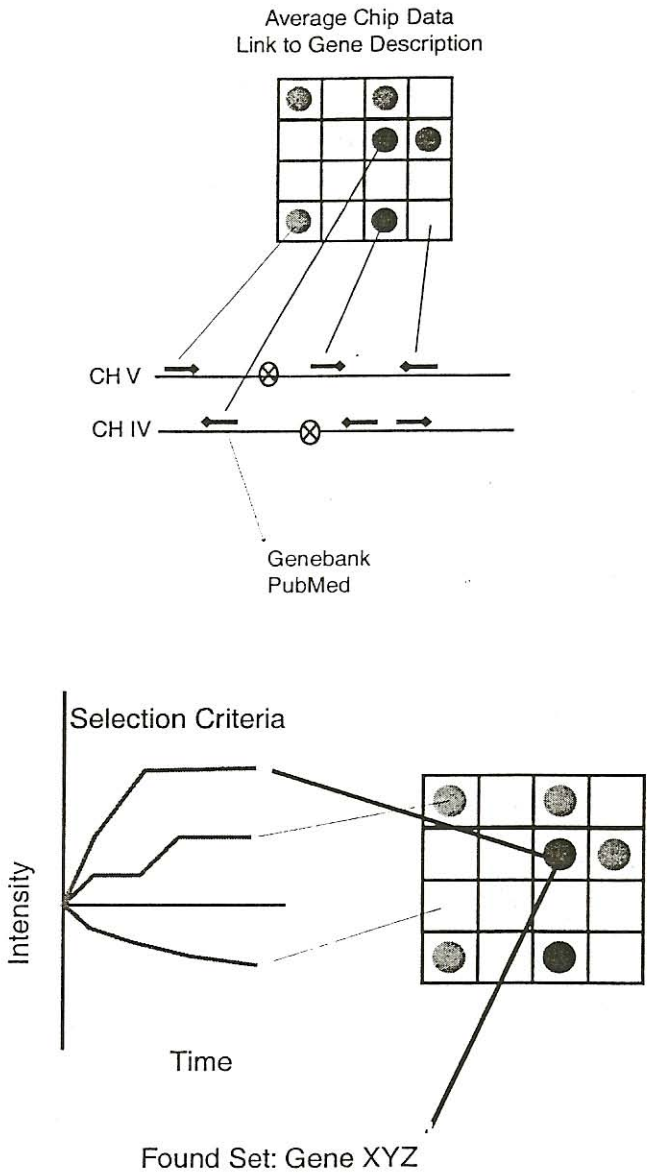


Figure 8 . The simplest type of data analysis - linking induction ratios to information about the genes represented in the array (above) and applying a simple database mining approach (below). In this representation, the selection criterion is for spots whose intensity increases more than three-fold. Spots that match this criterion represent the found set and can then be examined to see if there are interesting relationships among the members of the set. In real life, the selection criteria for the found set would be more complex.

∞ - - ∞

Perhaps the most challenging problem in the analysis of genome-wide microarray data is the development of suitable computational tools. The large amount of data produced by gene arrays must be processed to generate comprehensible and meaningful output.

The simplest type of analysis is to link information on the abundance of various mRNAs, as represented by the induction ratio at a particular spot, to information about the gene's identity, sequence and call-out in Genebank (Figure 8).

If each of the descriptors of a gene, and the measured mRNA levels under various conditions, are entered into a database, then simple database mining approaches can be used to analyze the expression data. These are generally Boolean queries in which ones ranks genes in terms of their induction ratios etc.

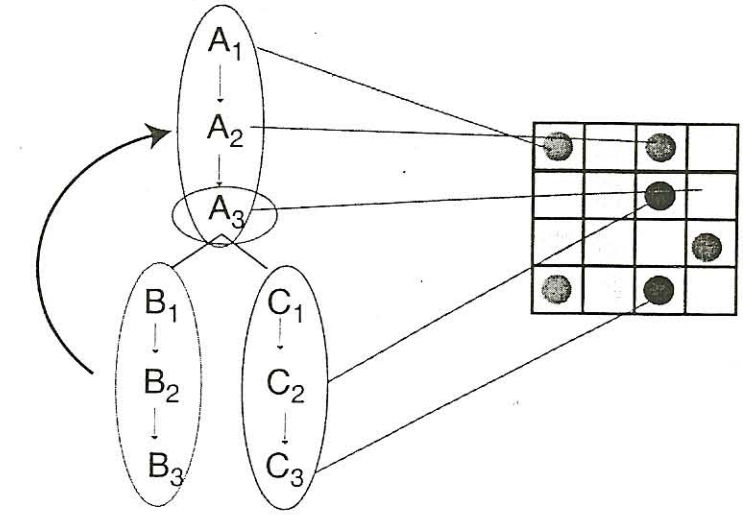
While these manipulation are simple in theory, they are complicated in practice by the fact that there is no universally accepted scheme for gathering together all of the information about a gene into tables with similarly named fields. Several labs have run into the problem that their data cannot be merged with data from other investigators because the descriptors of the data are dissimilar. One way to solve this is to agree on a universal standard for gathering and storing information about a gene (the YPD database from Proteome Inc. is a noteworthy example). A more likely alternative is for the data to be "self-describing" through the incorporation of metadata. Structured languages such as SGML and XML can be used to exchange both the data itself and the metadata that describe the data fields.

A more advanced way to analyze microarray data is to use pattern recognition techniques to find similarities. An early application of pattern recognition to microarray data is the clustering analysis of Eisen et al (PNAS 95, 14,863-68). Eisen et al. found that correlation coefficients were an effective measure of similarity among plots of abundance v. time for mRNAs from serum stimulated human fibroblasts (Figure 9). The correlation coefficient captured the extent to which genes were co-expressed without bogging down on differences in the overall level of expression. Unsupervised clustering of the correlation coefficients was then used to group genes whose pattern of induction (or repression) was similar. The clustering relationship was then plotted using a dendrogram in which the branch lengths reflected the degree of similarity (Figure 9).

The most significant finding from this analysis was that genes clustered , on the basis of expression, into discrete sets that appeared to reflect gene function. In almost every case, highly related genes were found close together in the dendogram. Another impressive example of this clustering was found when S. cerevisiae cultures were synchronized in the cell cycle and then released. Clustering linked together sets of genes known to be highly co-regulated (such as the histone genes) and discovered previously unrecognized similarities between other genes (figure 10).

However intriguing, the problem with these clustering and database mining approaches to array data is that they examine only correlation and do not take into account causation. For example, in a metabolic pathway, enzymes are ordered in substrate-product relationships. To capture this information , we need to map the data onto a representations of gene function. One large-scale approach, currently under way , is to determine the effects of mutating all 6200 yeast genes (Figure 11). It is planned that this "genetic footprint" be linked to microarray data.

The next step in the analysis of DNA microarray data will be map the expression information onto cellular pathways. For basic metabolic pathways this process is aided by our confidence in the overall order and structure of the pathway. However, for most other aspects of cellular physiology, including signal transduction, the situation is much less clear. In these cases, the representation of the pathway must be allowed to evolve as new data is collected. One approach is to design a database in which genome-wide data can be mapped to a dynamic representation of cellular physiology (Figure 12). Of particular interest would be cases in which the pathway information reveals details in the expression data that were not otherwise visible (or vice versa) .



● Map MicroArray data to static view of pathway

MicroArray Data → *Mapping* → Pathway

● Dynamically regenerate pathway based on MicroArray data

MicroArray Data ⇄ *Dynamic Linking* ⇄ Pathway

Figure 12. Analyzing expression data from arrays by linking it to information about cellular pathways. The key to this approach will be the development of methods for dynamic linking.

∞ - - ∞

Simplifying the Data

One difficulty in using microarray data is that it is very complex. In higher eucaryotes we anticipate that tens of thousands of pieces of data will be needed to fully describe a cell's transcriptional state. A rational basis is needed for reducing the variables to a more manageable number.

Clustering analysis has shown that many genes are tightly co-regulated. Thus, it should be possible to replace the database features representing each gene in a cluster with a composite feature that describes the entire cluster. More precisely, if we plot the gene array

data into an n-dimensional hyperspace with n equal to the number of genes, we see that the axes are not normal. However, it is possible to find a series of orthonormal eigenvectors to describe the gene array space. The space described by the eigenvectors will be identical to the original space but will have reduced dimensionality.

In image processing, similar decompositions of image space are possible but rarely useful. The feature vectors have no direct physical meaning when combined (how does one interpret an eigenvector made up of sharpness and contrast?). In the case of DNA arrays however, the set of features has a clear and unambiguous derivation: it is the set of all possible genes. Thus, an eigenvector made up of combinations of features that are genes might be interpretable as reflecting an underlying pattern of co-regulation.

Whether such a decomposition of image space is possible and useful remains to be seen.

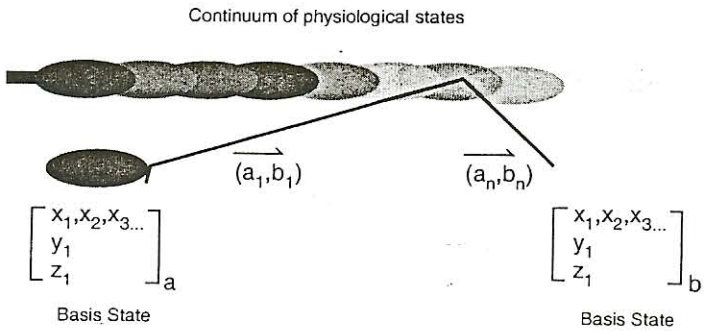


Figure 13 Decomposing the continuum of physiological states (as represented by the large number of possible of mRNA expression profiles) into a subset of discrete basis or eigenstates, that, when combined, compactly describe the original data.

∞ - - ∞

Summary

The steps of generating, hybridizing and analyzing a DNA microarray are tightly linked. It is important that the overall process be optimized as a whole. The better the DNA array the easier the scanning. The better the scanning, the more meaningful and reliable the data.

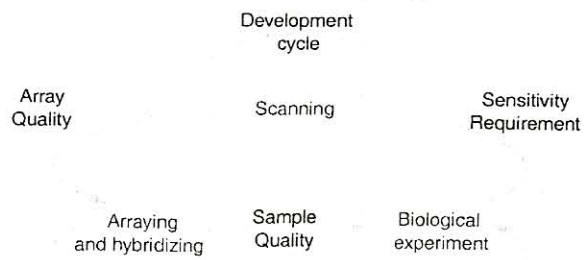


Figure 14. The coupling of arraying, scanning and experiment into a process that must be optimized overall. The goals of the optimization are high sensitivity and reliability and low variability.

∞ - - ∞

Further Reading

(Eisen et al., 1998; Iyer et al., 1999; Schena et al., 1996; Smith et al., 1996; Spellman et al., 1998)
Pat Brown's web site at the Stanford Genome Center ()

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A *95*, 14863-8.

Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson, J., Jr., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P. O. (1999). The transcriptional program in the response of human fibroblasts to serum [see comments]. Science *283*, 83-7.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A *93*, 10614-9.

Smith, V., Chou, K. N., Lashkari, D., Botstein, D., and Brown, P. O. (1996). Functional analysis of the genes of yeast chromosome V by genetic footprinting. Science *274*, 2069-74.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell *9*, 3273-97.