Technical Report for Phase II SBIR Spatially Defined Oligonucleotide Arrays

INTRODUCTION

The goal of the Human Genome Project is to sequence all 3 billion base pairs of the human genome. Progress in this has been rapid; GenBank® finished 1994 with 286 million bases of sequence and grew by 24% in the first quarter of 1995. The challenge to the scientific community is to understand the biological relevance of this genetic information. In most cases the sequence being generated for any single region of the genome represents the genotype of a single individual. A complete understanding of the function of specific genes and other regions of the genome and their role in human disease and development will only become apparent when the sequence of many more individuals is known.

Access to genetic information is ultimately limited by the ability to screen DNA sequence. Although the pioneering sequencing methods of Sanger et al. (15) and Maxam and Gilbert (11) have become standard in virtually all molecular biology laboratories, the basic protocols remain largely unchanged. The throughput of this sequencing technology is now becoming the rate-limiting step in both large-scale sequencing projects such as the Human Genome Project and the subsequent efforts to understand genetic diversity. This has inspired the development of advanced DNA sequencing technologies (9). Incremental improvements to Sanger sequencing have been made in DNA labeling and detection. High-speed electrophoresis methods using ultrathin gels or capillary arrays are now being more widely employed. However, these methods are throughput-limited by their sequential nature and the speed and resolution of separations. This limitation will become more pronounced as the need to rapidly screen newly discovered genes for biologically relevant polymorphisms increases.

An alternative to gel-based sequencing is to use high-density oligonucleotide probe arrays. Oligonucleotide probe arrays display specific oligonucleotide probes at precise locations in a high-density, information-rich format (5,4,12). The hybridization pattern of a fluorescently labeled nucleic acid target is used to gain primary structure information of the target. This format can be applied to a broad range of nucleic acid sequence analysis problems including pathogen identification, polymorphism detection, human identification, mRNA expression monitoring and de novo sequencing.

In this review, we briefly describe the method of light-directed chemical synthesis to create high-density arrays of oligonucleotide probes, the method of fluorescently labeling target nucleic acids for hybridization to the probe arrays, the detection of hybridized targets by epi-fluorescence confocal scanning and the data analysis procedures used to interpret the hybridization signals. To illustrate the use of specific highdensity oligonucleotide probe arrays, we describe their application to screening the reverse transcriptase (*rt*) and protease (pro) genes of HIV- I for polymorphisms and drug-resistance conferring mutations.

LIGHT-DIRECTED CHEMICAL SYNTHESIS

High-density oligonucleotide arrays are created using light-directed chemical synthesis. Light-directed chemical synthesis combines semiconductor-based photolithography and solid-phase chemical synthesis (3). To begin the process (Figure IA), linkers modified with photochemically removable protecting groups (12) are attached to a solid substrate. Light is directed through a photolithographic mask to specific areas of the synthesis surface, activating those areas for chemical coupling (Figure 1B). The first of a series of nucleosides (MeNPOC-dT in this instance) harboring a photo-labile protecting group at the 5' end (12) is incubated with the array, and chemical coupling occurs at those sites that have been illuminated in the preceding step (Figure IC). Next, light is directed to a different region of the substrate through a new mask (Figure 1D), and the chemical cycle (with MeNPOC-dC in this instance) is repeated (Figure IE). The process is repeated (Figure IF). Using the proper sequence of masks and chemical steps, a defined collection ofoligonucleotides can be constructed, each in a predefined position on the surface of the array.

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

HIGH-DENSITY PROBE ARRAYS

Since the synthesis areas are defined by the photolithographic process, highdensity arrays can be formed. Table I shows the relationship between the size of each synthesis site and the density of synthesis sites (a synthesis site being the region in which a homogeneous set of probes is synthesized). Production scale instruments are currently synthesizing arrays with 90-μm synthesis sites. This corresponds to approximately 20,000 synthesis sites in a standard 1.28-cm^2 synthesis area. Developmental instrumentation has demonstrated synthesis at a 10-μm resolution.

To support the production of increased numbers of arrays, instrumentation has been developed to synthesize multiple arrays on a single large substrate (wafer). After synthesis, wafers are diced to yield individual arrays. These arrays are packaged in individual injection-molded flow cells, making them easier to handle (see Figure 2). In addition, arrays synthesized by these methods have improved uniformity and quality and use significantly fewer resources to synthesize each array.

COMBINATORL&L SYNTHESIS OF A PROBE MATRIX

Light-directed synthesis provides a powerful and efficient method for generating random access molecular diversity in a spatially defined format. The location and composition of products depends on the pattern of illumination and the order of chemical coupling reagents (see Reference 5 for a complete description). In Figure 3 we illustrate the synthesis of all 256 4-mers in 16 chemical steps. Using the combinatorial synthesis strategy described in Figure 3, the set of all 4^k length k oligonucleotides (k-mers) can be generated in 4^k synthesis cycles. For example, the set of all 15-mers can be synthesized in 60 cycles. Using the current process, this synthesis can be completed in less than 10 h. Additional examples of the relationship between probe length, number of chemical synthesis steps and number of synthesis sites are given in Table 2.

Resolution	Synthesis Site Density
500 um	400 sites/cm^2
200 um	2500 sites/cm^2
100 um	10000 sites/cm^2
50 um	40000 sites/cm^2
20 um	250000 sites/cm^2
10 um	1000000 sites/cm^2

The same relationship between probe length and number of synthesis cycles applies to any set of probes. For example, since all 15-mers can be synthesized in 60 chemical steps, any random subset of 15-mers can therefore be synthesized in 60 or fewer steps. In particular, given any set of 15-mers, it is possible to define a set of at most 60 photolithographic masks to synthesize the set of 15-mers. The strategy for defining the masking patterns is given in Fodor et al. (5). More generally, any set of probes of length k or less (probes in different synthesis sites can be different lengths) can be synthesized in, at most, 4k chemical steps in any desired arrangement on the chip surface.

HYBRIDIZATION AND DETECTION

Hybridization of target nucleic acids to an oligonucleotide array yields sequence information. The nucleic acid to be interrogated, the target, is labeled with a fluorescent reporter group and incubated with the array. If the target nucleic acid has regions complementary to probes on the array, then the target will hybridize with those probes. Under a fixed set of hybridization conditions, e.g., target concentration, temperature, buffer and

salt concentration, and so on, the fraction of probes bound to targets will vary with the base composition of the probe and the extent of the target-probe match. In general, for a given length, probes with high GC content will hybridize more strongly than those with high NY content. Probes matching the target will hybridize more strongly than probes with mismatches, insertions and deletions.

Hybridization of the target to the array is detected by epi-fluorescence confocal scanning (Figure 4) (4). The array is inverted in a temperature-controlled flow cell. The target solution is introduced to the flow cell and allowed to hybridize with the probes on the surface of the array. The confocal system then scans the array, measuring fluorescent signal from target nucleic acid bound at the surface. This system provides strong background rejection from the unbound target, the glass and other parts of the system, a large dynamic range (1:1000), high resolution over a broad area and the ability to collect both kinetic and equilibrium data. The scanning resolution is chosen so that the synthesis sites of the array are oversampled, i.e., between 25 and 100) data points are collected from each synthesis site, corresponding to a 5 x 5 to 10 x 10 array of pixels. The collection of data forms an "image" of the array. Image processing software, GeneChip™ software (Affymetrix, Santa Clara, CA, USA), is used to segment the iniage into synthesis sites and integrate the data over each synthesis site. A robust benchtop version of this detection instrument, the GeneChip scanner (Affymetrix), has been developed, (Figure 5).

Probe Length	Chemical Steps	Number of Possible Probes
4	16	256
8	32	65 536
10	40	1 048 576
15	60	1 073741824

POLYMORPHISM SCREENING

The oligonucleotide arrays call be designed and used to rapidly and efficiently screen characterized genes polymorphisms. The genetic diversify of a population call be explored, and the relationships between genotype and phenotype for specific genes and groups of genes can be discovered. In addition, probes can be used as a simple diagnostic.

The significant instability of, internal probe-target mismatches, relative to perfect matches (3), is used to design arrays of probes capable of rapidly discriminating differences between nucleic acid targets. For example, to determine the identity of the base X in the target:

5' – GTATCAGCATXGCATCGTGC
we would use the following four probes:
3' AGTCGTAACGGTAGC
3' AGTCGTACCGGTAGC
3' AGTCGTAGCGGTAGC
3' AGTCGTATCGGTAGC

The probe with the highest intensity would indicate the identity of the unknown base. This concept can be extended to examine long nucleic acid targets and detect polymorphisms/mutations relative to a characterized consensus sequence. Given a consensus sequence, a set of four probes can be defined for each nucleotide in the target as described above. Thus, to screen 1000 nucleotides for polymorphisms/mutations would require 4000 probes. A 1.28-cm^2 array designed with 100 µm synthesis sites will have about 16000 probes and could screen 4 kb of sequence. The is approach has been applied to survey the *rt* and *pro* genes of HIV-1 for drug-resistance mutations and is described below.

HIV PROBE ARRAY

Currently, all of the approved HIV- I therapeutics are targeted against the RT enzyme. These include AZT (Zidovudine) ddI, ddC and d4T. In addition, many of the drugs under development by the pharmaceutical industry target either RT or the protease (PRO) enzyme. Significant debate has arisen concerning the efficacy of the approved and investigational drugs in inhibiting HIV-1 proliferation and the value of drug therapy to

patients. A significant portion of efficacy issues can be attributed to the acquisition of drug resistance HIV-1, governed by its ability to rapidly mutate while retaining pathogenicity.

A number of HIV-1 mutations have been cataloged that confer PRO or RT drug resistance to the approved antivirals. Depending upon the state of disease when therapy is initiated, AZT, the first approved HIV-1 antiviral may become ineffective after 3-12 months. Data from clinical specimens indicate that during this time period, a number of mutations arise in the *rt* gene that mediate this drug resistance (6-8). The first four characterized mutations were Asp 67 → Asn, Lys 70 → Arg, Thr → 215 Phe/Tyr and Lys 219 → Gin. When all four mutations occur together, the viral isolate is more than 100 times less sensitve to AZT (14). Most recently, several of the mutations associated with new therapeutics and the distribution of the mutations throughout the *rt* and *pro* genes, there is significant value in being able to rapidly and cost effectively screen these genes for mutations.

HIV ARRAY DESIGN

The array design described above was applied to the HIV-1 *rt* and *pro* genes. The sequences analyzed consisted of 1040 bases of the Clade B Consensus Sequence, including the last 18 bases of *gag*, all 297 bases of *pro* and the first 723 bases of *rt* (codons 1-241). The 1040 nucleotides in the array survey the positions at which all known resistance-conferring mutations are located. The array contains 18,495 93-x 95-μm synthesis sites (Figure 7).

The assay consists or amplification of clonal DNA, in vitro transcription and label incorporation, hybridization and scanning, (Figure 6). Nucleotide determinations were based on comparing the intensities of each set of four oligonucleotide probes interrogating each of the 1040 bases. Analysis tools are provided to view the intensity data in graphic and numeric formats, (Figure 8). Even with this simple first generation array design and base-calling algorithm, the method has proved to generate correct base calls with high confidence (unpublished). Future developments will focus on improving the design of the chip and the base-calling algorithm.

CONCLUSIONS

Light-directed chemical synthesis has been used to generate miniaturized, high-density arrays of oligonucleotide probes. Application-specific oligonucleotide probe array designs have been developed to rapidly screen known genes. These probe arrays are then used for parallel nucleic acid hybridization analysis, directly yielding polymorphism information from genomic DNA sequence. Dedicated instrumentation and software have been developed for array hybridization, fluorescent detection and data acquisition and analysis. The methods have been applied to detect resistance-conferring mutations in the *rt* and *pro* genes of HIV-I demonstrating their effectiveness.

Other applications of oligonucleotide probes arrays include, bacterial classification, mRNA expression monitoring and de novo sequencing. For bacterial identification, probes specific to individual species can be synthesized in a single array. Incubation of the array with labeled target nucleic acids from a single species generates a hybridization pattern that is unique to that species. A similar method can be used to identify the presence or absence of specific cDNAs in a cDNA library.

Hybridization can also be used to determine the sequence of unknown DNA (de novo sequencing, sequencing by hybridization [SBH]) (1,2,11). In SBH, the sequence of an unknown target nucleic acid is reconstructed from the hybridization data. SBH on probe arrays has been demonstrated in small-scale experiments (3,16). These experiments, in combination with numerous simulation studies, have shown that de novo sequencing is an application with significant potential. There are, however, important challenges to be met before it can be broadly implemented. These include generating signal from GC-rich and AT-rich probes in the same experiment, being able to effectively distinguish perfect matches from highly stable mismatches and resolving ambiguities due to repeated sequences. High-density oligonucleotide probe array technology shows significant promise in sequencing (3).

The Human Genome Project and related efforts have undertaken the formidable task of identifying and determining the sequence of all of the human genes, and it is only through more efficient access to genetic information that the true benefit of the Human Genome Project will be realized. High-densityoligonucleotide probes arrays should provide a basic platform for analyzing the genetic variation in the human genome enabling de novo

REFERENCES

1. Bains W. and G.C. Smith. 1988. A novel method for nucleic acid sequence determination.J. Theor. Biol. 135:303-307.
2. Drmanac R., 1. Labat, I. Brukner and R. Crkvenjakov. 1989. Sequencing of megabase plus DNA by hybridization: theory of the method. Genomics 4:114-128.
3. Fodor, S.P.A., R.J. Lipshutz and X. Huang. 1993. DNA sequencing by hybridization, p. 3-9. Proceedings of The Robert A. Welch Foundation 37th Conference on Chemical Research - 40 Years of the DNA Double Helix. Robert A. Welch Foundation, Houston.
4. Fodor, S.P.A., R. Rava, X.C. Huang, A.C., Pease, C.P. Holmes and C.L. Adams. 1993. Multiplexed biochemical assays with biological chips. Nature 364:555-556.
5. Fodor, S.P.A., J.L. Read, M.C. Pirrung, L. Stryer, A.T. Lu and D. Solas. 199 1. Light-directed, spatially addressable parallel chemical synthesis. Science 251:767-773.
6. Kozal, MJ and T.C. Merigan. 1993. HIV resistance to dideoxynucleoside inhibitors. Infect. Dis. Clin. Practice 2:247-253.
7. Land, S., C. McGavin, R. Lucas, and C. Birch. 1992. Incidence of zidovudine-resistant human irrinnuriodeficiency virus isolated from patients before, during, and after therapy. J. Infect. Dis. 166:1139-1142.
8. Larder, B.A. and S.D. Kemp. 1989. Multiple mutations in HIV-1 reverse transcriptase confer high-level resistance to zidovudine (AZT). Science 246:1155-1157.
9. Lipshutz R. and S.P.A. Fodor. 1994. Advanced DNA sequencing technologies. Curr. Opin. Struct. Biol. 4:376-380.
10. Lysov Y.P., V.L. Florentiev, A.A. Khorlin, K.V. Khrapko, V.V. Shik and A.D. Mirzabekov. 1988. DNA sequencing by hybridization with oligonucleotides. A novel method. Dokl. Acad. Sci. USSR 303:1508-1511.
11. Maxam A.M. and W. Gilbert. 1977. A new method for sequencing DNA. Proc. Nail. Acad. Sci. USA 74:560-664.
12. Pease, A.C., D. Solas, E.J. Sullivan, M.T. Cronin, C.P. Holmes and S.P.A. Fodor. 1993. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc. Nail. Acad. Sci. USA 91:5022-5026.
13. Richman, D.D. 1995. Protease uninhibited. Nature 374:494-495.
14. Richman, D.D., C.K. Shih, 1. Lowy, J. Rose, P. Prodanovich, S. Goff and J. Griffin. 1991. Human immunodeficiency virus type I mutants resistant to nonnucleoside inhibitors of reverse transcriptase arise in tissue culture. c Natl Acad Sci. USA 88:11241-11245
15. Sanger F., S. Nicklen and R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74:54635467.
16. Southern E., U. Maskos and R. Elder. 1992. Hydridization with olignucleotide arrays. Genomics 13:1008-1017.

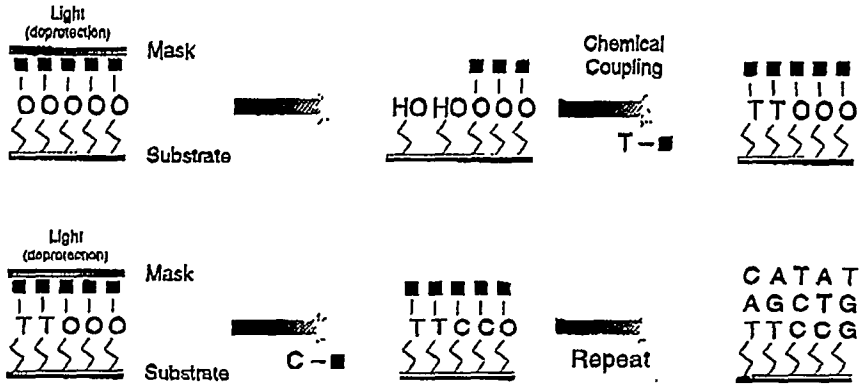

Figure1. High-density oligonucleotide probe array synthesis.

Figure 2. High-density oligonucleotide probe array package.

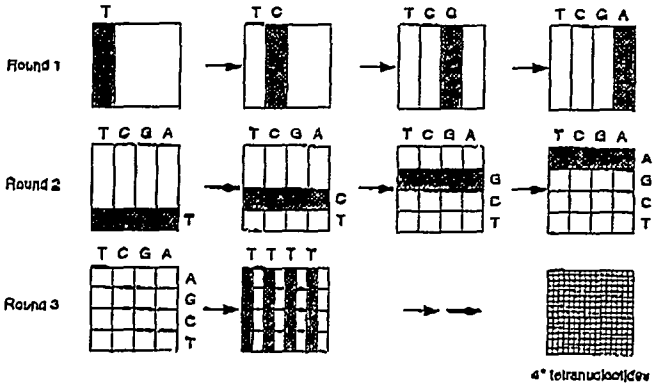

Figure 3. Combinatorial synthesis of all 256 4-mers. In round 1, mask 1 activates one-fourth of the substrate surface for coupling with the first of four nucleosides in the first round of synthesis (MeNPOC-dT). In cycle 2 of round 1, mask 2 activates a different quarter of the substrate for coupling with the second nucleoside (MeNPOC-dC). The process is continued to build four regions of mononucleotides. The masks of round 2 are perpendicular to those of round 1, and each synthesis cycle generates four new dinucleotides. The process continues through round 2 to form 16 dinucleotides. The masks of round 3 further subdivide the synthesis regions so that each coupling cycle generates 16 trimers. The subdivision of the substrate is continued through round 4 to form the tetranucleotides (256 possibilities).

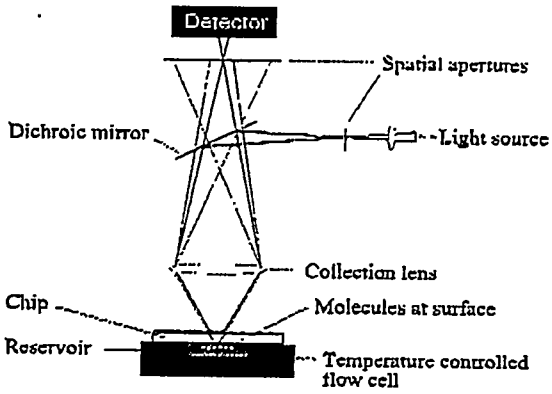

Figure 4. Detection format. Probe-target hybridization is detected by epi-fluorescence confocal scanning.

Figure 5. GeneChip scanner detection instrument.

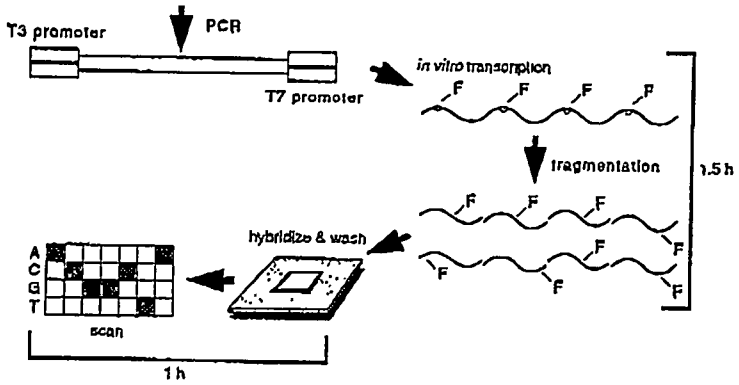

Figure 6. Sample preparation consists of amplifications, labeling, fragmentation, hybridization and scanning.

Figure 7. HIV-1 oligonucleotide array image.

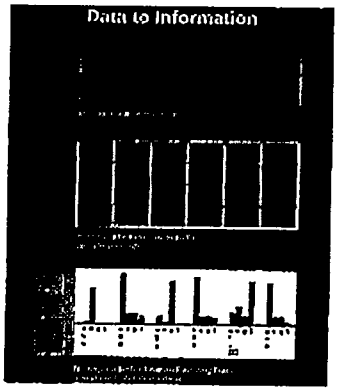

Figure 8. GeneChip software. Separate windows are available to display the initial image, the integrated intensities and the resulting base calls.