DNA VARIATION AND ITS ANALYSIS BY HYBRIDISATION TO OLIGONUCLEOTIDE ARRAYS.

EM SOUTHERN

Introduction

DNA analysis is now widely used in medical and forensic science. As an indicator for the presence of an infectious agent, as a means of detecting mutation associated with an inherited disease gene, or as an individual fingerprint, DNA analysis surpasses all other known tests. The reasons for this are severalfold. Unlike specific proteins, such as the blood group or leucocyte antigens, DNA is found in almost all cells. So that DNA analysis can be carried out on blood, semen and even spittle and hair. In recent years methods have been developed which allow analysis to be carried out on DNA from just a few cells, giving detection sensitivities not available for most molecules. Furthermore, the same reagents and methods can be used with any DNA sample, simplifying laboratory procedures. Finally individual variation in DNA is far greater than in any single phenotypic character.

DNA variability

Bacteria, viruses and simple eukaryotes such as yeasts have small compact genomes – most of the nuclear DNA codes for proteins. By contrast, in man it is estimated that less than 5% of the DNA codes for protein. The rest comprises a mixture of sequences, most of which appear to have no function relevant to fitness and hence are collectively known as "junk" DNA.

Much of non–coding DNA is found in the introns which interrupt the coding sequences. It has been speculated that these introns are the sites of recombination which, over evolutionary time, lead to new combinations of exons, providing new proteins with new functions, a process known as exon shuffling. Although they may be functional in this sense, there is much evidence to suggest that the base sequence of the DNA in the introns is not important. There is also a good deal of DNA between the genes. Apart from short sequences which bind proteins used to regulate gene activity there are no functions associated with this large fraction of the DNA. The non–coding DNA contains many sequences which are repeated in many copies throughout the genome. Among the most abundant are the so–called Alu sequences. These make up around 6% of the DNA by weight and in some regions of the genome occur once every 1–2Kb. The sequence of the Alu repeat is related to that of a small RNA which is present in all cell types. Many other dispersed repeats are related to transcribed sequences, suggesting that they were introduced into the genome by retrotransposition – a process in which RNA is copied to DNA which is then inserted into the genome. Although each family of repeated sequences has a characteristic consensus sequence, there is a great deal of variation between members, indicating that mutation is not selected out at the same rate as it is in the coding sequences. Other repeated sequences are made up of tandem repeats

of a short motif. These sequences are of two main types. The major satellites occurs as large blocks in particular regions of chromosome, most notably in the heterochromatic regions close to the centromeres. The basic unit repeated in the major satellites varies from two base pairs up to quite long sequences.

The minisatellites and microsatellites (also known as VNTRs) used in fingerprinting, are scattered around the genome. They comprise repeats of 2–30 base pairs. The feature which makes these sequences so useful as genetic markers and in fingerprinting is the sequences variability which is found in a high proportion [Jeffreys *et al.*, 1985]. Variability is caused by shrinkage and expansion by losing or gaining repeat units. Such sequences and their variations are easy to analyse. Sequence specific restriction endonucleases cut the DNA flanking the repeats, but not the repeat itself. Gel electrophoresis to separate the fragments by size, followed by molecular hybridisation to show the position of the bands reveals the pattern of fragment sizes characteristic of the individual. Such multilocus probes give a pattern rich in individual information. However, in some cases, it is preferable to use single locus probes, particularly for the analysis of small samples which must be amplified by the PCR to provide enough DNA for analysis. For this purpose it is necessary to isolate individual examples of a mini – or microsatellite and sequence the flanking DNA to obtain the information needed to make primers for the PCR. It is easy to find many individual examples by probing a collection of clones of genomic DNA with a mini – or microsatellite sequence. Several methods can be used to analyse the products after amplification by the PCR, but most commonly, for the micro and mini–satellites, gel electrophoresis is used to estimate the size. Many examples of the application of minisatellite analysis are cited in this symposium.

In addition to the micro – and minisatellites, the other major source of variation in the DNA is that due to point mutation – base substitution deletions and insertions. This kind of variability is also very great; between chromosome there is roughly one base difference for every 100–1000 bases, so that there will be roughly ten million differences between distantly related individuals. As compared with the micro and minisatellites, however, the number of types at any given site (the number of alleles) is relatively small [Botstein *et al.*, 1980]. Most often there are just two types in the population, and it is usually the case that one is more frequent than the other. Nevertheless, there is much interest in analysing point mutations because they are a major cause of genetic disease. Many cancers and inherited diseases are now known to be caused by point mutations in critical genes [Cooper and Krawczak, 1989]. Analysis is important first for finding the gene causing the disease. This may involved analysing many kilobases of DNA to compare wild type with mutant genes in many different individuals. Once the identity of the gene has been confirmed, methods for detecting mutation in individuals must be developed. The nature of the analysis used to detect mutation varies from one gene to another, as each has its own spectrum of mutation. Some diseases, such as sickle cell disease, are caused by only one mutation in one gene, a T to A transversion in the β-globin gene in this case. In others, very many different point mutations are

found. For example, around 300 point mutations have been found in the CFTR gene, causing cystic fibrosis [Zielenski *et al.*, 1991]. The need to analyse point mutation is obvious and many methods have been developed for this purpose [Cotton, 1993]. The methods can be divided into two main types. One type depends on gel electrophoresis to detect differences in sequence which in one way or another affect the mobility of fragments. The others depend on molecular hybridisation with oligonucleotide probes. An advantage of this approach is that it is possible to see both mutant and wild type in a single, simple analysis. The technique most commonly used is to amplify the relevant gene and fix the PCR product to a nylon membrane, which is then hybridised separately with the allele specific oligonucleotides (ASOs)[Conner *et al.*, 1983]. A homozygous wild type or mutants will test positive with only one probe, whereas a heterozygote will be positive with both. An advantage of the latter method is that large numbers of samples may be analysed together. However, each probe must be used separately, so that many analyses would be required to give an individual fingerprint. We and others have been developing a system in which multiple oligonucleotides can be used simultaneously. The oligonucleotides are made *in situ* or applied to a solid flat surface, of glass or plastic, so that they are covalently bound and available for molecular hybridisation. We have developed three radically different types of array. They are all made by the same basic procedure. We first apply a twenty atom linker to the surface of the glass plate. Oligonucleotides are synthesised using phosphoramidite chemistry, the first residue being coupled to a primary hydrxyl group at the end of the linker.

We normally apply the coupling reagents through channels formed by clamping a template to the surface of the glass plate. As the reagents pass through the channels coupling takes place, creating rectangular patches or lines of coupled bases on the surface. In one kind of array we make lines, each line comprising an allele specific oligonucleotide. These arrays can be used to test several samples simultaneously by applying the samples as lines orthogonal to the lines of oligonucleotides [Maskos and Southern, 1993]. Hybridisation takes place where the lines cross. This method can be scaled up to analyse many different alleles – up to 200 on a 200 mm plate – using multiple samples. For forensic purposes, a twenty allele system, each with two variants, would provide $2^{20} \simeq 10^6$ different genotypes. However, the discriminating power of the system would depend on the frequency of each allele in the population.

A second type of array can be used to "scan" a length, reading a short window of bases at each point in the sequence [Southern *et al.*, in preparation]. Such arrays are made by applying the first base to be synthesised (the complement of the first base in the target sequence) in a circular patch on the glass surface. The template is then offset by a fraction, say one tenth, of a diameter and the second base applied. The process is repeated. After 10 steps, the first ten bases of the sequence are represented as a decanucleotide. A further step produces a decamer representing bases 2–11 and so on to the other end of the sequence. Hybridisation of the target sequence then shows every position down the sequence. Variants are seen as lost signal in set of

oligonucleotides over the region of the sequence difference. Such arrays could be used in forensic applications by analysing sequences where there are may base variants over a short distance, such as the mitochondrial D–loop or the HLA region. We are using these arrays to develop methods for analysing disease genes which show a high frequency of different mutation; for example the more than 300 known mutations causing cystic fibrosis.

A third types of array provides a tool that can be used to fingerprint any nucleic acid, whether its sequences is known or not. Oligonucleotides of a given length are made as a complete set on the surface of the array [Southern, Maskos and Elder, 1992].

Ideally, the length of oligonucleotides should be octamers or longer, to achieve reasonable yields of hybridisation at room temperature. However, a complete set of octanucleotides is 65,536 in number. Smaller sets can provide a great deal of information. We have made arrays of the type $N_3X_2N_3$ where N represents a defined base and X a mixture of all four. This set contains all 65,536 octanucleotides sequences but the size of the array is reduced to 4096 cells [Bains and Smith, 1988; Bains, 1991]. An array that could be particularly appropriate for analysing individual "fingerprints" comprises all sequences of the formula N_3CGN_3. CG is a rare doublet in mammalian DNA, but is known to have a high rate of mutation to TG. The reason for this is that the C in CG is often methylated and 5-methyl cytosine deaminates to T, which is not recognised as an abnormal base by the repair systems. The "general" arrays analyse every constituent oligonucleotide of the target sequence, whereas arrays of the type N_3CGN_3 will "see" only those oligonucleotides conforming to this pattern. However, this will permit longer target sequences to be analysed.

After hybridisation with a radioactive or fluorescently tagged target sequence, an image is collected, using the PhosphorImager for radioactivity or a CCD camera for fluorescence. Both devices produce a digitised image. Such images can be displayed and manipulated on the VDU screen, but most importantly for the present discussion, they can be compared by numerical methods. Thus the image produced by one sequence can be subtracted from that produced by another. Oligonucleotides held in common in the two sequences disappear leaving only those which are different. Thus two sequences each one kilobase in length would give an image with almost a thousand "spots". If they differed by a single base change, the difference would affect only eight spots in an octamere array.

Model experiments using small arrays of octanucleotides have shown the power of this approach: subtraction did indeed simplify the images and produce a pattern that was easy to interpret. A most important feature of the overall approach was that the final assignment could be assessed quantitatively and the probability of its correctness readily estimated [Southern, Maskos and Elder, 1992].

In conclusion, hybridisation to oligonucleotides offers a range of opportunities in the field of forensic analysis of DNA sequences. It has sensitive discrimination of point

mutations; it is readily automated in a way that enables many samples and loci to be analysed simultaneously; powerful statistical methods can be used to assign confidence limits to the data.

References

1. BAINS, W. (1991). Hybridization methods for DNA sequencing. *Genomics* **11:** 294–301.

2. BAINS, W., AND SMITH, G. C. (1988). A novel method for nucleic acid sequence determination. *J. theor. Biol.* **135:** 303–307.

3. BOTSTEIN, D., WHITE, R. L., SKOLNICK, M., AND DAVIS, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32:** 314–331.

4. CONNER, B. J., REYES, A. A., MORIN, C., ITAKURA, K., TEPLITZ, R. L., AND WALLACE, R. B. (1983). Detection of sickle cell β^s globin allele by hybridization with synthetic oligonucleotides. *Proc. Natl. Acad. Sci. USA* **80:** 278–282.

5. COOPER, D. N., AND KRAWCZAK, M. (1989). The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. *Hum. Genet.* **85:** 55–74.

6. COTTON, RG, (1993)
Current methods of mutation detection. *Mutation Research* **285**: 125-144.

7. DRMANAC, R., LABAT, I., BRUKNER, I., AND CRKVENJAKOV, R. (1989). Sequencing of megabase plus DNA by hybridization: Theory of the method. *Genomics* **4:** 114–128.

8. JEFFREYS, A. J., WILSON, V., AND THEIN, S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature* **314:** 67–71.

9. U. MASKOS AND E.M. SOUTHERN, (1993)
A novel method for the analysis of multiple sequence variants by hybridisation to oligonucleotides. *Nucleic Acids Research*, **21**: 2267–2268.

10. U. MASKOS AND E.M. SOUTHERN, (1993)
A novel method for parallel analysis of multiple mutations in multiple samples. *Nucleic Acids Research*, **21**: 2269–2270.

11. SAIKI, R. K., WALSH, P. S., LEVENSON, C. H., AND ERLICH, H. A. (1989). Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *Proc. Natl. Acad. Sci. USA* **86:** 6230–6234.

12. E.M. SOUTHERN, U. MASKOS AND J.K. ELDER, (1992)
Analysis of Nucleic Acid Sequences by Hybridization to Arrays of Oligonucleotides: Evaluation using Experimental Models.
Genomics **20**: 1675–1678.

13. ZIELENSKI, J., ROZMAHEL, R., BOZON, D., KEREM, B.-S., GRZELCZAK, Z., RIORDAN, J. R., ROMMENS, J., AND TSUI, L.-C. (1991). Genomic DNA sequence of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene. *Genomics* **10:** 214–228.