GENE DETECTION WITH SYNTHETIC OLIGONUCLEOTIDE SEQUENCES

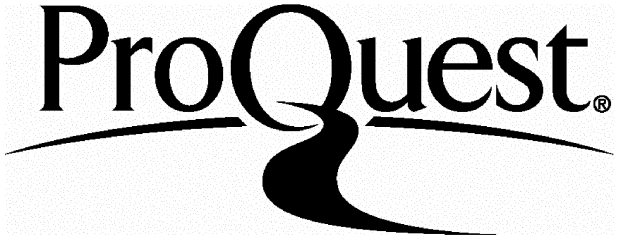
BY

JOHN BERNARD VINCENT

A thesis submitted for the degree of PhD at the University of London

The work reported in this thesis was carried out whilst registered at the University College London Medical School, University of London.

June 1994



ABSTRACT

This thesis describes experiments to test the hypothesis that it is possible to detect coding regions in genomic DNA by using short synthetic oligonucleotides. Two types of sequences were targeted. The first consisted of sequences complementary to sites for rare-cutter restriction enzymes, which are often clustered in the CpG rich islands located adjacent to, or within, coding regions. The second were sequences with complementarity to consensus sites for control and regulation within or around coding regions, such as splice sites and transcription factor binding sites.

Southern hybridization experiments were first carried out to test the hypothesis that short oligonucleotides, based on G/C rich regions and on consensus splice site sequences, could be used as hybridization probes to detect cosmids or fragments of cosmids containing CpG islands or splice site junctions. The cosmid vector pWE15 which contains two NotI sites and several clones containing the genes for human proteolipid protein, calcitonin/calcitonin gene related peptide-α (CGRP), glutathione-S-

transferase and NADH-ubiquinone oxidoreductase were used as model systems for testing this hypothesis.

Secondly, the polymerase chain reaction was used to test the hypothesis that short oligonucleotides based on rare cutter sites could be used as specific PCR primers using subclones of the cosmid vector pWE15 and phagemid pSL1180 as model systems, or in a less specific manner to amplify DNA in well characterised clones such as those containing the PLP, calcitonin/CGRP, glutathione-S-transferase and NADH-ubiquinone oxidoreductase genes. In addition the hypothesis that this method could also be used on total human genomic DNA, so that clones enriched for either CpG islands or for coding regions could be obtained, using rare cutter site, splice site and transcription factor site oligonucleotides, was also tested. Several methods including a "TA" cloning strategy were employed in order to generate mini-libraries of the amplification product for subsequent evaluation. Degenerate oligonucleotides with just their 3' ends based on the splice site and translation start site consensus sequences and with cloning sites at their 5' ends were also used on the model clones to test whether coding regions could be identified.

Thirdly, experiments were carried out to demonstrate that short oligonucleotides based on rare cutter sites could be used as PCR primers

for sequencing directly into CpG islands/coding regions in cloned DNA. Subclones of the four genes mentioned above were used as models to test this. Finally, experiments were carried out to test the hypothesis that the ligation of primer/linkers to rare-cutter restricted cosmids followed by direct PCR-sequencing could be used to obtain direct sequence from putative CpG islands in cloned genes. Several cosmid clones including the glutathione-S-transferase and NADH-ubiquinone oxidoreductase genes were used for this.

The experiments showed that some of the hypotheses concerning G/C rich sequence detection could be confirmed. When genomic DNA was used as template, G/C rich oligonucleotides as short as 8-mers could prime PCR amplification and enable "TA" clones to be produced which were enriched 66-fold for CpG rich sequences. In addition, conditions necessary for direct and specific amplification using G/C rich oligonucleotides as short as 7-mers with vector constructs as model target DNA were identified. However, PCR using G/C rich oligonucleotides was not capable of identifying CpG islands within cosmid clones. G/C rich 8-mer oligonucleotides may also be used in certain situations for directed sequencing within cloned genes and could thus be used as random or directed primers in a large volume sequencing project. Primer/linkers for rare-cutter restriction sites may also be used for sequencing into CpG islands within cosmids. Oligonucleotides

that have their 3' ends complementary to the splice site consensus sequence can also prime amplification (Degenerate Oligonucleotide Primed-PCR) from some splice sites in some cloned genes, although with low success rate.

In conclusion, the development of methods to enrich for CpG islands in genomic DNA was successful, but identifying CpG islands in cosmid clones was not. However a degree of success was achieved in the direct sequencing of CpG islands within clones. The work with splice site sequences was less successful, and it must be concluded that other methods should be employed for the identification of coding regions within cloned DNA.

ACKNOWLEDGEMENTS

I wish firstly to thank all those who helped with the benchwork involved in this thesis, in particular members of Georg Melmer's group at the Molekulare Genetike laboratory at the University of Bochum who sequenced some of the TA clones, also Sharon Spencer from Molecular Toxicology for providing me with the cosmid cosGSTrp7 and its subclones and to Paul Brickell for allowing me to use the cosmids cosCT1 and cosCT2 and their subclones. Much credit is also due to Robin Sherrington for his suggestions and for reading this thesis. I would also like to thank Sanober Shaikh for giving me the encouragement I needed to write this thesis. Above all I wish to thank Georg Melmer for starting me off on the project and to Hugh Gurling for seeing it through and for all their help and advice.

I would like to dedicate this work to the memory of my grandfather, Ted Hughes, who died on 1st May 1992.

ABBREVIATIONS

amp: ampicillin
ATP: adenosine triphosphate
bp: base pair
cDNA: complementary DNA
cM: centimorgan
CpG: cytosine-phosphate-guanine dinucleotide
dATP: deoxyribosyladenine 5'triphosphate
dCTP: deoxyribosylcytosine 5'triphosphate
dGTP: deoxyribosylguanine 5'triphosphate
dTTP: deoxyribosylthymine 5'triphosphate
ddATP: dideoxyribosyladenine 5'triphosphate
ddCTP: dideoxyribosylcytosine 5'triphosphate
ddGTP: dideoxyribosylguanine 5'triphosphate
ddTTP: dideoxyribosylthymine 5'triphosphate
DMSO: dimethyl sulphoxide
DNA: deoxyribonucleic acid
DOP-PCR: degenerate oligonucleotide primed PCR
dsDNA: double stranded DNA
DTT: dithiothreitol
EDTA: ethylenediamine tetra acetic acid
EST: expressed sequence tag
HTF islands: HpaII tiny fragments islands
IMS: industrial methylated spirits
IPA: isopropyl alcohol
IPTG: isopropyl-beta-D-thiogalactopyranoside
KAN: kanomycin
Kb: kilobase
LB: Luria broth
LMP: low melting point
LM-PCR: ligase mediated PCR
LUA: Luria agar

mA: milliamps

Mb: megabase

mins: minutes

mg: milligrams

ml: millilitres

mM: millimolar

mmol: millimoles

mRNA: messenger RNA

μg: micrograms

μl: microlitres

μM: micromolar

μmol: micromoles

ng: nanograms

nmol: nanomoles

oligo: oligodeoxyribonucleotide

PCR: polymerase chain reaction

RNA: ribonucleic acid

RNase: ribonuclease

SDS: sodium dodecyl sulphate

SSC: standard sodium citrate

SSPE: standard sodium phosphate EDTA

STS: sequence tagged site

TAE buffer: Tris-acetate-EDTA buffer

TBE buffer: Tris-borate-EDTA buffer

Tris: tris(hydroxymethyl)aminomethane

TMAC: tetramethylammonium chloride

V: volts

W: Watts

X-GAL: 5-bromo-4-chloro-3-indolyl-beta-D-galactoside

STANDARD CODE FOR NUCLEIC ACIDS:

A: Adenine

C: Cytosine

G: Guanine

T: Thymine

U: Uracil

I: Inosine

Y: pYrimidine (C or T)

R: puRine (G or A)

W: Weak (A or T)

S: Strong (G or C)

M: aMino (A or C)

K: Keto (G or T)

B: not A (C,G or T)

D: not C (A,G or T)

H: not G (A,C or T)

V: not T or U (A,C or G)

N or X: aNy (A,C,G or T)

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

CHAPTER 1: GENERAL INTRODUCTION

1.1 REVERSE GENETICS

The detection and cloning of genes responsible for specific heritable disorders requires either the isolation and characterization of the protein encoded by the mutant gene, or the direct cloning and sequencing of the gene based on linkage data. Once sequence information is available for the disease protein, synthetic oligonucleotides can be constructed based on codon usage for the amino acids and these synthetic oligonucleotides can be used to screen genomic or cDNA libraries for the disease gene. Alternatively monoclonal antibodies raised against the (partially) purified disease protein can be used to screen expression libraries, for example λgt11 cDNA libraries in *Escherichia coli* (Glover, 1985) or eukaryotic cell expression libraries (Kuhn et al, 1984; Littman et al, 1985).

Such strategies have proved successful, particularly for heritable metabolic disorders, where the error can be pinpointed to a particular step in a metabolic chain, either by the excess or deficiency of a certain metabolite or product. However for many genetic disorders the aberrant gene product is unknown. In these cases it is not possible to identify the responsible gene by "forward" genetic strategies. An alternative strategy is undertaken for the identification of such genes, called "reverse genetics". Reverse genetics, or "positional cloning", relies on genetic linkage and association analysis to localize a potential gene responsible for a definite phenotype to a particular region of the chromosome.

Linkage analysis determines whether a disease locus is positioned near a marker sequence of known chromosomal location, by analyzing the degree of meiotic recombination that occurs

between the two. Since the greater the distance between two loci, the greater the number of recombination events between them and thus the degree of recombination can be used as a measure of genetic distance. The unit used to quantify recombination is Θ, which is the fraction of recombination. A recombination fraction of $\Theta = 1\%$ is equivalent to approximately one centimorgan (1cM), which is the unit of genetic distance. If two loci are on different chromosomes and therefore not linked, the recombination fraction is 50%. Genetic distance and physical distance are not directly proportional, as frequency of recombination for a stretch of DNA can vary depending on chromosomal position and on the sex of the individual (Drayna et al, 1984; Hartley et al, 1984). However on average 1cM is equivalent to $1x10^6$ base pairs (bp). Distances less than this can be computed using a "coefficient of disequilibrium" based on linkage disequilibrium data.

For linkage analysis polymorphic markers are of considerable value. The more polymorphic the marker the more linkage information can be extracted, resulting in a more accurate and detailed map. Restriction fragment length polymorphisms (RFLPs) have been widely used in linkage analysis, as they are easily detected by digestion with a specific restriction endonuclease. This produces fragments of different lengths which can be detected by Southern hybridization analysis. More recently polymorphic repeat elements within DNA have become the preferred "currency" of markers for linkage analysis. In particular, markers containing polymorphic dinucleotide repeats have become very widespread across the genome (Weissenbach et al, 1992; NIH/CEPH Collaborative Mapping Group, 1992). Such markers are often very polymorphic and are easily identifiable using PCR-amplification based techniques (Weber and May, 1989; Litt and Luty, 1989).

There are several different approaches to linkage analysis. One method is the "candidate" gene approach, which relies on the availability of cloned "candidate" genes, which may

conceivably play a role in the disease. A polymorphism within the candidate gene can be analysed, either for linkage by following the polymorphism through pedigrees which have a high prevalence of the disease and calculating the frequency of recombination between the candidate gene and the disease locus, or for association between genotype and disease phenotype, comparing a "population" of affected individuals with a normal "population".

Positional cloning of a gene can follow a number of strategies. However, linkage between the disease gene and random markers is usually established first. This can be a laborious process, because there are 24 different chromosomes containing roughly three billion bases. However, in many cases the search for disease genes through positional cloning has been aided by the discovery of some rare individuals with the disease phenotype who possess chromosomal abnormalities. These cytogenetic aberrations may result in disrupted chromosomes receiving an extra portion of DNA (translocation), a deletion of DNA or a reduplication. The detection of these abnormalities often provides the first clue as to the whereabouts of the disease gene, directing attention towards a particular part of a chromosome. The identification of the genes for Duchenne muscular dystrophy, chronic granulomatous disease, retinoblastoma and others have been greatly assisted by the discovery of such cytogenetic abnormalities (Worton et al, 1984; Francke, 1984; Baehner et al, 1986; Cavanee et al, 1983). Positional cloning techniques can then be employed to detect a disrupted gene directly or further families can be studied by linkage of the disease phenotype to a particular marker or markers within this region. Once linkage has been established, allelic association data to establish linkage disequilibrium between mutation and marker can also be used to pinpoint the gene locus. Successive clones linked to a greater or lesser extent to the disease mutation are obtained by chromosome "walking" and "jumping" (Rommens et al, 1989). Usually a large amount of the DNA surrounding the marker needs to be cloned before the gene can be localized. The isolation and characterization of the gene responsible for cystic fibrosis has been achieved

entirely by positional cloning methods (Riordan et al, 1989), thus demonstrating the power of the technique. Searching for genes by use of positional cloning, though powerful, is a labour-intensive and lengthy process. For instance, whilst the gene for Huntington's chorea was localized by linkage analysis using RFLP's to chromosome 4 using the polymorphic marker G8 (D4S10) (Gusella et al, 1983) the gene itself took nearly a decade to be isolated despite extensive collaborative efforts (Pritchard et al, 1991). Some regions of the genome have been found to be extremely difficult to clone, as was the case in the search for the cystic fibrosis gene (Rommens et al, 1989). Recent improvements in the mapping of the human genome and the availability of clones covering large portions of the genome should greatly assist the search for disease loci.

1.2 THE HUMAN GENOME PROJECT

The first proponent of a large scale collaborative human genome initiative with the eventual aim of sequencing the entire human genome was the cancer researcher Renato Dulbecco in 1986, in an editorial in Science. He favoured the view that, rather than trying to identify individual genes responsible for cancer in a piecemeal approach, it would be far better to sequence the entire genome first. In this way all potential cancer genes could be identified and studied. The idea quickly caught on and was soon taken up by the U.S. Department of Energy and the U.S.A. National Institute of Health. In October 1990 the Human Genome Project implemented a set of five year goals. These goals included the following:-

1. The completion of a contiguous human genetic map, with markers spaced on average between 2 and 5 cM apart and identified by a sequence tagged site (STS).
2. The generation of physical maps of all chromosomes using STS's at approximately 100kb intervals.
3. The generation of contiguous clones each covering over 2Mb for much of the genome

4. The improvement of current methods of DNA sequencing and the development of new methods of sequencing, costing below $0.50 per base.

5. The sequencing of over 10Mb of human DNA in large stretches.

While one school of thought argues for the complete sequencing of the human genome, another believes that the project should concentrate foremost on the sequencing of coding regions, which represent only 3% of the genome but contain a major proportion of the information. Since the prediction of coding regions from genomic sequences is not always feasible, particularly when small exons are involved (Fickett, 1982), the sequencing of expressed regions via cDNAs is of unquestionable importance. Sequenced tagged sites (STSs) are at present one of the main "currencies" of genome mapping (Olsen et al, 1989). They are short unique sequences from clones which have been physically or genetically mapped, which can be amplified by PCR using unique flanking primers. Short sequences from cloned cDNA species, or expressed tagged sites (ETSs), can also be used as markers for the genome, serving the same purpose as STSs, but with the added bonus that such markers also indicate the presence of an expressed gene (Adams et al, 1991).

1.3 MAPPING THE HUMAN GENOME

1.3.1 MAPPING BY LINKAGE

The development of a detailed genetic linkage map of the human genome will assist the localization of many disease genes. At a recent gene mapping conference more than 2500 polymorphic markers and approximately 1800 expressed sequences were mapped onto the genome (Human Gene Mapping 11, 1991). More recently the Genethon project in France has developed a new linkage map of the genome based entirely on newly characterized dinucleotide repeats which are highly polymorphic and thus highly informative (Weissenbach et al, 1992). This initiative has added at least 2000 new markers for the genome and greatly

increased the power available for linkage analysis of genetic diseases. A linkage map of the genome with an average resolution of 3-5cM, as advocated by the Human Genome Project, is now close to completion.

1.3.2 PHYSICAL MAPPING

Once a disease gene has been localized to a particular region by linkage analysis, the disease gene has to be pinpointed. The development of fine resolution physical maps of the genome will complement the genetic linkage maps and aid the isolation of the disease gene. The densest map possible would be the entire genomic sequence.

Physical maps can be either cytogenetic or molecular. Cytogenetic maps order loci along chromosomes according to their positions relative to visible banding patterns, as determined by either *in situ* hybridization or by using somatic cell hybrids (Eubanks et al, 1992) induced by either cell fusion or by radiation. With the technology (Cox et al, 1990) high doses of x-rays are used to break human chromosomes of interest into fragments prior to cell fusion with rodent cells. The further apart two markers are on a stretch of chromosome, the higher the chances that the radiation will break the DNA between them. Radiation hybrid mapping provides a method for ordering markers at a resolution of 0.5Mb.

Another, complementary strategy envisages establishing a "contig" map of the entire genome. This is a collection of contiguous or overlapping clones. The use of yeast artificial chromosomes (YACs) (Burke et al, 1987), which have a much greater cloning capacity than the cosmid, along with pulsed field electrophoresis for separating large fragments of DNA, has made this strategy much more attractive. The usefulness of a contig map of YAC and cosmid clones would be enhanced by the addition of a small amount of sequence information

from each clone using the aforementioned sequenced tagged sites and expressed tagged sites, as proposed by Olsen et al (1989).

The pulsed field gel electrophoresis technique was originally developed by Schwarz and Cantor (1984) and with modifications (Carle and Olsen, 1984) allows the separation and resolution of DNA fragments several megabases long. This means that it is possible to construct long range restriction maps using rare cutter restriction enzymes, on to which sequences or clones of interest can be mapped. Portions of the genome have now been mapped in this way (Smith et al, 1987). The mapping of rare cutter restriction sites is also of great value in the search for genes and is discussed later on.

The availability of highly informative genetic markers spanning the genome with a high density should make the detection of linkage to disease genes much less problematic. The availability of ordered clones covering the entire genome with landmarks such as STS's, ETS's and rare cutter restriction sites mapped on to such clones will bypass much of the need to "chromosome walk" to the gene and thus greatly facilitate the identification of disease genes, once linkage has been established.

1.4 STANDARD PROCEDURES FOR GENE DETECTION

The most commonly used strategy for the identification of coding sequences involves screening short segments of cloned DNA for sequences that are conserved through evolution. This is usually achieved by performing inter-species cross hybridizations of the cloned fragments to so-called "zoo blots", which are Southern blots containing restriction digested genomic DNA from a wide range of evolutionary diverse organisms. Many single copy sequences that are conserved between species appear to represent genes and can be detected

in this way (Rommens et al, 1989; Monaco et al, 1986; Page at al, 1987; Call et al, 1990), whilst low copy number sequences found between genes tend not to be conserved. Cloned fragments shown to contain conserved sequences can then be used to isolate cDNA clones by hybridization to cDNA libraries. The cDNA clones, which are reversed transcribed DNA copies of mRNA species, can then be sequenced and analyzed for open reading frames.

There are a number of drawbacks associated with the cDNA approach to gene detection:-

1. The ability to detect cDNA species within a library depends on whether the gene is expressed in the tissue used to generate the library and the abundance of that particular mRNA species within the tissue. Housekeeping genes are transcribed at a high level and so a considerable proportion of cDNAs in a library represent such genes. The expression status of a non-housekeeping gene will depend not only upon tissue and cell type used for constructing the library, but on the developmental stage of the tissue. It is possible to construct a cDNA library from different tissues and to weight the library in favour of the less abundant mRNA species by subtractive hybridization with cDNAs from another library (Schmid et al, 1987; Fargnoli et al, 1990; Duguid and Dinauer, 1990; Travis and Sutcliffe, 1988) and by normalization, which results in all sequences in the library being present in approximately equal proportions (Patanjali et al,1991; Koch, 1990). However this involves much painstaking work.

2. Genes which belong to large gene families may not be easily distinguishable by their cDNA sequence alone and so screening for a particular member of that gene family may result in the unintentional cloning of other members. The nematode, for instance, has about 100 collagen genes and it would be a formidable task to clone them all as cDNAs.

3. cDNAs do not contain much of the important regulatory sequences which lie upstream and downstream (5' and 3') to the coding regions. In order to obtain these it is often necessary to use the cDNA clone to screen a library for its equivalent genomic clone. Also cDNAs have

been spliced to remove intronic sequences, which also contain important regulatory sequences. To perform single stranded conformational polymorphism (SSCP) studies which detect base pair changes using genomic DNA, it is usually necessary to know the whereabouts of the splice junctions in cDNA and to know some of the surrounding intronic sequence.

1.5 EXON TRAPPING

Several groups have developed alternative strategies for isolating mammalian genes from genomic DNA which involve using a retroviral construct to trap exons by RNA splicing (Reilly et al, 1990; Buckler et al, 1991; Auch and Reth, 1990; Duyk et al, 1990). The procedure involves the "shotgun" cloning of fragments of genomic DNA or cloned genomic DNA into a specially constructed retroviral vector. The cloning site is situated either between a viral donor and acceptor site from the HIV-1 tat gene (Buckler et al, 1991), or between a donor and acceptor site from the rat preproinsulin gene (Auch and Reth, 1990). Alternatively the clone is inserted downstream of a human beta globin donor site (Duyk et al, 1990). The cloned fragments are then transfected into COS cells where the retroviral DNA is transcribed. Recombinant molecules that contain a functional splice acceptor and donor site, cloned in the correct orientation, may undergo RNA splicing. RNA is then harvested from the cytoplasm and reverse transcribed into cDNA. Using PCR primers specific for the sequences flanking the retroviral donor and acceptor sites, the "trapped" exons are then amplified. The amplified "trapped" exons can then be detected on an agarose gel and be cloned into a sequencing vector (such as Bluescript) to create an "exon library". "Trapped" exons can then be sequenced.

The procedure of Duyk et al (1990) follows the same principle, although it is much lengthier. A single functional acceptor splice site is required in the recombinant molecule, as a positive splicing event joins the "trapped" exon to the gene encoding the α complementing factor of

E.coli β-galactosidase, thus generating a fusion protein with no β-galactosidase activity. This protocol requires several rounds of transfection to increase the titre of the recombinant molecules, but requires no PCR amplification as positive clones can be detected by β-Gal blue/white selection. Positive clones can be directly sequenced. "Trapped" exons can be used to identify potential genes and to create "transcription maps" of large areas of the genome.

There are a number of drawbacks to this line of investigation:-

1. Once an exon has been "trapped" by these methods, further analysis is required in order to demonstrate that the sequence represents a gene, for example by using it to probe Northern blots or to screen cDNA libraries.
2. The procedure will not be able to recover all genes. Firstly, not all genes contain introns and secondly some splicing events are temporally regulated or tissue specific and so may not occur in the packaging cell lines used for exon trapping.
3. DNA sequences are not propagated equally well in retroviral vectors, so that a library of clones may not be representative of the exons in the original cloned genomic DNA.
4. There is much uncertainty as to whether the splicing machinery would be able to detect all mammalian splice sites.
5. Cryptic splice events may occur, creating false positives at an estimated frequency of between 1:5 and 1:10.
6. The techniques involved are time consuming and labour intensive.
7. The lack of convenient restriction enzymes in the vicinity of an exon may limit the chances of the exon being trapped.
8. Only relatively small amounts of cloned DNA can be studied at a time.

The foregoing list of problems makes it clear that this strategy is still far from becoming an efficient and widely accepted laboratory technique for the isolation of new genes.

1.6 IDENTIFICATION OF HTF AND CPG ISLANDS

Another strategy for identifying new genes is to search for CpG rich islands and HTF (HpaII tiny fragments) islands. The dinucleotide CpG is relatively under-represented in the genomic DNA of vertebrates, occurring at between 20 and 25% of the frequency expected from the base composition (Josse et al, 1961; Swartz et al, 1962). Not only is CpG depleted in the genome, but it is also spread unevenly throughout the genome (Smith et al, 1983; Lennon and Fraser, 1983; Adams and Eason, 1984). Stretches of DNA with CpG frequency close to that expected have been shown to be associated with genes and in particular the 5' promoter regions (Tykocinski and Max, 1984; Cooper and Gerber-Huber, 1985; Bird, 1986). These regions are called CpG islands. In particular, housekeeping genes all appear to have CpG islands (Gardiner-Garden and Frommer, 1987). The association of CpG islands with coding regions has lead to the idea that the CpG island is somehow involved in the transcriptional regulation of genes. It has been postulated that the reason for CpG depletion in the genome is that 5-methylcytosine (^{5m}C) has a high tendency to mutate to thymine by deamination (Coulondre et al, 1978). If this is the case, then the presence of CpG rich regions adjacent to genes suggests that these regions must have a high selective importance (Cooper and Gerber-Huber, 1985), thus corroborating the theory that CpG islands play some regulatory role for gene expression.

Human genomic DNA *in vivo* is packaged as chromatin, which is the general term for the combined structure of the DNA together with the histone proteins, some RNA and acidic proteins (reviewed by Bradbury et al, 1981). At the subchromatin level the DNA is wrapped around a histone complex at intervals of about 200bp, forming nucleosomes. Transcriptionally active chromatin is usually in an extended open conformation. Cytogenetic analysis of chromosomes using banding techniques have revealed that the chromatin exists in several

different forms or "flavours" during metaphase (reviewed by Holmquist, 1992). The two main flavours are termed G-bands and R-bands respectively, depending on whether the region stains positive or negative with trypsin/Giemsa staining. R-bands can also be revealed by reverse banding by heat denaturation/Giemsa staining or by reverse of quinacrine banding. R-bands are also characterised by a high proportion of Alu repeats, high G+C content and contain many actively transcribed housekeeping genes, whereas G-bands are rich in L1/KpnI repeats and contain many tissue specific genes (Goldman et al, 1984). R-bands have been subdivided into four "flavours"; two flavours of T-bands, which are very G+C rich and two flavours of Alu-rich bands. *In situ* hybridization studies have shown that CpG islands are clustered predominantly within R-bands, with the highest concentration occurring within the T-bands (Craig and Bickmore 1994). Chromatin structure at CpG islands differs from bulk chromatin in several respests, for instance a high proportion of CpG islands have nucleosome regions and in nucleosomes that are associated with CpG islands histone H1 is depleted and H3 and H4 are highly acetylated (Tazi and Bird, 1990).

The CpG dinucleotide, which in the majority of the genome is extensively methylated at the cytosine 5' position, is highly under-methylated in these so called CpG islands adjacent to genes. The methylation status of CpG islands can be analyzed in genomic DNA by digesting the DNA either with the methylation sensitive restriction enzyme HpaII or with its methylation insensitive isoschizomer MspI, followed by Southern hybridization analysis of the fragments. HpaII and MspI both cut at CCGG sites. If HpaII digests within a CpG island it results in tiny fragments, hence unmethylated CpG islands are also known as HTF (HpaII tiny fragments) islands (Cooper et al, 1983). This method has been used to show that the methylation status of CpG in some genes is tissue specific (Silva and White, 1988). Evidence suggests that paternally and maternally derived genomes may have different patterns of expression during development (Solter, 1988). This is known as gene imprinting and several

genetic disorders are known to result from inheriting either maternal or paternal alleles (for instance Prader-Willi and Angelman syndromes). Little is known about the mechanism of imprinting, however differential CpG methylation and condensation of the chromatin have been shown to occur at imprinted genes.

As mentioned above the depletion of CpG in the genome is thought to be due to the high rate of mutation of 5-methylcytosine to thymine by deamination (Coulondre et al, 1978). CpG islands are hypomethylated and thus mutation by deamination does not occur. Since the estimated number of CpG islands is 45,000 (Antequera and Bird, 1993) and their size varies between 200 and 1400 bp, averaging at 982 bp (Larsen et al, 1992), CpG islands are thought to represent about 1% of the genome, assuming an average of 1000bp per island. CpG islands have an average G+C content of 65% compared to 45% in the total genome and show no depletion of the CpG dinucleotide. Restriction enzymes that cut at sites with one or more CpG are much more likely to cut within CpG islands than in bulk DNA. Lindsay and Bird (1987) predicted that 74% of sites cut by CpG cutting enzymes that recognize six bases (eg. SacII, BssHII and EagI) and 89% of those that recognize eight bases (eg. NotI and AscI) should be located within CpG islands. Such enzymes are also known as rare-cutters because they cut genomic DNA relatively infrequently. By searching through databases of sequences it has been shown that the proportion of NotI sites within CpG islands rather than bulk DNA is around 84% (Kusuda et al, 1990), very close to the predicted frequency.

As a result of this knowledge strategies have been developed for cloning potential CpG islands and hence genes, using such restriction sites as markers (Smith et al, 1987; Kusuda et al, 1989; Frischauf, 1989). The cloning of fragments generated by NotI digestion is considered of particular importance for "chromosome walking" towards a disease gene locus, once it has been localized to a particular chromosomal region by linkage analysis. NotI sites are on average 1Mb apart and since they are more often than not situated within CpG islands and

hence adjacent to genes, they serve as very useful markers for long range physical mapping (Barlow and Lehrach, 1987) and for searching for disease genes (Estivill et al, 1987; Pohl et al, 1988). Many of the restriction enzymes that cut at CpG sites are methylation sensitive and therefore do not cut well at sites where CpG is methylated, such as in the bulk genomic DNA, but do digest properly at the unmethylated sites in the CpG islands. Thus detection of such sites in total genomic DNA is very strong evidence for the presence of a CpG island. Rare cutting enzymes can also be used to screen in genomic DNA cloned into YAC, cosmid and phage vectors for the presence of CpG islands (Lindsay and Bird, 1987).

The tendency for rare-cutter restriction sites to cluster within CpG islands thus provides a useful method for mapping and cloning potential genes. However, searching for clones that contain CpG islands can be a fairly random process, requiring the isolation and preparation of DNA from each individual clone. Another more selective approach involves screening cosmid libraries using short oligonucleotides (eg. octamers) based on rare-cutter restriction sites, as hybridization probes (Estivill and Williamson, 1987; Melmer and Buchwald, 1990; Melmer et al, 1990). Clones that contain one or more rare-cutter sites can thus be selected and used to link DNA fragments separated on pulsed field gels and are therefore invaluable for completing long range restriction maps. Since there is a high probability that these clones also contain coding regions, they may also be analysed further and used to isolate cDNA clones from libraries.

The identification of genes through their respective CpG islands has a number of limitations. For instance it is not clear what proportion of genes possess a CpG island and for those that do, a proportion of them have islands either at the 3' end, within the gene itself and either in exons or in introns (Gardiner-Garden and Frommer, 1987). CpG islands that lie adjacent to a gene may be some considerable distance from the transcribed sequence. Thus whilst the

presence of a CpG island is a fairly strong indicator for the presence of a gene, the actual location of the gene is not always obvious.

1.7 CONSENSUS SITES ASSOCIATED WITH CODING REGIONS

One of the drawbacks associated with the cDNA cloning approach for identifying new genes is that the derived sequence is devoid of many of the regulatory sequences which lie upstream of the gene and in intronic DNA. Thus in order to investigate the expression and control of a gene it is necessary to analyze the genomic equivalent of the cDNA as well as flanking genomic DNA over large distances. To search for coding regions within a large genomic clone requires a large amount of subcloning and DNA sequencing, followed by detailed sequence analysis. Computer programs have been devised to search sequences for coding regions by looking for open reading frames (Gribskov and Devereux, 1991). Since three nucleotides code for each amino acid, a coding unit consists of three consecutive bases and is called a codon. A DNA sequence can be translated in to peptide sequence in six different ways; the first by taking the first base of the sequence as the first base of the initial codon, the second by taking the second base as the first base of the codon and the third taking the third base as the initial base of the first codon. The same can be done starting from the other end of the sequence, using the complementary DNA strand, giving a total of six reading frames. The reading frame ends when a codon encodes for "stop" (for example TAG). Searching for open reading frames can, however, give dubious results.

Another way of targeting coding regions within the genome is to use sites which follow a consensus sequence. There are number of such sequences associated with genes, all of which play an important role in regulation and transcription (Figure 1.1. shows a model gene showing the relative positions of the various consensus sequences). In many of these signals there is a great deal of variation between different sites, even though they are recognised by

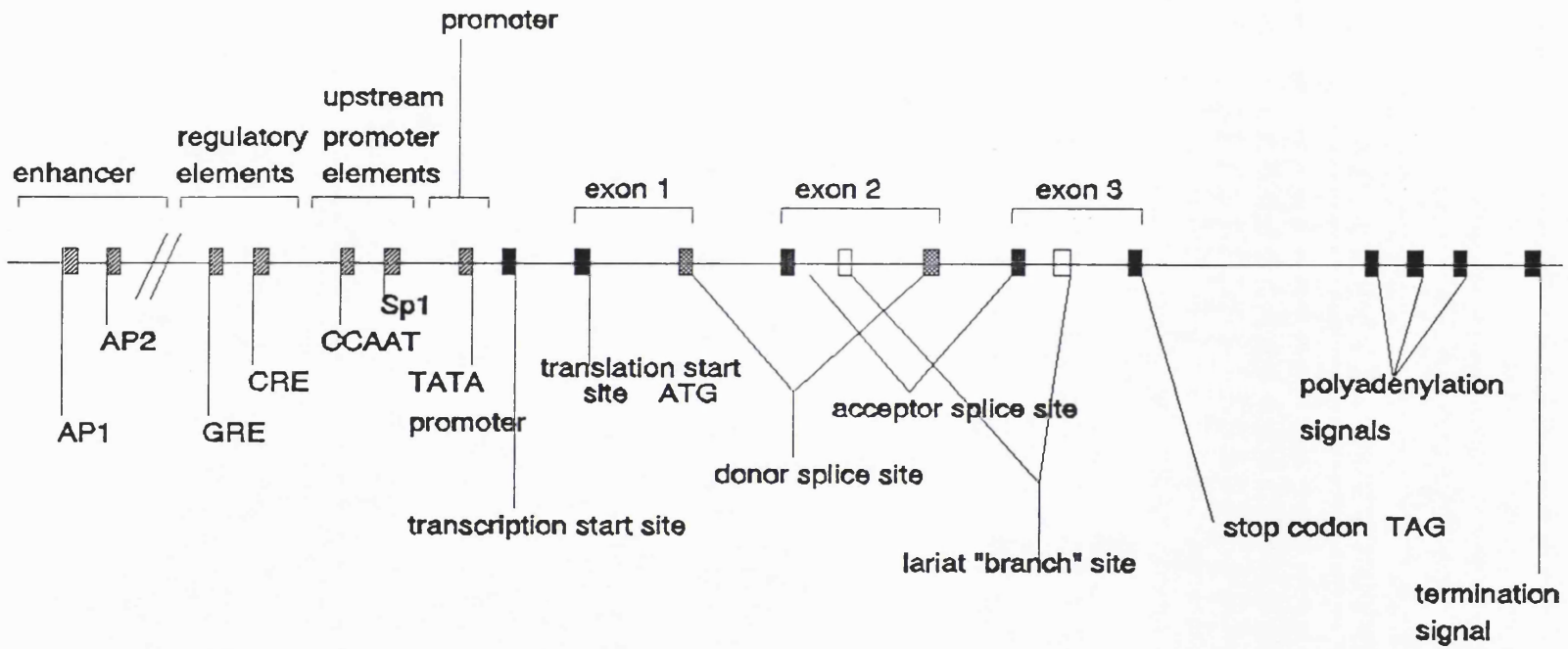


Figure 1.1: Schematic representation of a gene showing relative positions of the various consensus signals for regulation, promotion, transcription and translation.

the same DNA binding protein. Several bases, however, remain strongly conserved and there are often strong preferences for the surrounding bases. By searching through computer databases it has been possible to compile lists of sequences for known regulatory signals and frequency tables or weight matrices have been computed which show the nucleotide likelihood for each position of a consensus site. There are tables available for eukaryotic promoters such as the CCAAT box, GC box, TATA box and cap site (Bucher, 1990). Tables are also available for enhancers and transcription factors such as SP1 and AP1 (Ghosh, 1990) and for transcription terminators (McLauchlan et al, 1985). Translation signals have also been tabulated, eg. translation initiation sites (Sargan et al, 1982), polyadenylation signals (McLauchlan et al, 1985) and also for donor and acceptor splice sites (Senapathy et al, 1990) (See Table 1.1).

Table 1.1 Frequency tables for consensus sites. Base frequency tables show the number of occurrences of each base at each position. Weight matrices are the calculated natural logarithms of the ratio of observed to expected frequency of each base at each position and thus are a better reflection of the likelihood of finding a particular base at each position.

i. Eukaryotic translation initiation signal (Sargan et al, 1982).

Base Frequency

	-7	-6	-5	-4	-3	-2	-1	1	2	3
A	50	27	27	19	86	36	34	102	0	0
C	20	15	32	65	5	42	52	0	0	0
G	6	29	12	6	11	6	11	0	0	102
T	19	24	31	12	0	18	5	0	102	0

Weight Matrix

	-7	-6	-5	-4	-3	-2	-1	1	2	3
A	0.33	-0.29	-0.36	-0.71	0.80	-0.07	-0.13	-0.97	-4.35	-4.35
C	-0.09	-0.37	0.31	1.02	-1.54	0.58	0.80	-3.85	-3.85	-3.85
G	-1.06	0.52	-0.44	-1.13	-0.52	-1.13	-0.52	-3.61	-3.61	1.70
T	-0.05	0.19	0.37	-0.58	-3.76	-0.17	-1.45	-3.76	1.56	-3.76

consensus

	A	G	t/c	C	A	C	C	A	T	G

ii. Splice donor site (Senapathy et al, 1990).

Base Frequency

	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1
A	11	11	10	8	11	10	11	11	7	8	25	3	100	0	27
C	29	33	30	30	32	34	37	38	39	36	26	75	0	0	14
G	14	12	10	10	9	11	10	9	7	6	26	1	0	100	49
T	46	44	50	52	48	45	42	43	47	51	23	21	0	0	10

Weight Matrix

	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1
A	-0.43	-0.43	-0.52	-0.74	-0.43	-0.52	-0.43	-0.43	-0.88	-0.74	0.39	-1.73	1.78	-7.12	0.47
C	-0.04	0.09	-0.01	-0.01	0.06	0.12	0.20	0.23	0.26	0.18	-0.15	0.91	-7.71	-7.71	-0.77
G	-0.26	-0.42	-0.60	-0.60	-0.71	-0.51	-0.60	-0.71	-0.96	-1.11	0.35	-2.90	-7.20	1.70	0.99
T	0.28	0.24	0.36	0.40	0.32	0.26	0.19	0.21	0.30	0.38	-0.41	-0.50	-7.85	-7.85	-1.25

consensus

-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1
t	t	t	t	t	t	n	c/t	t/c	t	a/c	C	A	G	G/A

iii. Splice acceptor site (Senapathy et al, 1990).

Base Frequency

	-3	-2	-1	1	2	3	4	5	6
A	32	60	9	0	0	59	71	7	16
C	37	13	5	0	0	3	9	6	16
G	18	12	79	100	0	35	11	82	18
T	13	15	7	0	100	3	9	6	50

Weight Matrix

	-3	-2	-1	1	2	3	4	5	6
A	0.13	0.76	-1.14	-7.65	-7.65	0.74	0.92	-1.39	-0.57
C	1.32	0.27	-0.68	-6.60	-6.60	-1.19	-0.09	-0.50	0.48
G	-0.78	-1.19	0.70	0.93	-7.98	-0.12	-1.28	0.73	-0.78
T	-0.55	-0.41	-1.17	-7.43	1.49	-2.02	-0.92	-1.32	0.80

consensus

	C	A/c	G	G	T	A	A	G	T/C

iv. Polyadenylation signal (McLauchlan et al, 1985).

Base Frequency

	1	2	3	4	5	6	7	8	9	10	11	12
A	24	19	24	94	90	0	94	94	94	34	26	22
C	19	25	28	0	0	0	0	0	0	21	22	35
G	10	17	15	0	2	0	0	0	0	26	22	12
T	41	33	27	0	2	94	0	0	0	13	24	25

Weight Matrix

	1	2	3	4	5	6	7	8	9	10	11	12
A	-0.76	-0.99	-0.76	0.60	0.56	-3.93	0.60	0.60	0.60	-0.41	-0.68	-0.85
C	0.42	0.69	0.81	-3.71	-3.71	-3.71	-3.71	-3.71	-3.71	0.52	0.56	1.03
G	0.14	0.67	0.55	-2.85	-1.46	-2.85	-2.85	-2.85	-2.85	1.10	0.93	0.33
T	0.65	0.43	0.23	-3.75	-2.37	1.48	-3.75	-3.75	-3.75	-0.50	0.11	0.15

consensus

	T/C	C/G/T	C/G/t	A	A	T	A	A	A	G/C	G/C	C/g

Frequency tables and weight matrices can be used for computer analyses of sequence data, to establish the whereabouts of genes. They have also been used to design degenerate oligonucleotides to be used for screening by hybridization to identify coding regions. Melmer and Buchwald (1992) have shown that it is feasible to use short and degenerate oligonucleotides corresponding to splice junctions as hybridization probes for the identification of cloned genomic fragments containing the corresponding splice sites. Three degenerate oligonucleotides, a 10-mer corresponding to the 5' splice junction consensus, a 9-mer and a 15-mer corresponding to the 3' splice junction, were tested by Southern hybridization to restriction digested subclones, on the known intron-exon boundaries of the cloned human proteolipid protein gene (PLP). All the predicted hybridizations were observed. The oligonucleotides were also tested against random genomic plasmid and cosmid clones. The

presence of genes in the clones that were selected was supported by Northern blot analysis and cross-hybridization analysis using "zoo" blots (Melmer and Buchwald, 1992).

Since most genes contain splice junctions, this procedure could be employed for the identification of genes in random clones, as an alternative to the cDNA cloning approach (Brenner, 1990). It would also be useful in positional cloning studies, by locating genes in areas in which a disease gene locus is thought to exist.

The same approach could also be used with oligonucleotides designed from other consensus sequences, such as the three previously mentioned. However, consensus sites which are found to occur only once in a gene would be less effective, as the probability of detection would be lower.

1.8 SEQUENCING THE GENOME

One of the longer term goals of the Human Genome Project is the sequencing of the entire human genome. The fulfilment of such a goal will have a major impact on biological science (Watson, 1990; ASHG Human Genome Committee Report, 1991). However, the sheer size of the human genome at an estimated 3 billion base pairs presents formidable obstacles for complete sequencing without major technical advances (Cantor, 1990, Pohl, 1992, Maddox, 1992). Brenner has suggested the strategy of first sequencing the expressed regions (Brenner, 1990). Whilst there has been some remarkable progress in the use of expressed sequence tags (Adams et al, 1991) this strategy has some inherent drawbacks (Little, 1991, Roberts, 1991, Burglin and Barnes, 1992). Theoretical proposals for fast but random sequencing using short oligonucleotides have been described by Studier (1989) who has proposed that direct sequencing of cosmid DNAs could be performed using a library of octamers, nonamers or decamers. Studier calculated that a random octamer would have on average 1.37 sites in a

typical cosmid of 45,000bp length (40,000bp of insert in 5000bp of vector) and a nonamer would have 0.343 sites. Thus by employing a library of such oligonucleotides of random sequence as primers for sequencing reactions and by using cosmid clones as template, it should be possible to sequence most portions of a target cosmid. Studier estimates that sequencing approximately 500bp at a time from 30 primers and their complements would generate between 20 and 25% of the cosmid sequence in 8 to 11 blocks of 1000bp each. These initial blocks of sequence could then be extended in both directions by directed priming. The switch from random to directed priming would limit the sequencing of large overlaps and would speed up the final stages, filling in the last few gaps. This strategy has been termed "high volume sequencing" and could easily be automated.

However, the generation of a complete oligonucleotide library increases in cost exponentially with the length of oligo. Costs can be cut either by using the shortest possible primers, or by rationalizing the library. Siemieniak and Slightom (1990) used a computer program to select useful nonamer primers for sequencing, in which certain rules for primer selection were followed, such as no polynucleotide repeats, no dinucleotide repeats and GC content between 45 and 60%. By carefully selecting primers for inclusion in a primer library, the number of nonamers required for an effective sequencing strategy could be reduced from 262,144 to 3342, thus greatly reducing the cost of producing such a library. This streamlining of the library would also decrease the number of redundant sequencing reactions.The entire genome could be sequenced this way, with an estimated 100,000 contiguous cosmids necessary to cover it.

In 1990 Szybalski proposed the idea of using a library of 4096 hexamers as "building blocks" for constructing sequencing primers, using ligation reactions to link the hexamer units together. More recently this idea has taken root and it has been shown that selected hexamers,

when mixed with target sequence and single-stranded binding protein in the correct proportions, can initiate efficient sequencing, without the need to ligate the hexamers together beforeh and (Kieleczawa et al, 1992). The single-stranded binding protein prevents any priming from single or even pairs of hexamers, so that only the target sequence is amplified. Thus a directed sequencing strategy is now feasible, based on a library of only 4096 (hexamer) primers. This approach is also compatible with fluorescence based sequencing methods (Hou and Smith, 1993).

Several other proposals for high volume sequencing have been put forward. The idea of sequencing DNA by the hybridization of template DNA to a complete set of fixed-length oligonucleotides immobilized in a 2-D array on glass plates has been introduced (Khrapko et al, 1989). The idea has been tried out successfully on very short DNA sequences (Khrapko et al, 1991) and work is underway to increase the power and efficiency of the technique (Mirzabekov, 1992). Since the thermostability of duplexes between DNA and matrix-immobilized oligonucleotides is dependent on concentration as well as the sequence of each individual oligonucleotide, matrices with adjusted concentrations of immobilized oligonucleotides results in similar melting temperatures for AT and GC-rich duplexes. Using a fluorescence microscope coupled with a close-circuit camera and computer, the results of hybridizations can be processed instantaneously. However, the system has several drawbacks, in particular its inability to cope with simple repeat sequences, thus limiting its usefulness for high volume sequencing strategies. Sequencing by hybridization with oligonucleotide matrix (SHOM) could also be applied to mapping, fingerprinting, mutation detection and diagnosis.

Various strategies for the automation of DNA sequencing such as Applied Biosystems' "Dye-deoxy" sequencing using the ABI 373A automated DNA sequencer (Prober et al, 1987; Blasband, 1992) and other improvements in the number of bases of DNA that can be

sequenced on one gel (Roemer et al, 1992; Ansorge, 1992) mean that it will soon be possible to read around 2Kb of sequence from one reaction.

1.9 HYPOTHESES TESTED IN THE THESIS

It is apparent that once linkage analysis has established the approximate location of a gene, existing methods to pinpoint the gene are somewhat cumbersome and labour intensive. This thesis aims to test hypotheses that existing methods of gene detection can be complemented by the development of new techniques using short synthetic oligonucleotides. The overall objective was to use several classes of oligonucleotide which were homologous to:-

(1) Sequences based on rare-cutter restriction sites, which have been shown to cluster in the CpG rich regions adjacent to genes.

(2) Sequences based on splice sites.

(3) Sequences based on other consensus sites for control and regulation in or around coding regions.

In order to identify such sequences in target DNA, preliminary work was necessary to determine operating conditions such as time and temperature. Once these were established then the following experimental approaches could be used:-

1. Southern hybridization based procedures.
2. Polymerase chain reaction based procedures.
3. Direct sequencing based procedures.

Other applications of the methods developed such as high volume sequencing and landmarks for genome mapping such as STS's and ETS's, are discussed in the relevant chapters.

CHAPTER 2: MATERIALS AND METHODS

2.1 SOURCES OF CHEMICALS

Ampicillin: Beecham Research labs (Brentford, UK)

Calf intestinal phosphatase: Boehringer Corporation Ltd

Chemicals: Sigma Chemicals Ltd

Ultra pure dATP, dCTP, dGTP, dTTP ,ddATP, ddCTP, ddGTP, ddTTP and 5-deaza-2'-dGTP: Pharmacia Chemicals Ltd

dsDNA cycle sequencing kit: Gibco BRL

DNA polymerase Klenow fragment: Boehringer Corporation Ltd

Fuji-RX film: Genetic Research Instruments

Hybond-N: Amersham International plc

Kanamycin: Sigma Chemicals Ltd

SeaKem GTG agarose: Flowgen Instruments Ltd

Ultra pure LMP agarose: Gibco BRL

pUC18/19 DNA: Boehringer Corporation Ltd

mp18/19 M13 RF DNA: Boehringer Corporation Ltd

Restriction enzymes: Boehringer Corporation Ltd; New England Biolabs Inc.

TA cloning kit: Invitrogen Corporation

T4 DNA ligase: Boehringer Corporation Ltd

T4 DNA polynucleotide kinase: Boehringer Corporation Ltd

Sequenase kit (version 2.0): United States Biochemicals Corporation

Oligonucleotides synthesis: Oswell DNA services

Random hexanucleotides (pN6): Pharmacia Chemicals Ltd

Ribonuclease A: Sigma Chemicals Ltd

Sequagel kit: Flowgen Instruments Ltd

Agar: Difco laboratories

Yeast extract: Difco laboratories

Tryptone: Difco laboratories

DNA thermal cycler: Cetus Perkin Elmer

[α-^{32}P] dCTP 3000 ci/mMol: Amersham International plc

[γ-^{32}P] ATP 3000 ci/mMol: Amersham International plc

[γ-^{35}S] dATP NE0034S: Dupont NEN

2.2 SOURCES OF OLIGONUCLEOTIDES

Oligodeoxyribonucleotides were either purchased from Oswell DNA Services, University of Edinburgh, or were synthesized at the Medical Molecular Biology Unit, University College and Middlesex School of Medicine, using an Applied Biosystems Model 381A DNA Synthesizer. The GC octamers were purchased as DNA linkers from Boehringer Corporation Ltd., New England Biolabs Inc. and Clontech.

2.3 SOURCES OF VECTORS, CLONES AND SUBCLONES

The cosmid vector pWE15 (Wahl et al, 1987) was purchased from Stratagene. The phagemid vector pSL1180 was purchased from Pharmacia Chemicals Ltd.. M13mp18/19 and pUC 18/19 were purchased from Boehringer Corporation Ltd.. The TA cloning vector pCR1000 was purchased from Invitrogen Corporation. The vector pCV108 (Lau and Kan, 1983) was from Dr.Y.F. Lau (Howard Hughes Medical Institute) and pMBgpt was constructed by Dr.M. Lu (University of Toronto). The cosmid clone cosGSTrp7 and the subclones p5HB8 and p5TS0.6 were a kind gift from Dr. Sharon Spencer, UCMSM. The calcitonin/CGRP cosmid clone cosCT1 and subclones pGemCal5, 7, 10, 12, 20A, 21 and 22 were kindly donated by

Dr. Paul Brickell, UCMSM. The PLP subclones 34H1 and 34S1 were a kind gift from Dr. J. Riordan, University of Toronto.

2.4 BUFFERS AND SOLUTIONS

Ampicillin: prepared as a 100 mg/ml stock solution in water and stored at -20 °C. Used at a working concentration of 45μg/ml.

Denaturation buffer: 1.5M NaCl, 0.5M NaOH

100x Denharts: 2% (w/v) bovine serum albumin, 2% (w/v) polyvinylpyrrolidone, 2% (w/v) Ficoll 400 made up in water.

DNA loading buffer: 10 x, 0.25% bromophenol blue, 0.25% xylene cyanol, 25% Ficoll (type 400) in water.

Chloroform: chloroform and isoamyl alcohol mixed at a ratio of 24:1.

Ethidium bromide: solution of 10mg/ml in water. Stored 4°C.

IPTG: 100mM solution of isopropyl-B-D-thiogalactopyranoside (23.8 mg/ml of water). Store -20 °C.

Ligation buffer: 10 x, 660mM Tris.Cl (pH7.6), 50mM $MgCl_2$, 50mM DTT, 6mM ATP. Store -20 °C.

Lysozyme buffer: 50mM glucose, 10mM EDTA, 25mM Tris.Cl (pH8.0). Lysozyme added to 4mg/ml.

Neutralization buffer: 1M Tris.Cl (pH7.6), 1.5M NaCl.

Proteinase-K buffer: 100mM NaCl, 10mM Tris.Cl (pH8.5), 25mM EDTA.

Phage lysis buffer: 2.5% SDS, 0.5M Tris.Cl (pH9.0), 0.25 M EDTA.

Phenol: dissolved at 65°C, saturated with water and then equilibrated with TE pH7.6.

Spermidine: 100mM solution in water. Store -20 °C.

SM buffer: 100mM NaCl, 10mM $MgSO_4$, 50mM Tris.Cl (pH7.5), 0.01% gelatin.

KCl, 10mM $MgCl_2$, 10mM $MgSO_4$, 20mM glucose.

SSC: 20 x, 3M NaCl, 0.3M tri-sodium citrate.

SSPE: 20x, 3.6M NaCl, 200mM NaH_2PO_4 (pH7.4) 20mM EDTA (pH7.4).

T4 polynucleotide kinase buffer: 800mM Tris.Cl (pH7.6), 120mM $MgCl_2$, 60mM dithiothreitol (DTT), made up in water and stored at -20 °C.

TAE: 40mM Tris.Cl, 1mM EDTA in water, pH adjusted to 8.0 with glacial acetic acid.

Taq polymerase buffer: 10 x, 100mM Tris.Cl (pH8.3), 500mM KCl, 15mM $MgCl_2$, 0.1% gelatin. Stored -20°C

TBE: 90mM Tris, 90mM Boric acid, 1.25mM EDTA, pH8.3.

TE: 10mM Tris.Cl, 1mM EDTA in water, pH 7.5-9.0 as stated.

X-Gal: 2% solution made up by dissolving 20mg of 5-Bromo-4-chloro-3-indolyl-β-D-galactoside in 1ml of dimethyl formamide. Store -20 °C.

2.5 MEDIA

Glucose/minimal medium plates: 1.5% (w/v) minimal agar, 1 x M9 salts, 1mM $MgSO_4$, 0.1mM $CaCl_2$, 1mM thiamine HCl, 0.2% glucose.

Luria Bertani medium (LB): 1% (w/v) bacto tryptone, 0.5% (w/v) bacto yeast extract, 1% (w/v) sodium chloride made up in water and adjusted to pH7.5.

L-agar: 1.5% (w/v) Difco technical agar in LB.

H-agar: 1.2% (w/v) Difco technical agar in 1% bacto tryptone, 0.8% sodium chloride.

H-top agar: 0.5% agarose in 1% (w/v) bacto tryptone, 0.8% sodium chloride.

SOC media: 2% Bactotryptone, 0.5% Bacto yeast extract, 10mM NaCl, 2.5mM 10 x M9 salts: per litre, 60g Na_2HPO_4, 30g KH_2PO_4, 10g NH_4Cl, 5g NaCl, stored at 4°C.

2 x TY: 1.6% (w/v) bacto tryptone, 1% (w/v) bacto yeast extract, 0.5% (w/v) sodium chloride made up in water.

All media were sterilized by autoclaving at 120°C for 20 minutes.

2.6 SOUTHERN BLOTTING AND HYBRIDIZATIONS

Southern blotting (Southern 1975) was carried out using a modification of the protocol described by Amersham International plc for use with Hybond-N filter membranes. Digested DNA was separated by electrophoresis through 0.6-1% TAE agarose gels at 50V. The gel was stained with ethidium bromide, photographed on a UV transilluminator and denatured with denaturization buffer for 1hr. Transfer to Hybond-N membrane was carried out overnight in 20 x SSC after which the DNA was covalently linked to the membrane by UV irradiation for 2 minutes. Membranes were prehybridized at 65°C in 20mls of 0.9M NaCl, 1% SDS and 50μg/ml autoclaved denatured salmon testis DNA for a minimum of four hours. Hybridization was carried out overnight at 65°C in 20mls of fresh solution including 10% (w/v) dextran sulphate as well as 100ng of denatured oligolabelled probe at $\sim 2 \times 10^6$ cpm/ml. The membranes were then washed to a stringency of 1 to 0.1 x SSC, 0.1% SDS at 65°C. The filters were exposed against autoradiographic film at -70°C using intensifying screens. In order to reprobe the filters were treated as recommended by the manufacturer.

2.7 SLOT BLOTTING AND DOT BLOTTING

For slot blotting approximately 2μl of DNA was added to 200μl of 15x SSC. The membranes (Hybond-N) were pre-soaked in 15x SSC and 3MM filters in 3x SSC. Either Schleicher and Schuell or BioRad apparatus was used, adding the DNA samples under vacuum. The apparatus was washed thoroughly between use. The Hybond membranes were baked at 80°C for 30 mins and UV illuminated for 3 mins (DNA-side face down), to fix the DNA to the membrane.

For dot blots the DNA was transferred directly via a drawn-out Pasteur pipette, drying each application with a hair-dryer in order to concentrate the DNA in as small an area as possible.

The DNA was fixed to the membranes in the same manner as the slot blots.

2.8 PLASMID AND COSMID VECTOR PREPARATIONS (LARGE SCALE)

This method follows the alkaline-lysis procedure of Birnboim and Doly (1979), with modifications (Maniatis et al, 1982).

Solution 1: 50mM glucose, 25mM Tris-Cl and 10mM EDTA.

Solution 2: 0.2M NaOH, 1% SDS.

Solution 3: 147.2g Potassium acetate, 57.5 mls Glacial acetic acid. The resulting solution contains 3M Potassium ions and 5M acetate ions.

5ml overnight cultures prepared from single bacterial colonies were used to inoculate 500mls of sterile LB containing the appropriate antibiotic and grown overnight at 37°C with vigorous shaking. The bacteria were pelleted at 4200 rpm for 30 minutes at 4°C in a Beckman J6-B centrifuge and then resuspended in 50mls of solution 1. The cells were lysed by adding 100mls of solution 2 and the bacterial debris and chromosomal DNA was precipitated with the addition of 50mls of solution 3. After mixing, the precipitate was pelleted by centrifugation at 4200 rpm for 15 minutes at 4°C. The supernatant was filtered through nylon gauze and the plasmid DNA and bacterial RNA precipitated by adding 120mls of propan-2-ol. After centrifugation at 6000 rpm in a Sorvall GS3 rotor, the pellet was washed with 70% ethanol and resuspended in 5mls of TE pH8.0 before being transferred to a pre-weighed universal and made to 9g with TE pH8.0. To this was added 10g caesium chloride and 1ml of 5mg/ml ethidium bromide. After an initial pre-spin at 3000rpm to separate out precipitated protein/ethidium bromide, the plasmid DNA was separated from the RNA and bacterial DNA according to density by centrifugation for 24-48hrs in a Beckman 70Ti rotor at 55000rpm at 20°C. The lower plasmid band was extracted from the gradient. An equal volume of water-saturated butan-1-ol was added and the two phases were mixed thoroughly by shaking. The

upper (organic) phase was removed and the process was repeated several times until all the pink colour (ethidium bromide) disappeared. The plasmid solution was made up to 10ml with TE pH8.0. Plasmid DNA was precipitated with 20ml of absolute ethanol at room temperature and pelleted by centrifugation at 3000rpm for 20 minutes in a Beckman J6-B centrifuge. After washing with 70% ethanol and drying the DNA was resuspended in 500μl TE and treated with 40U/ml RNase A for 15 minutes at 37°C. This was then phenol extracted again to remove the RNase and precipitated in ethanol.

2.9 PHAGE DNA PREPARATION (LARGE SCALE)

Solutions:10% (w/v) maltose filter sterilized, 1M $MgSO_4$ (autoclaved),3M Sodium Acetate pH5.2, PEG 6000 50% (w/v) mix overnight to dissolve fully.

The host bacteria (LE392) was grown overnight in 10mls of LB media with 10mM $MgSO_4$ and 0.2% maltose. The cells were pelleted by centrifugation and resuspended in 10mls of 10mM $MgSO_4$ and stored at 4°C for a maximum of three weeks. The optical density at 600nm was measured (OD 1 = 8 x 10^8 bacteria/ml). The titre of the phage lysate was also determined. In a total volume of 200μl containing 10mM $MgSO_4$ ~ 10^{10} bacteria were incubated with 10^7 phage at 37°C for 20 minutes. After the adsorbtion of the phage, the mixture was added to 500mls of LB, 10mM $MgSO_4$ and 0.2% maltose (in a smooth sided flask) and shaken vigorously overnight at 37°C. The flask was shaken for a further 30 minutes with 1ml of chloroform and the cell debris settled for 30 minutes. The lysate was removed leaving the chloroform behind and centrifuged at 8K and 4°C for 10 minutes (Sorval GS3 rota). The cleared lysate was decanted into a clean flask and bacterial DNA and RNA were digested with 20μl of pancreatic DNase (20mg/ml) and 10μl pancreatic RNase (20mg/ml) by incubation at 37°C for 30 minutes. Sodium chloride was added to a final concentration of 1M dissolved and

left on ice for 30 minutes after which 100mls of 50% PEG 6000 solution was added, mixed thoroughly and left on ice overnight to precipitate the phage. The phage and other debris were then pelleted at 8K and 4°C for 10 minutes (sorval GS3 rota). The phage were then extracted with 3mls of SM buffer and spun at 8K for 2 minutes at RT (sorval HB4). The supernatant was collected and the remaining pellet was extracted once more with 1.5mls of SM buffer. To 0.5mls of the combined supernatants 5μl of 10% SDS and 10μl 0.25 M EDTA were added and the phage coats ruptured by incubation at 68°C for 15 minutes. This was extracted with an equal volume of phenol, then phenol chloroform and finally chloroform. The phage DNA was ethanol precipitated in the usual way.

2.10 PLASMID DNA MINI-PREPARATION

This method also follows the alkaline-lysis procedure of Birnboim and Doly (1979) with adaptions by Ish-Horowitz (Maniatis et al, 1982).

Solution 1: 50mM glucose, 25mM Tris-Cl and 10mM EDTA.

Solution 2: 0.2M NaOH, 1% SDS.

Solution 3: 147.2g Potassium acetate, 57.5 mls Glacial acetic acid. The resulting solution contains 3M Potassium ions and 5M acetate ions.

5ml of medium containing the appropriate antibiotic was inoculated with a single bacterial colony selected from an agar plate. This was incubated overnight at 37°C, shaking at 225rpm. 1.5ml was transferred to an Eppendorf tube and centrifuged for 1 min in a microfuge. Having removed the medium, the pellet was resuspended in 100μl of ice-cold solution 1 and stored for 5 min at room temperature. 200μl of freshly prepared solution 2 was added and mixed by inverting several times. The tube was then stored on ice for 5 min. 100μl of ice-cold solution 3 was then added and mixed by vortexing for 10 sec with the tube in an inverted position. The tube was then stored on ice for 5 min and then centrifuged at 4°C in a

microfuge for 5 min. The supernatant was transferred to a fresh 1.5ml Eppendorf tube and an equal volume of phenol/chloroform (50:50) added. After mixing by vortexing the tube was centrifuged for 2 min and the supernatant transferred to a fresh tube. To this was added 2 volumes of ethanol at room temperature. After centrifuging for 5 min the supernatant was removed and the pellet washed in 1ml 70% ethanol. This was re-centrifuged and the supernatant removed. The pellet was resuspended in 20μl of TE buffer (pH8.0) containing DNase-free pancreatic RNase (20μg/ml).

2.11 DNA LIGATIONS

After restriction endonuclease digestion of vectors, an aliquot was run on a 0.8% TBE agarose gel to confirm that restriction of the DNA had gone to completion. The restriction endonuclease was inactivated at 65°C for 10 minutes. If the restricted ends were incompatible the plasmid vector was precipitated with propanol and ammonium acetate (Maniatis et al 1982), otherwise the vector was treated with calf intestinal phosphatase (15-20U) for 60 minutes at 37°C. Phenol and chloroform extracted and precipitated with absolute ethanol. Inserts were separated and purified in low melting point agarose in TAE, the correct band being visualised on a UV transilluminator and excised. 100ng of insert was ligated to 20ng of vector in 1 x ligation buffer with 1U of T4 DNA ligase and the agarose diluted to 0.1%. Ligations were carried out overnight at RT.

2.12 PREPARATION OF COMPETENT E.COLI CELLS

Bacteria were grown in 50 mls LB, 10mM $MgSO_4.7H_2O$ and 0.2% glucose to mid logarithmic phase (OD 550nm of 0.4 for JM83 Rec- and 0.3 for TG2). The cells were then left on ice for 10 minutes and then gently centrifuged at 1500g for 10 minutes at 4°C. The pelleted cells were resuspended in 0.5mls of the above solution at an ice cold temperature. To this 2.5mls of 36% glycerin, 12% PEG (MW7500), 12 mM $MgSO_4$.7H2O added to LB

and sterilized by filtration, was added and mixed well without vortexing. The cells were aliquotted and stored at -80°C for up to 3 months.

2.13 TRANSFORMATION OF DNA INTO COMPETENT E.COLI

5ng of ligated vector was added to 100μl of competent cells and left on ice for at least 1hr. The bacteria were then heat shocked at 42°C for 3 minutes and left on ice. For plasmids an equal volume of 2 x LB was added and the cells incubated for 1hr to allow expression of the antibiotic resistance. The bacteria were briefly pelleted (15 seconds) and 1/10 and 1/100 dilutions were prepared in LB and spread onto an LB agar plate with the appropriate antibiotic and incubated overnight at 37°C. For pUC plasmid vectors X-Gal and IPTG were added to the agar to allow blue (non recombinants) white (recombinants) selection.

2.14 M13 CLONING PROCEDURE

TG2 bacteria were grown in 10 mls 2xTY medium overnight, shaken at 37°C. Using 2ml of the overnight culture 40ml of 2xTY medium was inoculated and grown to mid logarithmic phase (OD=0.3). The cells were then left on ice for 10 minutes and then gently centrifuged at 1500rpm for 2 minutes at 4°C. The pelleted cells were resuspended in 20ml of ice-cold sterile 50mM $CaCl_2$. The cells were left on ice for 20 min and then gently centrifuged for 2 min. The cells were resuspended in 4ml of cold $CaCl_2$. The cells were aliquoted and either used or stored at -80°C for up to 3 months. To 0.3ml of cells in a sterile 15ml tube on ice 5μl of ligation mix (10ng of M13 vector to 50ng of insert) were added and stored on ice for at least 1 hr. The cells were then heat shocked for 3 min at 42°C and then returned to ice. 200μl of cells, 40μl of X-Gal and 40μl of IPTG were added to molten H-top at 42°C. This was mixed by rolling and poured directly onto pre-warmed LUA or H-plates. Several controls were performed: a cut and religated vector at 10ng/plate, uncut vector at 1ng/plate, cut vector at 10ng/plate and untransformed competent cells.

2.15 PREPARATION OF SINGLE-STRANDED M13 DNA FOR SEQUENCING

100ml of 2xTY medium was inoculated with 1ml of an overnight culture of TG2 cells and 1.5ml aliquoted into 7ml bijou tubes. Each tube was then inoculated from a single clear plaque using sterile cocktail sticks and was shaken in a 37°C incubator for 6-7hr. The culture was then transferred to a 1.5ml Eppendorf tube and spun down for 5 min in a microfuge. 1.2ml of the supernatant were carefully removed and transferred to a fresh tube and re-centrifuged. To the supernatant 240μl of 20% PEG/ 2.5M NaCl were added. The tube was shaken and left to stand for 15 min. After centrifugation for 5 min the supernatant was removed and discarded. The tube was re-centrifuged briefly and the remainder of the supernatant was carefully removed with a drawn-out Pasteur pipette and by wiping the sides of the tube with a tissue. The viral pellet was resuspended in 200μl of TE buffer. After a phenol extraction 20μl of 3M NaOAc and 500μl of ethanol were added. The DNA was left to precipitate at -20°C overnight. After centrifugation for 10 min the pellet was washed in 1ml of cold ethanol. The tube was then drained and the pellet allowed to dry. The viral DNA was then redissolved in 30μl of TE buffer and stored at -20°C prior to sequencing.

2.16 ^{35}S SEQUENCING USING SEQUENASE™

This method of sequencing DNA involves standard dideoxynucleotide chain-termination (Sanger et al, 1977) and involves strand extension using the superior polymerase enzyme, Sequenase™ Version 2.0, which is a cloned genetic variant of the T7 DNA polymerase without the 3'-5' exonuclease activity of the wild-type enzyme (Tabor and Richardson, 1987 and 1989). For the sequencing of single stranded M13 clones, approximately 1μg of DNA was used. 0.5 pmol of -40 primer was annealed to the template in 40mM Tris-HCl (pH7.5), 20mM $MgCl_2$ and 50mM NaCl, in 10μl volume, by heating to 65°C for 2 min, then allowing the tube to cool gradually to room temperature. The tube was then placed on ice and the following added:- 1μl of 0.1M dithiothreitol (DTT), 2μl of diluted labelling mix (with 1.5μM

dGTP, 1.5μM dCTP and 1.5 μM dTTP), 0.5μl of [α-^{35}S]-dATP (10μCi/μl) and 2μl of the diluted Sequenase enzyme (in 8.9mM Tris-HCl, pH7.5, 4.4mM DTT and 0.44mg/ml BSA, diluted 1:8 in dilution buffer). This was then mixed thoroughly and then incubated at room temperature for 2-5 min. To 4 tubes labelled G, A, T and C, 2.5 μl of each termination mix was added. The ddG mix contained 80μM dGTP, 80μM dATP, 80μM dCTP, 80μM dTTP, 8μM ddGTP and 50mM NaCl. The ddA mix was the same except for 8μM ddATP instead of ddGTP and similarly 8μM dTTP in the ddT mix and 8μM dCTP in the ddA mix. The tubes were pre-warmed to 37°C for at least 1 min and then 3.5 μl of the labelling mix were added to each tube. The tubes were mixed and briefly spun before incubating at 37°C for 3-5 min. 4μl of stop solution (95% formamide, 20mM EDTA, 0.05% of Bromophenol Blue and Xylene Cyanol FF) were then added to each tube. The reactions were heated to 80°C for several min prior to loading. 3μl of each reaction were then run on a 0.4mm 6% denaturing gel (Sequagel mix) at 60W. Gels were fixed in 10% methanol/10% acetic acid for 15 min and dried at 80°C under vacuum, prior to autoradiography.

2.17 DOUBLE STRANDED PLASMID SEQUENCING

Plasmid DNA to be sequenced directly was prepared by the large scale preparation or more convenient mini preparation. 5μg of plasmid DNA were added to an equal volume of freshly prepared 0.4M NaOH and left at room temperature for 10 minutes. Denatured DNA was precipitated by addition of 0.1 volume 3M sodium acetate and 4 volumes ice-cold absolute ethanol and left at -70°C for 20 minutes. The DNA was recovered by centrifugation in a microfuge for 10 minutes, washed in 70% ethanol and air dried. The pellet was resuspended in 35μl of water and 7μl used in the sequencing reaction (see 2.15).

2.18 TA CLONING

The protocol of Mead et al (1991) was followed with adaptions by Invitrogen. The lyophilized

pCR™1000 vector was resuspended in 44 μl of TE buffer to a final concentration of 25 ng/μl. The PCR products were ligated directly into the vector without undergoing any purification procedures. The ligations were performed in 10μl volume, with 1μl of 10x ligation buffer (included with kit), 2μl of vector (50ng) and 1μl of PCR product and 1μl of T4 DNA ligase (4 units). Ligations were performed overnight at 12°C. The Eppendorf tubes containing the ligations were then spun down briefly, then placed on ice. For each ligation 50μl of competent *Escherichia coli* INVαF' cells (included in the kit and stored at -70°C) was thawed and 2μl of 0.5M β-mercaptoethanol added. 1μl of ligation reaction was then added and mixed gently. This was incubated on ice for 30 minutes, then placed in a 42°C waterbath for 60 seconds and then replaced on ice for 2 mins. To each reaction 450μl of pre-warmed SOC media were then added and was then placed in a shaker-incubator at 37°C and 225rpm for 1 hour. 25μl from each transformation were then plated out on LB agar plates containing kanamycin (50μl/ml) and with 50μl of X-Gal diffused into the surface. Plates were inverted and incubated at 37°C overnight or longer. Completely white colonies tend to be vector rearrangements and genuine recombinants tend to retain a degree of β-galactosidase activity, so white colonies with a tint of blue were selected. Selected colonies were then mini-prepped in 1.5ml LUB containing kanamycin (50μg/ml). DNA was single-stranded and sequencing was performed using ^{35}S and Sequenase™, with forward and/or reverse primers.

2.19 PCR SEQUENCING

This method of sequencing also employs standard dideoxy chain termination (Sanger et al, 1977) and also uses PCR cycling to amplify the signal (Murray, 1989). Using the dsDNA cycle sequencing kit marketed by Gibco BRL approximately 1pmol of the sequencing primer is end-labelled in 5μl volume with 1μl of [$\gamma^{32}P$] ATP (10μCi/μl) and 1μl of T4 polynucleotide kinase (1 U/μl) in 60mM Tris-HCl (pH7.8), 10mM $MgCl_2$ and 200mM KCl. The labelling

mix was incubated at 37°C for 30 min and the reaction was then terminated at 65°C for 5 min. This was then added to the pre-reaction mix, which contained approximately 50fmol of template (which is about 100ng for a 3Kb vector such as pCR1000), 2.5 units of Taq DNA polymerase, 4.5μl of 10x Taq sequencing buffer (300mM Tris-HCl, pH9.0, 50mM $MgCl_2$, 300mM KCl and 0.5%(w/v) W-1), in 36μl volume. To 4 tubes labelled G, A, T and C, 2μl of each termination mix were added. Each termination mix contained 50μM of each, dATP, dCTP, dTTP and 7-deaza-dGTP. Mix-A contained 2mM ddATP, mix-C 1mM ddCTP, mix-G 0.2mM ddGTP and mix-T 2mM ddTTP. 8μl of pre-reaction mix were transferred to each tube with gentle mixing. This was then overlayed with silicon oil. The PCR apparatus was pre-set to 95°C prior to loading. The tubes were then cycled at 95°C for 30 sec, 55°C for 30 sec and 70°C for 60 sec for 20 cycles followed by 95°C for 30 sec and 70°C for 60 sec for 10 cycles. 5μl of stop solution (95% formamide, 10mM EDTA pH8.0 and 0.1% each Bromophenol Blue and Xylene Cyanol) was then added. The samples were heated to 90°C prior to loading and running on a 6% denaturing polyacrylamide gel (as with 2.15). Annealing temperatures for sequencing with octamer primers was reduced to between 32 and 45°C (see Chapter 7).

2.20 POLYMERASE CHAIN REACTION

PCR was performed as follows, with adaptions as described in the results chapters. All reactions were performed using a Perkin-Elmer-Cetus thermal cycler. Taq DNA polymerase was used with 10x incubation buffer either from Amersham or Boehringer. The incubation buffer contained 100mM Tris-HCl (pH8.3), 15mM $MgCl_2$, 500mM KCl and 1mg/ml gelatine. dNTPs were used at 250μM. reactions were usually performed in 25μl volume and overlayed with mineral oil to prevent evaporation. The cycle regime was carried out as described in the results chapters and a final extension at 72°C was performed for 10 min. The product was analyzed on 1% agarose gel and visualized with ethidium bromide incorporation by polaroid photography, or if radiolabelled on 5% polyacrylamide gel and autoradiography.

2.21 5' END LABELLING OLIGONUCLEOTIDES

Oligonucleotides (25ng) were end labelled by T4 polynucleotide kinase in the presence of 120μCi of [γ^{32}P] ATP and 1 x kinase buffer in a total volume of 30μl, in accordance with the procedure of Maxam and Gilbert (1980). After 1hr at 37°C, the labelled probe was separated from the unincorporated isotope by Sephadex G25 (medium) spun column.

2.22 RANDOM PRIMER LABELLING OF PROBES

The method of Feinberg et al (1984) was used to label purified fragment DNA in low melting point agarose. 100ng was denatured by boiling for 7 minutes and quenched on ice before 10μl of labelling buffer, 30μCi of [α^{32}P] dCTP and 2 units of Klenow fragment were added to a final volume of 50ul. The reaction was incubated at 37°C for 1-2hrs. 50ul of TE buffer were added and the unincorporated radionucleotide removed by centrifugation (Beckman GRI) 2500rpm for 5minutes through a Sephadex G50 medium column. The incorporation was measured by Cerenkov counting.

2.23 HYBRIDIZATION OF SHORT OLIGONUCLEOTIDES

All oligonucleotides used in hybridization experiments were 5' end labelled with [γ^{32}P]ATP as described in 2.21 and were hybridized under various conditions, as specified in Chapter 3. Hybridization of short highly degenerate oligos based on splice site consensus sequences was performed as described by Melmer and Buchwald (1992), in 6xSSC, 10mM sodium phosphate pH6.5, 5x Denhardt's solution, 0.2% SDS, 40μg/ml tRNA and 40μg/ml poly(A) and 30% formamide, at room temperature for oligos 469 and 493 and at 37°C for oligo 406. Filters were washed down to a stringency of 2xSSC, 0.1%SDS, at 23°C for 469, 493 and at 40°C for 406.

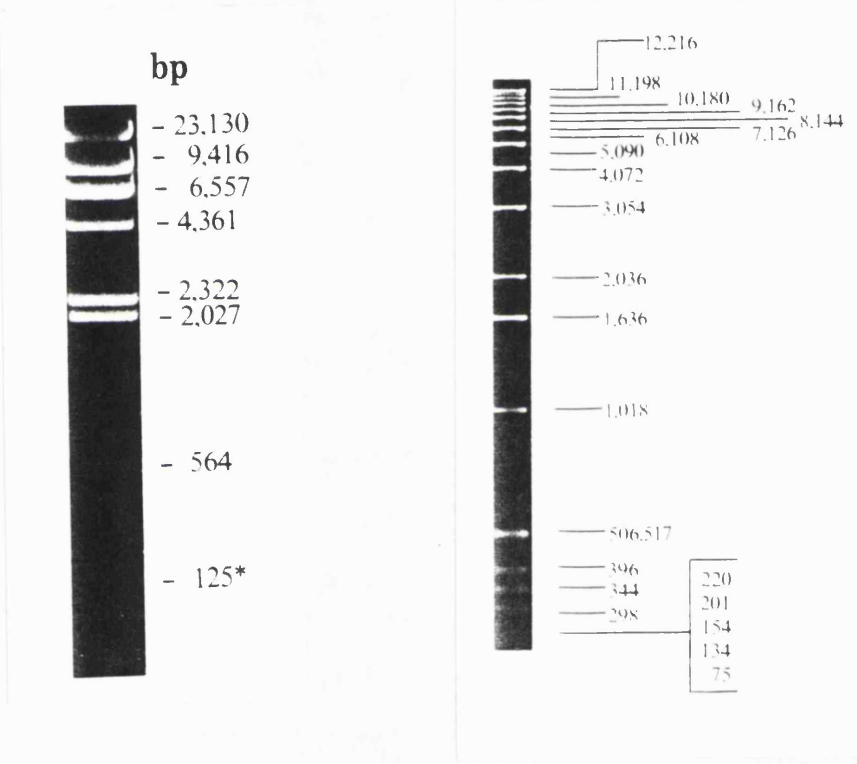
2.24 SIZE MARKERS FOR AGAROSE GELS

Size markers for all agarose gels were either λ/HindIII ladder or 1Kb ladder (Gibco BRL).

Band sizes are listed below:

λ/HindIII: 23,130, 9,416, 6,557, 4,361, 2,322, 2,027, 564, 125

1Kb: 12,216, 11,198, 10,180, 9,162, 8,144, 7,126, 6,108, 5,090, 4,072, 3,054, 2,036, 1,018, 517, 506, 396, 344, 298, 220, 201, 154, 134, 75



(Reproduced from the Gibco/BRL catalogue, 1994)

2.25 NORTHERN BLOTTING AND HYBRIDIZATION

RNA gel electrophoresis was performed according to the procedures of Lehrach et al (1977) and Goldberg (1980). Gel-running buffer (10x) was prepared, consisting of 0.2M morpholinopropanesulphonic acid (MOPS) (pH7), 50mM NaOAc and 1mM EDTA (pH8.0). The gel was prepared by melting the agarose and cooling to below 60°C before adding the gel-running buffer (to 1x) and formaldehyde to 2.2M. The samples (about 10μg RNA) were prepared in 1x running buffer, with 50% (v/v) formamide and 17.5% (v/v) formaldehyde and were incubated at 55°C for 15 min prior to loading. 2μl of loading buffer was added (50% glycerol, 1mM EDTA, 0.4% bromophenol blue, 0.4% xylene cyanol). After electrophoresis the gel was rinsed in DEPC treated water several times, and then denatured in 50mM NH_4OAc/ 10mM NaCl for 45 min. The gel was then neutralized in 0.1M Tris-HCl (pH7.5) for 45 min. The gel was then soaked in 20x SSC for 1hr before blotting onto Hybond in the same manner as with Southern blotting. Northern hybridization was performed using oligolabelled DNA probes at 68°C in 7% SDS, 0.5M PO_4.

CHAPTER 3: HYBRIDIZATION USING SHORT OLIGONUCLEOTIDES

3.1 INTRODUCTION

Rare cutter restriction enzymes that are used for the generation of long range pulsed-field maps cut mammalian genomic DNA infrequently because they usually contain in their recognition sites the unmethylated dinucleotide CpG (Brown and Bird, 1986). This dinucleotide which is under-represented in the human genome, is present in excess in specific regions of the genome known as CpG islands (Bird, 1986, 1987). As described in the introduction DNA segments containing CpG islands also have a high probability of containing or being adjacent to coding sequences (Bird, 1986; Lindsay and Bird, 1987; Gardiner-Garden and Frommer, 1987). Thus selection of clones containing such sites would be useful, not only for linking fragments in long range restriction maps (Poutska and Lehrach, 1986; Smith et al, 1988) but also for the detection of coding regions (Rommens et al, 1989).

A rapid method for the detection of such clones involves the use of a number of 8-mer oligonucleotides based on rare cutter restriction sites, as probes for hybridization (Estivill and Williamson, 1987; Melmer et al, 1990). A clone identified in this manner which hybridizes with several of these oligos has a high probability of containing a CpG island and thus may contain a coding region.

A similar method has also been developed for the detection of coding regions by hybridization of short and degenerate oligonucleotides based on consensus sequences for splice junctions (Melmer and Buchwald, 1992). Since most genes contain introns, the consensus regions around splice junctions are very suitable targets for detecting genes.

3.1.1 AIMS

Preliminary work for the testing of the hypothesis that short oligonucleotides could be used for hybridization against coding regions was carried out, in order to understand more about their capabilities for annealing to target DNA. Optimum annealing conditions were first established for hybridization of the short oligonucleotides so that they could then be used as a basis for PCR.

3.1.2 PROBLEMS USING SHORT OLIGONUCLEOTIDES FOR HYBRIDIZATION

i. Hybridization temperature: temperature for oligo annealing is not only much lower, but also much more sensitive than for larger oligos. For instance a change in temperature of a few degrees could significantly affect duplex stability and if too low might allow annealing to non-complementary target sequence and if too high, might prevent annealing to the correct target. Such duplex instabilities are employed in allele-specific oligomelting experiments, where short oligos corresponding to the wild type and to the mutant sequence are annealed to the target sequence and then the mismatched oligo is melted off at a slightly higher temperature (Tybjoerg-Hansen et al, 1990). The formula of Suggs et al (1981) is often used as a rule of thumb for oligonucleotide hybridization:

$$T_d = \{4x(G+C) + 2x(A+T)\}$$

where T_d is the temperature at which 50% of oligonucleotides dissociate from the DNA. This formula has been determined empirically and holds true for oligos as short as 11-mers (Wallace et al, 1979), but whether this formula holds true for oligos as short as 8-mers is not known. The predicted T_d for a GC 8-mer using this formula is 32°C, so hybridization would need to be performed 5-8°C below this temperature (Suggs et al, 1981). In the experiments performed by Melmer et al (1990), hybridization with the GC 8-mers was performed at 27°C.

ii. Hybridization stringency: stringency of the hybridization would need to be increased to prevent non-specific hybridization by the formation of stable heteroduplex. Stringency can be increased in several ways, namely by increasing hybridization temperature or decreasing salt concentration.

iii. Self-annealing. With the palindromic nature of rare-cutter enzyme sites one would expect oligos based on these sequences to undergo self annealing.

iv. Degenerate oligos: much larger concentrations of oligo would be required to balance out the degeneracy, in order to achieve a sufficient oligo:target sequence ratio.

3.2 TEST HYBRIDIZATIONS OF OLIGONUCLEOTIDES TO MODEL SYSTEMS

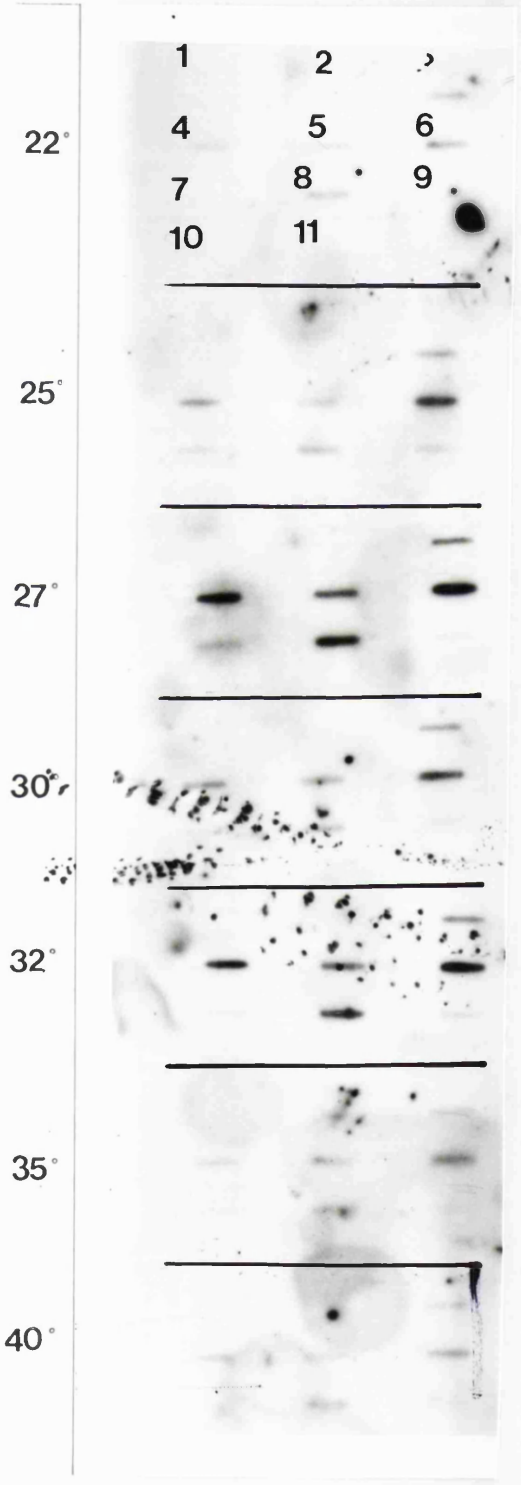
3.2.1 HYBRIDIZATION TEMPERATURE OPTIMIZATION

Slot blot filters were prepared using varying amounts of DNA from the pWE15 vector, which contains two NotI sites (Wahl et al, 1987). Also the vector pMBGPT (constructed by M. Lu, Toronto), which contains one NotI site, was used. As controls the pWE15 vector digested by NotI to destroy the sites, the vector pCV108 from which pWE15 was derived and which contains no NotI site (Lau and Kan, 1984) and the unrelated vector pUC13 were used.

Each filter was prehybridized for 30 mins at 2°C below the hybridization temperature and then hybridized with 5'-end labelled NotI 8-mer (5'GCGGCCGC 3') for one hour at the specified temperature (23, 25, 27, 29, 31 and 33°C) using a refrigeration controlled circulating waterbath (Heto Lab Equipment). All filters were washed down at room temperature (~20°C).

Hybridization occurred up to 33°C, beyond the temperature predicted by the formula of Suggs et al (1981), see Fig. 3.1.

Figure 3.1 Slot blot hybridization of NotI 8-mer oligonucleotide to model vectors at different temperatures. Lanes 1-6 contained SalI digested pWE15 at 0, 10, 50, 100, 200 and 500ng. Lane 7 had 500ng of NotI digested pWE15. Lane 8 had 500ng of EcoRI digested pMBgpt. Lane 9 had 500ng of NotI digested pMBgpt. Lane 10 had 166ng of EcoRI digested pUC13 and lane 11 had 500ng of EcoRI digested pCV108.

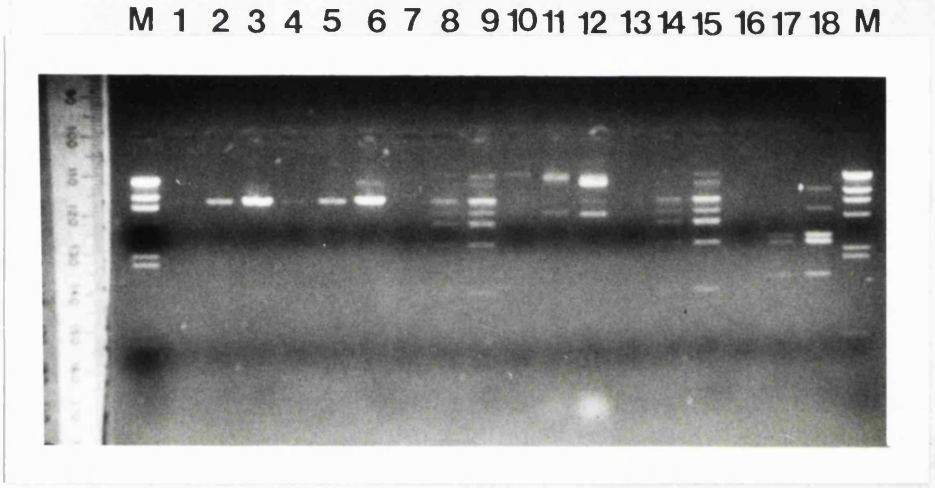


In order to demonstrate that the control pWE15 was being effectively digested by NotI and that the hybridization could discriminate between the NotI target sites and other sites- such as EagI sites- within pWE15 (with high homology to the NotI 8-mer) and to further study the effect of temperature on hybridization, a similar experiment was performed using the vectors as before, either linearized with SalI or EcoRI or digested with NotI or EagI. The DNA was run on an agarose gel to show whether the digestion had worked efficiently (see Fig. 3.2) and then Southern blotted onto nylon filters. A number of filters were produced in this way, all using the same digestion samples. As before the end-labelled NotI 8-mer was used as probe, hybridizing at temperatures between 8°C and 42°C (8, 21, 23, 27, 30, 32, 35 and 42°C) for 16 hours and then washing the filters at 2xSSC at room temperature.

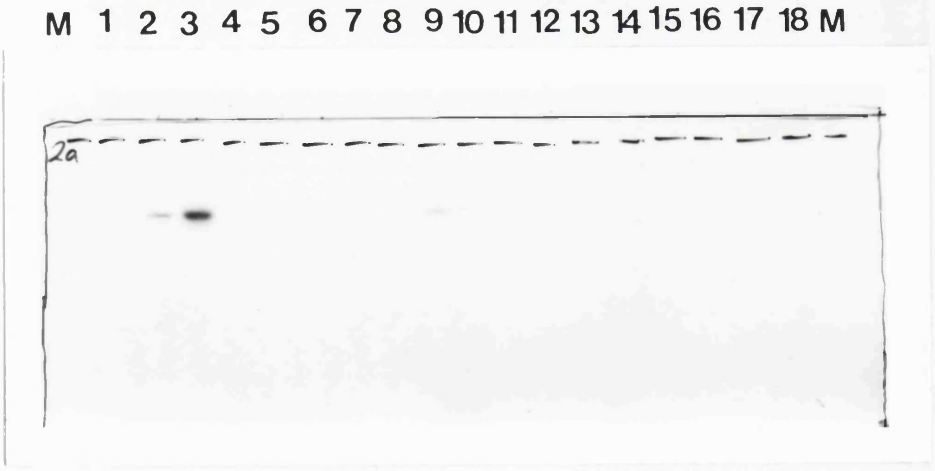
Hybridization occurred at all temperatures used, although specificity was low and background hybridization high at temperatures below 27°C. Specificity and strength of signal was optimum for the 35°C hybridization (3°C higher than the calculated T_d).

Figure 3.2 a. Model vector digests, b. Southern hybridization with NotI 8-mer oligo at 27-27.5°C. Lanes numbered as follows: 1-3 pWE15/SalI digests at 10, 100 and 500ng, 4-6 pWE15/NotI digests at 10, 100 and 500ng, 7-9 pWE15/ EagI digests at 10, 100 and 500ng, 10-12 pCV108/EcoRI digests at 10, 100 and 500ng, 13-15 pCV108/EagI digests at 10, 100 and 500ng, 16-18 pUC13/EcoRI at 3.3, 33 and 167ng, M is λ/HindIII ladder (see 2.24).

a.



b.



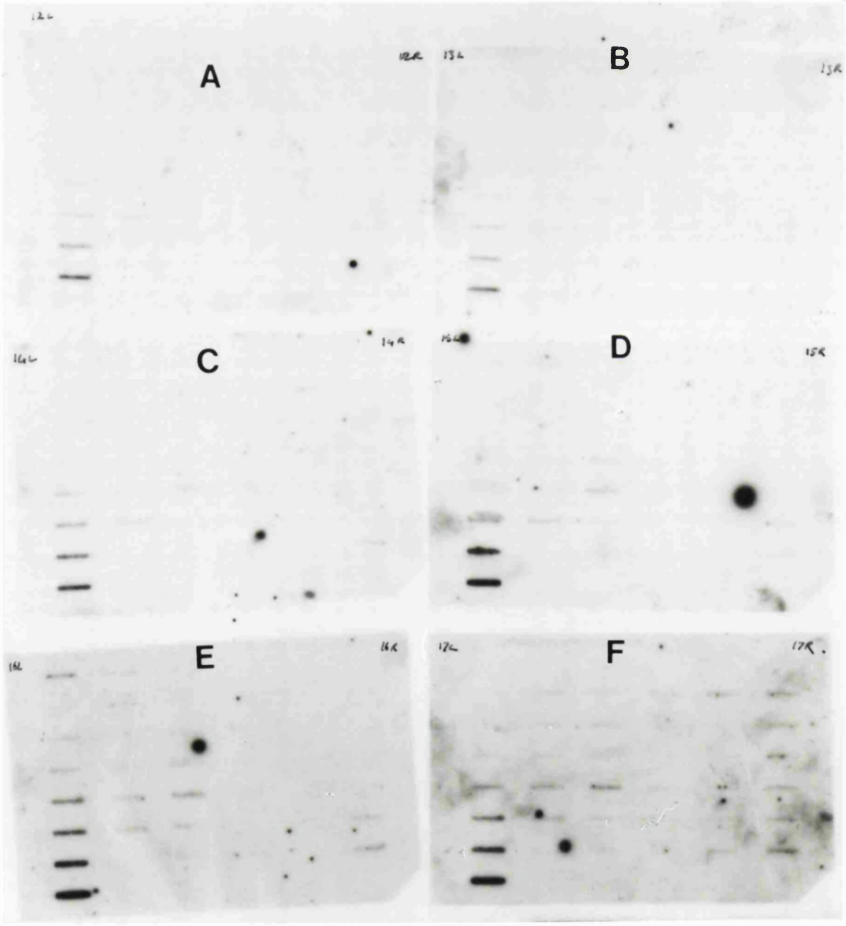
3.2.2 OPTIMIZATION FOR DURATION OF HYBRIDIZATION

Theoretically hybridization times for short oligos should be much shorter than with longer oligos. Kinetic theory would suggest that short oligos anneal much faster, since rate of hybridization of oligo to immobilized DNA follows first-order kinetics and the rate constant is a function of the length and complexity of the oligo (Wetmur and Davidson, 1968; Alwine et al, 1977). The rate of hybridization is also proportional to oligo concentration and because short synthetic oligos are available in much higher molar concentrations than with cloned DNA probes, hybridization times should be much shorter.

Slot blot filters of the various linearized vectors (pWE15, pCV108, pMBGPT and pUC13) at concentrations from 1 to 200ng were used for hybridization with the NotI 12-mer (5'NNGCGGCCGCNN 3'-see Melmer and Buchwald, 1990). The 12-mer was end-labelled and hybridized to the filters at 32°C in 5 ml of 6xSSC, 10mM PO_4 pH6.5, 5x Denhardt's solution, 0.2% SDS, 40μg/ml TRNA and poly A. Hybridizations were performed for varying durations- 5, 20, 60, 90 and 120 mins. Filters were washed at 4xSSC, 0.1% SDS, at 37°C.

Although effective hybridization occurred in just 5 mins, maximum signal was obtained from the 2 hour hybridization (see Fig. 3.3). Even longer hybridizations may increase the signal further.

Figure 3.3 Time course hybridization of NotI 12-mer oligonucleotide to slot blots of model vectors. Hybridization was performed in 6xSSC, 10mM PO_4, 5x Denhardt's solution, 0.2% SDS, 40μg/ml tRNA and poly A, except for filter F, for which 0.9M NaCl, 10% dextran sulphate and 1% SDS was used. Hybridization lengths were A. 5 mins, B. 20 mins, C. 60 mins, D. 90 mins, E. 120 mins and F. 120 mins. The filters were washed for 10 mins in 4xSSC, 0.1% SDS. The first column on each filter contains titrations of the vector pWE15, linearized with SalI, from 0 to 200ng. The second column has titrations of pWE15 digested with NotI, thus destroying the NotI sites, from 0 to 50ng. The third column has titrations of vector pMBgpt linearized with EcoRI, and the fourth linearized with NotI, thus destroying the single NotI site, from 0 to 25ng. The other 2 lanes contained titrations of vectors pCV108 and pUC13 which have no NotI sites.



3.3 HYBRIDIZATION EXPERIMENTS USING DEGENERATE CONSENSUS OLIGONUCLEOTIDES ON MODEL SYSTEMS

Initial studies on the hybridization of degenerate splice site oligos were performed (Melmer and Buchwald, 1992) using as template a number of subclones of the human proteolipid protein (PLP) gene (see Fig. 3.4). PLP is an X-linked gene with seven exons for which all exons and intron/exon boundaries have been sequenced (Diehl et al, 1986). For the Melmer and Buchwald study three oligos were designed (two for the 3'splice site and one for the 5' splice site) based on the computer matrices from consensus sites as compiled by Shapiro and Senapathy (1987):

i. 3'ss 5'YYY YYY YYY YNC AGG (406)

ii. 3'ss 5'YYY YYN YAG (469)

iii. 5'ss 5'NNW GGT RWG T (686)

Based on the predicted base utilization of splice sites (Shapiro and Senapathy, 1987), the highly degenerate Oligo i. (4096-fold) should detect about 4.2% of all acceptor splice sites in the human genome. The shorter but less degenerate (256-fold) 9-mer, Oligo ii, should increase the level of detection of acceptor splice sites to around 36%. The 10-mer 5' splice site oligo (Oligo iii) should detect around 18% of donor sites. Oligo i should only hybridize to the 3' splice site of intron 1, oligo ii (which is less specific than i) to the 3' splice sites of introns 1, 5 and 6 and oligo iii to the 5' splice sites of introns 4, 5 and 6. Against Southern blots of restriction digests of the various PLP subclones, only these expected hybridizations were observed, thus suggesting that these oligos can be used as probes to identify candidate splice sites in genomic DNA.

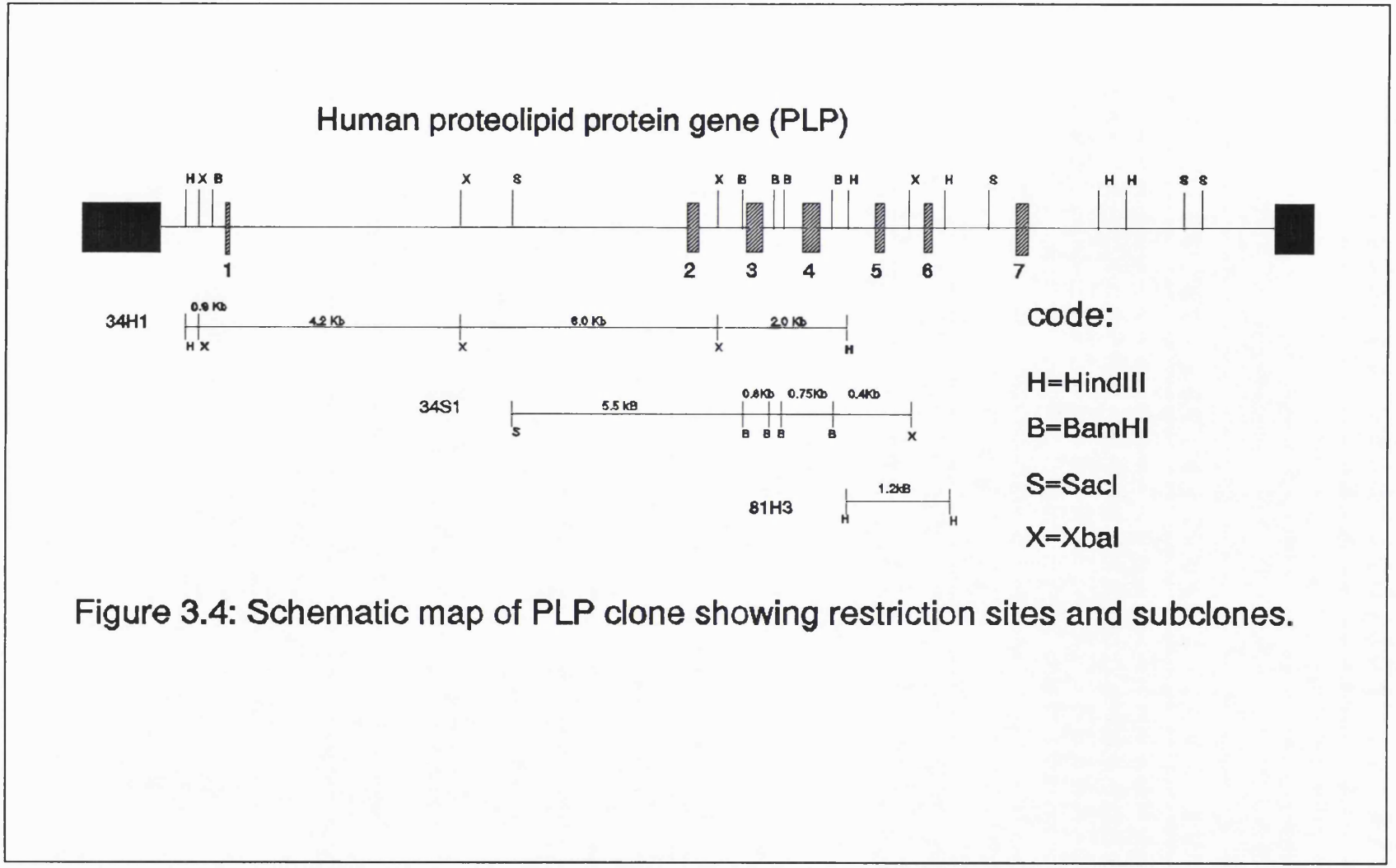


Figure 3.4: Schematic map of PLP clone showing restriction sites and subclones.

3.3.1 SELECTION OF MODEL TARGET DNA

Further to these experiments conducted by Melmer and Buchwald (1992) to identify genes by hybridization of oligos corresponding to consensus splice site sequences, several other model genes were selected to test the hypothesis. These were the cosmid clone cosGSTrp7 containing the glutathione-S-transferase π and NADH-ubiquinone oxidoreductase genes (Cowell et al, 1988; Spencer et al, 1992) and the cosmid clone cosCT1 containing the calcitonin/αCGRP gene (Broad et al, 1989). Both clones have been well characterized and the genes studied and all splice sites junctions have been sequenced.

3.3.2 SELECTION OF HYBRIDIZATION OLIGONUCLEOTIDES

The 3' splice site oligos chosen were those used in the experiments by Melmer and Buchwald (1992):-

406 5' YYY YNC AGG 3' (i.)

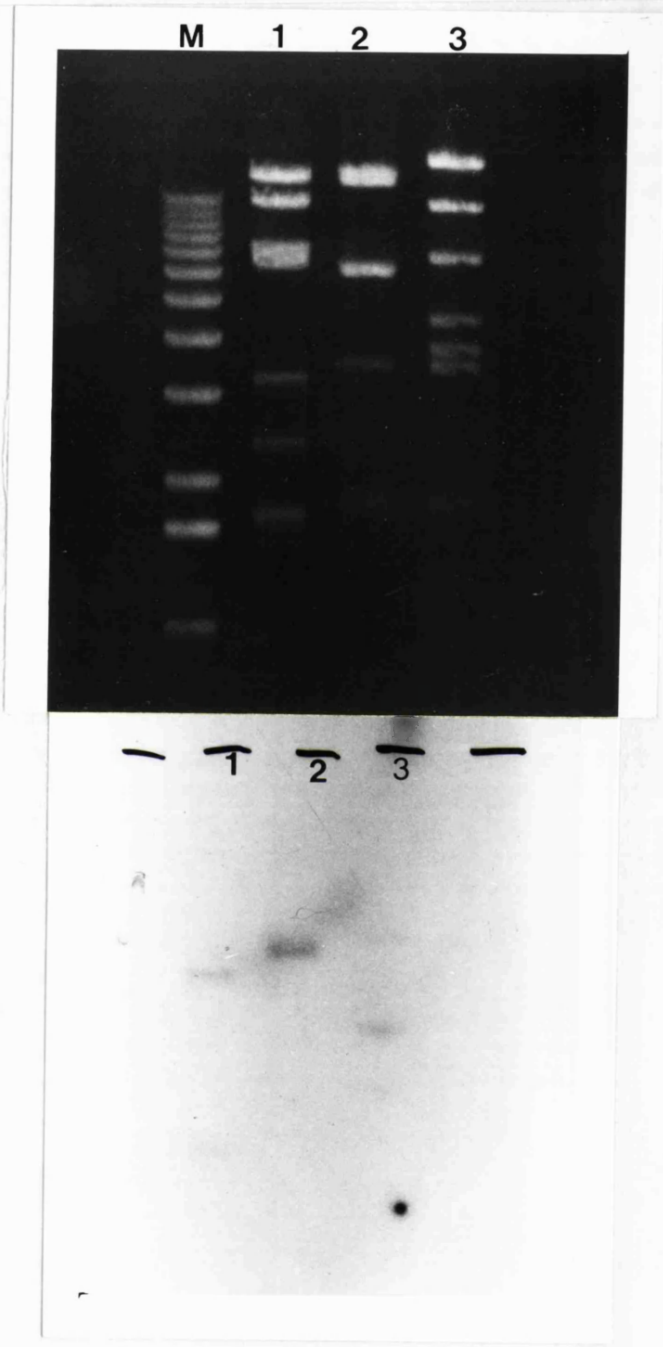
469 5' YYY YYN YAG 3' (ii.)

493 5' NN WGG TRW GT 3' (iii.)

3.3.3 SOUTHERN HYBRIDIZATION OF OLIGONUCLEOTIDES TO RESTRICTION DIGESTED COSGSTRP7 DNA

BamHI, HindIII and BglII digests of cosGSTrp7 were run on 0.8% agarose gels and Southern blotted (see Fig. 3.5). A restriction map of cosGSTrp7 is shown in Fig. 3.6 and shows the positions of the two coding regions relative to the restriction fragments. The oligos were 5'end labelled and hybridized to the Southern blots at room temperature overnight (and at 37°C for oligo 406), in 6xSSC, 30% formamide, 5x Denhardt's solution, 0.2% SDS, 40μg/ml TRNA, 40μg/ml polyA.

Figure 3.5 Restriction digests of cosmid cosGSTrp7 and hybridization to oligo 493. Lane M is 1Kb ladder, lane 1 BamHI, lane 2 HindIII and lane 3 BglI digests.



The filters were washed down to 2xSSC, 0.1% SDS stringency at 23°C (40°C for oligo 406) and exposed to X-ray film. Oligo 406 showed no hybridization under these conditions and so was rehybridized in 6xSSPE, 5x Denhardt's solution, 0.5% SDS, at room temperature. The

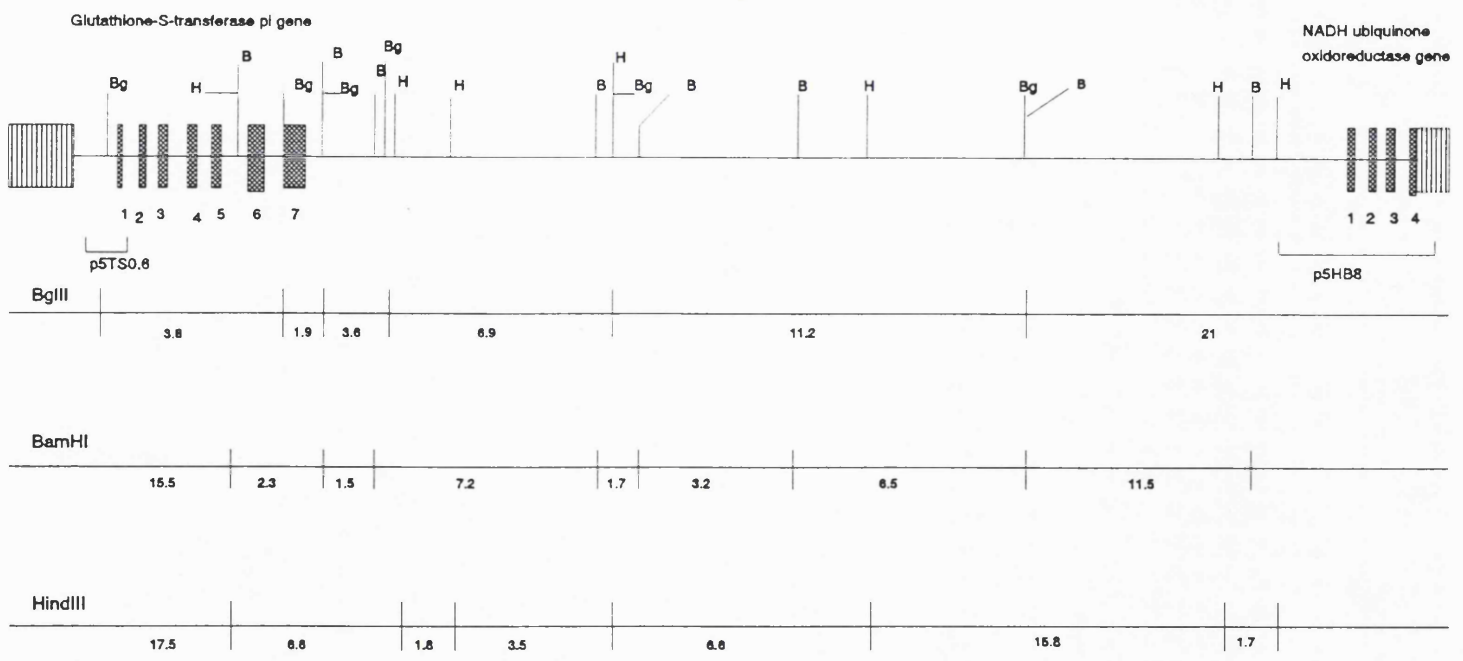


Figure 3.6: Schematic map of cosmid cosGSTrp7 showing the position of the glutathione-s-transferase and NADH-ubiquinone oxidoreductase coding regions and subclones p5TS0.6 and p5HB8, and restriction map.

resulting band-hybridizations observed are shown in Table 3.1 below:

3'splice site 406	3'splice site 469	5'splice site 493
BamHI 15.5+11.5Kb	BamHI 15.5Kb	BamHI 15.5+11.5Kb
HindIII 17.5Kb	HindIII 17.5Kb✓✓	HindIII 17.5Kb
BglII 20Kb	BglII 11.2Kb	BglII 11.2+6.9Kb

Table 3.1 Positive hybridization of consensus splice site oligos to cosGSTrp7 digests. ✓✓ = very strong signal.

Tables 3.2 and 3.3 show the homologies of the oligos to the splice site sequences of glutathione-S-transferase π gene and NADH-ubiquinone oxidoreductase gene. As can be seen from Table 3.2, expected hybridizations should occur with oligos 406 and 469 against acceptor 2. Oligo 493 shows 100% homology only with donor 4, but has 9 out of 10 bp homology with donor 6. Thus hybridization should occur between oligos 406, 469 and 493 and the BglII 3.8Kb fragment, as well as the BamHI 15.5Kb and HindIII 17.5Kb fragments. However, only the latter two of these expected hybridizations are observed, along with hybridization to several other bands (BamHI 11.5 Kb and BglII 11.2 and 6.9Kb fragments). For hybridizations with the NADH-ubiquinone oxidoreductase gene, the first four exons of which occur right at the 3' end of the cosmid insert (see Fig. 3.6), as can be seen from Table 3.3, expected hybridizations should occur with oligo 469 against acceptors 1 and 3, but 406 shows at best only 13 out of 15 bp homology (acceptor 2). Oligo 493 shows 100% homology with donor 3. Thus hybridization should occur between oligos 469 and 493 and the BglII 20Kb fragment, as well as the BamHI 15.5Kb and HindIII 17.5Kb fragments. The latter two of these hybridizations are observed (see Table 3.1).

Table 3.2 a.

406	5'yyy yyy yyy ync agg 3'
acceptor 1	**AGc Gcc tcG GgG** agg
acceptor 2	ctc cct ccc cgc agg
acceptor 3	**Gcc tcc ccc Aac agC**
acceptor 4	**ccA ccc AAc** ccc agg
acceptor 5	**GtG GtG** tct Ggc agg
acceptor 6	**GGc ctG** ccc tgc **agA**

b.

469	5'y yyy yny ag 3'
acceptor 1	c **tcG GgG** ag
acceptor 2	t ccc cgc ag
acceptor 3	c ccc **Aac** ag
acceptor 4	c **AAc** ccc ag
acceptor 5	**G** tct Ggc ag
acceptor 6	**G** ccc tgc ag

c.

493	5'nn wgg trw gt 3'
donor 1	ac **CAg** tga gt
donor 2	cg agg tag **gA**
donor 3	ct **GCg** taa gt
donor 4	ct tgg tga gt
donor 5	ct **aTg** tgt **gA**
donor 6	cc agg tga **gC**

Table 3.2 Splice sites in glutathione-S-transferase π gene and their homologies to degenerate splice site oligos, a. 406, b. 469 and c. 493. Bold typed capitals indicate where mismatches with the oligo are expected.

Table 3.3 a

406	5'yyy yyy yyy ync agg 3'
acceptor 1	ctt tGt ctc cct agA
acceptor 2	cct Att ctG tcc agg
acceptor 3	GtG GGc ccc tgc agg

b.

469	5'y yyy yny ag 3'
acceptor 1	t ctc cct ag
acceptor 2	t ctG tcc ag
acceptor 3	c ccc tgc ag

c.

493	5'nn wgg trw gt 3'
donor 1	ca Cgg tga gG
donor 2	gg agg tga gA
donor 3	gc agg tgt gt

Table 3.3 Splice sites in NADH-ubiquinone oxidoreductase and their homologies to the degenerate splice site oligos, a.406, b. 469 and c.493. Bold typed capitals indicate where mismatches with the oligo are expected.

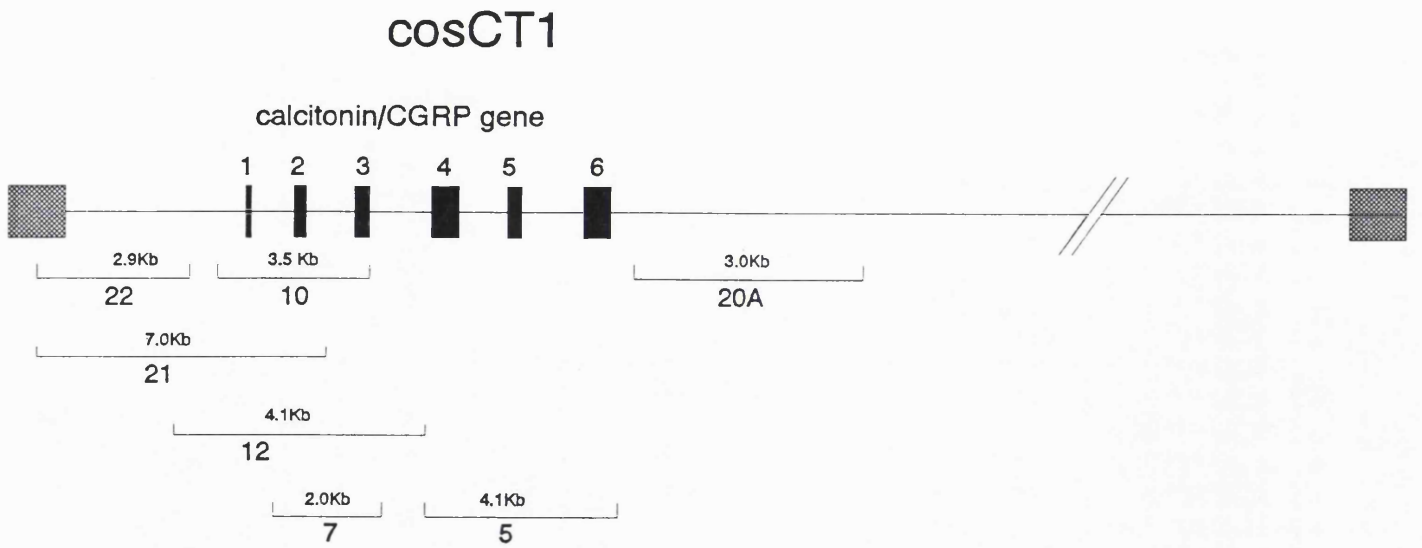
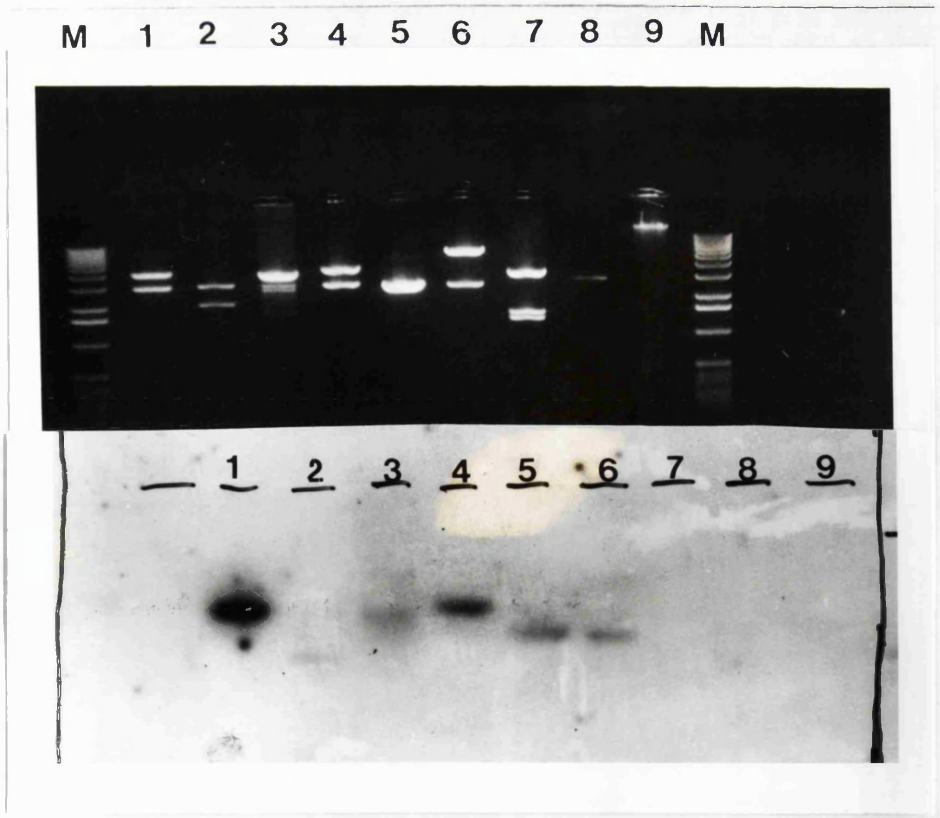


Figure 3.7: Schematic map of cosmid cosCT1, showing the calcitonin/CGRP coding region and the relative positions of various subclones in the vector pGem1.

3.3.4 SOUTHERN HYBRIDIZATION OF OLIGONUCLEOTIDES TO RESTRICTION DIGESTED COSCT1 SUBCLONES

Restriction digests of cosCT1 were prepared, but the cosmid DNA was found to be too degraded. A number of subclones covering most of the coding region were used instead (see Fig. 3.7). The inserts were excised from the parent vector (pGem1) and run on a 1% agarose gel and Southern blotted (see Fig. 3.8).

Figure 3.8 Restriction digests of pGem subclones of the calcitonin cosmid cosCT1 and hybridization to oligo 406. The lanes are as follows: Lane 1. subclone pGemCal5/PstI (4.1Kb insert), 2. pGemCal7/PstI (~2.0Kb), 3. pGemCal10/BamHI (3.5Kb), 4. pGemCal12/XbaI/EcoRI (4.1Kb), 5. pGemCal20A/BamHI (3.0Kb), 6. pGemCal21/BamHI (~7Kb), 7. pGemCal22/PstI (1.4 and 1.5Kb), 8. pGem1/BamHI (vector-2.9Kb), 8. cosCT1, M 1Kb ladder.



The oligos 406, 469 and 493 were hybridized to the blots as described in 3.3.3. The resulting hybridizations observed are summarized in Table 3.4, below:-

3'splice site 406	3'splice site 469	5'splice site 493
GemCal5 ✓✓	GemCal5, v	GemCal5
GemCal7	GemCal7, v	GemCal12
GemCal10	GemCal10, v	GemCal20A
GemCal12✓	GemCal12, v	GemCal21, v
GemCal20A	GemCal20A	GemCal22
GemCal21 v	GemCal21, v	cosCT1
(cosCT1)	GemCal22, v	
	cosCT1	

Table 3.4 shows to which cosCT1 subclones hybridization has occurred. ✓=strong hybridization, ✓✓=very strong hybridization, ()=weak hybridization and v=hybridization to vector.

Table 3.5 indicates the various donor and acceptor splice junctions found in the calcitonin/αCGRP gene and shows where mismatches will occur with the consensus splice site oligos 406, 469 and 493.

Table 3.5 a.

406	5'yyy yyy yyy ync agg 3'
acceptor 1	ttc ttc cct tgc agA
acceptor 2	ctt ccc ctc cac agg
acceptor 3	tGt ttt ccc tgc agC
acceptor 4	Atc ctG cAA Atc agA
acceptor 5	ttt ctt cct ctT agg

b.

469	5'y yyy yny ag 3'
acceptor 1	c cct tgc ag
acceptor 2	c ctc cac ag
acceptor 3	t ccc tgc ag
acceptor 4	**G cAA Atc** ag
acceptor 5	t cct ctt ag

c.

493	5'nn wgg trw gt 3'
donor 1a	cc agg tga gc
donor 1b	tc agg tat **At**
donor 2	tc agg taa gA
donor 3	cc agg tga gG
donor 4	cc **CCA aaG At**
donor 5	ga agg tga **Ct**

Table 3.5 Splice sites in the calcitonin/αCGRP gene and their homologies to degenerate splice site oligos a. 406, b. 469 and c. 493. Bold typed capitals indicate where mismatches with the oligo are expected.

From the map of cosCT1 and its subclones (Fig. 3.6) and Table 3.5 it is possible to see to which subclone inserts the oligos ought to hybridize. 406 should hybridize to the acceptor 2 splice site, which is present in subclones GemCal7, 10 and 12. 469 should hybridize to acceptors 1, 2, 3 and 5 which are present in GemCal5, 7, 10, 12 and 21. 493 does not correspond exactly to any donor sites, although it has 90% homology with donors 1a, 1b, 2, 3 and 5 and may thus hybridize weakly to the subclones. However, as can be seen from Table 3.4, many other hybridizations occur.

3.3.5 SOUTHERN HYBRIDIZATION OF OLIGONUCLEOTIDES TO PLP SUBCLONE RESTRICTION DIGESTS

Given the low success rate using the three splice site oligos to detect splice junctions within cosGSTrp7 and cosCT1 subclones, it was decided to repeat the hybridizations using some of the PLP subclones, to see whether the results published by Melmer and Buchwald were a realistic reflection of the capabilities of the method. The positions of the PLP exons are shown relative to a restriction map and the relevant subclones in Fig. 3.4. The insert of the subclones 34H1, 81H3 and 34S1 were excised by restriction digestion (with HindIII, HindIII and SacI, respectively). 34H1 was further digested with XbaI and 34S1 with BamHI. The resulting fragments were run on a 1% agarose gel (see Fig. 3.9) and Southern blotted. Oligos 406, 469 and 493 were hybridized as in 3.3.2 and 3.3.3. Table 3.6 shows to which fragments the oligos hybridized to. Table 3.7 indicates the various donor and acceptor splice junctions found in the PLP gene and shows where mismatches will occur with the consensus splice site oligos 406, 469 and 493.

Figure 3.9 Restriction digests of proteolipid protein PLP subclones and hybridization to oligo 493. Lanes are as follows: 1. 34H1/XbaI/HindIII, 2. 81H3/HindIII, 3. 34S1/SacI, 4. 34S1/SacI/BamHI and M= 1Kb ladder.

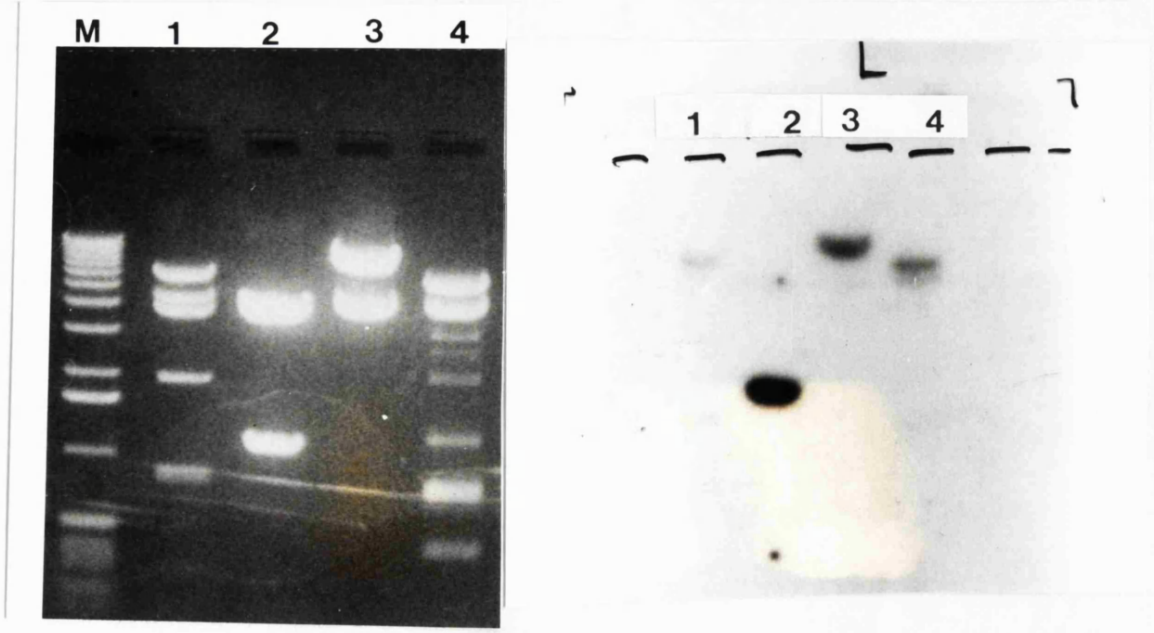


Table 3.6 Positive hybridization of oligos to PLP fragments. () indicates weak signal, v= hybridization to vector band.

Subclone digest	3'splice site 406	3'splice site 469	5'splice site 493
34H1 (HindIII/XbaI)	0.6, (2), 4.2, 6Kb	6Kb	6Kb, (0.9Kb)
81H3 (HindIII)		1Kb✓✓ (v)	1Kb✓✓
34S1 (SalI)	7kb	7Kb✓✓ (v)	7Kb
34S1 (SalI/BamHI)	5.5Kb	5.5, 0.8Kb, (v)	5.5Kb, (0.8Kb)

Table 3.7 a.

406	5'yyy yyy yyy ync agg 3'
acceptor 1	ttc ttc ttc ccc agg
acceptor 2	Acc tGt tAA tgc agg
acceptor 3	Gtc AAt cAt ttt agT
acceptor 4	tGt Gtc ttA ctt agg
acceptor 5	tcA ttt tcc tgc agT
acceptor 6	tct Gtt ccc tgc agC

b.

469	5'y yyy yny ag 3'
acceptor 1	c ttc ccc ag
acceptor 2	t tAA tgc ag
acceptor 3	t cAt ttt ag
acceptor 4	c ttA ctt ag
acceptor 5	t tcc tgc ag
acceptor 6	t ccc tac ag

c.

493	5'nn wgg trw gt 3'
donor 1	at Ggg taa gt
donor 2	at GTg taa gt
donor 3	ca agg tga TC
donor 4	ta tgg tga gt
donor 5	tg agg tga gt
donor 6	cc tgg tga gt

Table 3.7 Splice sites in the PLP gene and their homologies to degenerate splice site oligos a. 406, b. 469 and c. 686. Bold typed capitals indicate where mismatches with the oligo are expected.

Oligo 406 shows full complementarity only to acceptor 1 and should therefore only hybridize to the 6Kb fragment from the 34H1 HindIII/XbaI digest, the 7Kb insert of 34S1 (SalI) and the 5.5 Kb fragment of the 34S1 SalI/BamHI digest. 406 hybridizes to these fragments and to several others. Oligo 469 should recognize acceptor 1, 5 and 6 and should therefore hybridize to the 6Kb fragment of 34H1 (H/X), the 1Kb insert of 81H3 (H), the 7kb fragment of 34S1 (S) and the 5.5kb fragment of 34S1 (S/B). Whilst 469 recognized these four bands as predicted, it also bound weakly to a ~0.8Kb 34S1 (S/B) fragment and to the vector band. Oligo 493 shows full complementarity with donors 4, 5 and 6 and so should hybridize with 34H1 (X/H) 2Kb, 81H3 (H), 34S1 (S), 34S1 (S/B) 0.4 and 0.75Kb. Whilst 493 correctly hybridized to 81H3 (H) and 34S1 (S), it also hybridized to the 6Kb 34H1 (H/X) and 5.5Kb 34S1 (S/B) fragments, but not to 34S1 (S/B) 0.4Kb and 34H1 (H/X) 2Kb.

3.4 DISCUSSION

The experiments discussed in 3.2 show that the short 8-mer oligo based on the NotI site is able to hybridize to target DNA over a fairly wide range of temperatures and at temperatures much higher than that predicted by the formula of Suggs et al (1981). The hybridizations were also able to discriminate between correct target site, present only in pWE15 and pMBgpt and sites of 6 and 7 base pair homology present in pWE15 (the positions of these sites are shown in Chapter 4.7) and pCV108.

Wallace et al (1979) showed that it is possible to discriminate perfect hybrids from ones containing a single internal mismatch with oligos 11-17 nucleotides long. From the evidence shown in 3.2 the same applies to 8-mer oligos with mismatches either internally or at the ends. According to Drmanac et al (1990) shorter oligos give better discrimination for hybridization because the relative decrease in hybrid stability with a single mismatch is greater than for longer probes (T_m for a 20-mer with 1 mismatch decreases by 5-7.5°C; 1-

1.5°C per 1% of oligo mis-paired).

The experiments using degenerate splice site oligos on model genes were less successful. The oligos were unable to distinguish the correct sites (with full complementarity to known splice sites) from other sites in the clones. This could be partly blamed on the low stringency of the hybridizations and washings. However, a number of expected hybridizations were not observed, notably the 3.8Kb BglII fragment of cosGSTrp7, which should hybridize with oligos 406, 493 and 469, and the 2Kb insert of 34H1, which should hybridize with 493. Other inconsistencies are also observed, such as the hybridization of oligo 406 to the cosGSTrp7 BamHI 11.5 fragment, but not to the corresponding fragments of the other digests. If it is assumed that the restriction map of cosGStrp7 is correct, then doubts over the reliability of such oligos are inevitable. Inconsistencies have since been observed in the BglII restriction map of cosGSTrp7.

The fact that the degenerate oligo will have a wide range of T_d- depending on which bases occur at the degenerate positions in any sub-species of the oligo- could be partly responsible for this unreliability. For instance oligo 406 will have T_d between 36 and 78°C, 469 will have T_d between 20 and 34°C and 686 will have T_d between 30 and 38°C, according to the formula of Suggs et al (1981). The ideal temperature for hybridization will depend on the sequence of the target splice sites, but since these sequences would be unknown in an experimental situation, the hybridization would have to be performed at an approximated average temperature. This hybridization at just one temperature may suit only a small fraction of the degenerate oligos and might encourage hybridization of other sub-species of the oligo under non-optimum conditions, thus producing background signal. However, this is insufficient to explain the lack of observation of several of the expected hybridizations. The hybridizations were performed under low stringency conditions (high salt concentration and at room

temperature) and should thus have promoted all the expected hybridizations, although rate of hybridization would probably be well below optimum at such relaxed stringency.

Many of the problems experienced using degenerate oligos could be overcome by substituting the degenerate positions in the oligos with base homologues that have ambiguous pairing abilities. For instance 5-fluorodeoxyuridine could be used to pair with A or G (Habener et al, 1988) and could be used to replace pyrimidine degeneracies (C or T). Deoxyinosine which can pair with A,T,G or C (although more so with A or C) could be used to replace either fully degenerate bases (N) or keto-bases (G or T)(Martin et al, 1985; Ohtsuka et al, 1985; Takahashi et al, 1985). Deoxyguanosine could be used at positions of two-fold degeneracy between purines G and A, because the G-T base pairing is thought to be one of the most stable mismatches (Millican et al, 1984). In this way the consensus sequences could be rationalized to allow oligonucleotides with much less base degeneracy. Such an oligonucleotide would be much easier to optimize hybridization conditions for. However, of these base homologues, only deoxyinosine is commercially available for incorporation into custom synthesized oligonucleotides.

Another type of base substitution has been shown to increase the stability of homoduplex without increasing heteroduplex stability. The incorporation of 5-bromodeoxycytosine or 5-methyldeoxycytosine in place of deoxycytosine bases in oligos has been shown to achieve this effect (Hoheisel et al, 1990). Thus ^{5Br}C or ^{5Me}C for C substitutions could be made at non-degenerate positions to increase the overall stability of an oligo for hybridization. Also hybridization in tetramethylammonium chloride (TMAC) has been shown to eliminate the preferential melting of A-T versus G-C base pairs (Wood et al, 1985), causing the stringency of hybridization to be a function of probe length rather than base content, thus making the melting temperature of degenerate oligos much more uniform amongst multiple heterogeneous

target sites for hybridization.

CHAPTER 4: PCR USING SHORT GC OLIGONUCLEOTIDES

4.1 INTRODUCTION

It has been shown that short and degenerate oligonucleotides can be hybridized in a specific manner to cloned DNA (Estivill and Williamson, 1987; Melmer et al, 1990; Melmer and Buchwald, 1990; Melmer and Buchwald, 1992). The ability to use such oligonucleotides as primers in PCR amplifications would have important benefits for screening complex genomes. A PCR based approach would greatly increase the sensitivity, requiring only small amounts of target DNA and would speed up the detection process by allowing simultaneous PCR amplification using many primers and or many target clones or genomes. The alternative using the hybridization approach often requires lengthy autoradiography. In this chapter the use of short GC rich primers based on rare-cutter restriction sites is investigated, using model templates, in order to identify the reaction conditions necessary for specific amplification.

4.11 THE POLYMERASE CHAIN REACTION

The polymerase chain reaction (PCR) is an *in vitro* method of producing large quantities of a specific DNA fragment and since its development by Saiki et al (1985), has revolutionized molecular biology (White et al, 1989; Erlich, 1989). PCR is the enzymatic amplification of a fragment of DNA using two oligonucleotide primers targeted at sequences flanking the fragment. The primers anneal to their complementary sites, one on each strand and a DNA polymerase extends the primers across the target fragment. Following this the resulting strands are separated by heat denaturation and the primers are allowed to anneal again, to both the new strands and the original template strands. A second enzymatic extension results

in the synthesis of a complete DNA fragment, between the two primer sites- an identical copy of the template DNA. Further cycles of heat denaturation, primer annealing and enzymatic primer extension results in the exponential amplification of the DNA fragment. The principle of PCR is summarized in diagram form in figure 4.1.

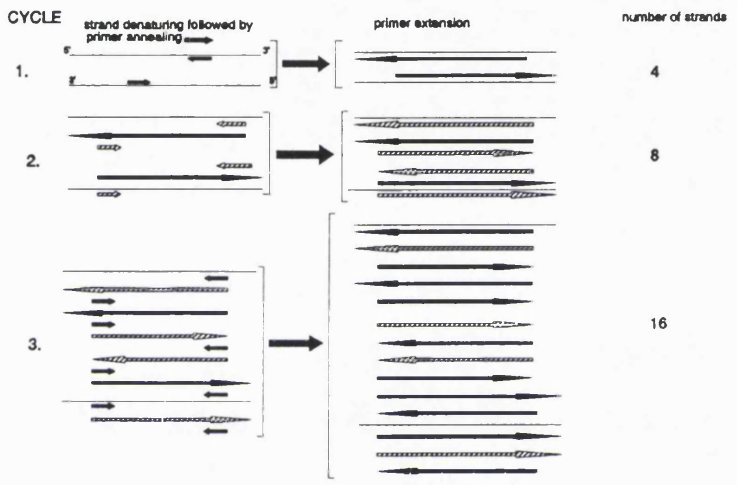


Figure 4.1: The principle of the polymerase chain reaction, showing the first 3 cycles.

The exponential amplification of a target sequence can result in a several million-fold increase of DNA. By replacing the Klenow fragment of DNA polymerase I from E.coli with the thermostable enzyme Taq DNA polymerase from the thermophilic bacterium Thermus aquaticus, it is no longer necessary to replace the enzyme in the reaction after each heat denaturation step (Saiki et al, 1988; Kogan et al, 1987) and so PCR can be easily automated. Taq DNA polymerase is active at much higher temperatures because the optimum activity is around 72°C, which allows the annealing and extension steps to be performed at higher temperatures, increasing the specificity of the reaction without harming the enzyme activity.

4.1.2 THE USE OF SHORT AND DEGENERATE OLIGONUCLEOTIDES AS PRIMERS

There are occasions when primer specificity is not required in PCR. For instance for the detection of new genes related to known gene families and the fingerprinting of cosmid or YAC clones (Orlandi et al, 1989; Kinzler and Vogelstein, 1989; Caetano-Anolles et al, 1991). Less specific amplification can be achieved either by decreasing the size of the primers, using primers containing degenerate positions or by decreasing the stringency during the reaction. This thesis aims to test the hypothesis that primers aimed at rare-cutter restriction sites and coding region-specific consensus sequences can be used to prime PCR amplification. In order to amplify a target DNA in a less specific manner through PCR, using sequence information derived from restriction sites or consensus sequences as a basis for primer design, various factors must be taken into consideration:-

A. PRIMER DEGENERACY

If primers with degenerate bases are used, their concentration in the PCR should be increased proportionately to counter the dilution effect. If, however, the primer concentration in the PCR is too high, it will inhibit amplification. Alternatively degenerate positions may be replaced by base analogues which can pair well with all four natural bases, such as deoxyinosine (Takahaski et al, 1985), or 2-amino-2'-deoxyadenosine (Dinh et al, 1985) and pyrimidine degeneracies (C or T bases) can be replaced by 5-fluorodeoxyuridine (Habener et al, 1988). Such analogues have been used successfully in hybridization oligonucleotides. The use of degenerate oligonucleotides as hybridization probes for the detection of genes related to known gene families has many limitations, in particular establishing hybridization conditions that can distinguish authentic signals from spurious signals is difficult and tedious. The use of degenerate primers coupled with PCR can help overcome many of these limitations and has been used to identify a number of related genes (Nurnberg et al, 1989).

Deoxyinosine has also been used successfully in degenerated PCR primers (Knoth et al, 1988; Patil and Dekker, 1990).

B. SHORT PRIMERS

Since restriction sites and consensus sequences can only provide a small amount of sequence information, PCR primers based on such sequences will be short and will require low annealing temperatures, possibly at temperatures at which the activity of Taq DNA polymerase is severely attenuated. If the temperature is increased to a level at which chain extension can occur, the primers may dissociate from the template DNA before they can be extended enough to form a stable duplex. Experiments using Taq DNA polymerase for sequencing reactions at varying temperatures shows that Taq polymerase retains some activity, even as low as 22°C (Innis et al, 1988), extending one base every four seconds, compared to 1.5 bases per second at 37°C and 24 bases per second at 55°C. Thus at 22°C an octamer would take only 96 seconds to extend to a 20'mer, which could exist as a stable duplex at the normal extension temperature of 72°C.

Experiments by Nevinsky et al (1990) studied the thermodynamic and kinetic properties of oligonucleotides of different lengths in the priming of DNA synthesis. Four different types of DNA polymerase from a wide variety of sources were used, although none from thermophilic bacteria. It was shown that the ability to prime chain elongation is not limited by primer length and that in fact mononucleotides can prime synthesis, although the efficiency of an oligonucleotide to act as a primer appears to become optimal at around 9-12 bases long. These findings are consistent with the assumption that the active cleft of DNA polymerases is of a size compatible to a chain 9-12 nucleotides long. This cleft size was similar for all four DNA polymerases studied, despite the diversity of sources from which the enzymes were derived. This is also consistent with DNA footprinting studies which have indicated that the

Klenow fragment covers 8 bases of primer and 19-20 bases of the template DNA strand (Joyce et al, 1986) and also with the minimal length for optimal priming of 8 nucleotides as established by Fisher and Korn (1981). If these facts follow for DNA polymerases from a wide variety of sources, it would be justifiable to at least begin experiments on the assumption that such primer lengths would be compatible with Taq DNA polymerase.

C. ANNEALING STRINGENCY

The stringency of primer annealing is directly dependent upon the salt concentration and temperature of the reaction. If the stringency is reduced enough a primer may be forced to anneal at non-complementary sites. Hadano et al (1991) took advantage of this in order to generate libraries from microdissected regions of a single chromosome, using a single PCR primer under low stringency conditions of low temperature and high salt concentration (22°C annealing, 75mM Kcl) to amplify randomly across the selected region.

The most obvious problem associated with the use of short oligonucleotides is that of specificity. PCR primers are generally selected on the basis that they will be complementary to only one site in the target genome. Given that the size of the genome is approximately 3 x 10^9 bases long, with two strands, a primer of 17 nucleotides long should, following the laws of probability, occur less than once in the genome ($4^{17} = 1.72 \times 10^{10}$). In general the literature concerning PCR recommends a minimum size of 17 bases and preferably 20 to 24 bases long to limit the chances of amplifying the wrong region (Sambrook et al, 1989). However, because PCR can only amplify between primers at most several kilobases apart, such specificity of primers may not be so important. In addition, when the target genome is much smaller, for instance the size of a cosmid clone (approximately 45000bp), the size of primer need only be between 8 and 9 bases to achieve specific priming. Studier (1989) calculated that an octanucleotide primer would have an average of 1.37 sites in a cosmid of 45000bp and a

nonamer would have 0.34 sites. This fact could have useful implications for mass sequencing projects (Siemieniak and Slightom, 1990) and is further discussed in Chapters 1 and 7.

4.2 SELECTION OF MODEL TEMPLATES

The cosmid cloning vector pWE15 which was designed for rapid genomic walking and restriction mapping (Wahl et al, 1987) was used as the basis for many of the PCR amplification experiments using short GC rich primers. The vector contains two NotI sites just 42bp apart, with a BamHI site in the middle. By using primers based on the rare-cutter site NotI, with the aim of amplifying between the NotI sites the vector pUC19 was digested with Sau3A and the four largest fragments of 955, 585, 341 and 258bp, were separated and cut out from a low melting point agarose gel, purified and then cloned into the BamHI site. In order to prevent the NotI oligos priming amplification of the entire vector genome rather than the insert, the resulting subclones, shown in Fig.4.2, were linearized prior to PCR amplification, by digestion with SalI. At a later stage it was realized that SalI cuts within the 955 bp insert, thus preventing amplification and so linearization was performed by digestion with SmaI instead. Use of these constructs as model template for PCR using short primers based on the NotI restriction site is also a good test for specificity, since pWE15 also contains an EagI site that has 7 out of 8 bp homology with NotI.

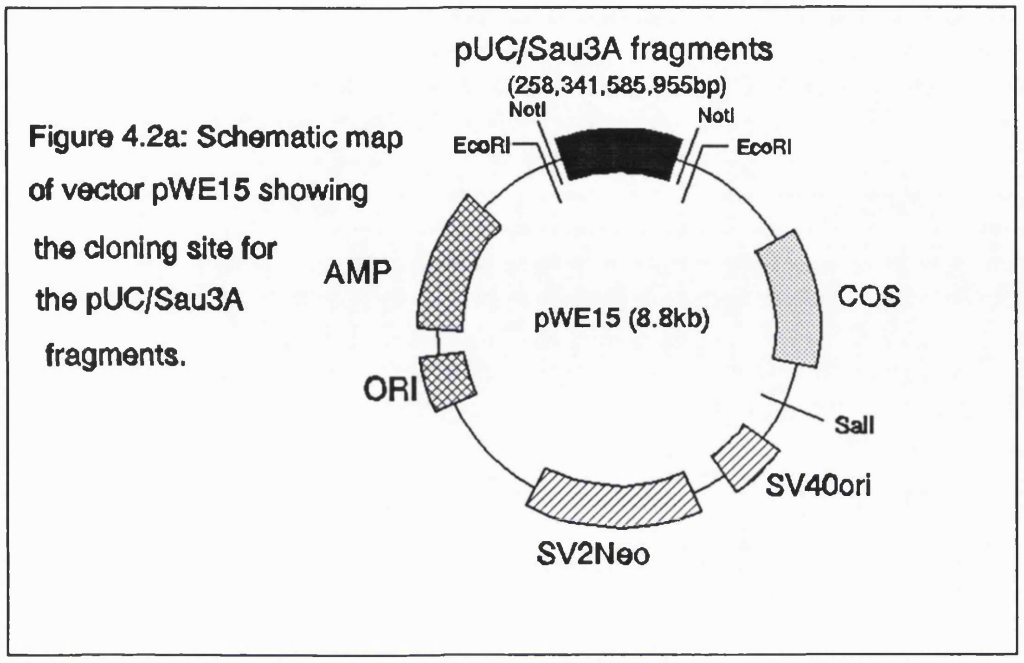
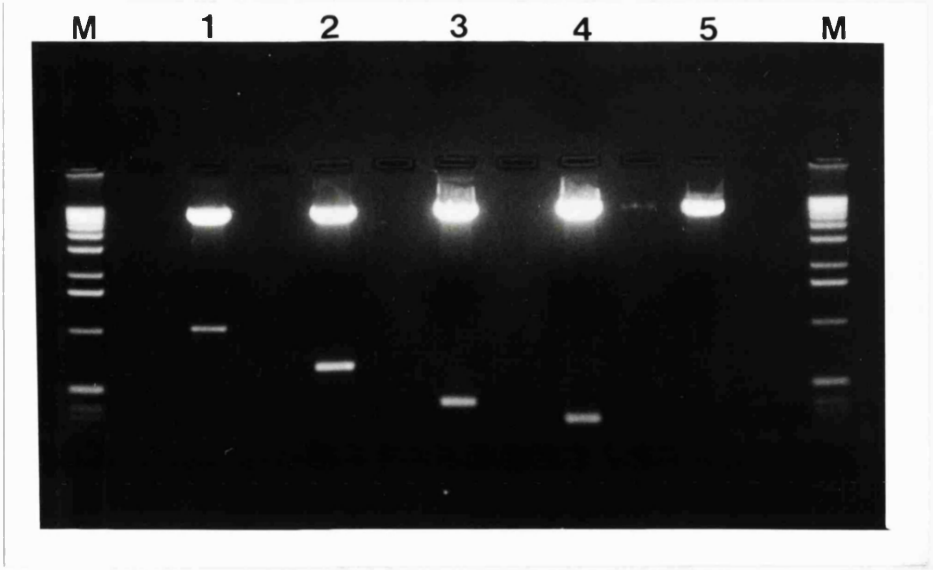


Figure 4.2a: Schematic map of vector pWE15 showing the cloning site for the pUC/Sau3A fragments.

Figure 4.2b Restriction digests of the four pWE15 subclones showing inserts excised with EcoRI. Lanes are marked as follows: 1. subclone pWE15/955, 2. pWE15/585, 3. pWE15/341, 4. pWE15/258, 5. native pWE15 and M= 1Kb ladder.



As a model to experiment with PCR amplification between two different octamers, the phagemid vector pSL1180 was selected. The vector, designed by Brosius (1989), has a large "superlinker" which includes sites for NotI, ApaI and BssHII; see figure 4.3.

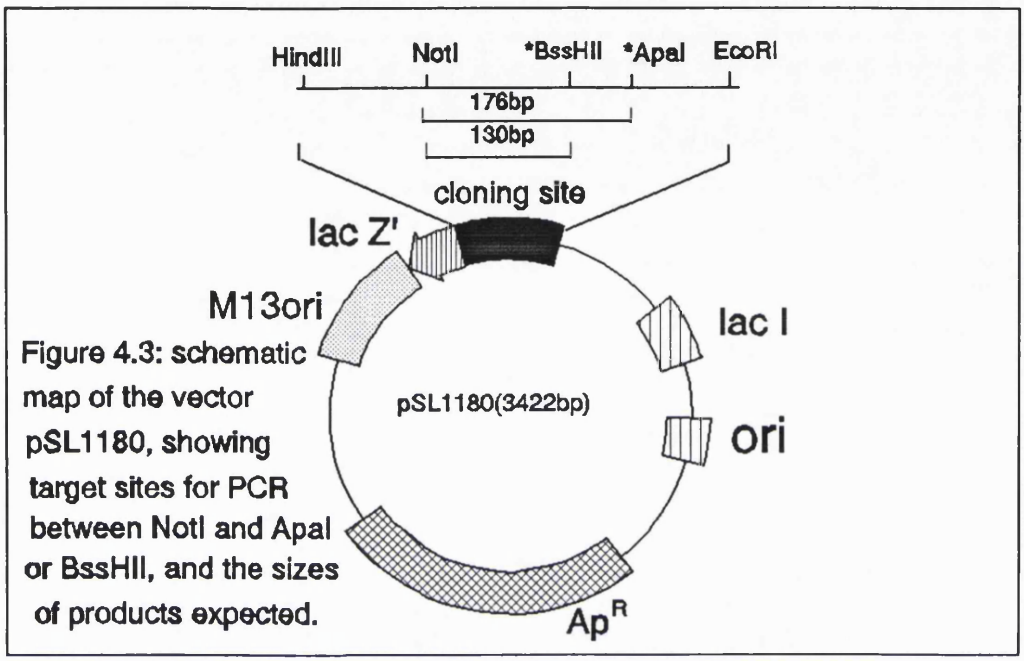


Figure 4.3: schematic map of the vector pSL1180, showing target sites for PCR between NotI and ApaI or BssHII, and the sizes of products expected.

4.3 PCR USING 8, 12 AND 16'MER PRIMERS BASED ON THE NOTI SITE

A number of factors were varied in the course of the experiments, the first being the PCR primer annealing temperature. The four pWE15 subclones were used as the template, with the native vector as a negative control. The template was linearized by digestion with SalI, in order to prevent amplification of the entire vector by primer extension from the reverse strand at the primer sites. This is likely since the primers used are highly palindromic. As described previously it was realized that SalI cut within the 955 subclone insert, thus preventing amplification, therefore this subclone was linearized by SmaI digestion instead. Amplification was attempted using three different sets of oligonucleotides each containing the

recognition site of the rare-cutting restriction enzyme NotI:

I 16mers (+) 5' ATT CGC GGC CGC ATT T 3'

(-) 5' ATT CGC GGC CGC ATA A 3'

II 12mer 5' NNGCGGCCGCNN 3

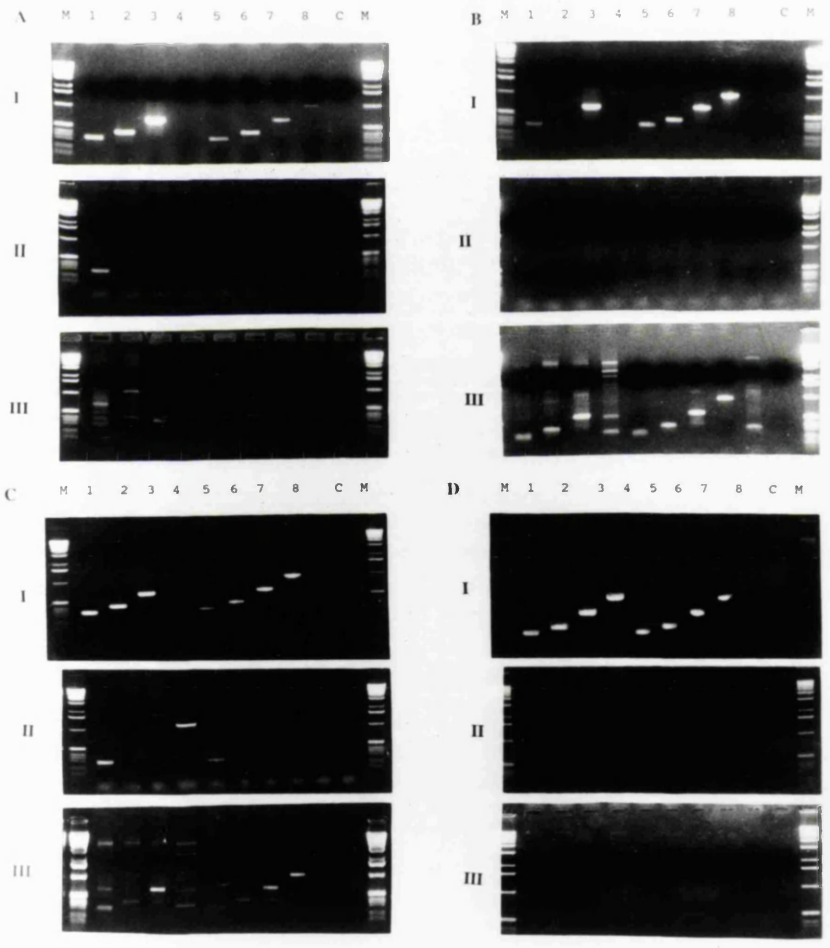
III 8mer 5' GGCGCCGC 3'

As described above, the sequence for the 16mers was based on the pWE15 sequence around the two NotI sites. Target sequences used were the four linearised pWE15 subclones and also the four inserts excised with EcoRI, thus retaining the NotI sites as shown in Fig.4.2a. All PCR amplifications were performed using a Perkin-Elmer-Cetus thermal cycler. The reactions were all carried out in 25μl volume, with constant conditions as follows :-

35 cycles of 94°C denaturation for 1 minute, annealing at temperature specified for 1 minute and 72°C extension for 1 minute, with a final 10 minute extension at 72°C. Standard PCR conditions were employed, namely 50mM KCl, 10mM Tris-HCl pH8.3, 1.5 mM $MgCl^2$, 250μM for each dNTP and 0.5 units of Taq DNA polymerase, 10ng of vector DNA or 1ng of insert DNA. The quantity of primer used was 20ng of 8-mer, 25ng of each 16-mer and 1μg of 12-mer. This excess of 12-mer was to compensate for the 256-fold degeneracy. The experiment was performed a number of times, varying the annealing temperature of the PCR between 25 and 65°C at 5°C intervals (see Fig.4.4).

As illustrated in Fig. 4.4, there is positive amplification between the NotI sites of the four pWE15 subclones across a wide range of temperatures with the 8mer oligonucleotide (30 to 55°C). There is amplification with the 12mer, but limited to conditions of approximately 45°C annealing temperature. PCR with the 16mers, which were based on the actual sequence around the pWE15 NotI sites, correctly amplified after annealing over a wide range of temperatures (<25°C to >65°C).

Figure 4.4 Effect of temperature on efficiency of PCR amplification. DNA (10ng) from the four different pWE15 subclones containing pUC19 fragments of different lengths (lane 1. 258 bp, 2. 341 bp, 3. 585 bp, 4. 955 bp) was amplified. The subclone inserts were excised by EcoRI digestion and 1ng of each was amplified (lane 5. 258 bp, 6. 341 bp, 7. 585 bp and 8. 955 bp. lane C shows PCR using native pWE15 as a negative control, and lane M shows 1Kb ladder as size marker. Three different sets of primers were used: I. a pair of 16-mers complementary to the sequence around the two NotI sites, +5'ATTC GCGG CCGC AATT and -5'ATTC GCGG CCGC ATAA (25ng per primer per 25μl PCR). II. a 12-mer 5'NN GCGG CCGC NN (1μg of primer per 25μl PCR). III. the 8-mer 5'GCGG CCGC (20ng of primer per 25μl PCR). Primer annealing steps were performed at 25°C (A), 35°C (B), 45°C (C) or 60°C (D).



4.3.1 SOUTHERN BLOTTING AND HYBRIDIZATION OF PCR PRODUCT WITH NOTI OCTAMER

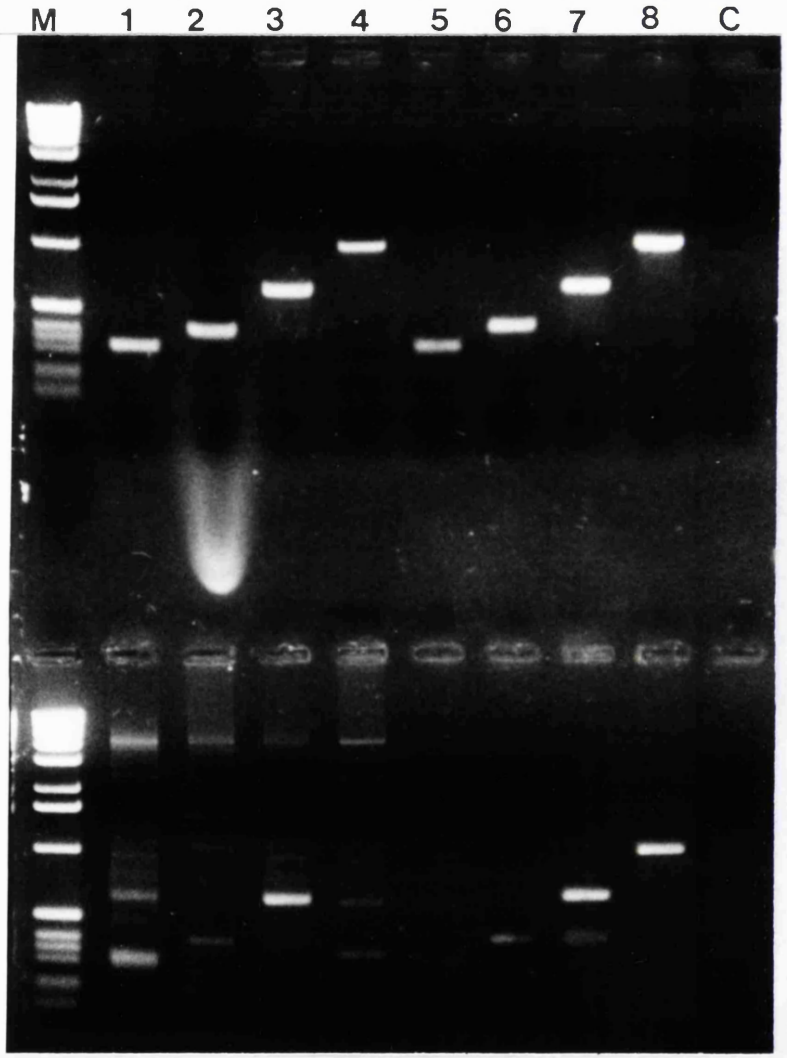
PCR product from amplification with the NotI 16mers and 8mer (under standard conditions and annealing at 45°C) was run on an agarose gel (as above) and then Southern blotted onto Hybond-N filter. The NotI octamer was end-labelled with [^{32}P-γ]-ATP using T4 polynucleotide kinase. Southern hybridization was performed overnight at 30°C in 6xSSC, 10mM PO_4, 5xDenhardt's solution, 0.5% SDS. The filter was washed at 30°C in 5xSSC, 0.1% SDS and exposed to Fuji RX X-ray film for 4 hours. Hybridization of the NotI octamer revealed that very specific amplification occurred with the 16mers, whilst slightly less specific amplification occurred with the 8mer as shown by a small amount of background product (see Fig. 4.5).

4.3.2 METHODS OF IMPROVING PCR SPECIFICITY

Other parameters were also investigated for their effect on PCR using the NotI 8mer and the pWE15 subclones. Titrations were carried out to investigate the effects of varying KCl and $MgCl_2$ concentration, the presence of 10% dimethyl sulphoxide (Filichkin and Gelvin, 1992; Shen and Hohn, 1992), the presence of 2% formamide and the use of tetramethylammonium chloride (TMAC) (Hung et al, 1990). Titration of template DNA (pWE15/585 clone) and primer (NotI octamer) was also performed. For these investigations, standard PCR was employed as in previous experiments and with annealing at 45°C.

Figure 4.5 PCR amplification on pWE15 subclones and EcoRI/excised inserts using I. NotI 16-mers and II. NotI 8-mers, annealing at 45°C (a) and Southern blot back hybridization using end-labelled NotI 8-mer as probe (b). Lanes are marked as follows: lanes 1-4 show PCR product from the follwing templates, 1. pWE15/258, 2. pWE15/341, 3. pWE15/585, 4. pWE15/955, lanes 5-8 show amplification from the EcoRI-excised inserts from the four subclones, lane C shows PCR product using native pWE15 as template, and M shows 1Kb ladder as size marker.

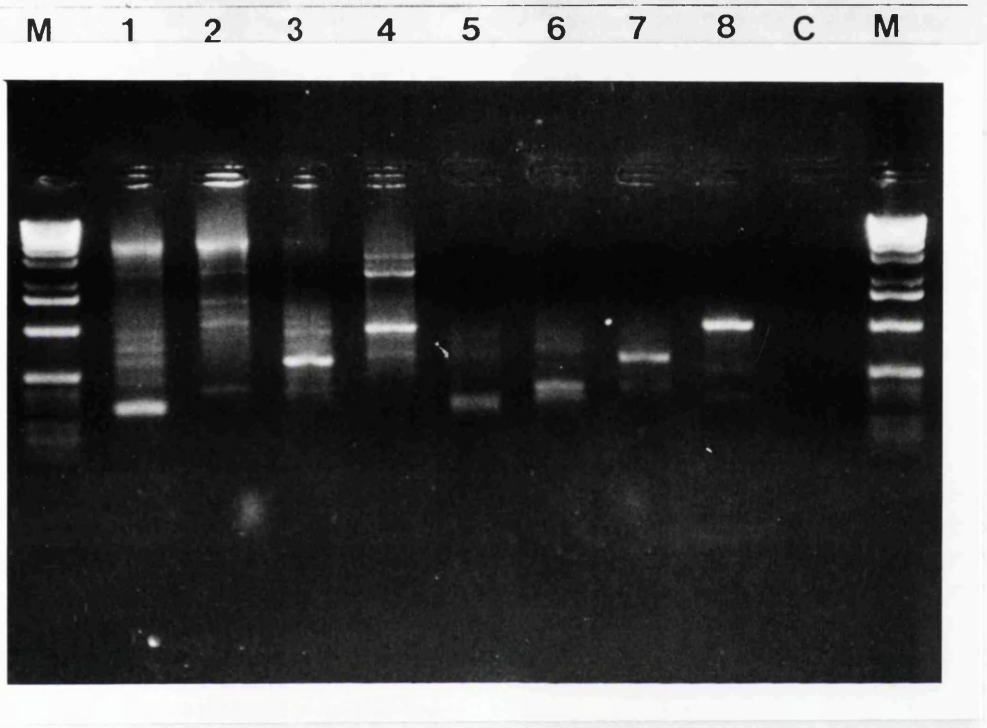
a.



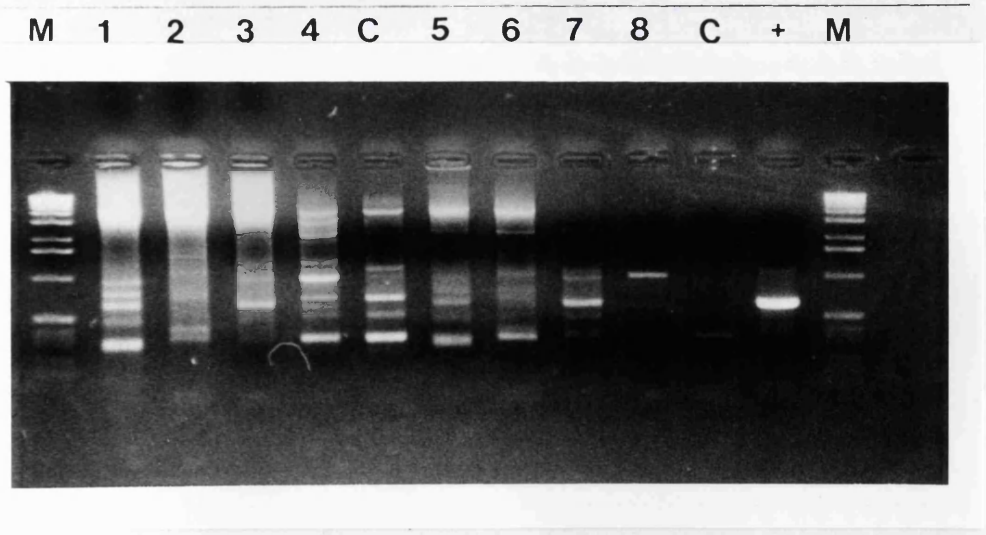
b.

Figure 4.6 PCR amplification using the NotI 8-mer on the four pWE15 subclones showing the effect of the addition of (a.) 10% DMSO, and (b.) 2% formamide, in both cases the annealing step was performed at 45°C. The lanes are marked as follows: lanes 1-4 show PCR product with the four pWE15 subclones 258, 341, 585 and 955 respectively, without DMSO or formamide, lanes 5-8 show PCR product for the subclones in the presence of DMSO/formamide. C indicates the use of native pWE15 as negative control, + shows the use of the pWE15/585 subclone versus the NotI 16-mers as a positive control and M is 1Kb ladder as size marker.

a.



b.



The ideal salt concentration for PCR with the NotI octamer (annealing at 45°C) was between 30 and 50mM KCl and 1.5 to 2.5mM $MgCl^2$ (results not shown). Addition of 10% DMSO or 2% formamide to the reaction mixture improved the specificity of the bands (see Figure 4.6a and b). Addition of different amounts of TMAC to the reaction did not improve the amplification (results not shown). Titration of the target DNA (pWE15/585 clone) and NotI octanucleotide (annealing at 45°C) suggested an optimum of 1 to 10ng (0.3 to 3fmoles) of template DNA and 25 to 100ng (9.5 to 38pmoles) of primer (results not shown).

4.4 PCR USING NOTI 6,7,10 AND 14-MER PRIMERS

Amplification with several other oligonucleotides based on the NotI sequence was investigated:-

IV 6mer 5' CGGCCG 3'

V 7mer 5' CGGCCGC 3'

VI 10mer 5' NNGCGGCCGC 3'

VII 14mer 5' GAATTCGCGGCCGC 3'

The 14mer, as with the 16mers, was based on the sequence of pWE15 at the NotI sites and also contains an EcoRI site at the 5' end. PCR was attempted with these primers using the pWE15 subclones as template, in 30mM KCl, 10%DMSO and annealing at various temperatures, with the other reaction conditions as standard.

Attempts to use primers shorter than 8 bases were mainly unsuccessful; no positive amplification was achieved using the hexamer, even though annealing temperature was reduced to as low as 20°C. Faint amplification product was observed, however, using the heptamer (Fig.4.7) annealing at 25°C. Positive yet faint amplification was achieved using the decamer (Fig.4.8). Surprisingly, however, no amplification occurred with the 14mer, despite efforts to vary conditions and annealing temperatures.

Figure 4.7 Amplification of the pWE15/585 construct with heptamer and hexamer primers. Lanes 1-6 show amplification using the heptamer 5'CGGCCGC (similar to the NotI 8-mer except that the 5' most G is omitted), lanes 7-12 show amplification using the hexamer 5'CGGCCG. Lane 13 shows a positive control with PCR using the NotI 16-mers against the pWE15/585 subclone. In lanes 1-3 and 7-9 PCR was performed in the presence of 10% DMSO, and lanes 4-6 and 10-12 were without DMSO. Lanes 1, 4, 7, 10 and 13 used the linearised pWE15/585 subclone as template (10ng). Lanes 2, 5, 8 and 11 used the EcoRI-excised band as template (1ng). Lanes 3, 5, 9 and 12 used native pWE15 as template (10ng), as negative controls. All reactions were performed in 30mM KCL and 2.5 mM $MgCl_2$, annealing at 25°C (other conditions as standard). 1kb ladder was run as a size marker (M). (NB the gel has run slightly unevenly, and thus the bands in lanes 4 and 13 should represent the same size PCR product).

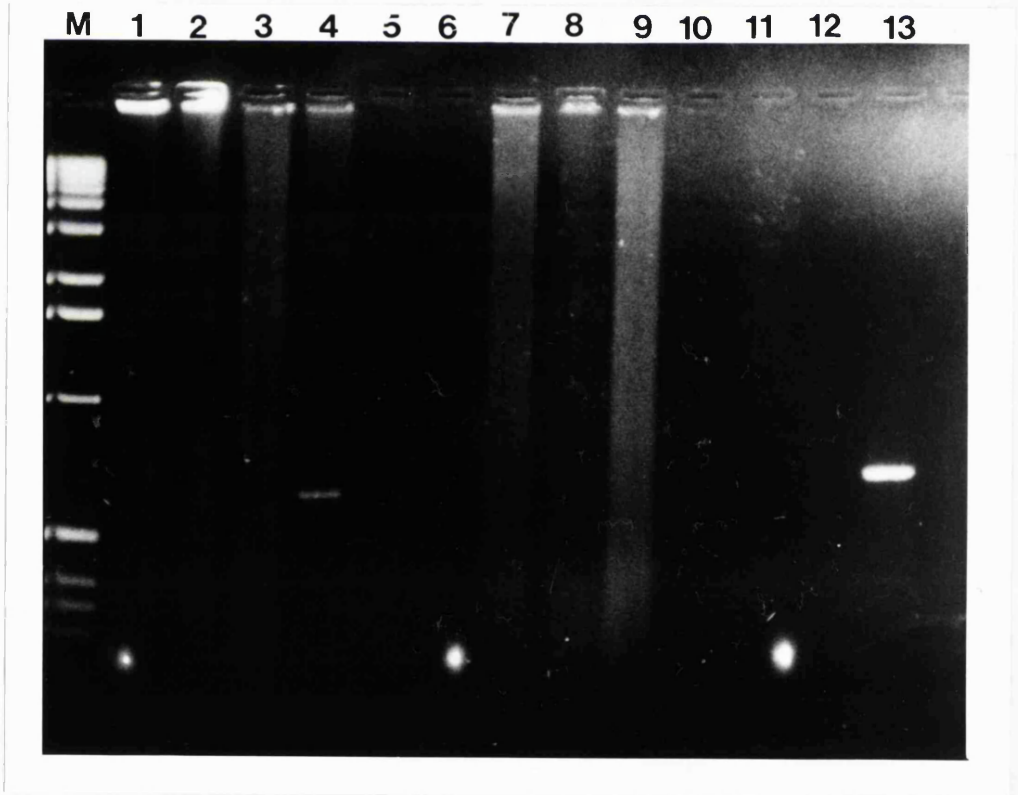
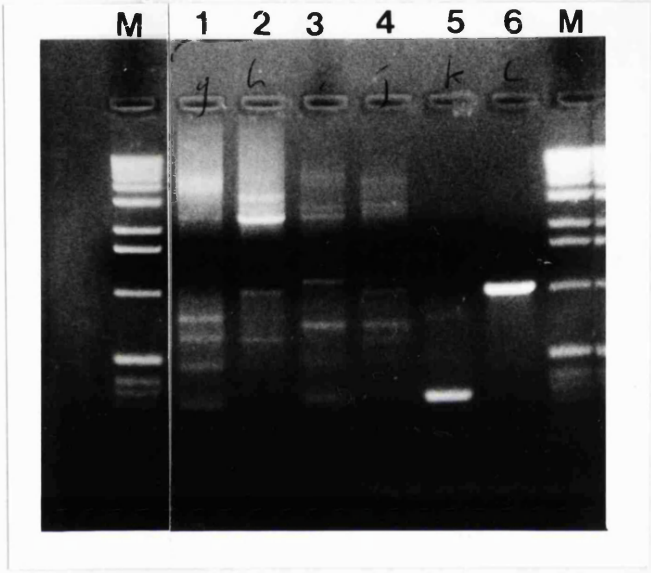


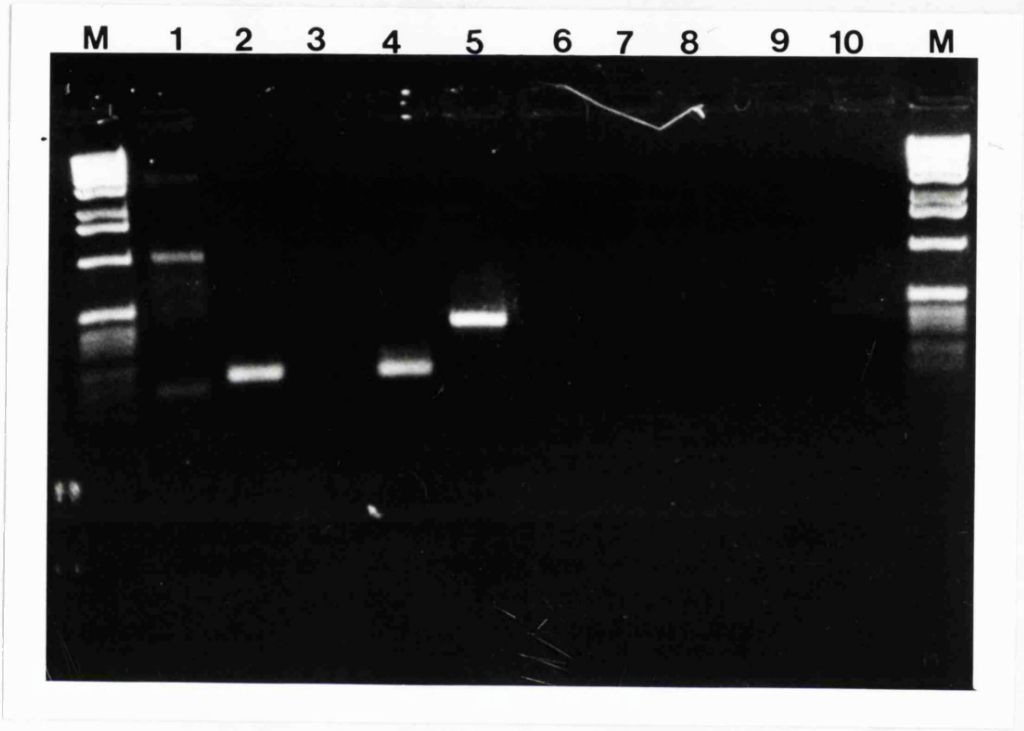
Figure 4.8 PCR amplification using NotI decamer. Lanes 1, 3 and 5 show amplification using the linearised subclone pWE15/258 as template. Lanes 2, 4 and 6 show amplification using subclone pWE15/955 as template. Lanes 1 and 2 use the NotI 8-mer as primer (50ng), lanes 3 and 4 use the NotI 10-mer 5'NN GCGG CCGC as primer (50ng), and lanes 5 and 6 use the NotI +/- 16-mers as primers (50ng). PCR was performed under standard conditions, annealing at 45°C. 1kb ladder was run as a size marker (M).



4.5 PCR USING 8-MER PRIMERS ON PSL1180

As a model for amplification between two different octamer primers, we used the vector pSL1180, which contains restriction sites for NotI, ApaI and BssHII in the multiple cloning site (see Figure 4.3). The NotI octamer was used as the 5' primer and as the 3' primer either BssHII (5'TGCGCGCG 3') situated 130bps away was used or ApaI (5'CGGGCCCT 3') situated 176bps away. PCR conditions used were as described above, but with 10% DMSO and 30mM KCl and annealing at 35°C. Successful amplification was achieved between the NotI and BssHII sites and between the NotI and ApaI sites of pSL1180 (Fig.4.9).

Figure 4.9 PCR amplification between 8-mers using pSL1180 as template. Lanes 1-2 used HindIII-linearised pSL1180(10ng)as template, lanes 3-4 used the HindIII/EcoRI-excised band (1ng) as template. Lanes 1 and 3 used 50ng of NotI 8-mer and 50ng of BssHII 8-mer, 5'TGCGCGCG, lanes 2 and 4 used 50ng of NotI 8-mer and 50ng of ApaI 8-mer 5'CGGGCCCT. Lane 5 shows a positive control of the NotI 16-mers versus the pWE15/341 subclone. Lanes 6-10 are repeats of lanes 1-5 but without the use of 10% DMSO in the reactions. The annealing step was carried out at 35°C. 1kb ladder was run as a size marker (M).



4.6 PCR ON GENOMIC DNA USING 8-MERS

Since it is possible to amplify specifically from model vector template using short G+C containing oligos, it became justifiable to test the hypothesis that such oligos could be used to amplify human genomic DNA. Since CG-rich regions tend to be clustered within regions known as HTF or CpG islands which are located adjacent to many coding sequences, it may be possible to use such oligos to enrich for such regions by PCR amplification. Oligo hybridization-based methods have previously been used to screen libraries using NotI and

other oligos (Estivill and Williamson, 1987) in order to select cosmid clones containing such sites. These have been of use as probes in long range restriction mapping projects using pulsed field gel electrophoresis (Estivill and Williamson, 1987; Melmer and Buchwald, 1990; Melmer et al, 1990). Since the probability of a NotI site and thus any other 8-mer CpG containing oligo being within a CpG island of a coding region is theoretically over 89% (Lindsay and Bird, 1987), the probability of two adjacent NotI sites within PCR range of each other being within a CpG island is over 98%. Similarly the probability of a CpG containing six-cutter being present in a CpG island is 74% and thus the probability of two adjacent six-cutters being present in a CpG island is over 93%. Thus a screening method based on the PCR amplification between two rare-cutter sites could be a very powerful method of enriching for coding regions.

4.6.1 AMPLIFICATIONS ON HUMAN GENOMIC DNA USING 8-MER OLIGOS

Amplifications were performed using 8-mer oligos based on NotI, SacII, SmaI, BssHII and ApaI sites (NotI: GCGGCCGC; SacII: GCCGCGGC; SmaI: GCCCGGGC; BssHII: CGCGCGCG; ApaI: GGGGCCCC) with 10ng of single stranded placental human DNA. PCR was carried out with reduced KCl concentration (30mM) and 10% DMSO and annealing at 47°C for 35 cycles and incorporating [^{32}P-α] dCTP into the reactions. The amplification product was run on 4% polyacrylamide gels and compared against an end-labelled 1Kb ladder. The resulting autoradiographs showed distinct patterns, or fingerprints, for each oligo. PCR amplifications were also performed using two oligos, for instance SacII versus SmaI. Some of the resulting bands matched those of the fingerprints produced using single primers, the other bands were assumed to represent the product of amplification between the two different primers.

As evidence that the PCR is preferentially amplifying CpG rich regions, the genomic DNA was digested prior to amplification using various restriction enzymes. Digestion with MspI or BssHII which are a CpG 4-cutter and a 6-cutter respectively, effectively blocks amplification. Digestion with HpaII, the methylation sensitive isoschizomer of MspI which only cuts at non-methylated sites, blocks amplification of about 50% of bands, as does digestion with an AT 6-cutter (DraI), although the reason for this is unclear. Digestion with TaqI, which cuts at TCGA, prevents most amplification, whilst digestion with AluI, which cuts at AGCT, has little effect at all.

Amplification was performed using the SacII 8-mer as primer on DNA from various different species, including rat, mouse, bovine, chicken, salmon, herring, bacterial and human DNA. A completely different fingerprint was produced for each species, only a few bands for the relatively small bacterial genome and several hundred bands for the much larger vertebrate genomes (see figure 4.10a). Whilst bacterial DNA has no HTF islands, it is much more CpG rich than bulk eukaryotic DNA, so this result is surprising. Also surprising are the strong amplification ladders produced by PCR on herring and salmon DNA. CpG islands in certain species of fish whilst possessing non-methylated islands of undepleted CpG, have low GC content (Cross et al, 1991). As a result one would have expected fewer bands. Fingerprints of unrelated humans appeared identical (figure 4.10b), suggesting that non-random amplification occurs and that there is little or no intra-species variation.

Fig. 4.10a PCR amplification of human genomic DNA from different individuals. PCR was performed in 30mM KCl, 2.5mM $MgCl_2$ and 10% DMSO, with other conditions as standard, incorporating 2μCi of ^{32}P-αdCTP into each reaction. Annealing was at 47°C for 2 mins. 1Kb ladder was used as size marker (M).

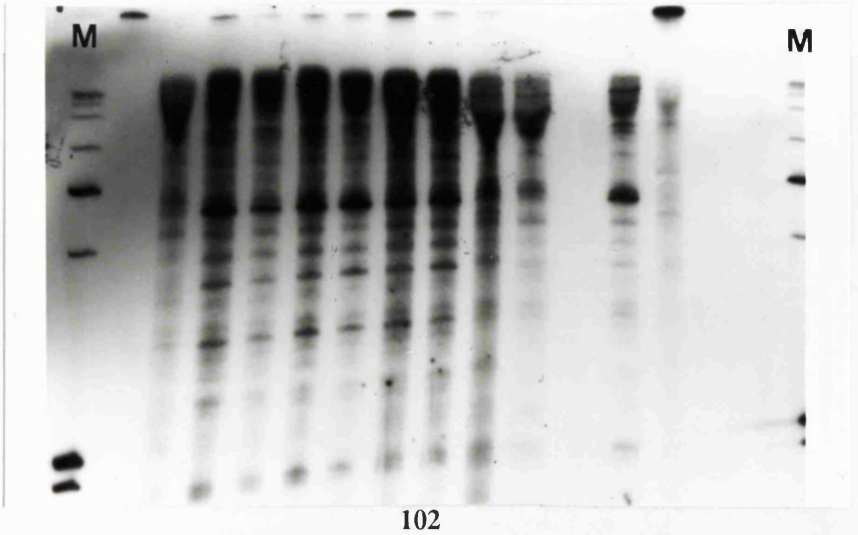
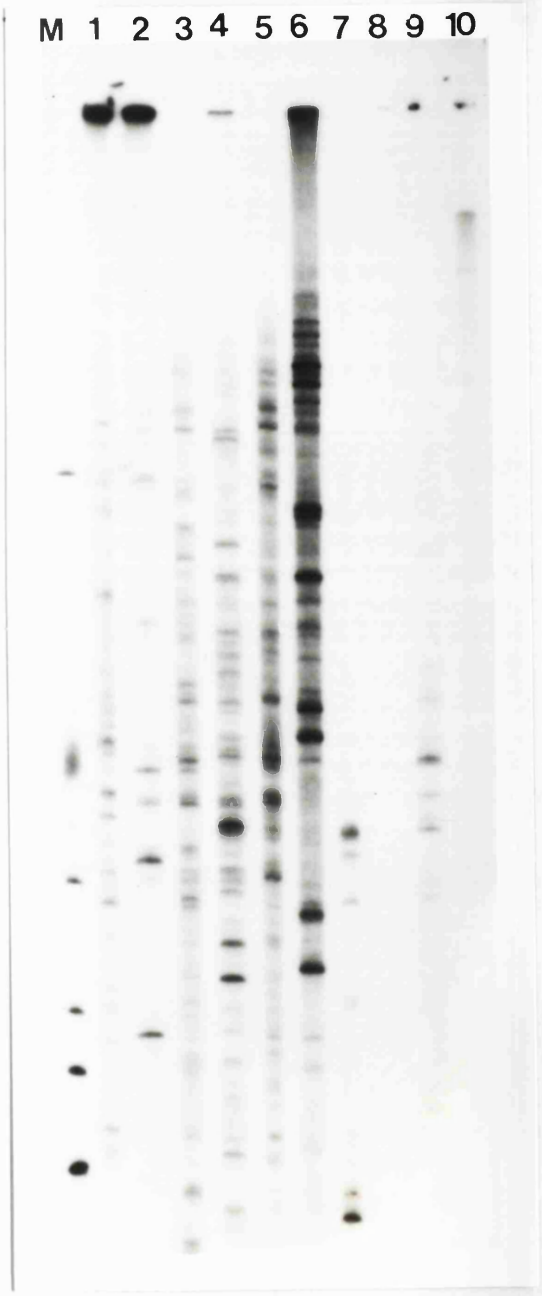


Figure 4.10 PCR amplification of genomic DNA from different species using short GC oligos. 2μg of oligo SacII 8-mer (5'GCCGCGGC 3') was used per reaction with 50ng of genomic DNA from the following species: lane 1. rat, 2. mouse, 3. bovine, 4. chicken, 5. salmon, 6. herring, 7. M. lysodeictus, 8. human (A), 9. human (B), 10. human chromosome 5 cosmid library DNA. PCR was performed in 30mM KCl, 2.5mM $MgCl_2$ and 10% DMSO, with other conditions as standard, incorporating 2μCi of ^{32}P-αdCTP into each reaction. Annealing was at 47°C for 2 mins. 1Kb ladder was used as size marker (M).



4.6.2 M13 CLONING AND SEQUENCING OF PCR PRODUCTS

Several attempts were made to clone the PCR products of genomic amplification using the short CG-rich oligos directly into vectors, in order to facilitate the sequencing of the PCR products and hence determine in a definitive manner whether or not the method is enriching for the CpG fragment of DNA. In particular a pUC/NotI vector was used for subcloning the PCR product from the NotI-based oligos. The fact that this was unsuccessful reflects on the fact that many restriction enzymes have difficulty cutting when the recognition sites are at, or near, the ends of DNA segments (see New England Biolabs catalogue, 1992, Appendix, 182). Attempts to blunt-end "shotgun" clone the PCR fragments also produced a very low yield of recombinants. This was most likely due to the fact that PCR using Taq DNA polymerase produces fragments with "ragged ends". For these reasons a "shotgun" cloning approach was employed that involved digesting the PCR product with the frequent cutter Sau3A, which produces cohesive ends and cloning the resulting fragments into the BamHI site of the vector M13. PCR was performed using the NotI, SacII, SmaI and BssHII octamers and the NotI 10-mer and 16-mers as primers. All amplifications were performed as before, with 10% DMSO and 30mM KCl and annealing the primers at 47°C. This temperature is 21°C below optimum for the NotI 16-mers and so should decrease the stringency so as to allow less specific annealing.

Although cloning was successful, efficiency was very low, with an average of just 7.4 recombinants per plate. The resulting recombinants were picked and grown and DNA was prepared, as described in the Methods section. The prepared DNA was used as template for ^{35}S sequencing using Sequenase™ with the "-40" primer. 13 clones were successfully sequenced and analyzed for CpG content. The sequences are shown in Appendix A. The sequences were also used to screen the Genbank database for homologous sequences or the presence of repeat DNA, using the BLAST program (Altschul et al, 1990). The results are

summarized in Table 4.1.

CLONE	PRIMER	bp	%G+C	O/E CpG
6a1	NotI 8'mer	166	60.8	0.865
6a2	NotI 8'mer	97	52.6	0.490
6b1	SacII 8'mer	128	40.6	0.851
6f3	NotI 10'mer	284	57.7	0.804
6f4	NotI 10'mer	101	58.4	1.098
6f5	NotI 10'mer	176	60.2	0.877
6g1	NotI 14'mer	148	46.6	0.124
6h1	NotI 16'mers	181	39.8	0.575
6h2	NotI 16'mers	125	59.2	0.647
6h4	NotI 16'mers	200	59.0	0.946
6h5	NotI 16'mers	135	39.3	0.000
6h6	NotI 16'mers	182	59.9	1.269

Table 4.1: sequence analysis of M13-cloned PCR product with short GC primers. Column 2 shows the primer used in the PCR. 4 other sequences were found to be identical to 6a1.

The criteria for CpG islands as stipulated by Gardiner-Garden and Frommer (1987), of having percentage G+C greater than 50% and observed/expected ratio for the CpG dinucleotide of over 0.6, were used to determine whether the sequences could be classed as CpG islands. However, because the sequences were relatively short with an average length of 161bp, it was not practicable to calculate a moving average using a 100bp window. The obs/exp CpG ratio was calculated as follows:

Obs/Exp CpG = {Number of CpG}/{Number of C x number of G} x total bp

Seven out of the twelve clones sequenced complied with both criteria and a further two qualified with one out of two criteria. If it is assumed that CpG islands make up 1% of total chromosomal DNA (Kusuda et al, 1990) this represents a 58 fold enrichment.

4.6.3 TA CLONING AND SEQUENCING OF PCR PRODUCT

Whilst Sau3A digestion and shotgun cloning of the PCR products from amplification with short CG oligos clearly works, the cloning efficiency is low and clones produced are rather short for determining whether or not the insert belongs to a CpG island. The template DNA may also be cloned, thus permitting a degree of background genomic sequences to be included. Recently a method was developed specifically for the cloning of PCR products. This method does not rely on either using primers with cloning sites at their 5' ends (Kaufman and Evans, 1990; Jung et al, 1990) or filling in the ragged ends using Klenow fragment (Williams, 1989). The method also does not require the "polishing" of the ragged ends using exonuclease, or the cutting back with the 3' to 5'exonuclease activity of T4 DNA polymerase to leave defined cohesive termini (Stoker, 1990; Aslanidis and de Jong, 1990). This method, known as TA cloning, takes advantage of the fact that Taq DNA polymerase has a 3' deoxyadenylation activity, often leaving the product with a protruding A base on the 3' end. A vector is used which is cut with HphI to leave a single 3' overhanging T base and is thus able to anneal and ligate double-stranded PCR fragments with the complementary overhanging A base (Mead et al, 1991). The ligation is supposedly some 50 times more efficient than straightforward blunt-end ligation. Even more recently other protocols have been developed which also take advantage of the TA pairing with PCR products. These methods involve T-tailing a cloning vector (particularly M13 or Bluescript) by addition of dTTP using Taq DNA polymerase (Marchuk et al, 1991) or ddTTP using terminal transferase (Holton and Graham,

1991).

The TA cloning vector pCR1000 (Invitrogen) was used to clone PCR product from the amplification of genomic DNA, using the NotI 8, 10 and 16-mers and AscI (GGCGCGCC) and SacII octamers. The cloning efficiency was much higher than with the M13-based approach already attempted, with an average of over 40 white colonies per plate. However, many of the white colonies were a result of vector recombination and so only colonies which produced a faint blue tint after 48 hours incubation were selected. DNA was prepared from these colonies and sequenced using M13 forward and reverse primers. Some of the sequences contained many band compressions and so PCR-coupled sequencing was used to "iron out" compressions.

The sequences obtained are presented in Appendix B. The sequences were analysed for CpG content as before and the Genbank database was used to screen the sequences with the BLAST program (Altschul et al, 1990) for homologies to known sequences. Also, all sequences showing homology to vector sequences and any sequences identical to each other were excluded from further analysis. The GC-primer sequence was not included in the analysis because the efficiency of priming may not be 100% and it cannot be assumed that the original template DNA shares exactly the same sequence as the primers. Thus the sequences chosen for analysis were confined to the sequence between primer sites. 35 out of 53 clones analysed complied with both citeria for CpG islands, representing an enrichment of 66 fold, assuming a level of 1% CpG islands amongst the entire human genome (Kusuda et al, 1990). This is similar to the enrichment value obtained with M13 cloning as described in Chapter 4.6.2. A further 21 clones were either false positives resulting from the vector pCR1000 recombining with itself, or appeared to have vector content within the sequence, as disclosed by the BLAST analysis. Three sequences were too poor to read, one sequence was excluded because

it was identical to another and one sequence was the non-recombinant vector. It is important to note that only five of the sequences that fulfil both criteria for CpG islands are over 200 bp long and can thus be regarded as true CpG islands, since the minimum size of a CpG island is 200 bp (Larsen et al, 1992). The results are summarized in Table 4.2.

Table 4.2: TA clones from PCR on human genomic DNA using short GC oligos.

CLONE	BASE PAIRS	%G+C	O/E CpG	HTF Criteria
TA2	210	60.3	1.466	2
TA4	236	45.8	1.058	1
TA5	104	62.5	1.090	2
TA8	179	58.1	1.136	2
TA9	198	57.6	0.433	1
TA12	124	56.5	1.231	2
TA13	118	67.8	1.199	2
TA14	190	53.7	1.319	2
TA15	236	46.2	1.117	1
TA16	164	54.3	0.995	2
TA17	150	64.0	1.368	2
TA18	144	53.5	1.360	2
TA20	170	55.3	1.479	2
TA21	102	52.0	1.162	2
TA22	148	59.5	1.549	2
TA23	143	57.3	1.219	2
TA24	197	57.9	0.990	2
TA26	139	51.1	1.552	2
TA28	196	58.2	1.237	2
TA29	152	57.2	1.536	2
TA30	175	53.7	1.111	2
TA31	117	67.5	1.500	2

TA32	179	57.0	1.129	2
TA33	107	68.2	0.922	2
TA35	156	54.3	1.130	2
TA36	157	40.8	0.943	1
TA37	116	57.0	1.173	2
TA38	95	56.8	1.441	2
TA39	104	55.8	0.989	2
TA40	234	50.8	0.997	2
TA41	177	54.8	0.903	2
TA42	211	46.9	0.173	0
TA43	148	60.8	1.096	2
TA44	212	62.7	1.044	2
TA45	209	50.7	0.293	1
TA46	121	58.7	1.056	2
TA47	316	49.4	0.327	0
TA48	220	52.7	1.099	2
TA49	351	51.0	0.395	1
TA50	211	57.8	0.965	2
TA51	369	51.6	0.425	1
TA52	159	60.4	0.837	2
TA53	204	37.3	0.424	0
TA54	142	39.4	0.182	0
TA55	417	43.4	0.306	0
TA56	352	44.0	0.352	0
TA57	169	40.8	0.571	0
TA58	174	49.4	0.607	1
TA59	176	51.5	0.230	1
TA60	108	51.1	0.807	2
TA61	133	51.9	0.796	2

TA63	191	41.9	1.441	1
TA64	69	47.8	1.062	1

Inserts from several of the TA clones were oligolabelled and used as probes against northern blots containing macrophage, B-cell and HeLa cell RNA. No clear bands were observed, other than those showing ribosomal RNA.

4.7 PCR AMPLIFICATION ON COSMIDS USING GC OLIGONUCLEOTIDE PRIMERS.

To test whether short GC oligos could be used in a non-specific manner to amplify CpG islands in cloned human DNA, two well characterized cosmid clones, cosGSTrp7 (containing the glutathione-S-transferase π and NADH-ubiquinone oxidoreductase genes) and cosCT1 (Calcitonin/Calcitonin gene related peptide) were used as model template. 1ng of template was used with the NotI 8, 10 and 16-mers, SacII, AscI and BssHII 8-mers. Amplification was performed as previously, with 30mM KCl and 10% DMSO, but in order to decrease the annealing stringency to encourage priming at sites with less than full complementarity, the first 4 cycles annealing was performed at 42°C for 2 min and continued at 47°C for a further 30 cycles. Amplification product is shown in Fig. 4.11. As has been discussed previously, an octamer should have approximately 1.34 sites per cosmid (Studier, 1989) and a heptamer 5.36 sites. Thus the chances of achieving successful amplification using octamers on cosmids is high, particularly if the stringency is relaxed to allow the oligos to prime at sites with say 7 or 6 complementary bases. Amplification at less than optimum conditions using the NotI octamer on pWE15 subclones has been shown to result in a number of product bands from a template only 9.5Kb long (see 4.3). Amplification of the cosmid clones cosGSTrp7 and cosCT1 with these oligo primers produces several distinct bands. The bands produced were cut out from a low melting point agarose gel and purified using MagicPrep Quick columns

(Promega). As before, the PCR product bands were cloned into the TA cloning vector and sequenced. The resulting sequences as shown in Appendix C were computer analyzed using the BLAST program for comparisons to databases of DNA sequences and compared to known sequences within the cosmid clones. The results are summarized in Table 4.3 and show that despite most of the clones obtained having CpG rich sequences, the method was unable to distinguish the HTF islands in the cosmid insert from the highly GC-rich DNA of the vector.

Figure 4.11 PCR amplification using short GC oligos on cosmids. Lanes 1-5 show PCR on 1ng of cosGSTrp7 DNA, lane 6 shows PCR on 1ng of cosCT1 DNA. 100ng of oligo was used per reaction as follows: lane 1. NotI 16-mer, lane 2. NotI 10-mer, lane 3. NotI 8-mer, lane 4. SacII 8-mer 5'GCCGCGGC, lane 5. AscI 8-mer 5'GGCGCGCC, lane 6. NotI 10-mer. PCR was performed in 30mM KCl, 2.5mM $MgCl_2$, annealing at 45°C for the first five and then at 47°C for 30 cycles. 1kb ladder was run as a size marker (M).

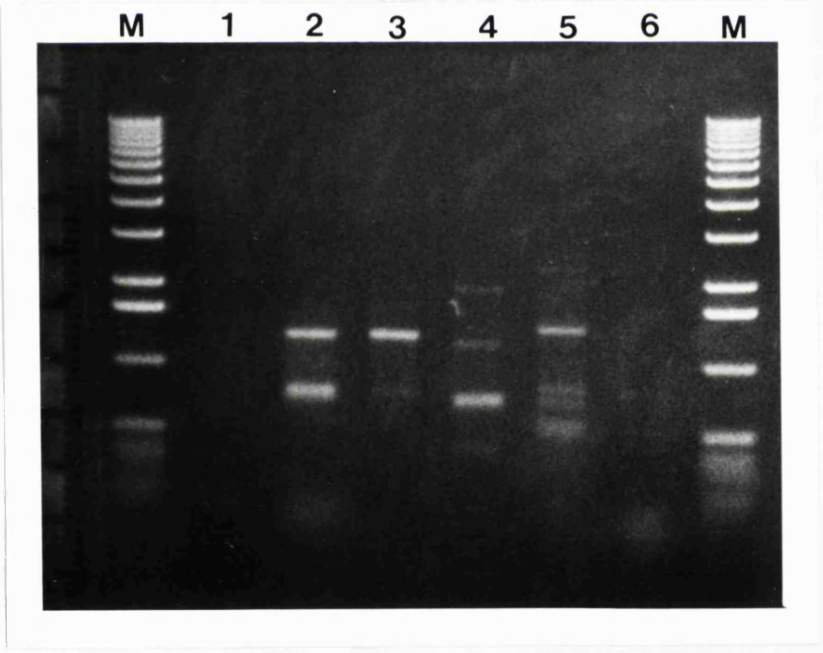


Table 4.3: TA clones from PCR on cosmid clones using short GC primers. The best match found using BLAST is shown, with BLAST score in parenthesis.

CLONE	PCR TEMPLATE	BLAST HOMOLOGY	BP	%G+C	O\E CpG
TA11(1)	cosGSTrp7	Broad bean DNA (109)	133	54.1	0.623
TA11(6)	cosGSTrp7	Hygromycin gene (491)	156	56.4	1.167
TA12(2)	cosGSTrp7	E.coli polA gene (823)	170	52.9	0.925
TA13(1)	cosCT1	E.coli glucosidase gene (345)	164	50.0	1.600
TA10(3)	cosGSTrp7	E.coli DNA (255)	154	45.5	1.540
TA13(3)	cosCT1	E.coli glucosidase gene (240)	87	50.6	1.465

4.8 DISCUSSION

The successful priming of polymerization by an oligonucleotide on target DNA is dependent on the rate of primer dissociation from the primer-template complex before polymerization is initiated. It also depends on the rate at which the DNA polymerase extends the primer until the primer-template complex becomes stable. Because the activity of Taq polymerase, which is now widely used in PCR, is much reduced at lower temperatures, primer extension rate is also reduced at such temperatures (Innis et al., 1988) and thus primer stability is reduced. On the other hand the rate of primer dissociation is dependent upon the temperature used and upon other factors such as oligonucleotide length, base composition and salt concentration.

The results of the experiments described previously confirm the hypothesis that it is possible

to prime DNA polymerisation effectively and specifically using very short oligonucleotides. For instance, using the NotI octamer to prime PCR on pWE15, the primer was able to discriminate against an extremely G/C-rich genome, including an EagI site and at least four other sites which have 7bp homology with the primer, at position 4228 (GCGGCCGG), 3902 (GCCGCCGC), 4994 (GCGGCTGC), 5821 (GCCGCCGC) and 7840 (GCGGCGGC) of the database sequence Z12112.em_sy. The correct match of the 3' nucleotide of a primer with the target sequence is of particular importance for successful priming (Huang et al., 1992). The NotI 8-mer has 3' nucleotides complementary to four of the above target sites on both strands and to the fifth on just one strand. Despite being within easy amplification range of each other, rather than amplifying between the above 5 sites, the NotI 8-mer preferentially amplifies between the two NotI sites. A large number of sites within pWE15 have 6bp homology with the NotI 8-mer.

Clearly, having attempted PCR with a hexamer and heptamer, the addition of a single extra base to an oligonucleotide of this size has a profound effect on its stability and its ability to form stable a duplex with the target DNA. Thus, whilst PCR with the octamer works with high efficiency, PCR with the heptamer was considerably less efficient. Previous reports have suggested that hexamers may be used efficiently for hybridization (Drmanac et al., 1990) and as primers for cDNA synthesis (Liu et al., 1989). However our own experience (Melmer et al, unpublished results) is that hexamers could be hybridized with some specificity only at very low temperature (11°C). However no meaningful amplification between hexamer primers was found. This conflicts with a previous report describing the use of primers as short as 5-mers for the detection of fragment length polymorphisms (Caetano-Anolles et al., 1991).

Hybridization of short oligonucleotides such as the octamer should give better discrimination, because the effect of a single base-pair mismatch should have a greater effect on the duplex

stability than it would for larger oligonucleotides. Theoretically this should also apply to primer annealing for PCR. On the other hand, it has been shown that the ability of oligonucleotides to prime synthesis is not limited to size and even mononucleotides can prime synthesis. However efficiency reaches an optimum at a length of 9-12 bases (Nevinsky et al., 1990). The fact that amplification was successful with octamers, but very difficult with smaller primers, may arise because the octamer is approaching this optimum. The high G/C content of the primers would increase the dissociation temperature and thus increase their stability and priming efficiency. The formula of Suggs et al (1981),

1. Td = 4{G+C} + 2{A+T},

provides a useful guide to melting temperature, where T_d = temperature of dissociation at which 50% dissociation occurs and G+C and A+T are the contributions from purine and pyrimidine bases pairing. A more accurate formula (see formula 2) that takes into account the effect of "nearest neighbour" thermodynamic values on duplex dissociation (Freier et al, 1986) is available on the computer program "OLIGO" (Rychlik and Rhoads, 1989).

2. Td = $\Delta H/\{\Delta S + R\ln(C/4)\}$ -273.15°C

ΔH and ΔS are enthalpy and entropy for helix formation. R is the molar gas constant and C is the oligo concentration. The formula of Suggs et al (1981) would suggest the NotI octamer has Td of 32°C. Hybridization experiments with the octamer (Chapter 3) show positive hybridization upto 40°C and in this chapter it is shown that the octamer can prime PCR up to 55°C, revealing this formula to be inaccurate. The formula described by Suggs et al. (1981) for DNA-DNA hybridization is an empirical one and does not take into consideration the dependence of melting temperature on oligonucleotide sequence interactions such as "nearest

neighbour" effect (Breslauer et al., 1986; Freier et al., 1986). The formula of Rychlik and Rhoads (1989) gives a Td 38°C for the NotI octamer. The polymerization experiments in this chapter have shown that this formula is also inadequate for explaining the behaviour of such short oligonucleotides. However, it is clear that DNA-DNA hybridization is governed by thermodynamic parameters while primer annealing in PCR has a kinetic component. An explanation for the adequacy of only transient annealing between primer and template being necessary is the possibility that the DNA polymerase is able to begin synthesis and once established an elongated primer-DNA structure stabilizes the complex (Wu et al., 1991). Whilst "OLIGO" attempts to optimize the primer annealing temperature, Wu et al. (1991) maximized the annealing temperature to give the highest temperature at which priming will occur, thus resulting in a formula giving generally higher primer annealing temperatures,

3. $T_p = 22 + 1.46\{2[G+C] + [A+T]\}$

For the NotI 8-mer this gives T_p as 45.5°C, which is closer to the observations made with short oligo PCR.

Whilst the experiments using the NotI 8-mer worked over a wide range of annealing temperatures, use of the 12-mer was much more limited with respect to temperature. It is possible that the large quantity of primer used in order to compensate for the four degenerate positions had an inhibitory effect on the amplification. The poor results using the NotI 10-mer is more surprising, but may be due to the fact that the primer concentration was too low to compensate for the two degeneracies.

The use of primers of palindromic sequence for PCR is ill advised, because primer self-annealing would occur and amplification would also be primed on both DNA strands. In the case of the pWE15 subclones, the PCR would amplify both the insert and the 8.8kb vector

and in some cases products from both were detectable. However, linearisation of the subclones prevents PCR completing the circle in the wrong direction and only allows amplification of the inserts. Octamers with just one A or T base, such as the ApaI and BssHII primers used, probably allow amplification in only one direction. Primer self-annealing explains why such large concentrations of primer are necessary in the PCR.

The PCR amplification of DNA with such short primers should have several applications. Firstly, it should allow the direct amplification of HTF islands, which are operationally defined as the clustering of recognition sites of rare restriction enzymes. Secondly, it should allow the amplification of DNA from other genomic landmarks such as consensus sequences for gene expression, regulatory proteins or gene splicing. Thirdly it might be used for the cloning of gene families based upon shared sequences. Lastly it might even be possible to develop the method for the high volume sequencing of cloned DNA (Studier, 1989; Szybalski, 1990).

The successful enrichment for genomic DNA containing CpG rich sequences as a "mini-library" of CpG islands could prove very useful for physical mapping of the genome. One of the main short term goals of the U.S. Human Genome Project is the establishment of Site Tagged Sequence (STS) maps for all human chromosomes at approximately 100 Kb intervals. Suitable STS's could be provided by this technique by obtaining sequence from both ends of the CpG rich TA clones and establishing the map positions by *in situ* hybridization or hybridization to ordered YAC contigs. These STS's would not only provide useful landmarks for physical mapping of the genome, but would also represent an instant map of possible coding regions. The rare-cutter restriction sites flanking the TA cloned sequences, at which amplification was primed, cannot be guaranteed to represent genuine sites in the genomic template DNA because the oligo-priming may have occurred at sites with less than 100%

homology. As stated earlier such sites were excluded from the sequence analyses for this reason. Nevertheless such clones, being CpG rich, would be abundant in rare-cutter sites and so would be useful for restriction site landmark mapping of the genome. The capacity of the CpG-rich TA clones that were isolated to hybridize to species of mRNA by northern hybridization could have been investigated more thoroughly, in order to give some indication of the usefulness of the clones for identifying coding regions.

CHAPTER 5: PCR AMPLIFICATION USING CONSENSUS SEQUENCE OLIGONUCLEOTIDES

5.1 INTRODUCTION

DNA containing CpG islands is characterized by the presence of rare-cutter restriction sites and seems to have a high probability of containing or being adjacent to genes (Bird, 1986, 1987; Lindsay and Bird, 1987). However, finding genes through CpG islands has certain limitations, namely that not all genes are located next to an island, or the CpG island may lie some distance from the transcribed sequence (Gardiner-Garden and Frommer, 1987).

Since most genes contain introns and hence splice junctions, the consensus sequence for these could be used to identify clones or regions of clones containing genes. One method that uses expression strategies to identify coding regions, "exon trapping", is discussed in Chapter 1. While these methods have been shown to identify known genes, they have limitations of being labour intensive and allowing only relatively small (20-50Kb) segments of the genome to be analyzed per transfection. The feasibility of using short oligos based on consensus sequences surrounding intron-exon junctions (Shapiro and Senapathy, 1987) has been demonstrated by Melmer and Buchwald (1992) and short oligos have been used successfully as PCR primers as described in Chapter 4). Because it is also possible to employ sequence specific primers with degenerate positions for effective PCR amplification as described by Compton (1990) and Heller et al (1992); the next step would be to attempt to use primers based on consensus sequences to test the hypothesis that amplification of coding regions can be achieved with PCR and consensus sequence primers.

The advantages of a PCR-based technique over a hybridization-based technique, as explored with the G/C-rich oligos are many fold. For instance the template DNA can be the size of a cosmid or a YAC clone and even entire genomes may be used in order to generate "libraries" of coding DNA. On the other hand in the hybridization approach the template needs to be cut into segments of convenient size and then either: (a) Cloned into a vector, transformed into E.coli, plated out on agar and then "lifted" onto nylon filters. or: (b) Separated by gel electrophoresis and immobilized onto a nylon filter. The hybridization approach can localize only one probe at a time, unless techniques involving probes labelled with different fluorochromes and computer imaging techniques are employed. In this case up to a dozen probes may be used simultaneously as described by Ferguson-Smith et al (1992) and Hoeltke (1993). On the other hand the PCR approach screens for two sites with the premise that the two sites must be within amplification range, a step which gives additional screening power to the procedure.

5.1.2 TEMPLATE AND PRIMER SELECTION

Initial studies on the hybridization of degenerate splice site oligos were performed (Melmer and Buchwald, 1992) using as template a number of subclones of the human proteolipid protein (PLP) gene (see Fig. 3.4). PLP is an X-linked gene with seven exons for which all exons and intron/exon boundaries have been sequenced (Diehl et al, 1986). In the study by Melmer and Buchwald (1992) three oligos were designed, two for the 3'splice site and one for the 5' splice site, based on the computer matrices from consensus sites as compiled by Shapiro and Senapathy (1987):

i. 3'ss YYY YYY YYY YNC AGG

ii. 3'ss YYY YYN YAG'

iii. 5'ss NNW GGT RWGT

Based on the predicted base utilization of splice sites (Shapiro and Senapathy, 1987), the highly degenerate Oligo i. (4096-fold) should detect about 4.2% of all acceptor splice sites in the human genome. The shorter but less degenerate (256-fold) 9-mer, Oligo ii, should increase the level of detection of acceptor splice sites to around 36%. The 10-mer 5' donor splice site oligo (Oligo iii) should detect around 18% of donor sites. Oligo i. should only hybridize to the 3' splice site of intron 1, oligo ii which is less specific than i, should hybridize to the 3' splice sites of introns 1, 5 and 6, and oligo iii to the 5' splice sites of introns 4, 5 and 6. With the use of Southern blots of restriction digests of the various PLP subclones, only the expected hybridizations were observed, thus indicating that these oligos can indeed be used as probes to identify candidate splice sites in genomic DNA.

PLP is thus a suitable model template for investigating the use of such oligos as primers for PCR amplification. However, as mentioned previously (4.1.2), there may be difficulties using such highly degenerated oligos for PCR. For this reason it was first decided to attempt PCR amplification using normal 20-mer primers targeted at the splice sites surrounding exons 3 and 4 of the PLP gene and then to use smaller primers (11-mers) with just four-fold degeneracy targeted at the same sites, before attempting PCR with the highly degenerated sequences above.

5.2 PCR AMPLIFICATION USING PRIMERS TARGETED AT PLP SPLICE SITES

5.2.1 TEST HYBRIDIZATION OF OLIGOS

Two pairs of 20-mer primers were designed to amplify across exons 3 and 4:-

Exon 3: iII3' 5'TGT CTA CCT GTT AAT GCA GG 3'

iIII5' 5'AAT CCT GAG GAT GAT CAC CT 3'

Exon 4: iIII3' 5'CCC ATG TCA ATC ATT TTA GT 3'

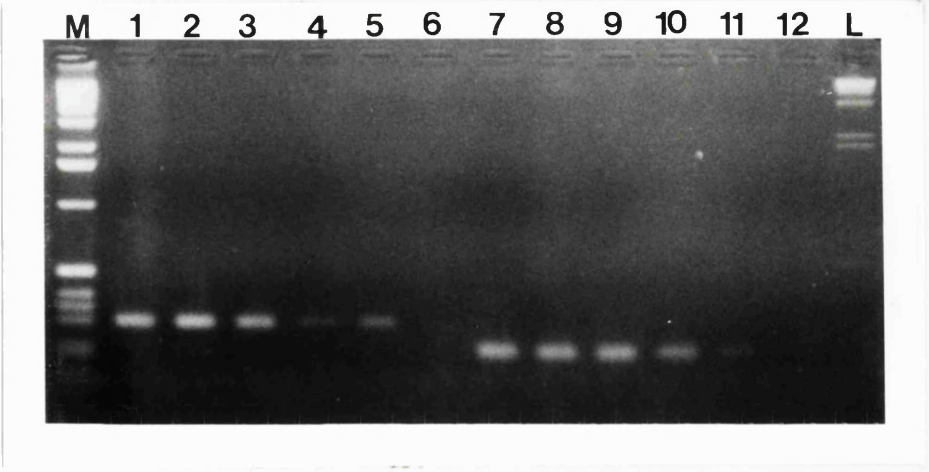
iIV5' 5'ACC CGT ACC CTA ACT CAC CA 3'

As a preliminary to PCR amplification experiments, the oligos were end-labelled and hybridized to Southern blots containing restriction fragments of the PLP subclones in order to check the specificity of the oligos and their annealing efficiency. Hybridization was performed at 50°C in 6.5x SSPE, 10 mM PO_4, 5x Denhardt's solution, 0.5% SDS, with tRNA and poly A, for 4 hours. The filters were washed in 4x SSPE, 0.1% SDS, followed by autoradiography. The oligos iII3' and iIII5' hybridized strongly to the appropriate fragments which were subclones 34H1 and 34S1, which both contain exons 2, 3 and 4. Oligo iIV5' also hybridized to these fragments, but less strongly, and iIII3' showed only very weak hybridization to the 34H1 insert.

5.2.2 PCR AMPLIFICATION USING SPLICE SITE 20-MERS

The PLP splice site 20-mers were used in PCR firstly with the PLP subclone 34S1 as template. Template concentration was titrated from 10 ng down to 1 pg with successful amplification. The products observed corresponded in size to those expected (299 bp for exon 3 and 201 bp for exon 4; see Fig. 5.1).

Figure 5.1 PCR amplification of PLP exons 3 and 4 using splice site 20-mers. PCR was performed on the PLP subclone 34S1, titrating the template DNA concentration, using primers iII3' and iIII5' (exon 3: lanes 1-6) and primers iIII3' and iIV5' (exon 4: lanes 7-12). Lanes 1 and 7 used 10ng of 34S1 as template, lanes 2 and 8 1ng, lanes 3 and 9 100pg, lanes 4 and 10 10pg, lanes 5 and 11 1pg and lanes 6 and 12 100fg. PCR was performed under standard conditions, annealing at 38°C. Lane M shows 1kb ladder and lane L shows λ/HindIII as size markers.



Standard PCR amplification was performed with 35 cycles of 94°C for 20 secs, 50°C for 2 mins and 72°C for 2 mins. Successful amplification was also achieved using total human genomic DNA as template. These primer pairs were thus suitable for use as positive controls and the individual primers for use as "anchor" primers paired with less specific oligos in PCR.

5.3 PCR AMPLIFICATION USING 11-MER SPLICE SITE PRIMERS WITH FOUR-FOLD DEGENERACY

A pair of primers were designed with the aim of amplifying exon 3 of the PLP gene with only 11 bases and with four-fold degeneracy This was in order to allow less specific amplification, possibly at other splice sites:-

Jss3': GTT AAT NCA GG **Td=28-30°C**

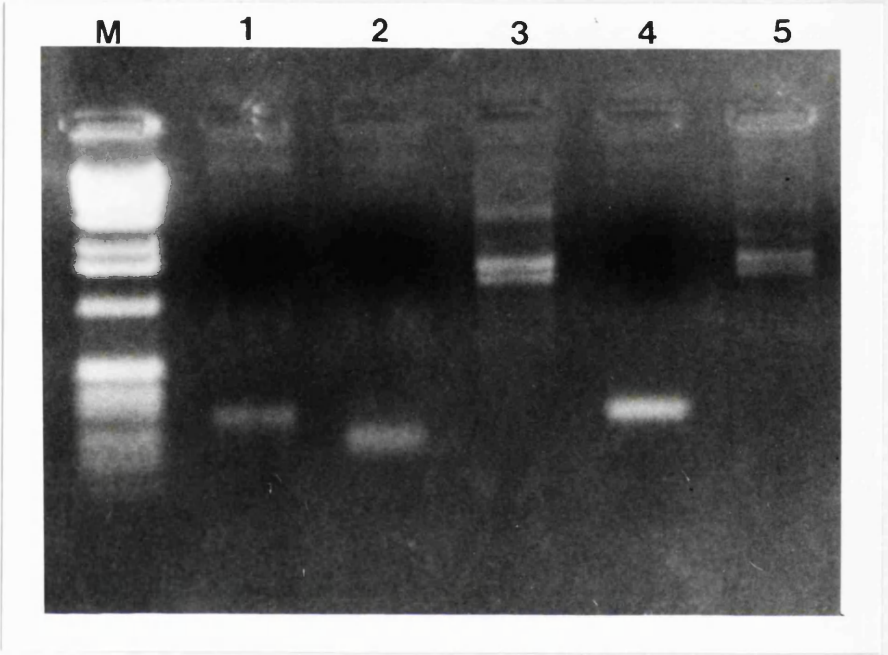
Jss5': GAT KAT CRC CT **Td=30-34°C** (Suggs et al, 1981)

Using the weight matrices and base frequencies for the acceptor and donor consensus sites (Senapathy et al, 1990), it is possible to calculate that Jss3' would match less than 0.001% of acceptor sites, and Jss5' about 0.0035% of donor sites. If an average gene has five exons and with around 100,000 genes in the human genome, we can estimate that Jss3' would detect about 5 acceptor sites and Jss5' about 17 donors. Hybridization experiments showed the oligos bound to the appropriate fragments of the PLP subclones at 30°C, although some background hybridization to vector DNA occurred.

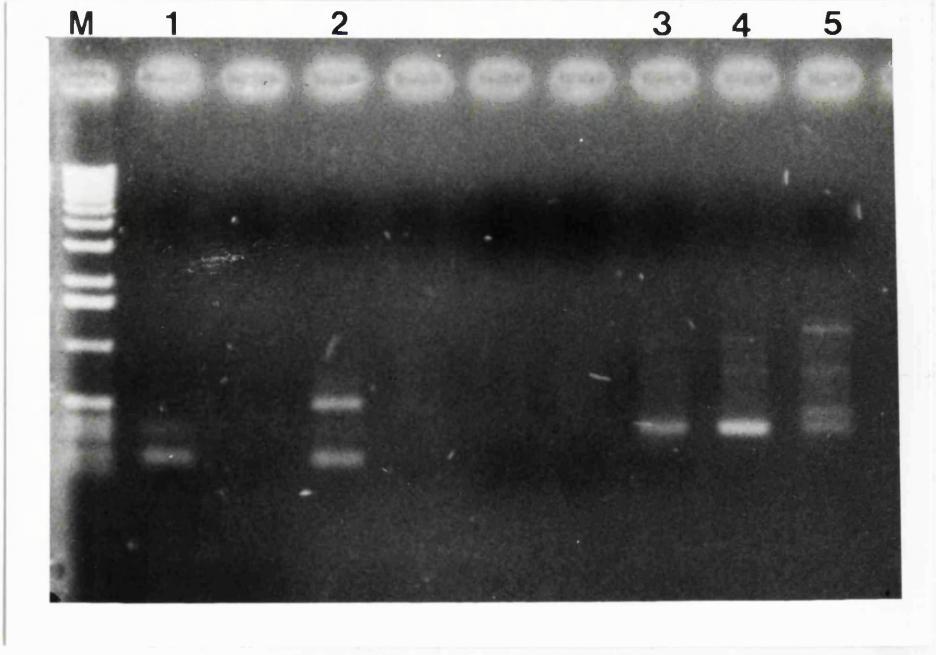
Firstly PCR was performed using 10 ng of the subclone 34H1 as template. The normal combinations of the splice site 20-mers were used to amplify exons 3 and 4 as positive controls. Amplification was also achieved using primers iII3' and iIV5' to produce a band of approximately 2 Kb, spanning exon 3, intron 3 and exon 4. The formula of Suggs et al (1981)

Figure 5.2 PCR amplification between splice site 20-mers and 11-mers. PCR was performed as described in the text (5.3).

a. Clone 34H1 was used as template (10ng) and primers used were as follows: lane 1. iII3' and iIII5', 2. iIII3' and iIV5', 3. iII3' and iIV5', 4. Jss3' and iIII5', 5. Jss3' and iIV5'. 1Kb ladder was run as a size marker (M).



b. Human genomic DNA was used as template (100ng) and primers used were as follows: lane 1. iIII3' and iIV5' (on 34H1-positive control), 2. iIII3' and iIV5', 3. iII3' and Jss5', 4. Jss3' and Jss5', 5. iIII3' and Jss5'. 1Kb was run as a size marker (M).



was used to calculate Td for the 11-mers, and these were used as a guide for the annealing temperature for the PCR. Amplification using the 11-mer Jss3' versus iIII5' and iIV5' respectively produced the expected bands (299 and ~2 Kb). PCR included 10% DMSO, 2.5 mM $MgCl_2$, with 10 ng of 20-mers and 100ng of 11-mer in 25 μl reaction volume, for 35 cycles of 94°C for 20 secs, 32°C for 2 mins and 72°C for 2 mins (see Fig. 5.2a).

Successful amplification was achieved using as primer pairs iII3' and Jss5', and Jss3' and Jss5', with 10 ng of 34H1 as template, with 10% DMSO, 2.5 mM $MgCl_2$, as before. Using these primer pairs with total genomic DNA as template (no DMSO, 1.5 mM $MgCl_2$ and annealing at 30°C) several distinct bands were produced, in addition to the expected exon 3 band (see Fig. 5.2b). A second round of amplification using the excised bands as template, improved the amount of product.

Because there are $4^{11} = 4.19 \times 10^6$ possible 11-mers probability predicts that a random 11-mer will occur 1430 times in a $2x(3 \times 10^9)$ bp genome (or 5720 times for an 11-mer with four-fold degeneracy) therefore the presence of extra bands could be a chance phenomenon. However the chances of two random 11-mers occurring within PCR range of each other are very slim, so it would appear to be a fair assumption that the successful bands may be a result of amplification at other splice sites of similar sequences in other genes. A control experiment using pairs of random 11-mers would have been useful to confirm or refute this. If a similar number of bands are produced with the random primers then this would suggest that the 11-mers used were not amplifying from splice sites in other genes.

5.4 AMPLIFICATION BETWEEN 20-MER AND 11-MER SPLICE SITE OLIGOS AND RARE-CUTTER 8-MERS

As a possible approach to combining both the criterion for regions of DNA containing CpG

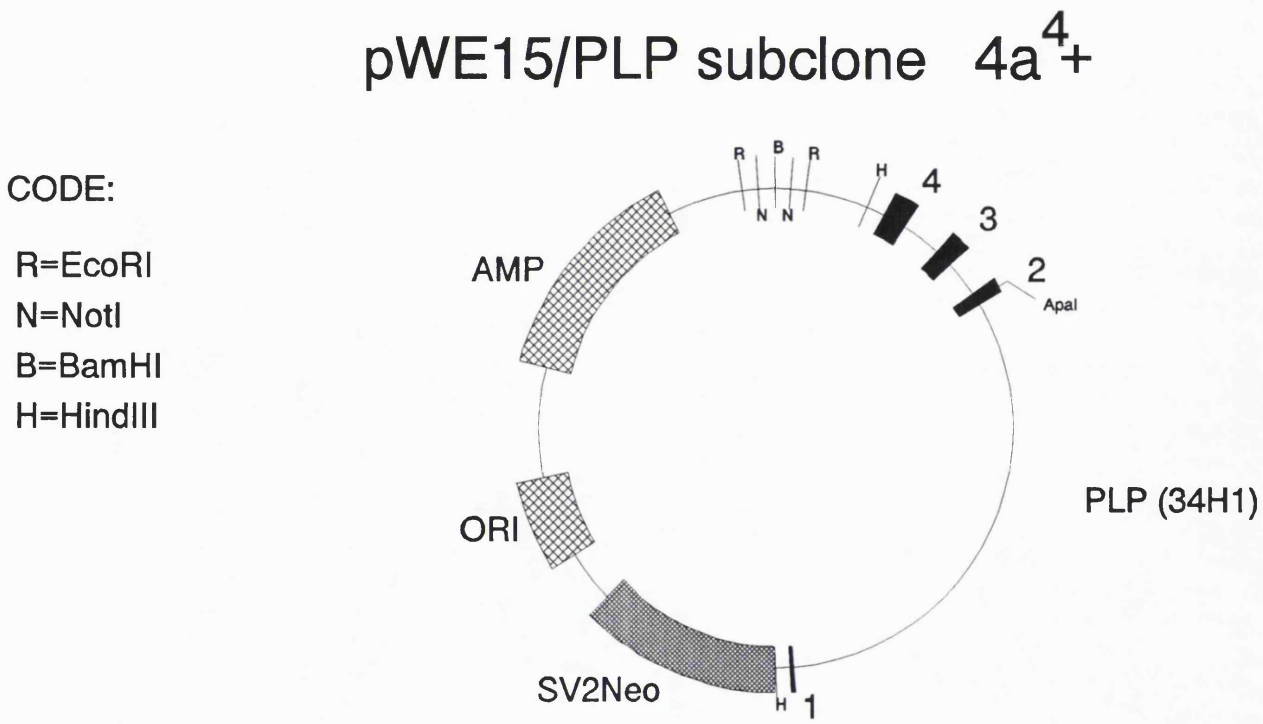


Figure 5.3: Schematic map of pWE15/34H1 subclone, showing the PLP coding regions and restriction sites.

rich sequences and for consensus sequences for splice sites, a model was developed that combined both features by subcloning the 14 Kb HindIII insert of the PLP subclone 34H1 into the vector pWE15 and by replacing a 2.2 Kb HindIII vector fragment adjacent to the NotI cloning site. The resulting subclone, $4a^4+$ (see Fig. 5.3) brings the NotI sites of pWE15 and exons 2, 3 and 4 of PLP into close proximity, thus allowing attempts at amplification between splice site and rare-cutter oligos. The orientation of the insert within the vector was determined by restriction digestion using enzymes with sites already mapped onto 34H1.

5.4.1 AMPLIFICATION BETWEEN SPLICE SITE 20-MERS AND THE NOTI 8-MER

A PCR amplification (under standard conditions, with 35 cycles of 94°C for 20 secs, 40°C for 2 mins and 72°C for 2 mins) using 10 ng of 4a4+ as template and the NotI 8-mer versus 20-mers iII3' and iIII3' respectively, produced several bands just visible on agarose gel by ethidium bromide staining and ultra-violet illumination (see Fig. 5.4a). The gel was blotted onto Hybond-N and hybridized firstly to the end-labelled iII3' oligo and then to the oligolabelled 34H1 insert. The iII3' probe hybridized strongly to the positive control product (iII3' versus iIII5', 299bp) and faintly to a ~800 bp band and fainter still to a ~1.7 kb band, both from the iII3' with NotI 8-mer product. This second band corresponds closely to the distance between the NotI site and the start of exon 3 on $4a^4+$ (see Fig. 5.4b). Hybridization with the 34H1 probe lit up both the above mentioned bands and also two bands, ~1 kb and ~600 bp, from the product of iIII3' versus NotI 8-mer. The smaller of these two bands corresponds roughly to the distance between the NotI site and the start of exon 4 on $4a^4+$ (see Fig. 5.4c). The 34H1 probe inexplicably fails to hybridize with the control product in lane 1.

Figure 5.4a PCR amplification of 4a4+ using NotI 8-mer and splice site 20-mers. PCR was performed on 10ng of the pWE15/PLP subclone 4a4+ with the following primer pairs: lanes 1. iII3' and iIII5' (positive control); 2. iII3' and NotI 16+; 3. iIII3' and NotI 16+; 4. iII3' NotI 8-mer; 5. iIII3' and NotI 8-mer; 6. just NotI 8-mer (negative control). 10ng of the 20-mer oligos were used and 25ng of the NotI oligos were used in the PCR. Amplification took place under standard conditions, with 35 cycles of denaturation for 20 secs, annealing at 40°C for 2 mins and chain extension at 72°C for 2 mins. Lanes marked M show 1Kb ladder as size marker.

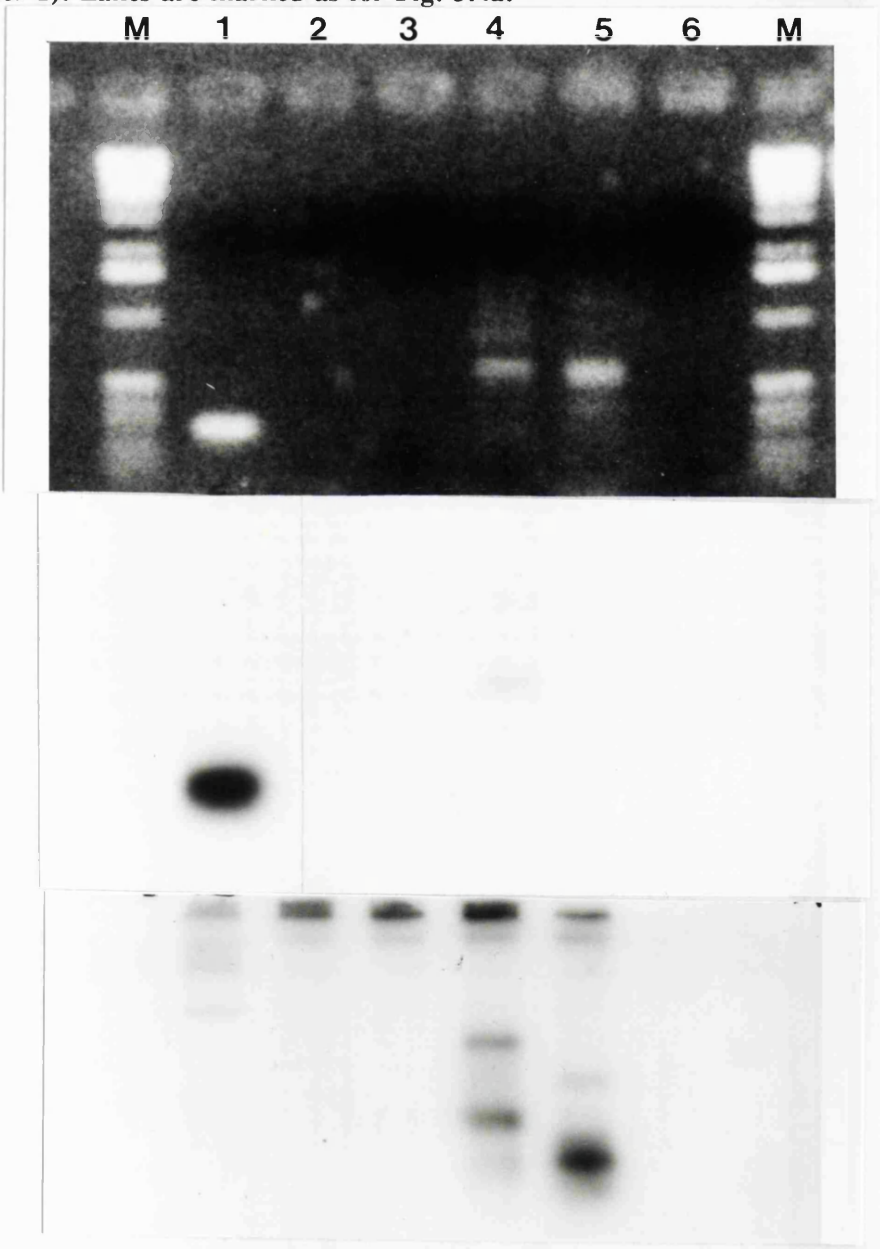
Figure 5.4b Back hybridization using iII3' as probe. The agarose gel shown in Fig. 5.4a was Southern blotted onto Hybond-N. Hybridization with end-labelled iII3' was performed in 6xSSC, 10mM PO_4 etc. at 50°C overnight. The filter was washed in 5xSSC, 0.1% SDS at room temperature for 10 mins. Lanes are marked as for Fig. 5.4a.

Figure 5.4c Back hybridization using 34H1 insert as probe. The stripped filter was hybridized to oligolabelled 34H1 insert under standard Southern hybridization procedure (see Chapter 2). Lanes are marked as for Fig. 5.4a.

a.

b.

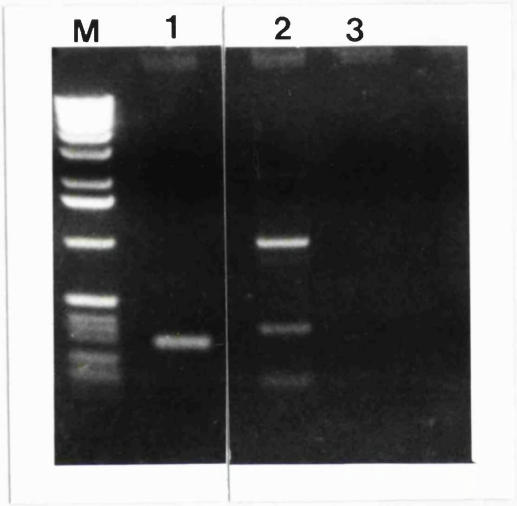
c.



5.4.2 AMPLIFICATION BETWEEN SPLICE SITE 11-MER JSS5' AND SMAI 8-MER

Successful amplification was also achieved between the splice site 11mer Jss5' and the SmaI octamer 5'GCCCGGGC 3', using subclone 34H1 as template, under standard conditions, with the primer annealing temperature at 32°C. A number of product bands were observed, most notably at ~1Kb, ~325bp and ~200bp (see Fig. 5.5).

Figure 5.5 PCR amplification on 34H1 using splice site 11-mer and SmaI 8-mer. Lane 1 shows amplification of 34H1 using iII3' and iIII5' as positive control. Lane 2 shows amplification between the SmaI 8-mer 5'GCCCGGGC 3' and oligo Jss5'. Lane 3 is a negatve control using just the SmaI 8-mer. 1Kb ladder is used as size marker (M). PCR conditions were as for Fig 5.4a, but annealing at 32°C.



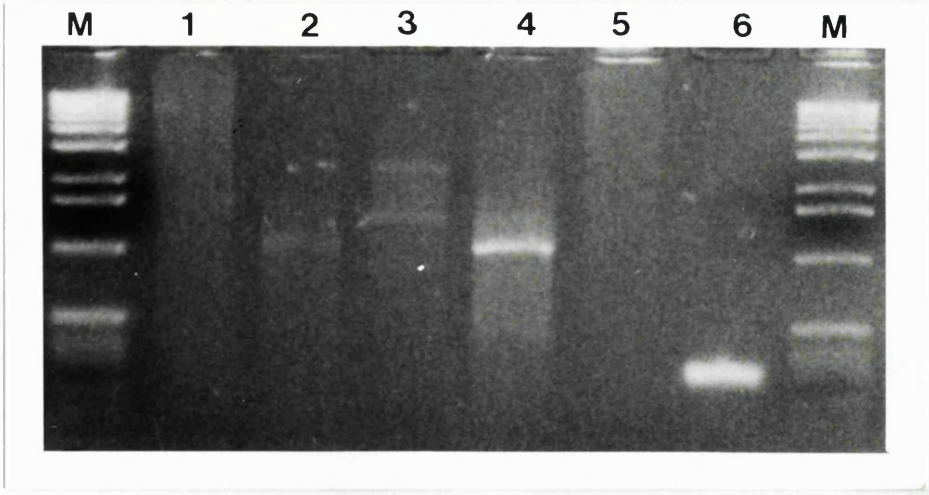
There are no SmaI sites in the vicinity of the PLP exon 3, let alone SmaI sites with full complementarity to the octamer used. However, assuming that correct priming is occurring with the 11mer splice site oligo and because the SmaI 8mer is annealing well below its optimum temperature, which is probably around 45°C, as described in the previous chapter, it is possible that at reduced stringency priming may have occurred from sites of less than full complementarity to the SmaI oligo. A site of 7/8 bp (GCCCGGCC) homology occurs at about 1.3Kb upstream to the Jss5' site. Although this does not correspond to any of the stronger bands produced, there are several faint bands in this region. A site of 6/8 bp homology (CCCCAGGC) occurs at about 1.1Kb upstream of the Jss5' site, also at 356bp (GCCAGGTC) and 119bp (GCCTGAGC). Sites of 5/8 bp complementarity occur at about

1.0Kb (TCCCTGGT), also at about 900bp (TCCCGAGT), 391bp (GCTCTGGT), 366bp (CGGGAAAG), 176bp (GCGCAGTC) and 48bp (GAGCGGGT). The prominent bands could correspond to amplification from some of the above sites, particularly those complementary at the 3' end of the oligo (mismatches at the 3' end would almost certainly prevent chain extension, Huang et al, 1992).

5.4.3 AMPLIFICATION BETWEEN THE SPLICE SITE 11-MER JSS5' AND APAI 8-MER

Using the PLP subclone 34S1 as template, amplification was achieved between the ApaI oligo, 5'GGGGCCCC 3', and the 5' splice site 11-mer Jss5'. PCR conditions were as standard except for the presence of 10% DMSO and 30mM KCl concentration and annealing at 40°C for 2 mins. A strong band of about 1 Kb was observed on agarose gel stained with ethidium bromide under uv-illumination (see Fig. 5.6).

Figure 5.6 PCR amplification on 34S1 using splice site 11-mer and ApaI 8-mer. Lanes 1-5 use the ApaI 8-mer 5'GGGGCCCC 3' (100ng) as primer. In addition lane 2 used 20-mer iIII5' (50ng), lane 3 20-mer iIV5' (50ng), lane 4 11-mer Jss5' (200ng), lane 5 highly degenerate splice site oligo 686 (200ng), and lane 6 20-mers iIII5' and iII3' as positive control (50ng each). PCR was performed as described in the text. Lanes marked M show 1Kb ladder as size marker.



The band appeared to correspond to amplification between an ApaI site with 100% complementarity to the ApaI 8-mer in exon II and the Jss5' site at donor 3 which is a region

of around 1.15 Kb. Amplification between the ApaI 8-mer and the 20-mer iIII5', which corresponds to the same splice site as Jss5', also produces a faint band of the same size.

5.5 AMPLIFICATION BETWEEN TWO HIGHLY DEGENERATE SPLICE SITE PRIMERS

Having attempted amplification between specific 20mers with complete success and using slightly degenerated 11mer oligos with some success, it was decided to attempt PCR using the highly degenerated oligos designed from the consensus matrices (Shapiro and Senapathy, 1987). To compensate for the degeneracy of the primers, much larger amounts were used in the reactions ($\sim$0.5-1μg). Also, because strong amplification products were not anticipated, α-^{32}P labelled dCTP was incorporated into the reactions. The products were separated on polyacrylamide rather than agarose gels and visualized by autoradiography.

5.5.1 SELECTION OF DEGENERATE PRIMERS FOR PCR

The following oligonucleotides were designed using the consensus matrices, as discussed previously. The first three were used in the hybridization experiments by Melmer and Buchwald (1992), as discussed in Chapter 3.3:

406 3'splice site 5'YYY YYY YYY YNC AGG 3'

469 3'splice site 5'YYY YYN YAG 3'

493 5'splice site 5'NNW GGT RWG T 3'

686 reverse 5'splice site 5'NAC WYA ACC WNN T 3'

730b translation start site 5'NNR TSA TGN KG 3'

732 reverse polyadenylation signal 5'NNTTTATTNNT 3'

433 SP1 transcription factor binding site 5'KRG GCG KRR Y 3'

712 AP1 transcription factor binding site 5'NST GAC TMA N 3'

713 liver transcription factor binding site 5'IHW MHV ACT SHN 3'

For PCR reverse strand oligos were required for the 5'splice site and poly A signal primers, in order to prime extension in the correct orientation for amplification of coding regions.

5.5.2 PCR ON MODEL TEMPLATE

Using the PLP 34H1 subclone 14 Kb insert as a model template, the 5' splice site 12-mer was used for PCR amplification versus a number of oligos, including the 3'splice site 15-mer and 9-mer. The reactions were performed in low salt concentration (30mM KCl and 1.2mM $MgCl_2$) and 10% DMSO and incorporating 0.4μCi of [α-^{32}P]-dCTP. 30 cycles of 95°C for 30 secs, 30°C for 2 mins and 72°C. PCR amplification was carried out with the 3' splice site 9-mer as well as several other oligos, including an 11-mer translation start site oligo and oligos for the SP1, AP1 and liver transcription factor binding. A large number of bands were observed, with many common bands between the reactions. This would suggest that low-specificity of priming had occurred. However, with the 3'splice site 15-mer just one strong amplification product band was observed at about 1.4 Kb and two faint bands at about 400 and 500 bp were observed (see Fig. 5.7). The 3'splice site 15-mer has 100% homology only with the acceptor splice site between intron I and exon II of PLP. The 5'splice site 12-mer, although having no 100% homology with any of the PLP donor splice sites, has strong (10 out of 12 bp) homology with the donor splice site between exon III and intron III. The 10 bp homology occurs at the 3' end of the primer. Correct annealing at the 3' end of a primer is crucial for successful primer extension and cannot occur where primers anneal with a mismatch at their 3' end (Huang et al, 1992). Table 5.1 shows the homologies of the two 3' splice site primers and the 5'splice site primer with the acceptor and donor splice sites present in the PLP 34H1 subclone. The 1.4 Kb amplification product corresponds closely to the expected band size for amplification between the PLP acceptor I and donor III (See Fig. 3.4

for PLP map).

Figure 5.7 PCR amplification on 34H1 using highly degenerate splice site primers. Lanes 1-9 used reverse splice site oligo 686. Opposite strand primers were as follows: lane 1 oligo 406, 2. 469, 3. 730b, 4. 737, 5. 433, 6. 712, 7. 713, 8. iII3', 9. blank. See 5.5.1 for key to oligos. Lane 10 had no primers. Lane M shows end labelled 1Kb ladder as size marker. 100ng of each primer was used under PCR conditions as described in the text (5.5.2).

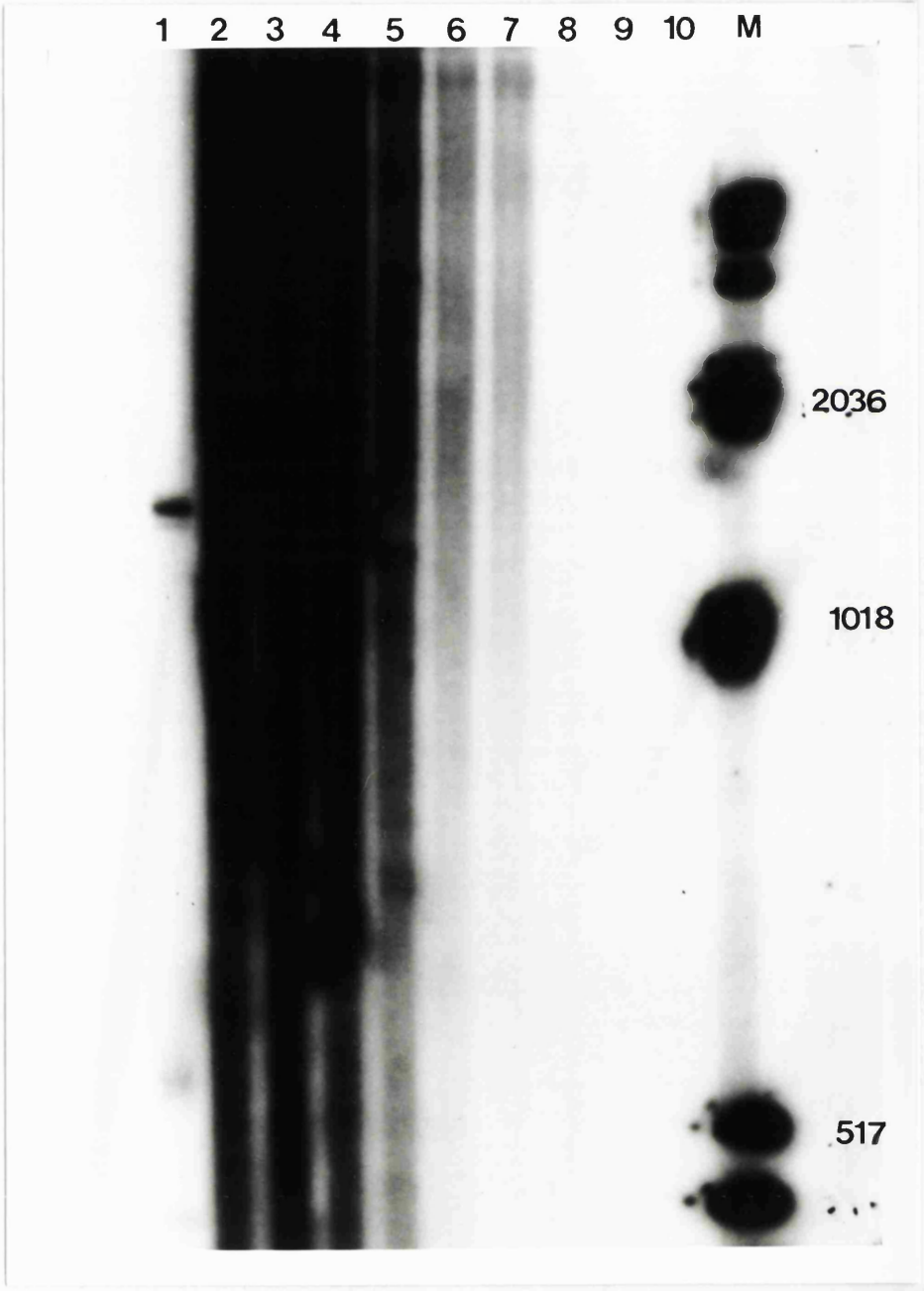


Table 5.1 Splice sites in PLP subclone 34H1 and homology of degenerate PCR primers. Bases shown in bold type represent mismatches with the oligo.

Oligo 406	5'YYY YYY YYY YNC AGG 3'
acceptor 1	ttc ttc ttc ccc agg
acceptor 2	acc tgt taa tgc agg
acceptor 3	gtc aat cat ttt agt

Oligo 469	5'Y YYY YNY AG 3'
acceptor 1	c ttc ccc ag
acceptor 2	t taa tgc ag
acceptor 3	t cat ttt ag

Oligo 686	5'NAC WYA CCW NNT 3'
reverse	5'ANN WGG TRW GTN 3'
donor 1	cat ggg taa gtt
donor 2	aat gtg taa gta
donor 3	aca agg tga tca
donor 4	gta tgg tga gtt

Whilst probability predicts that the degenerate 3'splice site 15-mer should have on average one target site per 262,144 bases and the 5'splice site 12-mer one target site per 32,768 bases, the 3'splice site 9-mer would have one target site per 1024 bases. Thus such a highly permissive primer, as the results here suggest, may be too ambiguous for use as a splice site-specific primer, even though as a hybridization primer it was shown to anneal to the correct fragments of the PLP subclones (Melmer and Buchwald, 1992).

Whilst the stringency in the amplification reactions was good for the 3'splice site 15-mer, it may have been too low for the 3'splice site 9-mer and the other oligos (which were 12 bases or shorter) and the possible annealing sites too frequent in the background template DNA. As

in the previous chapter, specificity of the PCR could be increased by increasing the reaction stringency, by maximizing annealing temperature and reducing salt concentration. However, in the previous chapter, a specific model template was available, with the primers homologous only to the two specified sites. The PLP model used for the splice site primers would probably not be specific enough as template for reaction-optimization experiments using such highly degenerated short oligos.

5.6 PCR ON GENOMIC DNA

Genomic DNA was used as template for PCR as described before with the splice site oligos as primers. When the amplification product was run on an agarose gel or polyacrylamide gel it showed a smear of bands across a large range of sizes (< 2Kb for PCR with 3'splice site 9-mer and 5'splice site 12-mer). Product smears were also obtained when using only one primer in the reaction. Thus it is unclear whether the PCR reaction is specifically amplifying coding regions. In order to resolve this matter several approaches were attempted.

5.6.1 SCREENING A CDNA LIBRARY WITH SPLICE SITE PCR PRODUCT

In an attempt to ascertain whether the smear of amplification product produced by PCR using the highly degenerate splice site oligos consisted of a coding region-rich fraction, a human foetal heart λgt10 cDNA library was screened using the random labèlled PCR product from a genomic amplification using the 3'splice site 9-mer (469) and the 5'splice site 12-mer (686). The labelled product hybridized to a large number of plaques. However it is quite possible that genomic amplification product using completely random primers would give a similar result, particularly in consideration of the fact that many cDNA clones contain DNA repeat sequences. In order to eliminate these possibilities , it was decided upon to select some of the positive plaques and to sequence them in order to analyse such sequences for the presence of either repeat DNA or coding DNA.

5.6.2 SUBCLONING AND SEQUENCING POSITIVE CLONES

12 plaques that had hybridized strongly with the labelled PCR product were selected and picked. After a second round of screening, individual plaques which gave a strong signal were selected and plugs taken. The phage were grown and DNA was prepared from the clones, as described in the Methods chapter. Insert DNA was PCR amplified using λgt10 primers. 5 of the product bands were selected and purified using Quik™columns (Stratagene). Due to the difficulties involved when blunt-end cloning PCR product (Taq polymerase produces DNA fragments with "ragged ends" through terminal transadenylation activity) it was decided to "shotgun" clone the product by digesting with frequent cutting enzyme Sau3A and subcloning the fragments into BamHI digested M13 vector. 9 recombinant plaques were selected and from them single stranded DNA was prepared for ^{35}S/Sequenase™ sequencing.

Only four different sequences were obtained. These sequences (shown in Appendix D) were analysed using the BLAST program, to search the Genbank database for homologous sequences (Altschul et al, 1990). Two sequences (2a1, 2b1) showed homologies with various eukaryotic genes (see Table 5.2a). The third sequence showed homology with the calmodulin gene from human, chicken, rat, xenopus and others and is therefore highly likely to represent coding DNA. The fourth sequence (2d2) showed homology with a considerable number of sequences in the database, from a wide variety of eukaryotic species and including a large number of sequences containing Alu repeats. Thus 2d2 is likely to contain a common repeat element.

Table 5.2a

Clone	bp	Maximum homology (BLAST score)
2a1	112	O.volvulus microfacial surface antigen mRNA (114)
2b1	202	Human TRPM-2 protein gene (134)
2c1	108	Human calmodulin mRNA (142)
2d2	135	Chicken 7S RNA sequence (168)

From these few results we can conclude that at least 75% of the clones (selected by screening with PCR product from the splice site amplification) had hybridized to PCR product consisting of coding sequences. It is difficult to speculate whether this represents an enrichment for coding regions in the PCR product, or whether the PCR product has a normal distribution of coding sequences (below ~5%) and only the coding fraction is hybridizing to clones in the cDNA library.

The PCR amplified insert of the first clone (2a1) was oligolabelled and used as a probe against a Southern blot of the splice site oligo/genomic DNA PCR product (see Fig. 5.8). The hybridization produced strong signal with several common bands in each lane for the genomic amplifications. This would suggest that the common primer in these three lanes, the 5'splice site 12-mer (686), is amplifying these bands alone, acting as both forward and reverse primer. When the insert (2a1) was used as a probe for Southern genomic digests, several distinct bands were seen, suggesting that the DNA segment is a unique sequence rather than a repeat sequence.

5.6.3 M13 CLONING OF CODING REGION PCR PRODUCT

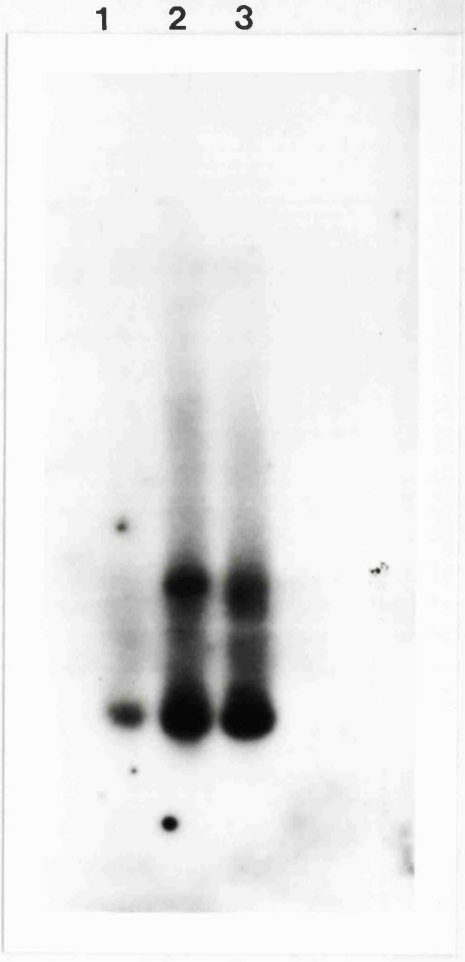
To test the hypothesis that the consensus site PCR product does itself represent a coding region-rich DNA fraction, PCR was performed in the following experiments :

i. The 11-mer and 15-mer 3'splice site primers (469 and 406) and the translation start site primer (730b) with the 12-mer 5'splice site primer (686) on genomic DNA template.

ii. The 11-mer and 15-mer 3'splice site primers (469 and 406) and the translation start site primer (730b) with the 12-mer 5'splice site primer (686) on DNA from a chromosome 5 cosmid library.

iii. The SP1 transcription factor binding site primer (433) with the 12-mer 5'splice site primer (686) on genomic DNA.

Figure 5.8 Southern hybridization of clone 2a1 versus a blot of PCR product using degenerate splice site primers on genomic DNA. PCR was performed on single stranded human total genomic DNA (10ng), using as primers oligos 406 and 686 (lane 1), 469 and 686 (lane 2), 730b and 686 (lane 3). Approximately 2μg of each primer was used. 30mM KCl and 2.5mM $MgCl_2$ was used in the reactions, and annealing at 30°C. The probe was prepared from the insert of λgt10 clone 2a1 by PCR amplification using λgt10-specific primers flanking the insert followed by random primer labelling. After Southern hybridization the filter was washed to a stringency of 1xSSC, 0.1% SDS.



Because of the difficulties encountered when blunt-ended "shotgun"cloning of the PCR product was attempted (Taq polymerase produces DNA fragments with "ragged ends" through terminal transadenylation activity) the product was digested with Sau3A and was cloned into the BamHI site of the sequencing vector M13. Only a small number of recombinant positive plaques were observed. Recombinants were picked and single stranded template DNA was prepared. Sequencing using ^{35}S and Sequenase™ was performed and the results were analysed

for homology to known sequences using the BLAST program (Altschul et al, 1990). The GCG programs COMPOSITION and MAP (Genetics Computer Group, 1991) were used for analysis of base content and presence of open reading frames using . The results are shown in table form (Table 5.2b). Sequences are shown in Appendix E.

Table 5.2b Sequence analysis of PCR product.

PCR	clone	bp	%G+C	O/E CpG	ORF's
i.	8b1	30	53.3	1.091	4
i.	8b2	45	61.0	1.071	2
ii.	8a1	127	56.7	1.078	3
ii.	8a2	82	57.3	1.640	1
iii.	7a1	135	42.9	1.290	3
iii.	7a3	65	56.9	0.867	4

Table 5.2b cont.

Clone	Max. Homologies (complementarity score)
8b1	C. briggsae gut esterase (83)
8b2	Human excision repair protein ERCCG mRNA (94)
8a1	NONE
8a2	E. coli ebg genes (127)
7a1	S. kluyveri plasmid pSKL DNA (108)
7a3	E. coli Asn t-RNA synthetase (291)

As can be seen the cloned sequences are far too short and too few for any meaningful analysis to verify whether they originate from coding regions. Despite the high CpG content and G+C content, which suggest the sequences may be part of CpG islands and thus possibly coding regions, the homologies detected using BLAST suggest that some of the sequences (8a2, 7a1 and 7a3) may originate from vector or host DNA. Also, for such short sequences, the presence of open reading frames has little or no significance.

5.7 DISCUSSION

In order to analyse the consensus site-PCR product in a realistic way, a much better approach to the cloning and sequencing of the PCR product is required. For instance it would be very useful to compare the sequence of the primers used to that of the primer annealing site. This would allow analysis of the primer annealing efficiency. In addition, if a model template DNA were used in which all the consensus sites had been characterized, it would permit quantification of the efficiency of the method of detection. The TA cloning system as used in Chapter 4 would have been useful for this, however the DOP-PCR method was chosen for the added advantages of primer stability and primer-cloning sites, as discussed in Chapter 6.

In order to ascertain whether or not the PCR amplification of genomic DNA using the degenerate splice site primers actually enriches for coding regions, a control experiment would need to be carried out in which random primers of similar length and degeneracy are used to PCR-amplify genomic DNA. The resulting products would then be sequenced, or used to screen a library, or hybridized against Southern blots or zoo blots, in order to establish whether a similar proportion of the product from the splice site primers and from the random primers is from coding regions.

Whilst it has been suggested that primers containing greater than 516-fold degeneracy would generate non-specific DNA fragments (Compton, 1990), other reports have shown that degenerate oligonucleotides with complexities of up to 8192 can be usefully employed as PCR primers to screen libraries (Heller et el, 1992). The primers used in these situations were longer (21 bases) than those used in the experiments in this chapter. Methods of effectively decreasing the degeneracy or increasing the stability of oligos by substituting certain base analogues have been discussed in Chapter 3 (3.4) and could equally be employed in oligos for PCR. The use of TMAC for hybridization was also discussed and may also be useful in

PCR situations with degenerate primers, since it has been shown to improve specificity of PCR (Hung et al, 1990).

CHAPTER 6: DEGENERATE OLIGONUCLEOTIDE PRIMED-PCR (DOP-PCR) WITH CONSENSUS SEQUENCE PRIMERS

6.1 INTRODUCTION

The lack of success with the short, highly degenerate consensus site oligonucleotides as primers for PCR on genomic DNA (Chapter 5) may have a number of causes, such as the instability of such short primers, a high amount of annealing to background/bulk genomic DNA and the difficulties in assessing the success or failure of the method. A different approach, also involving the use of model templates, was decided upon. In addition to generating libraries of coding regions from whole genomes, the use of primers based on consensus sites with or near to coding regions for PCR-amplification could also be used to screen cloned genomic DNA for genes. As with the idea of screening genomic clones by hybridization with consensus oligos (see Chapter 3), if successful this would speed up the process of searching for coding regions as discussed in Chapter 1.

The model templates chosen were the same as those used in Chapters 3, 4 and 5, namely cosGSTrp7 containing the glutathione-S-transferase π gene and the 5' region of the NADH-ubiquinone oxidoreductase gene (Cowell et al, 1988; Spencer et al, 1992), cosCT1 and cosCT2 containing the calcitonin/αCGRP gene (Broad et al, 1988) and the PLP subclones 34H1 and 34S1 (Diehl et al, 1986). Unsuccessful attempts were made to PCR amplify coding regions from the above clones, using the following primer pairs:

406 (3' splice site 15-mer)/ 686 (reverse 5'splice site 12-mer)

406/ 732 (reverse polyadenylation signal 11-mer)

730b (translation start site 11-mer)/ 686

433 (SP1 transcription factor binding site 10-mer)/ 686

Oligonucleotide sequences are listed in 5.5.2

Because of the inability to PCR amplify sufficient product (ie. visible on agarose gel by ethidium bromide staining) using these oligos, it was decided to attempt a technique known as degenerate oligonucleotide primed-PCR (DOP-PCR), as described by Telenius et al (1992). DOP-PCR was designed to give general amplification of target DNAs at frequently occurring priming sites, without restrictions due to DNA complexity or species specificity. It involves the principle of priming from short sequences specified by the 3' end of the oligonucleotide used, during initial low annealing temperature cycles of the PCR. Since these short sequences occur frequently, amplification proceeds at many sites simultaneously. The annealing of the specified 3' sequence is stabilized by the adjacent six degenerate bases. At the 5' end of the oligo is a further specified sequence, which includes restriction sites for cloning the PCR product.

6.2 PRIMER SELECTION

The splice site consensus matrices of Senapathy et al (1990) were used as a basis for the 3' end of the oligos. The forward primers were designed with HindIII and AsuII restriction sites and the reverse primers with BamHI and SacII sites. The primers first selected for use against the model cosmids were:

101 reverse 5'splice site 5'CCG CGG ATC AWC YYA CCD VG 3'

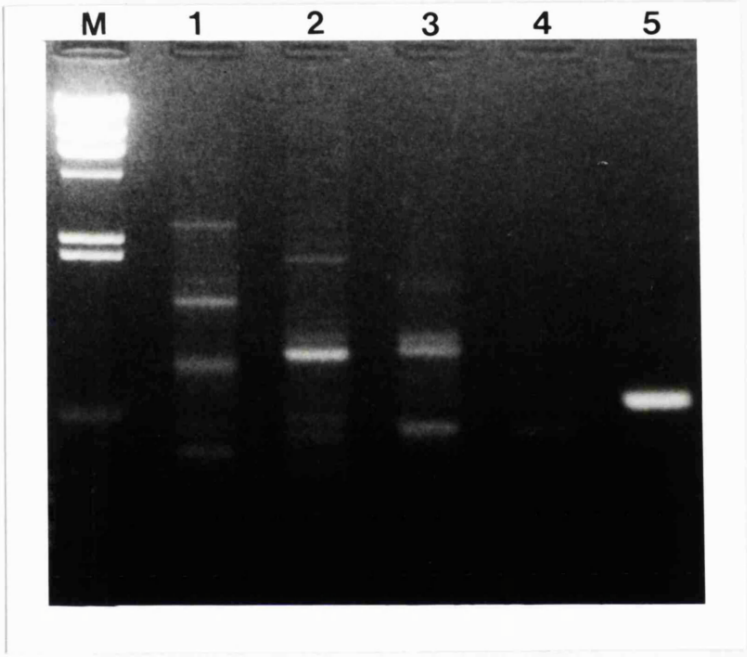
102 3'splice site 5'TTC GAA GCT TYY YYY NCA G 3'

6.2.1 AMPLIFICATION OF CLONES BY DOP-PCR

PCR was performed under standard conditions and with six preliminary rounds of low

temperature annealing to template (cosGSTrp7, cosCT1, cosCT2 and 34H1) at 24°C for 5 mins and slowly increasing the temperature to 72°C over another 5 mins, followed by 30 cycles annealing at 51°C. Using primers 101 and 102, a number of product bands were observed on agarose gel through ethidium bromide staining (see Fig. 6.1).

Figure 6.1 Amplification of exons in cosmid clones by DOP-PCR. Oligos 101 and 102 were used (100ng of each), and 1ng of template: in lane 1, 34H1, lane 2 cosGSTrp7, lane 3 cosCT1, lane 4 cosCT2. Lane 5 shows a positive control using subclone pWE15/585 and primers NotI 16+/-. λ/HindIII ladder is shown as a size marker (M). 60mM KCl was used in the PCR. Six preliminary cycles were carried out, denaturing at 95°C for 30 secs, annealing at 23°C increasing to 72°C over 5 mins, and extending at 72°C for 2 mins, followed by 30 cycles annealing at 49°C for 2 mins.



The PCR product was purified directly using PCR purification columns (Promega), then digested with HindIII and BamHI and cloned into M13mp18 and 19. Recombinants were selected and grown and single stranded DNA prepared. The clones were sequenced using the ^{35}S/Sequenase method, with M13 primers.

6.2.2 SEQUENCE ANALYSIS

Where the primers have amplified up known sequences, analysing the sequence of the PCR

product allows the comparison of the DOP-PCR primer sequence versus the target sequence. This should enable an evaluation of the efficiency of the primers; for instance how far along from the 3' end of the primer correct annealing occurs and also whether or not the primers are amplifying from known splice sites in the target sequence, or from sites which fit the consensus sequence but where splicing is not thought to occur (ie. cryptic splice sites).

Sequences were analysed using the BLAST program (Altschul et al, 1990), in order to indicate whether the sequence is a true recombinant and to localize the sequence within the cosmid. Sequences are shown in Appendix F. The results are summarized in Table 6.1

Table 6.1: Sequence analysis of cloned DOP-PCR products.

SEQUENCE	PCR TEMPLATE	BLAST SCORE	BLAST HOMOLOGY
PLP2	PLP 34H1		NONE
CT11	cosCT1	320	X02330 Hum mRNA calcitonin
CT12	cosCT1	417	X03145 Hum Kpn1 repeat element
CT13	cosCT1	714	M19503 Hum L1 repeat element
CT15	cosCT1	381	X52235 Hum L1 repeat element
CT16	cosCT1		NONE
CT17	cosCT1	487	M22334 Hum L1 repeat element
GST21	cosGSTrp7	237	X15673 Hum PTR2 mRNA repeat
GST22	cosGSTrp7	598	X15674 Hum PTR5 mRNA repeat
GST24	cosGSTrp7	231	S53175 NADH-ubiquinone oxid.
GST25	cosGSTrp7	134	L22403 Hum DNA repeat region
GST26	cosGSTrp7	472	X15673 PTR2 mRNA repeat
PLP2 (reverse)	PLP 34H1	103	L18873 viral expression

Sequence CT11 corresponds to the correct sequence of the calcitonin/CGRP gene, stretching

from the acceptor splice site of exon 4 to a region of intron 4 that shows strong homology to the donor splice site consensus sequence (a putative cryptic splice site). The PCR primers and their priming sites in cosCT1 are shown below:

oligo 102 5' TTCGAAGCTTYYYYYNCAG (acceptor)

cosCT1 GTATGTGTTTTCCCTGCAG

oligo 101 5' CCGCGGATCCAWCYYACCDVG (donor)

cosCT1 AAATTCCTCAGTCTTACCTGG

As can be seen, the primers show strong homology with their target sites at the 3' ends of the oligos (102 has 11bp homology from the 3' end, and 101 has 10bp homology), thus demonstrating the principle behind DOP-PCR. Clone GST24 shows strong homology to the published sequence for the NADH-ubiquinone oxidoreductase gene but the sequence occurs within intron 1. After analysis of experiments using oligos 101 and 102, it was realized that primer 101 was too degenerate at its 3' end. In an attempt to improve the efficiency of the system, a second reverse 5'splice site primer was designed, with the most conserved part of the consensus sequence at the 3'most position:

107 reverse 5'splice site 5'GTC CGC GGA TCC NAW CYY ACC 3'

DOP-PCR was performed using primers 102 and 107, using as template the clones 34H1, cosGSTrp7, cosCT1, cosCT2, cosD5/3 and cosD5/11, under similar conditions as used previously and was also repeated with the initial annealing temperature raised to 37°C, in order to increase the stringency. CosD5/3 and D5/11 contain the gene for the dopamine receptor DRD5, which is intronless. These clones were included as controls since the gene

has no splice sites and therefore DOP-PCR should theoretically not amplify in this gene. PCR product was cloned, as before, into M13mp18 and mp19 and sequenced. The sequences are shown in Appendix G. BLAST analysis of the sequences was carried out to identify whereabouts on the clones the PCR product originated and whether the primers were identifying splice sites. The results are summarized in Table 6.1b and 6.1c.

Table 6.1b: M13 clones from DOP-PCR using oligos 102 and 107 on cloned DNA and BLAST results showing sequence of maximum homology and BLAST score.

SEQUENCE	PCR TEMPLATE	BLAST SCORE	BLAST HOMOLOGY
mp18x1a	PLP 34H1	355	M15026 Hum myelin PLP
mp18x1b	PLP 34H1		NONE
mp18x2b	cosGSTrp7	305	X080895 Hum Glutathione-s-transferase pi gene
mp18x4b	cosCT2	630	M27155 E. coli Arg U-tRNA synth. gene
mp18x6b	cosD5/3	103	M93661 Rat Notch 2 mRNA
mp18x3b	cosGSTrp7	100	M67495 S. sulfataricus ribosomal RNA gene
mp18x2a	cosGSTrp7	88	J01787 S. dysenteriae trp d gene
mp18x4a	cosCT2	511	M27155 E. coli Arg U-tRNA synth. gene
mp18x5a	cosd5/3	250	L09795 Hum chromosome 4 STS4
mp19x1a	PLP 34H1		VECTOR
mp19x2a	cosGSTrp7	101	X06827 Rat porphobilinogen deaminase mRNA
mp19x2b	cosGSTrp7	400	X15675 Hum pTR7 mRNA for repeat
mp19x3a	cosGSTrp7	119	X62878 S. cerevisiae CDC-60 gene
mp19x3b	cosGSTrp7	101	X06827 Rat porphobilinogen deaminase mRNA
mp19x4a	cosCT2	122	M94130 C.elegans transformer protein

Sequence mp18x1a showed homology to the PLP genomic sequence, with the primer site for oligo 102 at a sequence with strong homology to the acceptor consensus situated about 150bp upstream of the first exon:

oligo 102 5' TTCGAAGCTTYYYYYNCAG (acceptor)

34H1 AAAGCCCTTTTCATTGCAG

The sequence extends to a BamH1 site 70bp upstream of exon 1 and the primer site for the donor oligo 107 was not established.

Sequence mp18x2b showed BLAST homology to the glutathione-s-transferase gene. The sequence extends from an internal HindIII site to the primer site for oligo 107, which is situated within intron 5:

oligo 107 5' GTCCGCGGATCCNAWCYYACC (donor)

cosGSTrp7 CACTTCAGCTGCCAACTCATC

The oligo has a mismatch with this sequence at the penultimate 3' base. This would imply that the oligo is functioning too non specifically. A higher annealing temperature may be required to increase the stringency.

Table 6.1c: M13 clones from DOP-PCR using oligos 102 and 107 on cloned DNA, annealing initially at 38°C and BLAST results showing sequence of maximum homology and BLAST score.

SEQUENCE	PCR TEMPLATE	BLAST SCORE	BLAST HOMOLOGY
mp18x2ii	cosGSTrp7	100	M67495 S. sulfataricus ribosomal RNA gene
mp18x3i	cosCT1	107	X52615 P. fluorescens cel B gene
mp18x3ii	cosCT1	631	X15943 Hum calcitonin/αCGRP gene
mp19x2i	cosGSTrp7	191	X15675 Hum pTR7 mRNA for repeat
mp19x2ii	cosGSTrp7	377	X15675 Hum pTR7 mRNA for repeat
mp19x4ai	cosD5/3	380	M13180 Epstein-Barr virus EBNA1 gene
mp19x4bi	cosD5/3		VECTOR

Sequence mp18x3ii showed homology to the calcitonin/CGRP genomic sequence, with the primer site for oligo 102 at a sequence with strong homology to the acceptor consensus sequence:

oligo 102 5' TTCGAAGCTTYYYYYNCAG (acceptor)

calcitonin GGGCCTCCTCTCCCCGCAG

This corresponds to a site 45bp into intron 3 of the calcitonin gene. Given that the DOP-PCR product was about 500bp, the reverse priming site may have been a cryptic donor site 492bp

downstream, also in intron 3.

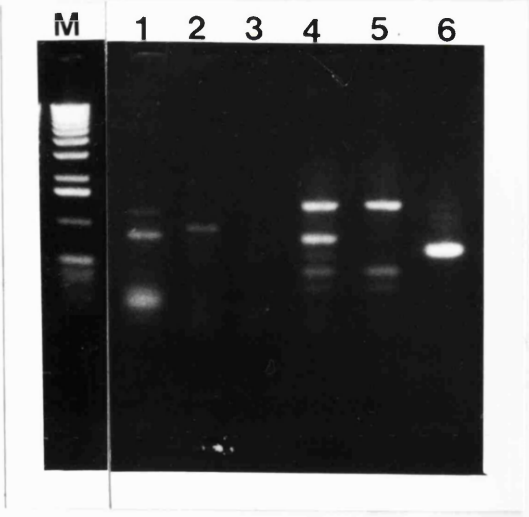
6.3 DOP-PCR BETWEEN TRANSLATION INITIATION AND 5' SPLICE SITE PRIMERS

DOP-PCR was attempted between reverse 5'splice site primer 107 and translation initiation site primer 103:

103 translation initiation site 5'TTC GAA GCT TAG YCR HVA TG 3'

under the same conditions as described in 6.2.1., using as template the clones 34H1, cosGSTrp7, cosCT1, cosCT2, cosD5/3 and cosD5/11. Oligo 103 should represent 4.7% of translation start sites, where the ten 3'-most bases cover the consensus sequence and 33.6% of sites with the eight 3'-most bases. Oligo 103 has strong homology with the translation start site of NADH:ubiquinone oxidoreductase (9bp at 3' end of oligo), PLP (6bp), calcitonin (6bp) and DRD5 (10bp). A number of product bands were observed on agarose gel through ethidium bromide staining (see Fig. 6.2).

Figure 6.2 DOP-PCR amplification of cloned DNA using translation initiation and 5'splice site primers. 100ng each of oligos 103 and 107 were used plus 1ng of the following clones: lane 1 34H1, lane 2 cosGSTrp7, lane 3 cosCT1, lane 4 cosCT2, lane 5 cosD5/3 and lane 6 cosD5/11. lane 6 shows PCR with pWE15/585 and primers NotI 16+/- as a positive control. 1KB ladder is shown as a size marker (M). 60mM KCl was used in the PCR. Six preliminary cycles were carried out, denaturing at 95°C for 30 secs, annealing at 24°C increasing to 72°C over 5 mins, and extending at 72°C for 2 mins, followed by 30 cycles annealing at 50°C for 2 mins. 1kb ladder was used as size marker (lane M).



6.4 DISCUSSION

The results in this chapter show only limited success for DOP-PCR in detecting splice sites and exons in cloned DNA. The only clear success was the amplification from the calcitonin/αCGRP exon 4 acceptor site across the exon to a site within intron 4. The primer site/ consensus sequence homologies ascertained through cloning and sequencing the DOP-PCR product and identifying the position of the PCR product using the BLAST program, reveal that the method is successful in that it identified and primed PCR from sites with strong homology to the splice site consensus sequences. However the majority of these putative splice sites do not correspond to the known splice sites of the cloned genes. A further problem with the technique was the frequency of BamHI and HindIII sites within the PCR product. Because the forward primer contained a HindIII site and the reverse primer a BamHI site for the purpose of cloning the PCR product into HindIII/BamHI digested M13, in several cases this prevented the successful cloning of the full amplified DNA segment. Although increasing the initial annealing temperature of the PCR product reduced the number of product bands visible on agarose gel (with ethidium bromide staining and u.v. illumination) it did not appear to improve the ability of the oligos to detect genuine splice sites. In addition, PCR controls should have been performed by using the individual primers singly, so that any product of amplification between just one primer could be identified and avoided. The presence of an excessive number of false splice sites within the PCR template seems to reduce the chances of DOP-PCR for detecting genuine splice sites and thus limits the usefulness of this method. It is known that there are other factors involved in the recognition of splice sites in vivo, other than the mere sequences immediately surrounding the donor and acceptor sites, for instance the lariat branch point which occurs within introns and recognition sites for various tissue specific splice factors. The method employed here may be more appropriate for PCR amplification from other gene regulatory sequences that are perhaps more unique within the vicinity of coding regions.

CHAPTER 7: DIRECT SEQUENCING OF COSMID SUBCLONES WITH CG OCTAMER PRIMERS

7.1 INTRODUCTION

Improvements in the efficiency of sequencing will help to increase the speed at which the Human Genome Project is completed. Several theoretical proposals using short oligos to prime at random and directed sites within cosmids for sequencing reactions, known as high-volume sequencing, have been put forward (Studier, 1989; Szybalski, 1990). Studier (1989) proposed that direct sequencing of cosmid DNAs could be performed using a library of octamers, nonamers or decamers. Studier calculated that an octamer would have 1.37 sites on average in a cosmid length 45-kb DNA fragment and a nonamer would have 0.343 sites. It has been suggested that GC-rich sequences and sequences with polynucleotide and dinucleotide repeats are unsuitable as short primers (Siemieniak and Slightom, 1990) and that primer polymerization efficiency of oligos shorter than nine nucleotides is very low (Kieleczawa et al, 1992). Contrary to this, the results presented in Chapter 4 confirm the hypothesis that primers at least as short as seven nucleotides can act as primers in a specific manner. In this chapter experiments were attempted using short and G+C-rich (octamer) oligos based on rare-cutter sites as primers in PCR coupled sequencing (Engelke et al., 1988; Innis et al., 1988; Ruano and Kidd, 1991) to test the hypothesis that targeted amplification and sequencing can be carried out, using a model vector as template. CpG islands often lie adjacent to or within coding regions (Gardiner-Garden and Frommer, 1987) and over 80% of *Not*I sites lie within CpG islands (Kusuda et al., 1990). Therefore methods using G+C-rich primers could be a useful means for targeting CpG islands and hence genes. The use of GC-composed octamers for PCR-sequencing into CpG islands of several well characterised

genes subcloned from cosmids is also reported in this chapter, as well as attempts to ligate primer/linkers to rare-cutter digested cosmid clones to enable direct sequencing into CpG islands.

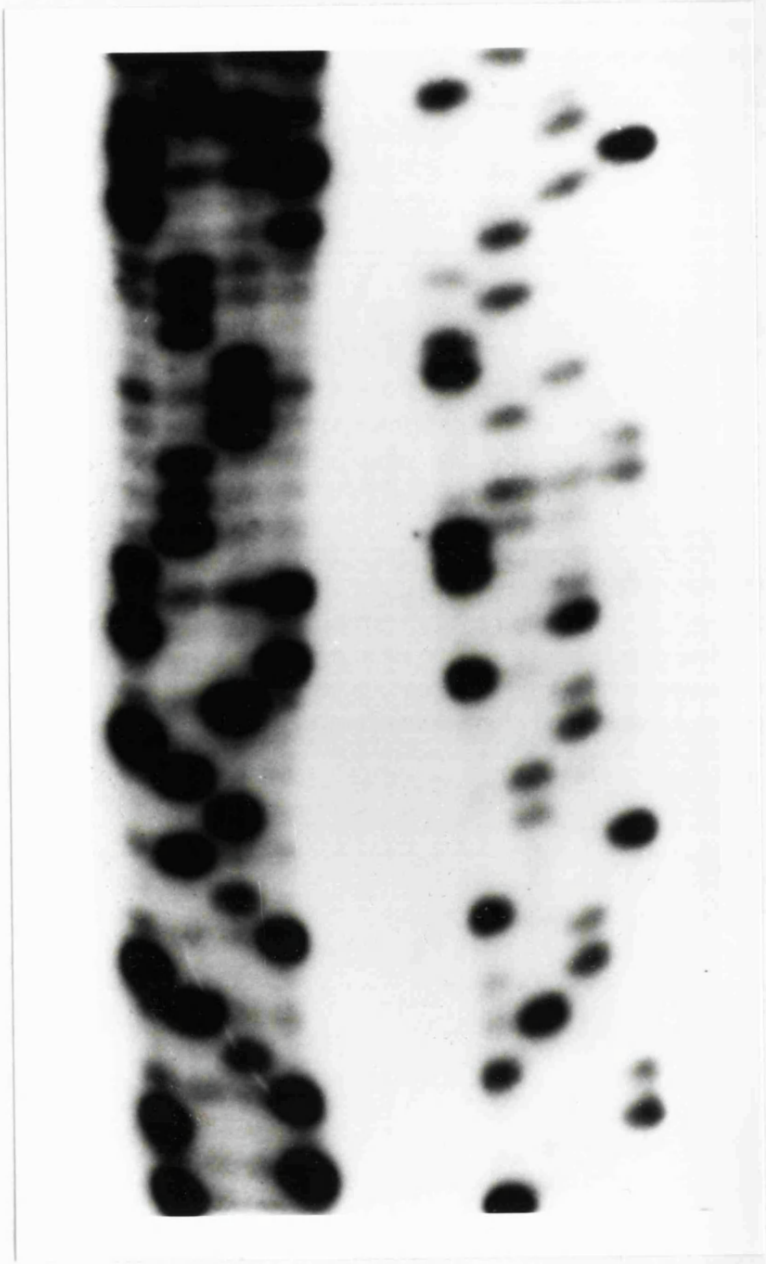
7.2 AMPLIFICATION AND SEQUENCING OF MODEL TEMPLATE USING OCTAMERS

The NotI octamer (5'GCGGCCGC) was used as the 5' primer, the 3' primer was ApaI (5'CGGGCCCT) situated 176 bp away (see Fig. 4.3 for map of pSL1180). The DNA was linearised by digestion with HindIII. PCR conditions used were as standard, but with 10% DMSO and 30mM KCl, with 10ng of template and 10ng each primer, 35 cycles of 95°C for 30 seconds, annealing at 35°C for 2 min and extending at 72°C for 2 min (also see 4.4 for PCR amplification of PSL1180 using octamers).

The product of amplification using the NotI and ApaI octamers on pSL1180 was excised from a low melting point agarose gel and purified on a PrimeErase Quik column (Stratagene). Sequencing was carried out using [γ-^{32}P] ATP end labelled primers (NotI and ApaI) with the dsDNA cycle sequencing kit (Gibco BRL) according to the manufacturers recommendations with the exception that the salt concentration in the reaction was decreased from 52 mM to 38 mM and the $MgCl_2$ concentration was decreased from 6.1 mM to 3.8 mM, with 10% DMSO. 35 cycles of amplification at 95°C for 30 seconds, 35°C for 2 min and 72°C for 2 min were carried out. 5μl of 95% formamide running buffer was added to each sequencing sample and 5μl of each was loaded onto a 6% polyacrylamide sequencing gel in TBE buffer. Sequencing of the 176 bp section of pSL1180 using this method is shown in Fig 7.1.

Fig. 7.1: PCR-sequence of pSL1180, using the *Not*I and *Apa*I octamers. Both sequences are C, T, A and G, from left to right. Reaction conditions are as described in the text.

ApaI(CGGGCCCT) NotI(GCGGCCGC)
C T A G C T A G



Correct amplification was achieved between both NotI and ApaI, and NotI and BssHII sites for the model template phagemid pSL1180 (see 4.5). Moreover, correct amplification and sequencing was achieved between the NotI and ApaI sites (see Fig.7.1). PCR-sequencing was more successful using the ApaI oligo than the NotI oligo, resulting in stronger bands. Difference in signal strength between the two oligos could be due to labelling specificity being different for the two primers. Non-specific bands may be the result of high G+C content and strong secondary structure in the template, particularly around some of the palindromic sequences.

7.3 PCR-SEQUENCING OF COSMID SUBCLONES USING OCTAMERS

The subclones used included the genes that encode the human glutathione-S-transferase π gene (p5TS0.6), the 51-kDa subunit of NADH-ubiquinone oxidoreductase (NUO)(p5HB8), the calcitonin/CGRP gene (pGemCal22, Broad et al, 1989) and the proteolipid protein gene (Diehl et al, 1986). Digestion of the PLP subclone 34S1 and separation of the larger XbaI fragment separates the target ApaI site from another ApaI site in the third exon (exon positions are marked I to VII). The octamer primers used were based on rare-cutter restriction sites NotI, BssHII, SmaI and ApaI which are present in the CpG islands of the above genes (see Table 7.1). To prevent amplification from both DNA strands the subclones were restriction digested as close to the priming site as possible.

Template DNA was purified using "Magic prep" columns (Promega, Madison, WI) and 500 ng was used per sequencing reaction. For each reaction 5 ng of primer was end-labelled in 1x"One-Phor-All" buffer (Pharmacia) with 5μCi [γ-^{32}P] ATP using T4 polynucleotide kinase (Boehringer Mannheim). This was then used for sequencing with a dsDNA cycle sequencing kit (Gibco BRL), incorporating 10% DMSO. 35 cycles were performed at 95°C for 30 sec, annealing at either 35°, 40° or 45°C for 2 min and extending at 72°C for 2 min.

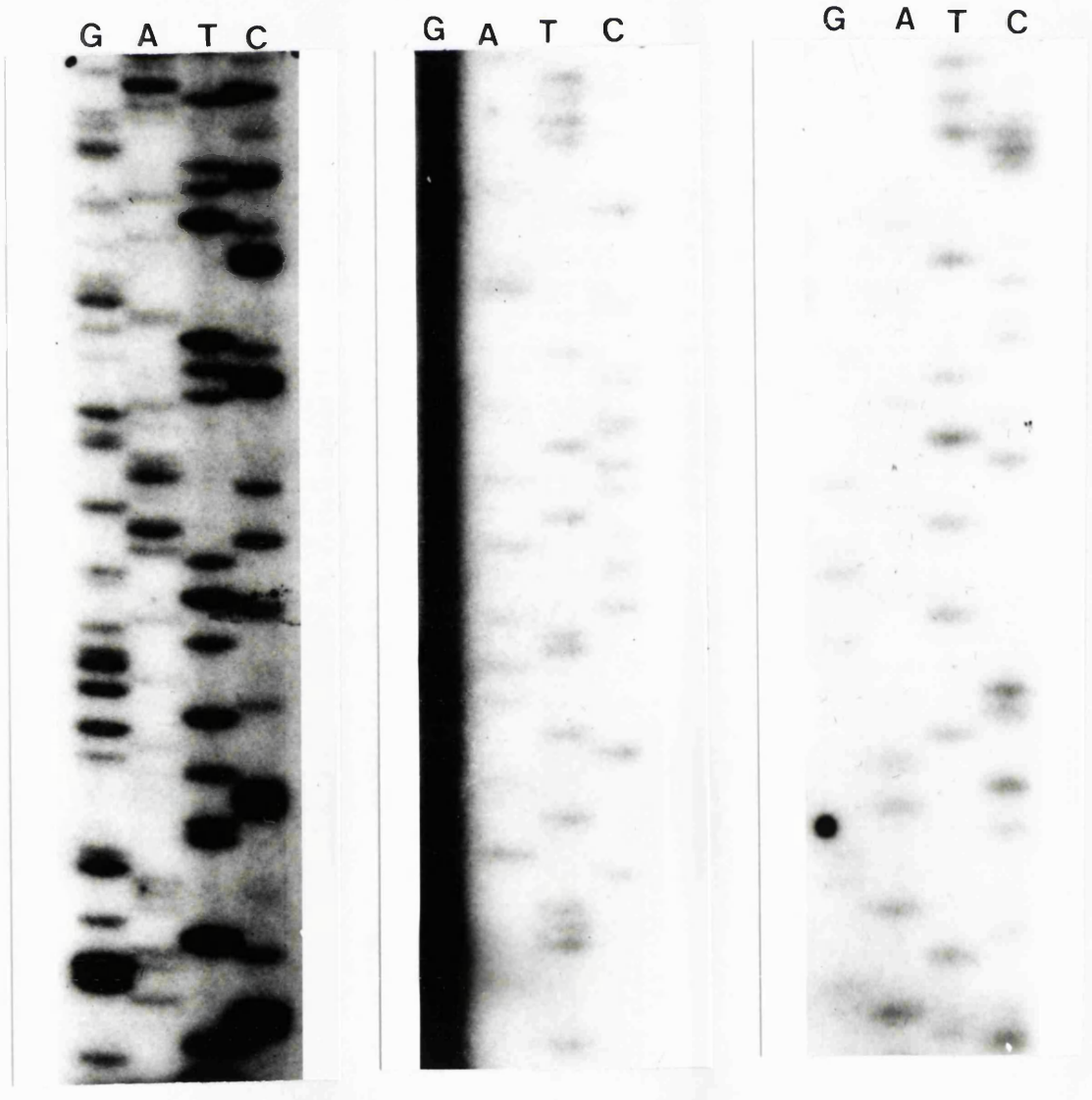
TABLE 7.1. Sequencing of cosmid subclones using octamers.

Cloned gene (encoded protein) (a)	Cosmid clone (b)	Subclone (c)	Restriction digest (d)	Distance from primer site (bp) (e)	8-mer primer (f)	Sequence at 45°C annealing (g)
Calcitonin/CGRP	cosCT1	pGemCal22	*Pst*I	12	*Bss*HII CGCGCGC G	-
Calcitonin/CGRP	cosCT1	pGemCal22	*Hind*III	41	*Bss*HII CGCGCGC G	-
NADH-ubiquinone oxidoreductase	cosGSTrp7	p5HB8	*Cfo*I	5	*Not*I GCGGCCG C	+
NADH-ubiquinone oxidoreductase	cosGSTrp7	p5HB8	*Sau*96I	27	*Not*I GCGGCCG C	+
NADH-ubiquinone oxidoreductase	cosGSTrp7	p5HB8	*Bam*HI	5	*Sma*I GCCCGGG C	++
Glutathione-*S*-transferase π	cosGSTrp7	p5TS0.6	*Nae*I	11	*Not*I GCGGCCG C	++
Glutathione-*S*-transferase π	cosGSTrp7	p5TS0.6	*Sma*I	5	*Not*I GCGGCCG C	+
Proteolipid protein	EMBL3-LP7	34S1*Xba*I	*Eae*I	15	*Apa*I GGGGCCC C	++

Table I. Lists the genes used (a), the subclones used as template (c) and their parent cosmids (b). Also shown are the restriction sites (d) used to cut the template on one side of the intended primer site and the distance from the primer site (e). The octamer primer used is shown (f), with sequence reading from the 5' end. Success or failure of PCR sequencing at 45°C annealing temperature is shown as either "-", where no sequence was achieved, "++" where correct sequencing was achieved from the intended primer site and "+" where sequencing was primed from a site other than that intended (g).

Using this strategy direct sequencing of the cosmid subclones was achieved using the octamer primers, annealing at 35°, 40° and 45°C. However, in some cases the primers appeared to prime sequencing not from the intended site of complementarity, but from sites of very close (6 or 7 out of 8 nucleotides) complementarity although still within the established CpG island of the gene. The best results were obtained by annealing at 45°C, at which temperature two out of the eight reactions failed and of the remaining six, three primed sequencing at the intended site (Table 7.1 and Fig.7.2). In several cases the oligos primed from sites with one or two mismatches at the 5' end, showing that 100% complementarity is not essential for successful sequencing. The smear in the G lane of the SmaI-primed sequence is probably a result of detergent contamination.

Fig. 7.2. Sequencing cosmid subclones using *Not*I (GCGGCCGC), *Sma*I (GCCCGGGC) and *Apa*I (GGGGCCCC) octamer primers, annealing at 45°C. All sequences are G, A, T and C, from left to right. Reaction conditions are as described in the text.



7.4 LIGASE MEDIATED PCR SEQUENCING OF COSMID CLONES

7.4.1 INTRODUCTION

As mentioned previously coding regions can often be localised by the presence of an adjacent CpG island. These CpG islands can themselves be recognised by the presence of rare-cutter restriction sites, which are generally clustered in such regions. By adapting the ligase-mediated PCR technique, it was possible to test the hypothesis that direct sequencing into CpG islands in cloned cosmid DNA can be carried out. It was planned to do this by restriction digesting the cosmid with both a rare-cutter enzyme and a vector-cutter enzyme and then to separate out fragments on an agarose gel. A linker-primer was then to be ligated to the rare-cutter cohesive end and PCR coupled sequencing was to be performed. If this hypothesis could be confirmed then this method could be used to speed up the search for CpG islands and thus coding regions within cloned DNA. Such a technique could also be employed for generating rare cutter-sequence tagged site maps for contig cosmid clones, covering a region of interest in the genome, without the drawback of requiring a repetitive element within amplification range of the rare cutter site.

7.4.2 DIRECT SEQUENCING OF MODEL COSMIDS BY LIGATION MEDIATED PCR-SEQUENCING

Cosmid clones used included the glutathione-s-transferase pi gene and the 51KDa subunit of NADH-ubiquinone oxidoreductase (cosGSTrp7), and a number of cosmid clones that had previously been identified as having NotI or BssHII sites (Melmer et al, 1990). The linker-primers used were based on rare-cutter restriction sites NotI and BssHII present in the CpG islands of the above genes. CosGSTrp7 and comid cH335 were digested with NotI and ClaI and cosmids cH335, cH324 and cH237 were digested with BssHII and NarI. The digested cosmids were then run on a 0.8% agarose gel to separate all bands. The bands were excised

and placed into the wells of a low melting point agarose gel onto which the DNA was run. DNA was cut from the LMP agarose. 40pmol of either NotI or BssHII linker was used in ligation reactions with the bands at 37°C overnight. This was then purified using "Magic prep" columns (Promega) to remove unligated linker/primer. The linker/primers used were:

NotI

5'GGATCCGTTTTCCCAGTGC 3'

3'CCTAGGCAAAAGGGTCACGCCGG 5'

BssHII

5'GGATCCGTTTTCCCAGTGG 3'

3'CCTAGGCAAAAGGGTCACCGCGC 5'

Approximately 20ng of template DNA was used per sequencing reaction. For each reaction 1pmol of primer was end-labelled in 1x"One-Phor-All" buffer (Pharmacia) using ^{32}P-γ-ATP with T4 polynucleotide kinase. This was then used for sequencing with a dsDNA cycle sequencing kit (Gibco BRL). 30 cycles were performed at 95°C for 30secs, annealing at 54°C for 1min and extending at 70°C for 1min. 5μl of 95% formamide running buffer (Gibco BRL) was added to each sequencing sample and 5μl of each was loaded onto a 6% polyacrylamide sequencing gel (Sequagel, Flowgen) in TBE buffer.

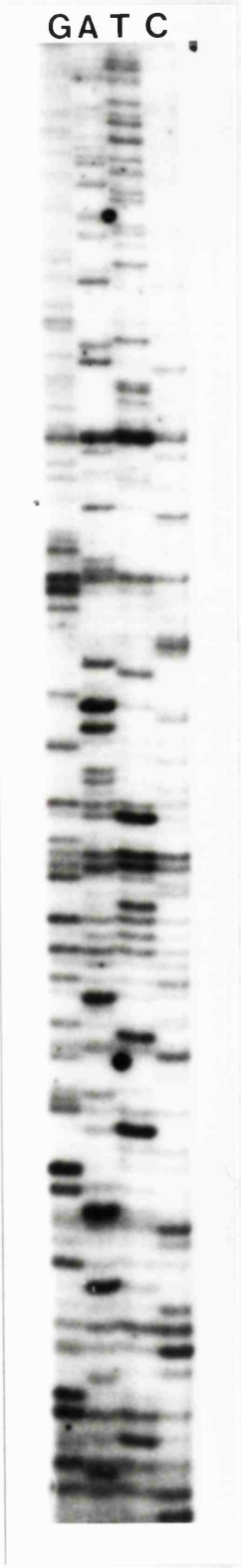
7.4.3 RESULTS

Correct amplification and sequencing was achieved for the model template cosGSTrp7, for sequence adjacent to a NotI site within the CpG island of the glutathione-S-transferase π gene, as shown in Figure 7.3. Amplification was achieved using another band from the same clone, but was unreadable, probably due to contamination with a third band. No sequence from the other clones was obtained.

Figure 7.3 Ligation mediated PCR sequencing from NotI sites. The sequence shown is from ligation mediated PCR sequencing on NotI/ClaI digested cosmid clone cosGSTrp7 using a NotI primer/linker:

5'GGA TCC GTT TTC CCA GTG C 3'

3'CCT AGG CAA AAG GGT CAC GCC GG 5'



7.5 DISCUSSION

7.5.1 SEQUENCING FROM RARE-CUTTER SITES

Siemieniak and Slightom (1990) reported the use of numerical methods to select useful nonamer primers for sequencing in which specific rules for primer selection were followed. These were the presence of GC nucleotide content between 45 and 60% and the exclusion of polynucleotide repeats and dinucleotide repeats. The success of the octamer primers used in the experiments described above suggest that such rules need not apply in the case of short primers and in fact the high GC content may affect the stability favourably. In order to prevent priming on both DNA strands as a result of the palindromic primers used, we cut the template to one side of the primer site by restriction digestion. For this to be feasible, a detailed and focused map of the area is necessary. Thus much mapping work would have to be carried out on each new cosmid. However, it seems that the primers favour amplification in one direction only, particularly if there is a mismatch at one end of the primer site. Thus in order to sequence around a rare-cutter site, a restriction map may not be needed. Another way to avoid the risk of sequencing from both strands would be to use primers targeted at sites for class-IIS restriction endonucleases, which are non-palindromic. The problem with this strategy is that only one five-cutter IIS enzyme cuts at GC-only sites (*Fau*I). The use of the ligase-mediated PCR method (Patel et al, 1991) for direct sequencing, by restriction-digesting cosmids with a rare-cutter and a vector-cutter and separating the fragments prior to primer-ligation and PCR-sequencing is a much more practical method. This was attempted successfully (7.4), although sequence was obtained only in a few cases, possibly because the template was not clean enough or band separation was not good enough. It is, however, quite possible, as has been shown here, to sequence directly into cosmids which may contain genes using short primers targeted at rare-cutter restriction sites.

7.5.2 SEQUENCING FROM CONSENSUS SITES IN OR NEAR CODING REGIONS

The use of short and degenerated oligonucleotides for the detection of splice site junctions has been reported previously (Melmer and Buchwald, 1992). If such little sequence information is required for designing effective but short sequencing primers, it may also be possible, with the aid of degenerate or deoxyinosine nucleotides in the primers to sequence directly from most restriction sites and from consensus sequences within or near coding regions such as splice sites and transcription factor binding sites etc. This might be used as an alternative to the cDNA sequencing strategy (Brenner, 1990; Adams et al, 1991). It would have the advantage of being independent of tissue expression and of the relative abundance of the corresponding mRNA. Sequencing from transcription factor-recognition sites would provide information about the sequences regulating gene expression, which is otherwise unobtainable using the cDNA approach.

7.5.3 CONCLUSIONS

Despite the reservations stated above over the use of short primers for direct sequencing of CpG islands, the results shown here suggest that G+C-rich octamers are indeed capable of efficient priming from rare-cutter restriction sites and also lend strong support to the use of short primers for a high density genome sequencing project.

CHAPTER 8: GENERAL DISCUSSION

The main aim of this thesis was to research further the use of short oligonucleotides for the detection of coding regions from amongst bulk DNA. Preliminary studies using hybridization of short oligos have been carried out (Estivill and Williamson, 1987; Melmer and Buchwald, 1990; Melmer et al, 1990) which have explored the possibilities of using oligos based on rare-cutter restriction sites for selecting clones for long range restriction mapping projects. Studies have also been performed in which short degenerate oligos based on splice site consensus sequences were hybridized to cloned DNA, specifically to sequences bearing splice sites (Melmer and Buchwald, 1992). Possibilities exist for the extension of these methods for the detection of CpG islands and coding regions, either in bulk genomic DNA or in uncharacterized DNA clones covering regions of the genome that are of interest with regards to genetically linked hereditary disorders. Whilst there are already methods for detection of CpG islands and coding regions, as discussed in Chapter 1, these methods can often be slow and labour intensive. The development of additional technologies would seem appropriate.

8.1 DETECTION OF CPG-RICH SEQUENCES USING SHORT OLIGOS

In Chapter 4 conditions were established for the amplification by PCR, with reasonable specificity, between two octadeoxyribonucleotides based on rare cutter sequences. PCR using such short primers may have a number of potential applications. For instance this method was used on human total genomic DNA under relaxed stringency conditions to generate a CpG island "mini-library" by cloning the resulting PCR product. Enrichment of over 60-fold was observed. This could possibly be increased if longer primers were used, with perhaps some

G/C degeneracies towards the 5' end of the primers. Such a library could be used to establish many Sequence Tagged Sites (STS's), with a high probability of being adjacent to or within coding regions. For a sequence to be useful as a STS, it should be short (200-500bp) with enough sequence data available from the flanks to design unique primers for single locus PCR amplification (Olsen et al, 1989).

Most mapping projects using STS's have so far involved the use of STS's derived from well characterized clones or sequences. In order to map the less well characterized regions of the genome it will be necessary to develop methods of generating large numbers of novel STS's from undefined DNA fragments specific to selected chromosomal regions. The PCR based method used here to develop CpG rich mini-libraries from genomic DNA could easily be applied to this task, perhaps using microdissected chromosomal fragments as template. Several other methods have been developed for generating STS's using Alu-Alu PCR on rodent/human radiation hybrids (Cole et al, 1991) and NotI/Alu PCR (Patel et al, 1991). Because the Alu primers are specific for the human repeat, rodent sequences are selected out. NotI/Alu-PCR which uses ligation of NotI linker/primer to NotI digested DNA, followed by NotI-Alu PCR has the added advantage that the sequences can then be used to identify CpG islands within the region. However only CpG islands that have a conveniently close Alu repeat nearby and in the correct orientation can be identified this way. Alu repeats are not distributed randomly across the genome and are found more frequently in light staining G-bands, whereas LINE repeats occur more in dark G-bands (Korenberg and Rykowski, 1988; Moyzis et al, 1989; Chen and Manuelidis, 1989). LINE PCR primers could be used for isolating sequences from dark staining G-bands and other rare-cutter linker/primers could be used to improve the efficiency of these methods. The PCR method developed in Chapter 4 could be used to detect CpG islands in any region of interest, regardless of the presence of repeat elements and has the advantage over the method sescribed by Patel et al (1991) by

virtue of being simpler and requiring fewer steps.

Interest in the use of CpG islands as gene markers is growing. A recent study (Larsen et al, 1992) showed that 46% of all first exons and 14% of all exons are CpG rich, suggesting that CpG island sequences could be used to identify transcripts since they often extend into exons. A CpG island library could provide an unbiased collection of DNA segments corresponding to the promoters of approximately 60% of human genes. CpG islands have previously been made by cloning the small fragments from the digestion of genomic DNA with HpaII. However this method fragments the CpG islands into small segments which are of little use. A recent article discusses an attempt to generate a CpG island library by removing methylated CpG regions by passing DNA through an affinity matrix that contains the methyl-CpG binding domain from the rat chromosomal protein MeCP2, leaving intact CpG islands (Cross et al, 1994). The method for generating CpG islands discussed in this thesis (4.6.3) would give larger segments than the HpaII method, but would not give whole intact CpG islands as with the affinity column method. In addition, the assessment of the CpG islands in Chapter 4 by sequencing part of the clones was inconclusive, since most of the sequences were under 200 bp- the minimum size for a genuine CpG island (Larsen et al, 1992). If only the sequences over 200 bp are taken into account, then the enrichment for CpG content is only 36-fold (rather than 66-fold, as mentioned in Chapter 4). This compares poorly to the enrichment of 80-fold claimed for the affinity column method (Cross et al, 1994). Further analysis of some of the TA clones obtained would have been useful. It might also have been appropriate to screen a cosmid library using several clones and to analyse the selected cosmids for conserved regions, in order to provide evidence of the effectiveness of this method. Also further attempts at using the TA clones for hybridzation against northern blots or zoo blots would have been useful.

Amplification between GC oligos was unable to identify CpG islands within cosmid clones, being incapable of discriminating between genuine CpG islands and the strongly CpG rich regions of the vector DNA. Amplification was not occurring between rare cutter sites within the vector since none are present, but was probably taking place at areas of high, but less than 100% homology, with the primers. Annealing stringency during the PCR was relaxed with the intention of allowing such low specificity primer extensions to occur within the CpG islands of the cloned genes because these did not have many rare cutter sites close enough for specific amplification. The cosmid cosGSTrp7 had single NotI sites in the CpG islands of both the glutathione-s-transferase and NADH-ubiquinone oxidoreductase genes. Cosmid cosCT1 had no NotI sites. In addition low stringency was necessary because the purpose of the experiment was to establish conditions which could be applied to any cosmid clone prior to any knowledge of rare cutter restriction sites and therefore required a degree of flexibility in annealing stringency. The method may be workable if the vector DNA could be excised prior to amplification. For the cosmids used in these experiments (cos202- Kioussis et al, 1987) this was not practicable, but may be possible for cosmid clones in other vectors such as pWE15 (Wahl et al, 1987) and SuperCos1 (Stratagene) and lambda clones. The method may be applied effectively to YAC's, whose vector DNA is a relatively minor component.

8.2 SEQUENCING DIRECTLY INTO CPG ISLANDS IN CLONED DNA

In Chapter 7 correct sequence from the CpG islands of the model cosmid clones was achieved by using direct approaches, either by:-

i. PCR-sequencing from rare cutter sites with octamer primers, having first cut the clone close by to allow primer extension in only one direction (subclones were used, but the method could be used on whole cosmid clones).

ii. Ligase mediated PCR-sequencing, by attaching a linker/primer to rare-cutter sites in the cosmid.

The first method has the drawback of requiring extensive restriction mapping before it can be used on a new clone, whereas the second only requires knowledge of the presence of a rare-cutter site within the clone and this could be achieved by selecting clones by screening cosmid libraries with octanucleotides based on rare-cutter sites, as has been done previously (Estivill and Williamson, 1987; Melmer and Buchwald, 1990; Melmer et al, 1990). This second method gave one correct sequence from cosGSTrp7 and another sequence assumed to be the result of two sequences from two NotI restriction fragments of the same size. The method needed refinement and could have benefited from the use of biotin labelled primer linkers. Streptavidin coated magnetizable particles could then be used to separate, clean and obtain single-stranded primer-linked DNA fragments prior to sequencing.

These methods cut out the labour intensive procedures involved in cloning the PCR products and would not produce a high background of sequences from the vector genome (see Chapter 4). Clone-specific STS's at CpG islands could be generated this way, either for cosmids or YACs. The STS would then be mapped back onto the clone and the position of the clone within the genome could be mapped using the STS. Sequence from the STS could then be used to design primers for sequencing the flanking regions which may contain coding regions.

8.3 IDENTIFICATION OF CODING REGIONS USING SHORT CONSENSUS SEQUENCE OLIGOS

Whilst the results of the hybridization experiments discussed in Chapter 3 show good discriminatory capabilities for short GC oligos against model target DNA, the use of the short oligos based on consensus splice sites appears to be less discriminatory. This is confirmed in Chapter 6 with the use of DOP-PCR, but points more to the inability of the oligos to discriminate between genuine splice sites and putative cryptic splice sites, which have sequence corresponding to the consensus but where no splicing occurs. Other problems with

degenerate oligo hybridization are outlined in Chapter 3, such as the effect of T_d for a degenerate oligo varying according to the target strand sequence on specificity. For instance an oligo with a run of pyrimidine degeneracies will have much lower T_d when it forms a duplex with a complementary strand with a run of A's rather than a run of G's. The chemical TMAC could be employed beneficially in these circumstances because it reduces the preferential melting of the A-T base pair over G-C (Wood et al, 1985). Ideally longer oligos would be used as hybridization probes, but because the consensus sequences are generally very short this is not possible. However it may be possible to increase the length of the oligos and thus increase stability by adding a string of deoxyinosine bases to the ends of the oligo. This would be preferable to adding four-fold degenerate positions (N) to the ends, since this effectively decreases probe concentration four-fold for each N. Other base analogues which give ambiguous base pairing could be used at "wobble" positions within the consensus part of the oligo, as is discussed in 3.4. This would have the effect of increasing probe concentration and should hopefully decrease background hybridization (Drmanac et al, 1990). Unfortunately only a few base analogues are available in commercially synthesized oligos and are very expensive.

In Chapter 5 attempts were made to PCR amplify coding sequences directly using the short degenerate splice site oligos that were used in Chapter 3. The amplification product generated using total human genomic DNA and "shotgun" cloned into M13 was unsuitable for any meaningful analysis, since the sequences were too short for open reading frame, analysis and the only concrete proof that the sequences represented coding regions would have required screening cDNA libraries, northern blots or zoo blots. A proper test of the procedure would have involved using a model gene or genes as template for the PCR (as was done in Chapters 3 and 6), with full analysis of the resulting PCR products carried out by cloning into a PCR vector (for instance the TA cloning procedure- see Chapter 4) followed by sequencing from

both directions. Sequence analysis would show conclusively the origin of the PCR product and comparison with the model gene sequence would show whether coding or non-coding DNA was being amplified. Attempts to amplify from model genes were made, but only weak or no product was observed. Positive amplification was observed by incorporating radioactive label into the PCR and a band corresponding approximately with the expected product size for amplification between two splice sites was identified. However the band was not detectable when PCR was performed without radioactive label and consequently cloning and sequencing was not possible. One of the main problems of this approach stems from the fact that the oligos are highly degenerate and consequently individual oligo sequences are at very low concentrations. In an effort to compensate for this the oligos were used at very high concentrations of up to 1μg per reaction, which probably affected the reactions adversely. Substitutions with base analogues, as discussed on the previous page, may have improved matters.

DOP-PCR, as discussed in Chapter 6, can be used to stabilize the oligos for effective use in PCR and at the same time tagging the product with restriction sites which can then be used for cloning. This would appear to resolve some of the problems encountered in Chapter 5 as discussed in the previous paragraph. "Clonable" PCR product was generated through DOP-PCR on several "model" cloned genes, cloned into M13 and sequenced. Sequence analysis of the product showed that a lot of background, non-coding DNA was being amplified. However in one case the 3' splice site DOP-PCR primer had primed amplification from a genuine splice site which was at exon 4 of the calcitonin/αCGRP gene. In several other cases priming had occurred from sites with strong homology to the consensus splice site sequences although not at recognized intron/exon or exon/intron boundaries. One of the model genes used as template acted as a "dummy", since it was known to be an intronless gene (Dopamine 5 receptor gene, DRD5)- although the presence of other coding regions within the cosmid

clone could not be ruled out. DOP-PCR product was observed from this gene and this indicates either the presence of splice site-like sequences or that the conditions were not quite right for specific amplification. Comparison of the cloned DOP-PCR sequences to the known target sequences showed four sequences at the primer site to have strong similarities with the consensus splice site and one with a mismatch at a crucial position which was at the second base from the 3' end of oligo107. This would suggest that primer annealing occurs at many splice-site like sequences and that amplification is occurring under low stringency conditions.

8.4 THE HUMAN GENOME PROJECT AND ALTERNATIVE STRATEGIES FOR THE DETECTION OF CODING REGIONS

Recently the U.S. Human Genome Project announced revised five-year research goals which include the development of efficient methods for the identification of genes and for the placement of known genes on physical maps or sequenced stretches of DNA (Collins and Galas, 1993). The observations made in this thesis in Chapters 3, 5 and 6 might suggest that splice sites are not suitable targets for the identification of coding regions using short oligos, since similar sequences seem to occur in non-coding regions. It may be well worth trying DOP-PCR using other consensus sites associated with coding regions. Other methods for isolating coding regions are probably now more appropriate. Exon trapping systems are now commercially available although many of the reservations expressed in Chapter 1 still apply. Another method that involves isolating human transcripts from human/rodent somatic cell hybrids can scan entire chromosomes or microdissected chromosome fragments but generally only constitutively expressed genes are isolated (Liu et al, 1989; Corbo et al, 1990). Direct screening of cDNA libraries with YACs enables fragments of several hundred Kb in length to be scanned at once (Wallace et al, 1990; Elvin et al, 1990). Recent advances have been made in producing normalised cDNA libraries which enrich for low copy cDNAs and deplete high copy cDNAs such as those encoding housekeeping genes (Patanjali et al, 1991). Such

libraries can be pooled from many tissue sources at many different stages of development in order to create a library representing all possible transcribed sequences. However, because of the complexity of the probes, the method produces high background signal, resulting in low reproducability. Another approach has used immobilized YACs to enrich for cDNAs (Lovett et al, 1991; Parimoo et al, 1991). This strategy has more recently been modified by biotinylating the human genomic DNA, having first digested it and added linker/primers for PCR amplification and hybridizing with cDNAs in solution (after preblocking repeat sequences) and then capturing the hybridized duplexes using streptavidin-coated magnetic beads (Tagle et al, 1992; Morgan et al, 1992; Korn et al, 1992; Tagle et al, 1993a). This technique was used to screen a pool of 6 YACs spanning the Huntington's Disease candidate region (Tagle et al, 1993b). Low abundance cDNAs from this region were enriched several thousand fold. One drawback is that either a full-length library must be screened with the resulting small inserts in order to get full-length transcripts or the eluted cDNAs captured must be preserved by cloning and the resulting clones sequenced.

The magnetic bead cDNA capture method constitutes a very powerful technique and if used with a pool of cDNA libraries representing all possible transcripts and with effective blocking of all repeat regions, could be used to identify all coding sequences within a given genomic region. Since the method is PCR based, it could also be extended to larger regions such as microdissected chromosomes or whole chromosomes. The approach is preferable to the Expressed Tagged Site approach of partially sequencing cDNAs from a library, for instance a brain library as carried out by Adams et al (1992) and then mapping the sequences. The magnetic bead method could be used for scanning brain cDNA libraries for each YAC in turn across areas on chromosomes of interest.

8.5 THE HUMAN GENOME PROJECT SEQUENCING STRATEGIES

As mentioned above the U.S. Human Genome Project have recently revised their five-year research goals. These include: (1) The development of efficient approaches to sequencing one-to-several Mb regions of DNA of high biological interest. (2) The development of technology for high throughput sequencing, focusing on systems which integrate all steps from template preparation to data analysis. (3) The development of sequencing capacity to allow sequencing at a collective rate of 50 Mb per year (Collins and Galas, 1993). Ideas for new sequencing technologies for large volume sequencing of the human genome are discussed in Chapter 7. Sequencing from octamers, which has been demonstrated as feasible using a PCR based approach here, would be very useful for bulk sequencing projects, since a library of all possible octamers (65,536) could easily be made. However the superior technology has now been developed whereby a library of hexamers (4096) is capable of sequencing any template (Kieleczawa et al, 1992). Correct sequencing requires the selection of three or four hexamers that will align on the template adjacent to each other. Addition of single stranded binding protein at the correct concentration prevents priming from a single hexamer or just two adjacent primers. Thus priming should be completely specific to the selected site. On an uncharacterized cosmid clone the procedure would be performed in a random manner, using pools of hexamers, until a proportion of the cosmid had been sequenced (20-25%, Studier, 1989) and then directed sequencing would be used to fill in the gaps. The method has already been adapted for use with "dye-deoxy" sequencing (Hou and Smith, 1993) compatible with the Applied Biosystems automated system. It seems that a computerised mass sequencing strategy using such approaches cannot be far off. A similar project has been developed in parallel to that of Kieleczawa et al (1992) involving libraries of hexamers or pentamers, stacking these "modular" primers along the template in the same way, but without the use of single stranded binding protein (Kotler et al, 1993). Another approach to large volume sequencing , known as the "Janus Strategy", has been developed and involves subcloning

DNA into Janus, an M13 vector, that will allow the sequencing of both strands from a single stranded template (Burtland et al, 1993). Adoption of these or similar techniques should bring the cost of a genome sequencing project down below $0.50 per base, in accordance with one of the original U.S. Human Genome Project short term goals.

SUMMARY

The hypotheses tested in this thesis have demonstrated the following:-

1. Short GC-rich primers could be used for PCR amplification and when human genomic DNA was used as template the products, when cloned and sequenced, were enriched 66-fold for non-CpG depleted sequences. However the length of sequences was not enough to prove that they had originated from CpG islands, nor was any evidence of the methylation status of the source DNA provided and no evidence was shown to suggest that the PCR product was derived from human DNA rather than from contamination by prokaryotic genomes. More powerful methods of isolating CpG islands have since been developed (Cross et al, 1994).

2. Short GC primers when used less stringently could not distinguish between the CpG islands of cloned genes and the non-CpG depleted vector genomes. Conventional methods for identifying CpG islands within clones using rare cutter restriction analysis would be appropriate.

3. GC-rich octamers were capable of priming PCR sequencing accurately in only 3 out of 8 instances. Ligation of primer/linkers to rare cutter sites prior to PCR sequencing could be a mor specific method of sequencing directly into CpG islands within clones, although only limited evidence was shown to support this.

4. Rare-cutter sites can be targetted by hybridization with 8-mer oligos capable of discriminating 7/8 and 8/8 complementarity. This method has been used to screen large numbers of clones (Estivill and Williamson, 1987; Melmer and Buchwald, 1990).

5. Targetting splice sites using oligos based on consensus sites had low success rate and a

high rate of false positive hybridization. Results using PCR with such oligos were inappropriate for analysis.

6. Amplification using DOP-PCR for targetting splice sites also had low success rate, partly due to the high occurrence of splice site-like sequences within the clones. Methods such as exon trapping or screening normalized cDNA libraries with clones would be the preferred method for detecting coding regions.

APPENDIX A: M13 CLONES OF PCR USING SHORT GC-OLIGOS ON HUMAN GENOMIC DNA

6a1.seq Length: 166 August 22, 1991 18:21 Check: 9232

1 tgtcgtggtc cgtcatcatc accggataat gccagcgtcg cgagtgcctc

51 ctgggcagtc aggctgttcg ccgacaccgc atccgtcagc tgctgatccc

101 aagctggcca gccatccgtc atccatatca ccacgtcaag ggtgacagca

151 ggctcataag acgccc

6a4.seq Length: 171 September 25, 1991 14:20 Check: 1162

1 gctactacta cggccataag tagccgctgg cagcgtcctc ctgggagcta

51 ggcggtctgc ggccagacgc cactgcctac gctgctgatc ccaagctggc

101 ctgccatccg tactccatat caccacgtca aagggtgaca gcaggctact

151 aagacgccca gggcctcact a

6b1.seq Length: 128 August 22, 1991 18:06 Check: 5238

1 ctccacctcc tcaccatcac gacctaacac ttcagatcca tcagttttct

51 ctaactgcca ctatctaagt catctatgcg acaacacaca atgggtcaca

101 tttacttcaa aaattcgatc catattaa

6f3.seq Length: 284 October 4, 1991 16:57 Check: 4566

1 ctctcaaaca tgcttgggta acggatgcac tgccccggcc attgtctctg

51 taggaaaggc ggccgcgcac atagaactgg agacgcactg cctgggccat

101 cgtctctgta ggaaggcgga catggcacat agaaccggag agtacgctcc

151 cgggccattg tctccgtaga aaggcggaca tggcacatag aaccggagat

201 gcaccgcccg ggccattgtc tcgtaggaaa ggcagacatg acacatagaa

251 ccgagatgca ctgcggccat tggcacttgt agga

6a2.seq Length: 97 August 23, 1991 14:53 Check: 6003

1 atcaaagcta agggtggagc caagtgggga gaccataagt gaaaagggga

51 gagtttggag cctgatcacc aagccgaaca aggagtgttc gtcgcct

6f4.seq Length: 101 October 4, 1991 17:00 Check: 8830

1 ccgggcgagg gcaggagtta ctgatagttt gacgcgtcga cgggcaggtg

51 cgcttgaagc attgctcaag accttggcaa acgtggagcc aacgaggtac

101 t

6f5.seq Length: 176 October 4, 1991 17:01 Check: 610

1 ccggctgaac tgtggcagga atccgggcgc tgggaagagt acggccctga

51 gttgcagcgc ttcaaggatc gtcatggccg cgacttctgc gcaggcccga

101 cccatgaaga agtgatcacc gacctgatgc gcaacgagtt gagcagctac

151 aaacggctgc ccctcaacct gtatca

6g1.seq Length: 148 August 23, 1991 14:58 Check: 3071

1 ccagcctggg caacatgcaa aaccccatct ctgctaaaaa aaaaaaaata

51 caaaaattag ctggcatgca cctgtaatcc tagctactct ggaagctgag

101 gcaggagaat cacttgaacc aggagacaga ggtggcggag ttagctga

6h1.seq Length: 181 October 4, 1991 16:41 Check: 3364

1 tagaggatca gtcaaccacc agggaataat ccttcatatt attatgcgtc

51 ttcaccaagc gctcctcaag ttctgaactg cttccagaga caccttatgt

101 tctatacatg caattacaac atcagggtaa ctcatagaaa tggtgctatt

151 aagcatattt ttacacgaat cagatcacga g

6h2.seq Length: 125 August 23, 1991 15:04 Check: 739

1 gcaggcagcc ttatcgagga ggccgatgtt gctctgctga ctgcccaccg

51 ctccaagggc ttggaattcc ccagcttgta ccgctcagcg atttccataa

101 cctgatcact aatggcggcc agcgc

6h4.seq Length: 200 October 4, 1991 10:40 Check: 887

1 ccggataatc cagcgtcgcg agtgcctcct gggagctcag gctggtcacc

51 agcatccgtc agctgctgat cccaagctgg ccagccatcc gtcatccata

101 tcaccacgtc aaggtgacag caggctcata agacgcccag ctcgcatagt

151 gcgttcaccg aatactgcgc aacaaccgtc ttccgagact gtcatacgcg

6h5.seq Length: 135 August 23, 1991 15:10 Check: 5351

1 aagatggatt aaagacttaa atggtagacc taaaaccata aaaaccctag

51 aagaaaacct gggcattaca tcaggacata ggcatgggca aggacttcat

101 gtctaaacac caaagcaaaa gccaaatgac aatgg

6h6.seq Length: 182 September 23, 1991 17:10 Check: 4904

1 gtggtccgtc atcatcaccg gataatgcca gcgtcgcgag tgcctcctgg

51 gcagtcaggc tgtcgccgac accgcatccg tcgtcgatcc aagctggcca

101 gccatccgtc atccatatca ccacgtcaag gtgacagcag gctcataaga

151 cgccccagct cgcatagtgc gttcaccgaa ta

APPENDIX B: SEQUENCES FROM TA CLONED PCR PRODUCTS USING SHORT GC OLIGOS ON HUMAN GENOMIC DNA

ta2.seq Length: 210 June 1, 1993 13:03 Type: N Check: 9311 ..

1 gctaccaata ctctctgccg ccgaaggtac actcgacgac tgtcacgcac

51 acgagcacga ggataccaac cgcgtcggct tctactcgtc acgcgtagtt

101 caggcctcac acttccagcc cacgcgatgt agcgccgcgg cctcatacct

151 cgctcacgta agtcctgttg cacgtcgacg tgctcgacgt ccggatagtc

201 gaggttctga

ta4.seq Length: 236 June 1, 1993 13:15 Type: N Check: 8443 ..

1 cgtgattaag gaatagattc caaacgaaag agcactaccc gaaacaaaac

51 aacaacaata aacaaattta cgggaagata aagactcctg cactaataat

101 tatttaacag gtacacggaa atcacatgac taagaggact tcgaaacgag

151 ggacctttc gttcccaaat ccccctggtc cctaggagtt tccgcgcagt

201 cttggttctc gatcaggccc agggcgcatg ggcacg

ta5.seq Length: 104 June 1, 1993 13:17 Type: N Check: 7172 ..

1 gaggcgagct accattaagg ccgcggctcc cgcgtcggca atgggagtag

51 cacttccgcg aagctgcgaa tcagtagctt ggggctggcc gaacactcga

101 gatc

ta8.seq Length: 179 June 1, 1993 13:25 Type: N Check: 7044 ..

1 cttcagctta cgccaatagc ccatagaaca tggaaatgac cggcggccga

51 tttaggcgta ggttatgcgc tgcgcgtgtg gtcgctctgt gactcggagg

101 tttgcgaccc agctagtccg aagtgcatgg ctctccggtc tatggccaat

151 tggaccggga acttggcgcg atgctagcg

ta9.seq Length: 198 June 1, 1993 13:29 Type: N Check: 2763 ..

1 accacagagc acagcagcct ctcctttaag cgctacgctt acagaccctt

51 acccagagga ctgcactgat gaccccactc actgaggagc tgaagggagg

101 cgcggccggt gtggccatac caagtgccac gcccagagag gcctgattga

151 tggttttgtg ccttcattaa tagccctcta ctaggctcgg ctacctcc

ta12.seq Length: 124 June 1, 1993 13:40 Type: N Check: 9131 ..

1 gaagattagc ccgattgcgc aaaggcatta ttgcggaggc aggcgctgat

51 tgtctgattg tcggcctgta cgcgactcga tatccagtga ctacgaagca

101 cgcagccgga gagtagcccg atgc

ta13.seq Length: 118 June 1, 1993 14:55 Type: N Check: 8687 ..

1 cctccggcgt ctcgtccgcg cgtttgcctc gttgcccgcg tgcagcgacg

51 cgatctcagg cttgtgtccg ggtcatacag tggtcggccg tcagccttct

101 agcctcccgg tcctagcc

ta14.seq Length: 190 June 1, 1993 14:59 Type: N Check: 7384 ..

1 cgagcctagg cccggcatac actcagcaca actccgaacc gaaccgcatt

51 agtaccagta tcgacaaaga cgacgagttt gaagcagata ggcgagtgtt

101 acaggtcgtg ttgtcatgct ccggcacttc gtattcacat tcggacccac

151 ggatactccg atgatgtaca ttacgcaccg gagtgacggc

ta15.seq Length: 236 June 1, 1993 15:03 Type: N Check: 6221 ..

1 cgtgcccatg cgccctgggc ctgatcgaga accaagactg cgcggaaact

51 cctagggcac agggggagtg tgaatgctat acgaggcacg ttacgtactg

101 gtcttagcga tcttgatctt cgtctacctc ttatataatt attagtgcag

151 gagtctttac tttcccgtaa atttgttatt gttgttgttt tgtttcgggt

201 agagcctttt cgtttggaat ctattcctta atcacg

ta16.seq Length: 164 June 1, 1993 15:10 Type: N Check: 9277 ..

1 cggagccaag gcccatggga tactctagct cgaaccgcat aagtacaagt

51 atcgacatag acgtgcacat gtgctgtcag gagtcggtgt ccattgctgt

101 gttgctatcg ctccggccat tcgtactgac tacattcgga cccgacggat

151 tcctcagctg catg

ta17.seq Length: 150 June 1, 1993 15:13 Type: N Check: 5383 ..

1 gcgctgcggg gtcctcgggt cacataccgg ctcgtcgcgg tggatcctct

51 tcgtcgagtg cgcgcaggta tgccatgcgc tcacacgtct acgtgacggt

101 agtcgagcgc ggccctcata gagttcgtcg atgctactgg ctcagtccgt

ta18.seq Length: 144 June 1, 1993 15:16 Type: N Check: 7791 ..

1 agtgcacacg atcacgacga cgcttaagtg acacggcagc aaaatgatct

51 gacagacact gacccttttg ggatacaccg acaatgggtt gaattagccg

101 gaatcgtcgt gtggggaagc gcgtcgaccg ttatcgcaac ctcc

ta20.seq Length: 170 June 1, 1993 15:22 Type: N Check: 1952 ..

1 gccgcggcta attaataatg cagaggtaat cccgggcggt gcgcactcgt

51 gaaatacgaa cggcgcccca gcctacaacg cgcctttaac accgcctatt

101 cgcttcaaag tgtgtcccct tgccatactc ggatcatcgg ttcggtgttt

151 acgctgacgt gatatccccg

ta21.seq Length: 102 June 1, 1993 15:24 Type: N Check: 1230 ..

1 gcgctcgtga tcattgatgc gatgtgatct ctgctgacat agagacagac

51 gattcggtca atgaagcctt cttcaaccag tcgagaacta ggccgaccgt

101 gc

ta22.seq Length: 148 June 1, 1993 16:18 Type: N Check: 8704 ..

1 ggagtgcgtt accgtgcgat tgtcacgtcg cgcacttgct aggtctacac

51 tggcgtaccg agcgattact acgctgcacg gtcagtgatg acgtgacgac

101 agtcgcgagc tgagcggacg aagtcagcct tgagcgcgcg tagctgaa

ta23.seq Length: 143 June 1, 1993 16:22 Type: N Check: 1459 ..

1 tgcacaccgg acactgtcgc acacagataa gattgatgct gtcgcgtgtc

51 gagcgattga tcgtcgaggc taggcgtcgt gtagcgtcga gcgtcgccgc

101 tgctcgttgg atcttgatca tccgctgacc ggtctactta gtg

ta24.seq Length: 197 June 1, 1993 16:25 Type: N Check: 9560 ..

1 cctccatcgg ctggcatctc cgccttaata cttcacgtgt gtttttgatg

51 tagtccggag agacccgcac cgtgaaccat accggtgtgg ccggccgcca

101 gggaaacgtg tgaggctcac ccccagatga tcacgtcagg agacccattc

151 ccagacattc gcatcgcgaa tttcctctcc gacgacagga gaccaca

ta25.seq Length: 203 June 1, 1993 16:29 Type: N Check: 9152 ..

1 cgagatagga caacggactt aagggaccgc gacgcgaacg tctagttgtc

51 aaattgcgcg agttccagat acaccggaga tgccgatggt catctcgata

101 acctaaggag agatcggcag agccaacaac catcgagaac taggccgttt

151 gttgcgacct acgccaccaa aaaaactatc gttcgtcgtc taagtgcgca

201 tgc

ta26.seq Length: 139 June 1, 1993 16:36 Type: N Check: 9040 ..

1 gaatccgcat tagtaccagt atccgacaaa ggacgacgac tttaacgaga

51 tacgcgagtg ttaaggtcgt gttgtcatcg ctccggcctt ccgtatttcc

101 agatttccgg accccacgga ttactcactc gcattgagg

ta28.seq Length: 196 June 1, 1993 16:43 Type: N Check: 502 ..

1 gcgatcgtag cgcggttcaa gggccaggtt aaccggtatc tggcctctcg

51 gtacgtgaag cctgatcgac tgcggcccag cgtttggagg ctgagtgctt

101 cgctggtgtg ggcgtcgcgt attggatgcg gatttagccg gcgggccagt

151 aaaggtacaa gatacccgaa tatccgcatt cgcattgcat tctagg

ta29.seq Length: 152 June 1, 1993 16:46 Type: N Check: 9586 ..

1 gcggaatcgc gcggtcgggc gcgggcgaaa tagaaggcca gcataaatcc

51 gcgcttaaac ggcctattgt taaagtgagt cgctttgtcc gatatggtat

101 aacgcgtcct agttatgtga gctgaattcc ccgcgcccac cgggcccgag

151 gt

ta30.seq Length: 175 June 1, 1993 16:51 Type: N Check: 7112 ..

1 gtacgtcgac tccttaggca gcccaggctt acatcagtca tgcttaccgg

51 cctcgctatc gttgttcgtt acctgtggct gaggactgtc gtgtacacgt

101 gcagatacag ctatgaacat gaatacgcca agcctgattc ctcagtagcc

151 tagcgggttc gcgtaatacc gtcgg

ta31.seq Length: 117 June 1, 1993 16:54 Type: N Check: 4069 ..

1 ctcgtcggct ggtgcctctg ctcgcgcgcc cgcgtagccg aggccgatcg

51 acaggacacg ccgttatgaa gccgtgccgt tcatagttga cgcgcatacg

101 gcgtgcggca gcgtctg

ta32.seq Length: 179 June 1, 1993 16:57 Type: N Check: 320 ..

1 cactttcaag ttccggatca tcgtcccagt aattacttcc gtcgtttgtg

51 gtcgtccgga gagacagcac cgtgaatcgg acgagacgcg ttcctgaggg

101 aagtccgagg agtcactcac cccagtagtc acgtcaggag acccccttccc

151 ctaagtcgca tccccctttcc tctccgacg

ta33.seq Length: 107 June 1, 1993 17:00 Type: N Check: 7863 ..

1 cagccagagc cgccttcgta ggtttggctc gtgggtctcg cgtcccagct

51 cgctgtgcca gcacctggcc caatccacga gaccgccacc gcacaacgcc

101 cgcagtc

ta35.seq Length: 156 June 1, 1993 17:06 Type: N Check: 8460 ..

1 ccaaaagacc ttccgttggt cctcgacacc gggtccggat acacccctcc

51 atgtccgttc ttaggtggag ggcgatccac gtcgggaacc tctacggaac

101 gttcagacga ctcgtaggac acgggtaaag ttgaaatcac aattgtccta

151 cacagt

ta36.seq Length: 157 June 1, 1993 17:09 Type: N Check: 7593 ..

1 gggcaagcgc ggattgggag aaccctgtct tagtcatgtg atttcgtgta

51 cctctaaata attattaggc aggagtctta tcttcccgta aattgttatt

101 gttgttgttg tttgtttcgg gagtcctctt gtttagaatc tattccttaa

151 tcacgag

ta38.seq Length: 95 June 1, 1993 17:13 Type: N Check: 2049 ..

1 ggctgacgtg tcggtcacat tgttccaatg ctcgctcatc atcgcgaact

51 tcttcaccac cggattgacg cggatgtgag ccggtaacgt ccacg

ta39.seq Length: 104 June 2, 1993 15:06 Type: N Check: 8134 ..

1 gcgagactcc atctcaaata aaagtcgcgg ccgcttgcag ttacccggat

51 gcgagcctcg agatagatag tavgtgccag gctcctagtc gctccaatgc

101 tggg

ta40.seq Length: 234 June 1, 1993 17:34 Type: N Check: 643 ..

1 gtgcacagac gacgagtatc gtcagaagca ccagcatcgg ggaaagtcta

51 gggacagggg gattcgggaa gaaaaggtcc cgcgtggcgt tgactgtctt

101 agtcacgtga cctcgtgtac ctcttccata atgatttgca ggagtcttga

151 gctgcccgta aatgtgctac tggtgttgct ttgtttcggg tagtcctctt

201 ggtttcgaat ctatttcctt taatcgcgag cacc

ta41.seq Length: 177 November 5, 1993 14:32 Type: N Check: 8074 ..

1 agacttgtcc tgtggtgaaa ggaaggaggg cgcttttatt cagcgcttcc

51 ctgtcttgat gctcgagatg caaatactgc cgccccatgg atttttttg

101 aactaggccg ggaccactgt cgccaggtag ccgaagaccg ctgcacgcga

151 ccggcaatac cttggccagc aggtcga

ta47r.seq Length: 163 November 8, 1993 17:10 Type: N Check: 5505 ..

1 agattggagc cctctctaat gtatgttaca aaatttctcc ctgggcttct

51 cggaggatta tggggtccac cttaaaaaag gcaaactcca gacactctgt

101 gaagtagaat tgaccacagt ttggaaccgg gtgccccaga aggtcactga

151 atctcacgct gtt

ta42.seq Length: 211 November 5, 1993 14:28 Type: N Check: 9251 ..

1 tatggcagaa cctgagagca tatcttacat ggcagcagac aagagagaag

51 cagaaacaag aaaaaggggt ttccccttat aaaaccatca gatctcatga

101 gacttattca ctaccacaag agcagcatgg gggaaaccac ccctcatgat

151 tcaattatct cccaccggtc ccaccacaca tgtggttatg gctcattcac

201 gagttggagg g

ta43.seq Length: 148 November 5, 1993 14:39 Type: N Check: 185 ..

1 aacagcgcac tgccggcatc ggccgcgaga tcgagtccac ggcgaaaccg

51 gcaaacaggg gataggccag cagcaacggt tctccagcgc aaccaacgac

101 agccgatgca gaacagcttc cgggataagg ccttcggatt gcttgagg

ta44.seq Length: 212 November 5, 1993 14:46 Type: N Check: 5477 ..

1 tcggaaacag gaccgccagg gcgccagcag cccggccgct gtcagcagcc

51 ttgccggaga atcgccaact caatacgaga gctccagctc gtcaccgtcg

101 aacaacaggc cgatcctgcc ccacgagctc ggccacattg gcgaacgcgc

151 cacaccgtac tcaacgcgcg aggtaactct ctggttcacc tgaatatcag

201 agcagccccg ac

ta45.seq Length: 209 November 5, 1993 14:54 Type: N Check: 6195 ..

1 tcctcaagac catcacgcag gaaaatcact ctcttcagca ggctgaaagc

51 ctttgttcgc ttggaaccac aggacgagcc ttcatccttg catctgaagg

101 gtgactttct aggcttgatg aagaagtgtc tgcaacaggc catacaagcc

151 cttgacccag agagcccagg cttaggcata actctccatc ggtgaagatg

201 agatgagca

ta46.seq Length: 121 November 5, 1993 16:15 Type: N Check: 4224 ..

1 ggcggatgtt tgcctacatc gccggttcac cttcgtcttt atcaagctgt

51 atggcgtgcc ggccgagcac tacggttggc tgttcggcac caatgcgccg

101 ggctttatcc tggtggccag a

ta48.seq Length: 220 November 8, 1993 17:05 Type: N Check: 5949 ..

1 cctcgcggtt gtagcccgtg acgtcgataa ggtcctggtt ggcgtaggta

51 atcgaactgc ttaagtcagt ggtcgagagg atgttggcgt ctggggcaat

101 gtcaacactt cgacccgtga ccggagattg atcttcatgg aagnntcggg

151 gttattgagg gagtcggcca ggggctccga ctctannacc ggttgagcaa

201 tactgatcca gctcatgttc

ta49f.seq Length: 207 November 8, 1993 17:32 Type: N Check: 1597 ..

1 cgaggtcact gtgaaaggac agaaaacaaa gggcatgtca tgaggtcatc

51 atggtggaga aggggttcaa aggataagcc tcggtctcct attcctcact

101 cctcaccgga attgcatcaa cctcaagatg ctgtggttta gaccttatgc

151 agcggcatac aggagagtca ggtgctgcat cgtgatacta ctgtgaacag

201 gaagcgg

ta49r.seq Length: 144 November 8, 1993 17:40 Type: N Check: 2104 ..

1 agtctcattc caccaaggat ccgctctgtc ccttctccta gcatctcacc

51 caacccttga ggctcaggtc ctttattctc cggctactct gaacctgagt

101 gaaatatcct gcttgccatt ccttgcagtg acccgccttg agta

ta50.seq Length: 211 November 8, 1993 17:51 Type: N Check: 9130 ..

1 aacagcgcac tgccggcatc ggccgcagga tcgagtccac ggcgaaaccg

51 gcaaacaggg gataggccag cagcaacgcg ttctccagcg caaccaacga

101 cagctgcaga acagcttgcg gtaagcttcc gattgcttga ggttggctgg

151 cgattctgcg attctgcttc tcgatcagat ttggcttgtg catgtgacca

201 gagtataggc c

ta51f.seq Length: 126 November 8, 1993 19:07 Type: N Check: 6166 ..

1 ggccgcggct gtgtgtgcgt gcatgcacgc acatgtatgc gtgtgctcta

51 gagatgtatg tgcacatgca gtgtgagtgc attttgcata ctcctggtgg

101 acagagctgc aggtgcgcgt gctgct

ta51r.seq Length: 243 November 8, 1993 19:16 Type: N Check: 6763 ..

1 ggctgtgtgt gcgtgcatgc aggccaatgt atgtcgagtg tgtctctaga

51 gatgtatgtg cacatgacaa gtggtgaggt ggcatctctc tggccatact

101 ccctgggtgg gaccaganag ccttggcaag agggtaggca gcggttggcc

151 tggtctggac tggccagctg gcataggaca cttgcaccag ccagagcacc

201 agcataagct ctgcttagct tgtgcagtgc tatcagtcag gag

ta52.seq Length: 159 November 8, 1993 19:25 Type: N Check: 5569 ..

1 atgtgccggg tgaggtgatg tctcgtggct gctcaacagg aggccccggc

51 acgatggtcc tcgccctcac acgtagcgct cagttacgta cacacatgtg

101 cctcatcaca cactgagcaa actccaccgg gtcacctcac gggaccagcc

151 cgaggccta

ta53.seq Length: 204 November 8, 1993 19:35 Type: N Check: 1068 ..

1 ccagacaccc cagggtagcc ttcacatctc aatattctta attatttgca

51 aagtcccttt tgccacttaa ggtaacacca caggtttgga gattaggagg

101 tggacatctt tcagggccat tcttatttaa aatgtactgt cgataagaat

151 agttaaaatg taaaatattt gaagggattt attctgaatc gaatatgagt

201 gacc

ta54.seq Length: 142 November 8, 1993 19:41 Type: N Check: 3008 ..

1 agcttcatcc atgtccctac aaaggacatg aactcatcct tttttatggc

51 tgcatagtat tccatggcgt atatgtgcca catttctaat ccagtctatc

101 atgtgacatt tgggttggtt tcaagtcttt gctattgtga at

ta55f.seq Length: 175 November 8, 1993 19:46 Type: N Check: 384 ..

1 aaggtgacta tctttcaagg aagtacagaa gattctacag ttctactgga

51 ctttctacta ctactactac tactgctact caaagcgtgc ttcctggtca

101 tgacgaggtg agcacagaat tgagggagtg ttagaaacct ttagcagctg

151 gcattgatgt accaggatgt tagaa

ta55r.seq Length: 242 November 8, 1993 19:52 Type: N Check: 8224 ..

1 atacaagcaa ggaacgaaaa tactcatacc ttacacctct caatggaaca

51 ctcaaatcta gaagtgtatt acagagttgt gcttgtcnga ctgtactgaa

101 aagccacaag gaccaggttg gctggtttgt ttacttttca gtctgtctct

151 aaccaataca cggtcctact gtgtgtgact gtgtaactca cccatctgcg

201 acggttcacc aacactgaac tgtctacact gtcagtgaga ag

ta56f.seq Length: 215 November 8, 1993 20:01 Type: N Check: 3479 ..

1 caacactaaa cccactgaaa ccaagccacc agcactccct tatctcttcc

51 agggaaagtg tgcagaagtg tctgcttatt tttagaagcc actcagccac

101 agaacaggtg tccataagaa gaggagtgat tgttcgtgtg gctaagtggc

151 gaggagagag tcacatccat gaggcctgat gtgccagctg tctagacctt

201 aatgaactac cttat

ta56r.seq Length: 137 November 8, 1993 20:07 Type: N Check: 5100 ..

1 ctgggaaaca tagcaagacc ctgattctac ttaaaaaaac atgggggaat

51 agacaagatg accaatcatc ataatgacct gagattattt cgtcaaacgg

101 aatctatata atgttcaaat cagaatctag cgaggcg

ta57.seq Length: 169 November 8, 1993 20:15 Type: N Check: 5540 ..

1 agcattcatc catgtcccta caaaggacat gaactcatcc ttttttatgg

51 ctgcatagta ttccatggcg tatatgtgca cattttctta atccagtcta

101 tcattgttgg acattggttt ggttcagtct tctatgtgat cggccgaacc

151 tctactggcg tagttataa

ta58.seq Length: 174 November 8, 1993 20:24 Type: N Check: 5094 ..

1 atgatgtagg ttgagaggat tgaacttaag ttcttttcct gactgattat

51 gaggcaacac atgtaccatt aaaatttctc agccggcgcg gtggctcaca

101 ctgtaatcca gcactttgga ggccgaggca ggcagatcga ggtcaggaga

151 tcgagacatc tgctacacag tgaa

ta59.seq Length: 134 November 8, 1993 20:27 Type: N Check: 6907 ..

1 tctgacatgc acacttcccc cttcacacac acaccataga gagcccgtgc

51 tgtagagggt tgtgctggta tgagggtgtt cttgtacaag gcaacaggga

101 agaagaggaa tcgaggccat cagactctgg agaa

ta60.seq Length: 108 November 8, 1993 20:30 Type: N Check: 7155 ..

1 ataaggggat ggccatggct aggtttatag atagnnggtg gttggtgtaa

51 atgagtggca ggagtcgagg aggttagttg tggcaataaa aatgattaag

101 gatactag

ta61.seq Length: 133 November 8, 1993 20:33 Type: N Check: 7012 ..

1 tcatacctgt aaaccagcgg tttgggaggc tgaggaggta ggtacatgag

51 gctaagagtt cgagaccagc ctggccaaca cagtgaaact ctgtcctact

101 aaaaatcgca aaatagccgc gcgtggtggc gag

APPENDIX C: SEQUENCES OF PCR ON COSMID CLONES USING SHORT GC OLIGOS

ta10x3.seq Length: 154 August 4, 1993 17:38 Type: N Check: 1435 ..

1 ttcgcatacc aacgaatggc gttaaagtat ggcaataaag ccttttttag

51 ctacgcttca tagccccagt gaatcggaat accgatggta tcgatatctt

101 tgccgtcgct ttcaggctgc gaataaccct tctcacctcg gccgataagc

151 tttt

ta11x1.seq Length: 133 August 4, 1993 17:52 Type: N Check: 7263 ..

1 gcagaaacac tagttttcct cccagaccac atggaggacc gaggaaggtc

51 ggatttgtgg tcttccgacg cattctcgaa aacctgcacc cttgcctgtc

101 ctcctagacc acaaggagga ccgaccgaga aaa

ta11x6.seq Length: 156 August 4, 1993 17:53 Type: N Check: 620 ..

1 atcctgcaac tccggatgcc tccgctcgaa gtagcgcgtc tgctgctcca

51 tacaagccaa ccacggcctc cagaagaaga tgttggcgac ctcgtattgg

101 gaatcccgaa catcgctcgc tccagtcaat gaccgtgtta tgcgccattg

151 tccgtc

ta12x2.seq Length: 170 August 4, 1993 17:51 Type: N Check: 6291 ..

1 gtcaacgtcg gaagaggtac tggaagaact ggcgctggac tatccgttgc

51 caaaagtgat tctggagtat cgtggtctgg cgaaggctaa atcgacctac

101 accgacaagc tgccgctgat gatcaacccg aaaaccgggc gtgtgcatac

151 ctcttatcac caggcagtaa

ta13x1.seq Length: 164 August 4, 1993 17:51 Type: P Check: 2686 ..

1 ccgcatcgga gttttaaacg aaccggtgac tttcactacc tgatcgtaqg

51 ttcaggatat cggatactgc gcttgcttgc gctccagcgc actgttttga

101 tgtcgtagtt cggatcgcct tttttatggt tcaggccatc gcgaatcgca

151 aacaccagtt tcgg

ta13x3.seq Length: 87 August 4, 1993 17:50 Type: N Check: 1928 ..

1 cggagtttta aacgaaccgg tgactttcac tacctgatcg tagttcagat

51 atccggatac atgcgcttgc ttccgactcc agcgcca

APPENDIX D: M13 SUBCLONES OF λGT10 CDNA CLONES IDENTIFIED BY HYBRIDIZATION OF PCR PRODUCT USING SHORT GC-OLIGOS ON HUMAN GENOMIC DNA

2a1.seq Length: 112 August 23, 1991 13:35 Check: 5332 ..

1 ctcagtgatc gccactcggc tccaactgct ggattacagc gtcagccgcg

51 cctggcagga ctatcttaat agcattgtgg ctcacatata taatgatctc

101 taaaaaaaaa aa

2b1.seq Length: 202 August 13, 1991 20:21 Check: 9294 ..

1 ttcgcggccg actaccttca gtgaaattaa gggtagtgtg atgtccatgt

51 aataaaacag gcttagaaca tcaattcctc tgccatttat gtgcttttat

101 tccccttcaa aatttttatt ggatatatat atttttcagt attcagactc

151 tgaatctgtc atacatcagt gctgtacgct gaagattctg tgctagtact

201 ag

2c1.seq Length: 108 August 13, 1991 16:58 Check: 8148 ..

1 ggagcagata ttgatggtca agtaaactat gaagagttgt acaatgatga

51 cagcaagtga agacttgtac agatgtgtta attctgtaca atgttattgc

101 tttcttgt

2d2.seq Length: 135 August 15, 1991 15:51 Check: 5797 ..

1 tttcttccag ggtaaaaagc aaaagaattc gcggccgctt ttttttttt

51 ttttttttt caatttacag aatttttatt gtaaacagaa gctcattact

101 agttattaca aattgaatca actgaactca agtta

APPENDIX E: M13 CLONES OF PCR PRODUCT USING DEGENERATE CONSENSUS SITE PRIMERS

7a1.seq Length: 135 June 20, 1993 13:04 Type: N Check: 6201 ..

1 agtaaaccag agagcgccaa tcattactct gggaaatgca aaaaccacaa

51 aaataataag ggcgataagg cctgcaactt tgccgaaata acgccacatg

101 aatgccgaaa aaataacgac gatcatgccg tgatc

7a3.seq Length: 65 June 20, 1993 13:02 Type: N Check: 9727 ..

1 cccatttcca gcatacactc gtccagcacg tccagacgtt cttcacgctg

51 ggagccacca tgatc

8a1.seq Length: 127 June 20, 1993 13:08 Type: N Check: 584 ..

1 gatcccgatg cgtccggacc acggtcatca gatgctggac gacctgaaga

51 agaaaaccaa cccaggttac tccgcaattg gtcgtctgaa aggcctggcc

101 gaagttcgcg gtgtctaact ggcgatc

8a2.seq Length: 82 June 20, 1993 13:13 Type: N Check: 2497 ..

1 tagttttgtt ggcgcatgac gtcgccggct agtcgccgcc ggaatttatc

51 ccgagtattc cgtggtcgtc gtgtcctgga tc

8b1.seq Length: 30 June 20, 1993 13:15 Type: N Check: 4559 ..

1 cgcgcctggc ctccattaca tttcttgatc

8b2.seq Length: 45 June 20, 1993 13:17 Type: N Check: 3306 ..

1 gatcctccac ctcctcacca tcacgaccta atcacttccc ggatc

APPENDIX F: SEQUENCES FROM DOP-PCR CLONED INTO M13

plp2.seq Length: 191 August 26, 1993 19:38 Type: N Check: 2684 ..

1 atccatccta ccagcccaag tatcgtccac ccttaggacc tgggacccaa

51 tgtgcatcct caatcgtgcc cttatctgtg gcactcttcc tctggaggaa

101 accacccagc ctccataaca ggcaattgtc ctgtggtcac tttatttgtt

151 cctaccttct cagggatttc atccctgagg aaggtaatat t

plp2r.seq Length: 89 August 26, 1993 19:40 Type: N Check: 2224 ..

1 aagctttttt tgcagggtgg ttgggttttc tttgcgtgct tgggtcagaa

51 tcatttgacc aattatgtgt gaccagatga gaggtgaat

ct11.seq Length: 178 August 26, 1993 19:44 Type: N Check: 6375 ..

1 atccaactca ccaggaagag gagcgagttt gcagcatcaa gtcctagagg

51 tgcttacctg ctccttagga cccctgttct tcctgcttag cttggtcaaa

101 gggaccaccc accatatctc tgtccagcta gcctagggtt aaggctattt

151 ctaagcatag gtcttagaca ctatgaag

ct12.seq Length: 202 August 26, 1993 20:19 Type: N Check: 7304 ..

1 atccatctta cctgggatgt aaggaatcct cttcaaggag aactacaaac

51 cactgctcaa ggaaataaaa gaggatacaa acaaatggaa gaacattcca

101 tgctcatggt aggaagaatc aatattgtga aatggccata ctgccaaggt

151 aatttacaga ttcaatgcca tcccatcaag ctaccaanac ttcttcacag

201 aa

ct13.seq Length: 210 August 26, 1993 20:21 Type: N Check: 620 ..

1 atcccttcct tgcaccttat acaaaaatca attcaagatg gattaaagac

51 ttaaatgtta gacctaaaac cataaaaacc ctagaagaaa acctaggcat

101 taccattcag gacataggca tgggcaagga cttcatgtct aaaacaccaa

151 ggcaatggca acaaagcaaa ctgacaatgg atctataact aagagctctg

201 cacagcaaga

ct15.seq Length: 80 November 8, 1993 17:13 Type: N Check: 5470 ..

1 gtgttatacc taaaaccata aaaaccctag aagaaaacct aggcattacc

51 attcaggaca taggcatggg caaggacttc

ct16.seq Length: 155 August 26, 1993 19:52 Type: N Check: 1498 ..

1 aagcttcttt ttcagtgacg agttttgttt caatattacg cacacgaata

51 caattctgaa gcgagatatc aaagggaaca ggtgaagcgc catttttaat

101 tgtcccagag ctatattctc ccattctaac cgttgaacca ttgactgtag

151 gcccg

ct16f.seq Length: 163 August 26, 1993 19:56 Type: N Check: 4623 ..

1 atccatctta ccgggccatc agtcaatggt tcaacggtta gaatgggaga

51 atatagctct gggacaatta aaaatggcgc ttcacctgtt ccctttgata

101 tctcgcttca gaattgtatt cgtgtgcgta atattgaaac aaaactcgtc

151 actgaaaaag aag

ct17.seq Length: 113 August 26, 1993 20:08 Type: N Check: 8648 ..

1 ggatccatct cacctgggat gtgaaggacc tcttcaagga gaactacaaa

51 ccactgctca aggaaataag agaggataca aacaaatgga agaacattcc

101 atgctcatgg tag

gst21.seq Length: 68 August 26, 1993 20:11 Type: N Check: 722 ..

1 atccaaccca ccaagaagan ntgcagtttc actcctaagc cngcgagacc

51 acgnacncac cagaagga

gst22.seq Length: 152 August 26, 1993 20:03 Type: N Check: 182 ..

1 atccatctca ccgagnncta gatacagagt gtccgttggt gcattcacaa

51 accctgagct agacacaggg tgctgattgg tgtatttaca aaccttgagt

101 tagatacaga gtgccgattg gtgtatttac aatccctgag ctagacataa

151 ag

gst24.seq Length: 237 August 26, 1993 19:59 Type: N Check: 9242 ..

1 atccaactca cctaggcctc ccaaagtgct gggattacag gcgtgacgac

51 tgcacccggt caaaagtttg ggtttcttct ggcctatata tggaccaagt

101 aggggcttgt ggaccttgga tattgtcctc agagcaatac cattctgcct

151 cttgcatcca atttgagctt cctcaggtcc tgccaacata tgtgtactga

201 gttactgagg tcaggtgagc cattacaata ggagaca

gst25.seq Length: 124 August 26, 1993 20:05 Type: N Check: 2195 ..

1 ggggacacag ccaaaccata tcacatacaa actccagaac ttaaatcgag

51 ttctctgagt tgagccgatc tggactgggt ttggaagttg tggtagaagt

101 ttaacgcatc tgaaaagaaa gctt

gst26.seq Length: 134 August 26, 1993 20:15 Type: N Check: 4783 ..

1 atccaactta cctcgaaggt ctgcagtttc actcctaagc cagcgagacc

51 acgaacccac cagaaggaag aaactccaaa cacatctgaa cattagaagg

101 aacaactcca gatgcgccac cttaagagct gtaa

APPENDIX G: SEQUENCES FROM DOP-PCR ON COSMID CLONES USING OLIGOS 103 AND 107, CLONED INTO M13

18x1a.seq Length: 86 August 27, 1993 18:41 Type: N Check: 3471 ..

1 aagctttttt cgcaggagaa gaggacaaag atactcagag agaaaaagta

51 aaagaccgaa gaaggaggct ggagagacca ggatcc

18x1b.seq Length: 103 August 27, 1993 15:15 Type: N Check: 6679 ..

1 aagcttgtgc tcagggtcac acagctaaga aatgacaagg tggagattta

51 aacctatgct tgtctggcta ggcactgtct tatatagtaa gagtgtgagg

101 atg

18x2a.seq Length: 43 August 27, 1993 18:44 Type: N Check: 6179 ..

1 aagcttccat gaatccgcat ccggagattg catcagcacc gca

18x2b.seq Length: 76 August 27, 1993 15:18 Type: N Check: 8648 ..

1 aagcttttgc aatcatctgg tgagagaacc cagcaaggat ggacaggcag

51 aatggaatag aggtaagttg ggatcc

18x3b.seq Length: 145 August 27, 1993 15:20 Type: N Check: 681 ..

1 aagcttcctt cccagaaagt ctgacaccca tgtctttagt ccagcggcca

51 ctagtcactt ttaactggcc gacagtgcct ggtatttagc ccccgaattc

101 taaggaaaga taggacagaa tagcaagcga aaggggtcca atggt

18x4a.seq Length: 146 August 27, 1993 18:50 Type: N Check: 2742 ..

1 aagcttyagg gcttcactgc gaaattcagg cgaatntgtt tacggggttt

51 tttactggtt gatactgttt ttgtcatgtg agtcacctct gactgagagt

101 ttactcactt agccgcgtgt ccactattgc tgggtgagat ggatcc

18x4b.seq Length: 139 August 27, 1993 15:24 Type: N Check: 7165 ..

1 ggcttcactg cgaaattcag gcgaatcgtg tttacggggt tttttactgg

51 ttgatactgt ttttgtcatg tgagtcacct ctgactgaga gtttactcac

101 ttagccgcgt gtccactatt gctgggtagg ttaggatcc

18x5a.seq Length: 86 August 27, 1993 18:58 Type: N Check: 8870 ..

1 cctcagcacc ttcagctgta cagtgagaat gaacagacct cctcacaagg

51 ctgctataag aattaacgag gctgggcatg gtggct

18x6b.seq Length: 79 August 27, 1993 18:33 Type: N Check: 9718 ..

1 aagctttttt tttaaggagc cagcaccata ctctcttctc tcacaggatc

51 attgttccat ctctggtaag tttggatcc

19x1a.seq Length: 91 August 27, 1993 15:06 Type: N Check: 7653 ..

1 ggatccaatc ctaccggctc cagatttatc agcaataaac cagccagccg

51 gaaggccgag cgcagaagtg gtcctgcaac tttatccgcc t

19x2a.seq Length: 101 August 27, 1993 15:08 Type: N Check: 9947 ..

1 ggatccatag aagcagagac catgatgaga acccatcttt aactctgggt

51 cgctagcaca ggggcagagt aggcagtcac tacatcttca gagcttgtga

101 g

19x2b.seq Length: 152 August 27, 1993 18:36 Type: N Check: 2580 ..

1 ggatcctaac ctaccagaag gaagaaactc caaacacatc tgaacattag

51 aaggaacaaa ctccagatgc gccaccttaa gagctgtaac actcaccgcg

101 aggtccacgg cttcattctt gaagtcagtg agagaccaag aacccaccaa

151 tt

19x3a.seq Length: 96 August 27, 1993 15:10 Type: N Check: 4177 ..

1 ggatccaaac tcaccctcac atcttggctc ctccagcccc ttctctgcaa

51 acatggaaca gtgggattat tataaatctc agacccgttt tgcccc

19x3b.seq Length: 93 August 27, 1993 18:38 Type: N Check: 5815 ..

1 ggatccatag aagcagagac catgatgaga acccatcttt aactctgggt

51 cgctagcaca ggggcagagt aggcagtcac tacatcttca gag

19x4a.seq Length: 115 August 27, 1993 15:12 Type: N Check: 5431 ..

1 ggatccaaac ccacctgtat agcccatatg cactttaaac atgagtccaa

51 aatggaatcc aaaatattag atgccctcgt tgccaccacc aggtttgctt

101 tcctctcaag ccact

18x2i.seq Length: 160 December 8, 1993 19:07 Type: N Check: 4419 ..

1 tccttcccag aaagtctgac acccatgtct ttagtccagc ggccactagt

51 cacttttaac tggccgacag tgcctggtat ttagcccccg aattgtaagg

101 aaagatagga cagaatagca agcgaaaggg gtccaatggt actcactgct

151 tggcgatagg

18x3i.seq Length: 161 December 8, 1993 18:38 Type: N Check: 5414 ..

1 ctggtgttgt tgcataaacg gctcaccgcc taactgatac atctgccgta

51 aaatccacgc ctgacggcta ccacgtagcc gatggcgagg agactttgaa

101 acgaagtgat taggtaatac agctgccagt aatgcagctt ccgaccggta

151 agttcggatc c

18x3ii.seq Length: 174 December 8, 1993 18:42 Type: N Check: 9737 ..

1 ttgcagcata cacaggaagg aggatcccga gaggtaggag agaacacact

51 ggccaggaat ccaacaggct gtgttgttca ccgggacctg gggcccagct

101 gttctcagcc tccaagggag acagaggtcc actgcagctg gaggtaccgt

151 ggttagacat aacaaaaggc tccg

19x2ii.seq Length: 135 December 8, 1993 18:45 Type: N Check: 7108 ..

1 ggatccaaac ccaccagaag gaagaaactc caaacacatc tgaacattag

51 aaggaacaaa ctccagatgc gccaccttaa gagctgtaac actcaccgcg

101 aggtccacgg cttcattctt gaagtcagtg agaga

19x4ai.seq Length: 154 December 8, 1993 18:48 Type: N Check: 7676 ..

1 ggatccgaac ctaccgtcca gcaaaaaggg ggacgaggaa tttgaggcct

51 ggcttgaggc tcaggacgca aatcttgagg atgttcagcg ggagttttcc

101 gggctgcgag taattggtga tgaggacgag gatggttcgg aggatgggga

151 attt

19x4bi.seq Length: 136 December 8, 1993 18:52 Type: N Check: 2419 ..

1 ggatccaaac ccacccaggc cctggccgaa agcctgctcc tacagccact

51 ccatcctcct ggaacatggc caagagtagt agctcccctg caaaaaaagc

101 ttggcgtaat catggtcata gctgtttctg tgtgaa

REFERENCES

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R. and Venter, J.G. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252, 1651-1656.

Adams, R.L.P., and Eason R. (1984) Increased G+C content of DNA stabilizes methyl CpG dinucleotides. Nucleic Acids Res. 12, 5869-5877.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

Alwine, J.C., Kemp, D.J. and Stark, G.R. (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl paper and hybridization with DNA probes. Proc. Natl. Acad. Sci. USA 74, 5350.

Ansorge, W., Voss, H., Wiemann, S., Schwager, C., Sproat, B., Zimmerman, J., Stegemann, J., Erfle, H., Hewitt, N. and Rupp, T. High throughput automated DNA sequencing facility with fluorescent labels at the European Molecular Biology Laboratory. Human Genome '92 Abstracts, 59.

ASHG Human Genome Committee Report (1991) The Human Genome Project: Implications for Human Genetics. Am. J. Hum. Genet. 49, 687-691.

Aslanidis, C. and de Jong, P. (1990) Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res. 18, 6069-6074.

Auch, D. and Reth, M. (1990) Exon trap cloning: using PCR to rapidly detect and clone exons from genomic DNA fragments. Nucleic Acids Res. 18, 6743-6744.

Baehner, R.C., Kunkel, L.M., Monaco, A.P., Haines, H.L., Conneally, P.M., Palmer, C., Heerema, N. and

Orkin, S.H. (1986) DNA linkage analysis of X chromosome- linked chronic granulamatous disease. Proc. Natl. Acad. Sci. USA, 83, 3398-3401.

Barlow, D.P. and Lehrach, H. (1987) Genetics by gel electrophoresis: the impact of pulsed gel electrophoresis on mammalian genetics. Trends Genet. 3, 167-171.

Bird, A.P. (1986) CpG-rich island and the function of DNA methylation. Nature 321, 209-213.

Bird, A.R.P. (1987) CpG islands as gene markers in the vertebrate nucleus. Trends Genet. 3, 342-347.

Birnboim, H.C. and Doly, J. (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. 7, 1513-1523.

Blasband, A. (1992) Fast DNA sequencing on the ABI 373A DNA sequencer. Human Genome '92 Abstracts, 58.

Brenner, S. (1990) The Human Genome: The Nature of the Enteprise, Ciba Found. Symp., 149, 6-12.

Breslauer, K.J., Frank, R., Blocker, H. and Marky, L.A. (1986) Predicting DNA duplex stability from the base sequence. Proc. Natl. Acad. Sci. USA 83, 3746-3750.

Broad, P.M., Symes, A.J., Thakker, R.V. and Craig, R.K. (1989) Structure and methylation of human calcitonin/ α-CGRP gene. Nucl. Acids Res. 17, 6999-7011.

Brosius, J. (1989) Superpolylinkers in cloning and expression vectors. DNA 8, 10.

Brown, W.R.A., and Bird, A. (1986) Long range restriction site mapping of mammalian genomic DNA. Nature 322, 477-481.

Bucher, P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol. 212, 563-578.

Buckler, A.J., Chang,D.D., Graw, S.L., Brook, J.D., Haber, D.A., Sharp, P.A. and Housman, D.E. (1991) Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. Proc. Natl. Acad. Sci. USA 88, 4005-4009.

Burglin,T.R. and Barnes,T.M. (1992) Introns in sequence tags. Nature 357, 367

Burke, D.T., Carle, G. and Olsen, M.V. (1987) Cloning of large exogenous DNA into yeast by means of artificial chromosome vectors. Science 236, 806-812.

Burland, V., Daniels, D.L., Plunket, G. and Blattner, F.R. (1993) Genome sequencing on both strands: the Janus strategy. Nucleic Acids Res. 21, 3585-3590.

Caetano-Anolles,G., Bassam,B.J. and Gresshof,P.M. (1991). DNA amplification fingerprinting using very short arbitrary oligonucleotide primers. Biotechnology 9, 553-557.

Call, K.M., Glaser, T., Ito, C.Y., Buckler, A.J., Pelletier, J., Haber, D.A., Rose, E.A., Kral, A., Yeger, H., Lewis, W.H., Jones, C. and Housman, D.E. (1990) Isolation and characterization of a zinc finger polypeptide gene at the human chromosome 11 Wilms' tumor locus. Cell 60, 509-520.

Calladine, C.R. (1982) Mechanics of sequence-dependent stacking of bases in B-DNA. J. Mol. Biol. 161, 343-352.

Cantor,C.R. (1990) Orchestrating the human genome project. Science 248, 49-51.

Carle,G.F. and Olsen,M.V. (1984) Separation of chromosomal DNA molecules from yeast by orthogonal-field alternation gel electrophoresis. Nucleic Acids Res. 12, 5647-5664.

Castleman, H., Hanau, L.H., Zacharias, W. and Erlanger, B.F. (1988) Z-DNA and loop structures by immunoelectromicroscopy of supercoiled pRW751, a plasmid containing left-handed helices. Nucleic Acids Res. 16, 3977-3996.

Cavanee, W.K., Dryja, T.P., Phillips, R.A., Benedict, W.F., Godbout, R., Gallie, B.L., Murphree, A.L., Strong, L.C. and White. R.L. (1983) Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. Nature 305, 779-784.

Chen, T.L. and Manuelidis, L. (1989) SINEs and LINEs cluster in distinct DNA fragments of Giemsa band size. Chromosoma 98, 309-316.

Cole, C.G., Goodfellow, P.N., Bobrow, M. and Bentley, D.R. (1991) Generation of novel sequence tagged sites (STSs) from discrete chromosomal regions using Alu-PCR. Genomics 10, 816-826.

Collins, F. and Galas, D. (1993) A five-year plan for the U.S. Human Genome Project. Science 262, 43-46.

Compton, T. (1990) Degenerate primers for DNA amplification. Innis, M.A., Gelfond, D.H., Sninsky, J.J. and White, T.J. (Eds) p39-45, PCR protocols: a guide to methods and applications. Academic Press, New York.

Cooper, D.N., Taggart, M.H. and Bird, A.P. (1983) Unmethylated domains in vertebrate DNA. Nucleic Acids Res. 11, 647-658.

Cooper, D.N. and Gerber-Huber, S. (1985) DNA methylation and CpG suppression. Cell Differ. 17, 199-205.

Corbo, L., Maley, J.A., Nelson, D.L. and Caskey, C.T. (1990) Direct cloning of human transcripts with HnRNA from hybrid cell lines. Science 249, 652-655.

Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. Nature 274, 775-780.

Cowell, I.G., Dixon, K.H., Pemble, S.E., Ketterer, B. and Taylor, J.B. (1988) The structure of the human glutathione S-transferase π gene. Biochem. J. 255, 79-83.

Cox, D.R., Burmeister, M., Price, E.R., Kim, S. and Myers, R.M. (1990) Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. Science 250, 245-250.

Cross, S.H., Charlton, J.A., Nan, X. and Bird, A. (1994) Purification of CpG islands using a methylated DNA binding column. Nature Genetics 6, 236-244.

Dickerson, R.E. (1983) Base sequence and helix structure variation in B and A DNA. J. Mol. Biol. 166, 419-441.

Dickerson, R.E., Bansal, M., Calladine, C.R., Dieckmann, S., Hunter, W.W., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H.C.M., Olson, W.K., Saenger, W., Shakked, Z., Sklenar, H., Soumpasis, D.M., Tung, C-S., Wang, A.H.J. and Zhurkin, V.B. (1989) Definition and nomenclature of nucleic acid structure parameters. EMBO J. 8, 1-4.

Diehl, H.J., Schaich, M., Budzinski, R.M. and Stoffel, W. (1986) Individual exons encode the integral membrane domains of human myelin proteolipid protein. Proc. Natl. Acad. Sci. USA 83, 9807-9811.

Drayna, D., Davies, K., Mandel, J.L., Camerino, G., Williamson, R. and White, R. (1984) Genetic mapping of the human X chromosome by using restriction fragment length polymorphisms. Proc. Natl. Acad. Sci. USA, 81, 2836-2839.

Drmanac, R., Strezoska, Z., Labat, I., Drmanac, S. and Crkvenjakov, R. (1990) Reliable hybridization of oligonucleotides as short as six nucleotides. DNA Cell Biol. 9, 527-534.

Duyk, G.M., Kim, S., Myers, R.M. and Cox, D.R. (1990) Exon trapping: a genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. Proc. Natl. Acad. Sci. USA 87, 8995-8999.

Duguid, J.R. and Dinauer, M.C. (1990) Library subtraction of in vitro cDNA libraries to identify differentially expressed genes in scrapie infection. Nucleic Acids Res. 18, 2789-2792.

Dulbecco, R. (1986) A turning point in cancer research: sequencing the human genome. Science 231, 1055-1056.

Ehrlich,H.A., ed.(1989).PCR Technology: Principles and Applications for DNA Amplification. Stockton Press, New York.

Elvin, P., Slynn, G., Black, D., Graham, A., Butler, R., Riley, J., Anand, R. and Markham, A.F. (1990) Isolation of cDNA clones using yeast artificial chromosome probes. Nucleic Acids Res. 18, 3913-3917.

Engelke, D.R., Hoener, P.A. and Collins, F.S. (1988) Direct sequencing of enzymatically amplified human genomic DNA. Proc. Natl. Acad. Sci. USA 85, 544-548.

Estivill, X., Farrall, M., Scambler, P.J., Bell, G.M., Hawley, K.M.F., Lench, N.J. and Bates, G.P. (1987) A candidate for the cystic fibrosis locus isolated by selection for methylation-free islands. Nature 326, 840-845.

Estivill, X. and Williamson, R. (1987) A rapid method to identify cosmids containing rare restriction sites. Nucleic Acids Res. 15, 1415-1425.

Eubanks, J.H., et al (1992) Localization of D5 dopamine receptor gene to human chromosome 4p15.1-p15.3, centromeric to the Huntingdon's disease locus. Genomics 12, 510-516.

Fargnoli, J., Holbrook, N.J., and Fornace, A.C.(Jr) (1990) Low-ratio hybridization subtraction. Anal. Biochem. 187, 364-373.

Feinberg, A.P. and Vogelstein, B. (1984) A technique for radiolabelling DNA restriction endonuclase fragments to high specific activity. Addendum Anal Biochem 137, 266-267.

Ferguson-Smith, M., White, J. and Albertson, D. (1992) Multicolour fluorescence in situ hybridization in human gene mapping and related projects. GNome News 11, 27-28.

Fickett, J.W. (1982) Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. 10, 5303-5318.

Filichkin, S.A. and Gelvin, S.B. (1992) Effect of dimethyl sulfoxide concentration on specificity of primer matching in PCR. Biotechniques 12, 828-830.

Fisher, P.A. and Korn, D. (1981) Biochemistry 20, 4560-4578.

Francke, U. (1984) Random X inactivation resulting in mosaic nullisomy of region Xp21.1-p21.3 associated with heterozygosity for ornithine transcarbamylase deficiency and for chronic granulamatous disease. Cytogenet. Cell Genet. 38, 298-307.

Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, J.N., Caruthers, M.H., Neilson, T. and Turner, D.H. (1986) Proc. Natl. Acad. Sci. USA 83, 9373-9377.

Frischauf, A.M. (1989) Construction and use of linking libraries. Technique 1, 3-10.

Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. J. Mol. Biol. 196, 261-282.

Genetics Computer Group (1991) Program Manual for the GCG package, version 7, April 1991. 575, Science Drive, Madison, Wisconsin, USA 53711.

Ghosh, D. (1990) A relational database of transcription factors. Nucleic Acids Res. 19, 1749-1756.

Glover, D.M. (1985) DNA cloning- a practical approach, Vol.1, published by IRL Press, Oxford.

Goldberg D.A. (1980) Isolation and partial characterization of the *Drosophila* alcohol dehydrogenase gene. Proc Natl. Acad. Sci. USA 77, 5794-

Gribskov, M. and Devereux, J. (1991) Sequence Analysis Primer, published by Stockton Press, New York.

Gusella, J.F. , Wexler, N.P., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C. et al (1983) A polymorhic DNA marker genetically linked to Huntingdon's disease. Nature 306, 234-238.

Habener, J.F., Chon, D.V., Dinh, B.L., Gryan, G.P., Ercolani, L. and Wang, A.H.J. (1988) 5'Fluorodeoxyuridines as an alternative to the synthesis of mixed hybridization probes. Proc. Natl. Acad. Sci. USA 85, 1735-1739.

Hadano,S., Watanabe,M., Yokoi,H. Kogi,M., Kondo,I., Tsuchiya,H., Kanazawa,I., Wakasa,K. and Ikeda,J.E.

(1991). Laser microdissection and single unique primer PCR allow generation of regional chromosome DNA clones from a single human chromosome. Genomics 11, 364-373.

Hartley, D.A., Davies, K.E., Drayna,D., White, R.L. and Williamson, R. (1984) A cytological map of the human X chromosome: evidence of non-random recombination. Nucleic Acids Res. 12, 5227-5285.

Heller, R.A., Song, K. and Freire-Moar, J. (1992) Rapid screening of libraries with radio-labeled DNA sequences generated by PCR using highly degenerate oligonucleotide mixtures. Biotechniques 12, 30-35.

Hoeltke, H.J. (1993) A new spectrum of nucleic acid detection systems. Boehringer Mannheim Biochemica Bulletin 1, 5-6.

Hoheisel, J.D., Craig, A.D. and Lehrach, H. (1990) Effect of 5-Bromo and 5-Methyldeoxycytosine on duplex stability and discrimination of the NotI octadeoxynucleotide. J. Biol. Chem., 265, 16656-16660.

Holliday, R. (1989) Untwisting B-Z DNA. Trends Genet. 5, 255-256.

Holton, T.A. and Graham, M.W. (1991) A simple and efficient method for direct cloning of PCR products using ddT-tailed vectors. Nucleic Acids Res. 19, 1156.

Hou, W. and Smith, L.M. (1993) Fluorescence-based DNA sequencing with hexamer primers. Nucleic Acids Res. 21, 3331-3332.

Huang, M.M., Arnheim, N. and Goodman M.F. (1992) Extension of base mispairs by Taq DNA polymerase: implications for single nucleotide discrimination in PCR. Nucleic Acids Res., 20, 4567-4573.

Human Gene Mapping 11 Abstracts (1992) Cytogenet. Cell Genet. 43

Hung, T., Mak, K. and Fong, K. (1990) A specificity enhancer for polymerase chain reaction. Nucleic Acids Res. 18, 4953.

Innis, M.A., Myambo, K.B., Gelfand, D.H., Brow, M.A.D. (1988). DNA sequencing with Thermus aquaticus DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. Proc. Natl. Acad. Sci. USA, 85, 9436-9440.

Josse, J., Kaiser, A.D. and Kornberg, A. (1961) Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J. Biol. Chem. 236, 864-875.

Joyce, C.M., Ollis, D.L., Rush, J., Steitz, P.A., Konigsberg, W.H. and Grindley N.D.F. (1986) UCLA Sym. Mol. Cell Biol., New Ser. 32, 197-205.

Jung, V., Petska, S.B. and Petska, S. (1990) Efficient cloning of PCR generated DNA containing terminal restriction endonuclease recognition sites. Nucleic Acids Res. 18, 6156.

Kaufman, D.L. and Evans, G.A. (1990) Restriction endonuclease cleavage at the termini of PCR products. Bio Techniques 9, 304-306.

Khrapko, K.R., Lysov, Y.P., Khorlyn, A.A., Shick, V.V., Florentiev, V.L. and Mirzabekov, A.D. (1989) An oligonucleotide hybridization approach to DNA sequencing. FEBS Let. 256, 118-122.

Kieleczawa, J., Dunn, J.J. and Studier, F.W. (1992) DNA sequencing by primer walking with strings of contiguous hexamers. Science 258, 1787-1791.

Kinzler, K.W. and Vogelstein, B. (1989). Whole genome PCR: application to the identification of sequences bound by gene regulatory proteins. Nucleic Acids Res. 17, 3645-3653

Kioussis, D., Wilson, F., Daniels, C., Leveton, C., Taverne, T. and Playfair, J.H.L. (1987) Expression and rescuing of a cloned human tumour necrosis factor gene using an EBV-based shuttle cosmid vector. EMBO J. 6, 355-361.

Knoth, K., Roberds, S., Poteet, C. and Tamkun, M. (1988) Highly degenerate, inosine containing primers specifically amplify rare cDNA using the polymerase chain reaction. Nucleic Acids Res. 16, 10932.

Koch, M.S.H. (1990) An "equalized cDNA library" by the reassociation of short double-stranded cDNAs. Nucleic Acids Res. 18, 5705.

Kogan, S.C., Doherty, M. and Gitschier, J. (1987) An improved method for prenatal diagnosis of genetic diseases by analysis of amplified DNA sequences. New Engl. J. Med. 317. 985-990.

Korenberg, J.R. and Rykowski, M.C. (1988) Human genome organization: Alu, Lines and the molecular structure of metaphase chromosome bands. Cell 53, 391-400.

Korn, B., Sedlacek, Z., Manca, A., Kioschis, P., Konecki, D., Lehrach, H. and Poustka, A. (1992) A strategy for the selection of transcribed sequences in the Xq28 region. Hum. Mol. Genet. 1, 235-242.

Kotler, L.E., Zevin-Sonkin, D., Sobolev, I.A., Beskin, A.D. and Ulanovsky, L.E. (1993) DNA sequencing: modular primers assembled from a library of hexamers pentamers. Proc. Natl. Acad. Sci. USA 90, 4241-4245.

Kuhn, L.C., McClelland, A. and Ruddle, F.H. (1984) Gene transfer, expression and molecular cloning of the human transferrin gene. Cell 37, 95-103.

Kusuda, J., Kameoka, Y., Takahashi, I., Fujiwara, H. and Hashimoto, K. (1989) A simple method for screening NotI linking clones. Nucleic Acids Res. 17, 8890.

Kusuda, J., Hirata, M., Yoshizaki, N., Kameoka, Y., Takahashi, I. and Hashimoto, K. (1990) Over 80% of *Not*I sites are associated with CpG rich islands in the sequenced human DNA. Jpn. J. Hum. Genet. 35, 277-282.

Larsen, F., Gundersen, G., Lopez, R and Prydz, H. (1992) CpG islands as gene markers in the human genome. Genomics 13, 1095-1107.

Lau, Y.F. and Kan, Y.W. (1983) Versatile cosmid vectors for the isolation, expression, and rescue of gene sequences: studies with the human α-globin gene cluster. Proc. Natl. Acad. Sci. USA 80, 5225-5229.

Lehrach, H., Diamond, D., Wozney, J.M. and Boedtker, H. (1977) RNA molecular weight determinations by

gel electrophoresis under denaturing conditions, a critical reexamination. Biochemistry 16, 4743-.

Lennon, G.G., and Fraser, N.W. (1983) CpG frequency in large DNA segments. J. Mol. Evol. 19, 286-288.

Lindsay, S. and Bird, A.P. (1987) Use of restriction enzymes to detect potential gene sequences in mammalian DNA. Nature 327, 336-338.

Litt, M. and Luty, J.A., (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. 44, 397.

Little, P. (1991) Complementary Questions. Nature, 352, 20-21.

Littman, D,R., Thomas, Y., Maddon, P.J., Chess,L. and Axel,R. (1985) The isolation and sequence of the gene encoding T8: a molecule defining functional classes of T lymphocytes. Cell 40, 237-246.

Liu, P., Legerski, R. and Siciliano, M.J. (1989) Isolation of human transcribed sequences from human-rodent somatic cell hybrids. Science 246, 813-815.

Lovett, M., Kere, J. and Hinton, L.M. (1991) Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. Proc. Natl. Acad. Sci. USA 88, 9628-9632.

Maddox, J. (1992) Ever-longer sequences in prospect. Nature 357, 13.

Maniatis T., Sambrook, J. and Fritsh, E.F. (1982) Molecular cloning (a laboratory manual) Cold Spring Harbor Laboratory.

Marchuk, D., Drumm, M., Saulino, A. and Collins, F.S. (1991) Construction of T-vectors, a rapid and general system for direct cloning of unmodified PCR products. Nucleic Acids Res. 19, 1154.

Martin, F.H. and Castro, M.M. (1985) Base pairing involving deoxyinosine. Implications for probe design. Nucleic Acids Res. 13, 8927-8938.

Maxam, A.M. and Gilbert, W. (1980) Sequencing end-labelled DNA with base-specific cleavages. Methods Enzymol. 65, 499-.

McLauchlan, A.D., Gaffrey, D., Whitton, J. and Clements, J. (1985)The consensus sequences YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. Nucleic Acids Res. 13, 1347-1368.

Mead, D.A., Pey, N.K., Herrnstadt, C., Marcil, R. and Smith, L.M. (1991) A universal method for the direct cloning of PCR amplified nucleic acid. Bio Technology 9, 657-663.

Melmer, G. and Buchwald, M. (1990) Use of short oligonucleotides to screen cosmid libraries for clones containing G/C rich sequences. DNA Cell Biol. 9, 377-385.

Melmer, G., Sood, R., Rommens, J., Rego, D., Tsui, L.-C. and Buchwald, M. (1990) Isolation of clones on chromosome 7 that contain recognition sites for rare-cutting enzymes by oligonucleotide hybridization. Genomics 7, 173-181.

Melmer, G. and Buchwald, M. (1992) Identification of genes corresponding to splice site consensus sequences. Hum. Mol. Genet. 1, 433-438.

Millican, T.A., Mock, G.A., Chauncey, M.A., Patel, T.P., Eaton, M.A.W., Gunning, J., Cutbush, S.D., Neidle, S. and Mann, J. (1984) Synthesis and biophysical studies of short oligodeoxynucleotides with novel modifications: a possible approach to the problem of mixed base oligodeoxynucleotide synthesis. Nucleic Acids Res. 12, 7435-7453.

Mirzabekov, A.D. (1992) Sequencing by hybridization to oligonucleotide matrix: development and application. Human Genome '92 Abstracts, 27.

Monaco, A.P. et al (1986) Isolation of candidate cDNAs for portions of Duchenne muscular dystrophy gene. Nature 323, 646-650.

Morgan, J.G., Dolganov, G.M., Robbins, S.E., Hiton, L.M. and Lovett, M. (1992) Nucleic Acids Res. 20, 5173-5179.

Moyzis, R.K., Torney, D.C., Meyne, J., Buckingham, J.M., Wu, J-R., Burks, C., Sirotkin, K.M. and Goad, W.B. (1989) The distribution of interspersed repetitive sequences in the human genome. Genomics 4, 273-289.

Murray, V. (1989) Improved double stranded DNA sequencing using the linear polymerase chain reaction. Nucleic Acids Res. 17, 8889.

Nevinsky, G.A., Veniaminova, A.G., Levina, A.S., Podust, V.N. and Lavrik, O.I. (1990). Structure-function analysis of mononucleotides and short oligonucleotides in the priming of enzymatic DNA synthesis. Biochemistry 29, 1200-1207.

NIH/CEPH Collaborative Mapping Group (1992) A comprehensive genetic linkage map of the human genome. Science 258, 67-86, 148-162.

Ohtsuka, E., Matsuki, S., Ikehara, M., Takahashi, Y. and Matsubara, K. (1985) An alternative approach to deoxyoligonucleotides as hybridization probes by insertion of deoxyinosine at ambiguous codon positions. J. Biol. Chem. 260, 2605-2608.

Olsen, M.V., Hood, L., Cantor, C. and Botstein, D. (1989) A common language for physical mapping of the human genome. Science 245, 1434-1435.

Orlandi,R., Gussow,D.H., Jones,P.T.and Winter,G. (1989). Cloning immunoglobulin variable domains for expression by the polymerase chain reaction. Proc. Natl. Acad. Sci. USA 86, 3833-3837.

Page, D.C., Mosher, R., Simpson, E.M., Fisher, E.M.C., Mardon, G., Pollack, J., McGillivray, B., de la Chapelle, A. and Brown, L.G. (1987) The sex-determining region of the human Y chromosome encodes a finger protein. Cell 51, 1091-1104.

Parimoo, S., Patanjali, S.R., Shukla, H., Chaplin, D.D., and Weissman, S.M. (1991) cDNA selection: efficient

PCR approach for the selection of cDNAs encoded in large chromosomal fragements. Proc. Natl. Acad. Sci. 88, 9623-9627.

Patanjali, S.R., Parimoo, S. and Weissman, S.M. (1991) Construction of a uniform-abundance (normalized) cDNA library. Proc. Natl. Acad. Sci. USA 88, 1943-1947.

Patel,K., Cox,R., Shipley,J., Kiely,F., Frazer,K., Cox,D.R., Lehrach,H. and Sheer,D. (1991). A novel and rapid method for isolating sequences adjacent to rare cutting sites and their use in physical mapping. Nucl. Acids Research 19, 4371-4375.

Patil, R.V. and Dekker, E.E. (1990) PCR amplification of an Escherichia coli gene using mixed primers containing deoxyinosine at ambiguous positions in degenerate amino acid codons. Nucleic Acids Res. 18, 3080.

Peticolas, W.L., Wang, Y. and Thomas, G.A. (1988) Some rules for predicting base-sequence dependence of DNA conformation. Proc. Natl. Acad. Sci. USA 85, 2579-2583.

Pohl, F. (1992) Reality of sequencing costs. Nature 357, 106.

Pohl, T.M., Zimmer, M., MacDonald, M.E., Bucan, M., Poutska, A., Volinia, S., Searle, S., Zehetner, G., Wasmuth, J.J., Gusella, J., Lehrach, H. and Frischauf, A.-M. (1988) Construction of *Not*I linking library and isolation of new markers close to the Huntingdon's disease gene. Nucleic Acids Res. 16, 9185-9198.

Pritchard, C. et al (1991) The end in sight for Huntingdon's disease ? Am. J. Hum. Genet. 49, 1-6.

Prober, J.M., Trainor, G.L., Dam, R.J., Hobbs, F.W., Robertson, C.W., Zagursky, R.J., Cocuzza, A.J., Jensen, M.A. and Baumeister, K. (1987) A system for rapid DNA sequencing with fluorescent chain terminating dideoxynucleotides. Science 238, 336-341.

Poutska, A. and Lehrach, H. (1986) Jumping libraries and linking libraries: the next generation of molecular tools in mammalian genetics. Trends Genet. 2, 174-179.

Reilly, J.D., Melhem, R.F., Lutz, C.M. and Edmonds, M. (1990) Transcription vectors that facilitate the identification and mapping of RNA splice sites in genomic DNA. DNA Cell Biol. 9, 535-542.

Riordan, J.R., Rommens, J.M., Kerem, B-S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, et al (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245, 1066-1073.

Roberts, L. (1991) Gambling on a shortcut to genome sequencing. Science 252, 1618-1619.

Roemer, S., Middendorf, L., Steffens, D., and Sutter, S. (1992) Rapid automated DNA sequencing using various gel sizes and matrices. Human Genome '92 Abstracts, 58.

Rommens, J.M., Iannuzzi, M.C., Kerem, B-S., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, J.L., Kennedy, D., Hidaka, N., Zsiga, M., Buchwald, M., Riordan, J.R., Tsui, L-C. and Collins, F.S. (1989) Identification of the cystic fibrosis gene: chromosome walking and jumping. Science 245, 1059-1065.

Ruano, G. and Kidd, K.K. (1991). Coupled Amplification and sequencing of human DNA. Proc. Natl. Acad. Sci USA, 88, 2815-2819.

Rychlik,W. and Rhoads,R.E. (1989). A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acid Res. 17(21), 8543-8551.

Saenger, W. (1984) Principles of nucleic acid structure. Spring-Verlag, NY.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., and Arnheim, N.(1985).Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230, 1350-1354.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239, 487-489.

Sambrook, J., Fritsh, E.F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sanger, F., Niklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74, 5463-5467.

Sargan, D.R., Gregory, S.P. and Butterworth, P.H. (1982) A possible novel interaction between the 3'end of 18S ribosomal RNA and the 5'-leader sequence of many eukaryotic messenger RNAs. FEBS-Lett. 147, 133-136.

Schmid, D.W. and Girou, C. (1987) Cloning of cDNA derived from mRNA of the electric lobe of *Torpedo marmarata* and selection of putative cholinergic-specific sequences. J. Neurochem. 48, 307-312.

Schwarz, D.C. and Cantor, C.R. (1984) Separation of yeast chromosome-sized DNA's by pulsed field gradient gel electrophoresis. Cell 37, 67.

Senapathy, P., Shapiro, M.B. and Harris, N.L. (1990) Splice junctions, branch point sites, and exons: sequence statistics, identification and applications to the genome project. Meth. Enzym. 183, 252-278.

Shen, W.H. and Hohn, B. (1992) DMSO improves PCR amplification of DNA with complex secondary structure. Trends Genet. 8, 227.

Siemieniak, D.R. and Slightom, J.L. (1990) A library of 3342 useful nonamer primers for genome sequencing. Gene 96, 121-124.

Silva, A.J. and White, R. (1988) Inheritance of allelic blueprints for methylation patterns. Cell 54, 145-152.

Smith, C.L., Econome, J.G., Schutt, A., Klco, S. and Cantor,C.R. (1987) A physical map of the *Escherichia coli* K12 genome. Science 236, 1448-1453.

Smith, C.L., Lawrence, S.K., Gillespie, G.A., Cantor, C.R., Weissman, S.M. and Collins, F.S. (1988) Strategies for mapping and cloning macroregions of mammalian genomes. Methods in Enzymology (M.M. Gottesman Ed.)

Academic Press, San Diego, Vol. 151, 461-489.

Smith, T.F., Waterman, M.S. and Sadler, J.R. (1983) Statistical characterization of nucleic acid functional domains. Nucleic Acids Res. 11, 2205-2220.

Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98: 503-517.

Spencer, S.R., Taylor, J.B., Cowell, I.G., Xia, C-L., Pemble, S.E. and Ketterer, B. (1992) The human mitochondrial NADH: ubiquinone oxidoreductase 51-KDa subunit maps adjacent to the glutathione-S-transferase P1-1 gene on chromosome 11q13. Genomics 14, 1116-1118.

Stoker, A.W. (1990) Cloning of PCR products after defined cohesive termini are created with T4 DNA polymerase. Nucleic Acids Res. 18,

Stollar, B.D. (1992) Immunochemical analyses of nucleic acids. Prog. Nucleic. Acid Res. Mol. Biol. 42, 39-77.

Studier, F.W. (1989) A strategy for high volume sequencing of cosmid DNAs: Random and directed priming with a library of oligonucleotides. Proc. Natl. Acad. Sci. USA, 86, 6917-6921.

Suggs, S.V., Hirose, T., Miyake, E.H., Kawashima, M.J., Johnson, K.I. and Wallace, R.B. (1981) Use of synthetic oligodeoxyribonucleotides for the isolation of specific cloned DNA sequences. In D.D.Brown (ed.), ICN-UCLA Symp. Dev. Biol. Using Purified Genes. Academic Press Inc., New York, Vol. 23, 683-693.

Swartz, M.N., Trautner, T.A. and Kornberg, A. (1962) Enzymatic studies of deoxyribnucleic acid. J. Biol. Chem. 237, 1961-1967.

Syzbalski, W. (1990) Proposal for sequencing DNA using ligation of hexamers to generate elongation primers (SPEL-6). Gene, 90, 177-178.

Tabor, S. and Richardson, C.C. (1987) DNA sequence analysis with a modified bacteriophage T7 DNA

polymerase. Proc. Natl. Acad. Sci. USA 84, 4767-4771.

Tabor, S. and Richardson, C.C. (1989) Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by in vitro mutagenesis. J. Biol. Chem. 264, 6447-6458.

Tagle, D.A., Valdes, J., Hinton L., Morgan, J., Bates, G., Macdonald, M., Gusella, J., Lehrach, H., Lovett, M. and Collins, F.S. (1992) Magnetic capture of expressed sequences using a pool of six YACs that span the 2.2 Mb Huntington disease candidate region. Am. J. Hum. Genet. 51, A22.

Tagle, D.A., Swaroop, M., Lovett, M. and Collins, F.S. (1993a) Magnetic bead capture of expressed sequences encoded within large genomic segments. Nature 361, 752-753.

Tagle, D.A., Valdes, J., Hinton L., Morgan, J., Bates, G., Macdonald, M., Gusella, J., Lehrach, H., Lovett, M. and Collins, F.S. (1993b) Magnetic capture of cDNAs using a pool of 6 YACs that span the HD candidate region. Cytogenet. Cell Genet. 62, 86.

Takahaski, Y., Kato, K., Hayashizaki, Y., Wakabayashi, T., Ohtsuka, E., Ikehara, M. and Matsubarak, (1985) Molecular cloning of the cholecystokinin gene by use of probe design. Proc. Natl. Acad. Sci. USA, 82, 1931-1935.

Telenius, H., Carter, N.P., Bebb, C.E., Nordenskjöld, M., Ponder, B.A.J. and Tunnacliffe, A. (1992) Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics 13, 718-725.

Travis, F.H., and Sutcliffe,J.G. (1988) Phenol emulsion-enhanced DNA-driven subtractive cDNA cloning: isolation of low abundance monkey cortex-specific mRNAs. Proc. Natl. Acad. USA 85, 1696-1700.

Travers, A.A. (1989) DNA conformation and protein binding. Ann. Rev. Biochem. 58, 427-452.

Tybjoerg-Hansen,A., Gallagher,J., Vincent,J., Houlsten,R., Talmud,P., Dunning,A.M., Seed,M., Hamsten,A., Humphries,S.E., Myant,N.B., (1990) Familial defective apolipoprotein B-100: detection in the United Kingdom

and Scandinavia, and clincal characteristics of ten cases. Atherosclerosis 80, 235-242

Tykocinski, M.L. and Max, E.E. (1984) CG dinucleotide clusters in MHC genes and in 5' demethylated genes. Nucleic Acids Res. 12, 4385-4396.

Wahl, G.M., Lewis, K.A., Ruiz, J.C., Rothenberg, B., Zhao, J. and Evans, G.A. (1987) Cosmid vectors for rapid walking, restriction mapping, and gene transfer. Proc. Natl. Acad. Sci. USA 84, 2160-2164.

Wallace, M.R., Marchuk, D.A., Andersen, L.B., Letcher, R., Odeh, H.M., Saulino, A.M., Fountain, J.W., Brereton, A., Nicholson, J., Mitchell, A.L., Brownstein, B.H. and Collins, F.S. (1990) Type 1 neurofibromatosis gene: identification of a large transcript disrupted in three NF1 families. Science 249, 181-186.

Wallace, R.B., Shaffer, J., Murphy, R.F., Bonner, J., Hirose, T. and Itakura, K. (1979) Hybridization of synthetic oligodeoxyribonucleotides to $\phi\chi$ 174 DNA: the effect of single base pair mismatch. Nucleic Acids Res. 6, 3543-3557.

Watson,J.D. (1990) The Human Genome Project: Past, Present and Future. Science 248, 44-49.

Weber, J.L., and May, P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. 44, 388-396.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G. and Lathrop, M. (1992) A second-generation linkage map of the human genome. Nature 359, 794-801.

Wesley, C.S., Ben, M., Kreitman, M., Hagay, N. and Eanes, W.F. (1990) Cloning regions of the Drosophila genome by microdissection of polytene chromosome DNA and PCR with nonspecific primer. Nucleic Acids Res. 18, 599-603.

Wetmur, J.G. and Davidson, N. (1968) Kinetics of renaturation of DNA. J. Mol. Biol. 31, 349.

White,T.J., Arnheim,N. and Ehrlich,H.A. (1989). The polymerase chain reaction. Trends Genet. 5(6), 185-189.

Williams, J.F. (1989) Amplifications 3, 19.

Wood, W.I., Gitschier, J., Lasky, L.A. and Lawn, R.M. (1985) Base composition independent hybridization in tetramethylammonium chloride: a method for oligonucleotide screening of highly complex gene libraries. Proc. Natl. Acad. Sci. USA 82, 1585-1588.

Worton, R.G., Duff, C., Sylvester, J.E., Schmickel, R.E. and Willard, J.F. (1984) Duchenne muscular dystrophy involving translocation of the dmd gene next to ribosomal RNA genes. Science 224, 1447-1449.

Wu, D.Y., Ugozzoli, L., Pal, B.K., Qian, J. and Wallace, R.B. (1991) The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. DNA Cell Biol. 10, 233-238.

ADDITIONAL REFERENCES

Antequera, F. and Bird, A. (1993) Number of CpG islands and genes in human and mouse. Proc. Natl. Acad. Sci. USA 90, 11995-11999.

Bradbury, E.M., Maclean, N. and Mathews, H.R. (1981) DNA, Chromatin and Chromosomes. Blackwells, Oxford.

Craig, J.M. and Bickmore, W.A. (1994) The distribution of CpG islands in mammalian chromosomes. Nature Genet. 7, 376-381.

Cross, S., Kovarik, P., Schmidtke, J. and Bird, A. (1991) Non-methylated islands in fish genomes are GC-poor. Nucleic Acids Res. 19, 1469-1474.

Goldman, M.A., Holmquist, G.P., Gray, M.C., Caston, L.A. and Nag, A. (1984) Replication timing of genes and middle repetitive sequences. Science 224, 686-692.

Holmquist, G.P. (1992) Chromosome bands, their chromatin flavors, and their functional features. Am. J. Hum. Genet. 51, 17-37.

Solter, D. (1988) Rev. Genet. 22, 127-146.

Tazi, J. and Bird, A. (1990) Alternative chromatin structure at CpG islands. Cell 60, 909-920.