CONSTRUCTION OF PROBES FOR HUMAN COLLAGEN GENE SEQUENCES

A thesis submitted to the University of Adelaide
for the degree of Doctor of Philosophy

by

David McKenzie Bird, B.Sc. (Hons.)

Department of Biochemistry,
University of Adelaide,
ADELAIDE. SOUTH AUSTRALIA.

JANUARY, 1984.

STATEMENT

This thesis contains no material which has been accepted for the award of any other degree or diploma by any University. To the best of my knowledge, it contains no material that has been previously published by any other person, except where due reference is made in the text.

David McKenzie Bird

ACKNOWLEDGEMENTS

I wish to thank:

Professor W.H. Elliott, for permission to work in the Department of Biochemistry, University of Adelaide. **Dr. J.R.E. Wells,** for his supervision and support.

Paul Krieg, Alan Robins and other members of the laboratory for creating a friendly and stimulating atmosphere in which to work, and for providing useful technical advice and critical discussion.

During the course of this work I was supported by a Commonwealth Postgraduate Research Award.

CONTENTS

CHAPTER I : INTRODUCTION

CHAPTER II : MATERIALS AND METHODS

CHAPTER III: ISOLATION OF A PUTATIVE COLLAGEN GENOMIC CLONE

CHAPTER IV : CHARACTERISATION OF GENOMAL CLONES

CHAPTER V: CONSTRUCTION OF A HUMAN FIBROBLAST cDNA LIBRARY

CHAPTER VI: ISOLATION AND CHARACTERISATION OF RECOMBINANTS CONTAINING HUMAN COLLAGEN GENE SEQUENCES

CHAPTER I

INTRODUCTION

I.1. INTRODUCTION

In general, biological systems have similar functional aims. To achieve these similar aims, superficially diverse systems frequently utilize similar structures. The "collagen helix" is one such structure, which, along with the α-helix and β-pleated sheet, is one of the fundamental "ordered" configurations that a protein may take.

This very rigid structure is widely found in both vertebrate (Braconnot, 1820) and invertebrate (Maser and Rice, 1962; Adams, 1978) integument, it is a major component of the organic matrix upon which the vertebrate skeleton is assembled (Glimcher and Krane, 1968), it is associated with tissue membranes (Goodman et al., 1955) and provides a number of enzymes, including acetyl cholinesterase (Hall and Kelly, 1971; Rosenberry and Richardson, 1977) and the C1q component of complement (Ried et al., 1972; Ried and Porter, 1976), with rigid stalks. In addition, ultrastructural analyses suggest that collagen-like molecules may exist in some unicellular organisms (discussed by Gross, 1980). Thus, in a strict sense, it is not correct to speak of "collagen" in the same manner as one would, for instance, say, "insulin". However, for convenience, those "collagen-containing" proteins whose primary role appears to be structural are collectively referred to as "collagen", and those "collagen-containing" proteins which appear to have a different role, for example as enzymes, are not classed as collagens.

The vertebrate collagens have been divided into five structurally distinct types (Bornstein and Sage, 1980) shown in Table I.1., and, unless otherwise mentioned, I

TABLE I.1.

Chain composition and tissue distribution of collagen types (Bornstein and Sage, 1980; [a]Trüeb *et al.*, 1980; [b]Brown *et al.*, 1978).

TYPE	TISSUE DISTRIBUTION	CHAIN COMPOSITION
I	Almost ubiquitous	$\alpha\ 1(I)_2\ \alpha\ 2(I)$ $\alpha\ 1(I)_3$ - type I trimer
II	Cartilage, vitreous	$\alpha\ 1(II)_3$
III	Same as type I, except absent in bone, very low in tendon. Predominant in distensible tissues	$\alpha\ 1(III)_3$
IV	Basement membrane	$\alpha\ 1(IV)_2\ \alpha\ (IV)^a$
V	Foetal membranes, placenta	Not clearly established. Either, or both, of the species $\alpha\ 1(V)_3$ and $\alpha\ 1(V)_2\ \alpha\ 2(V)$ may exist. Another form containing $\alpha\ 3(V)^b$ chain(s) also occurs.

shall use the collective definition and this classification of collagen.

In addition to these five types, however, a number of other collagen species recently have been identified, including endothelial collagen (EC), isolated from cultured bovine aortic endothelial cells (Sage et al., 1980), type VI collagen, isolated from human (Furuto and Miller, 1980 and 1981) and bovine (Jander et al., 1981 and 1983) placenta, high and low molecular weight collagens from chick hyaline cartilage (Reese and Mayne, 1981), long-chain (LC) or type VII collagen, found in human chorioamniotic membranes (Bentz et al., 1983) and the 7S collagenous bridging fragments associated with basement membranes (Risteli et al., 1980; Madri et al., 1983). It is likely that other species may be discovered, especially now that gene probes for collagen sequences are available (Section I.3.).

Although much is known about the structure and biosynthesis of collagen (Section I.2.), comparatively little is known about how this relates, mechanistically, to its function. Whilst it is clear that, in general, the role of collagen is related to its tensile properties, and that tissues which have different mechanical requirements (e.g. compressive resiliance in articular cartilage, tensile strength in tendon, partial elasticity in skin) have different collagen types, it would be naïve to try to understand the function of collagen solely by studying its structure alone. Collagen is but one component of connective tissue, which contains, in varying amounts and spatial arrangements, collagen, elastin, proteoglycan,

glycosaminoglycan, laminin, minerals, fibronectin, other glycoproteins and, likely, other, as yet unidentified, components. The particular function of any one of these components may be obscured by the function of the overall composite.

Perhaps the most powerful technique available to analyse the functional contribution of each of the elements of connective tissue, including collagen, to its overall function, is the use of genetics. Unfortunately, the metazoans most amenable to genetic manipulation, *viz.*, *Drosophila* and *Caenorhabditis elegans*, appear to have collagens which are different from the interstitial collagens (i.e. types I, II and III) of vertebrates at least (Monson *et al.*, 1982; Kramer *et al.*, 1982). Direct mutation of the mouse α 1(I) collagen gene using retroviral insertion into the germ line has recently been reported (Schnieke *et al.*, 1983), but this method is probably a technically unwieldy way of generating specific mutants, as the site of retroviral insertion is likely random (although it is possible that the site of insertion may be biased towards transcriptionally active DNA).

An alternative approach is to identify naturally occurring mutations within a population. Superficially, this might appear to be a daunting task. However, for some species, notably humans, this has already been done; the genetic diseases of connective tissue represent a bank of mutations in a variety of genes whose products are associated with either the structure or synthesis of connective tissue. Furthermore, although these mutations are likely to represent only a small number of the changes possible in

the connective tissue proteins or their synthetic or degradative pathways, they are likely to represent many of the changes in regions of functional importance (excluding those which may be lethal during early embryogenesis); changes which have little or no effect on function do not cause a disease.

The work presented in this thesis focusses on this approach to analyse functionally important structures of human type I collagen. In this "Introduction", therefore, the literature on the structure and biosynthesis of collagen is discussed, although not in great detail as the area has been summarised well in a number of recent reviews (Bornstein and Sage, 1980; Harwood, 1979; Bornstein and Traüb, 1979; Fessler and Fessler, 1978). Rather, this review considers, (1) the structure and expression of collagen genes, (2) the genetic diseases of collagen and (3) the specific aims of this project.

I.2. THE STRUCTURE AND BIOSYNTHESIS OF COLLAGEN.

Early ultrastructural analysis of skin and tendon collagen (Hall *et al.*, 1942) revealed a pattern of regular cross-striations, which, for many years was considered to be a defining characteristic of all collagen. However, as different classes have been discovered, it has been realised that the higher order structure which gives rise to the banded appearance under the E.M., *viz.*, a "quarter-stagger" of collagen molecules into long fibrils, is only found to any large degree for interstitial collagen; other collagen types have different higher order structures.

I.2.i. Structure of Types I, II and III.

Interstitial collagen molecules are rod-like structures composed of three peptide chains (see Table I.1.) called "α-chains", each of which is twisted approximately one complete left turn per three residues; the three chains are wound into a very stable right-handed helix. The long, central region [for example, 1014 amino acids per chain for calf type α 1(I) (see Hofmann et al., 1978)] has glycine as every third amino acid, and a large proportion of lysines and prolines which are post-translationally hydroxylated (see below). Short globular domains flank the helical region [for sixteen amino acids at the chick α 1(I) amino terminal (Hörlein et al., 1978) and either twenty-five (Fietzek et al., 1972; Rauterberg et al., 1972) or twenty-six (Fuller and Boedtker, 1981) at the chick α 1(I) carboxy terminal].

The α-chains are covalently cross-linked both to each other and to adjacent collagen molecules via reactive aldehydes enzymatically generated from peptidyl lysine and hydroxylysine residues (reviewed by Tanzer, 1973) and, for type III collagen, intramolecular (Chung and Miller, 1974) and intermolecular (Cheung et al., 1983) disulphide cross-links. These cross-links stabilise the triple helices into long, regular fibrils with the collagen molecules axially staggered with respect to one another by about 67 nm (Hodge and Schmitt, 1960) and a lateral packing which appears to involve 8 nm diameter microfibrils, perhaps composed of four molecules (Parry and Craig, 1979). The fibrils further aggregate into large fibres that are visible under the light microscope.

I.2.ii. Structure of Other Collagens

Although all collagens are composed predominantly of the unique triple helical structure described above, not all contain a single uninterrupted stretch of "glycine-X-Y" sequence. Other vertebrate and some invertebrate collagens have one or more globular domains interspersed throughout the helix, thereby giving rise to a range of different higher order structures.

Type IV

Schuppan et al. (1980), who partially sequenced the mouse α 1(IV) chain, found both large and short discontinuities in the triple helix, the latter of which resemble the single amino acid substitutions in the collagenous chains of Clq, known to result in a distinct bend in the protein (Ried, 1979). The α-chains are considerably larger than those found in interstitial collagen and it is possible that they are not proteolytically processed (see below).

Direct visualization of type IV collagen using rotary shadowing techniques (Timpl et al., 1981) has revealed a continuous four-armed matrix, although a minor species has five arms (Madri et al., 1983). This organisation accounts for the amorphous appearance of collagen in basement membranes.

Type V

Little is known about the structure of this type. The α-chains are of similar lengths to the interstitial types, but their amino acid composition suggests that they may contain non-collagenous domains, perhaps resembling those in type IV.

Type VI

Although not sequenced, this class resembles types I, II and III in that it probably consists of a continuous triple helix joining two globular domains (Odermatt et al., 1983). However, its constituent α-chains are approximately half the length of those in types I, II and III and its higher order structure is novel. Rotary shadowing electron microscopy (Furthmayr et al., 1983) has revealed a tetramer arranged as two dimers, each formed by lateral association of collagen molecules in anti-parallel fashion, with a 30 nm stagger, cross-linked at their ends to generate scissors-like structures. This tetramer is likely the basis of a filamentous form of type VI collagen.

Invertebrate Collagens

Collagen has been identified in a large range of invertebrate species (see Table 1 in Adams, 1978) by ultrastructural means. For species of most phyla, its structure appears to resemble vertebrate interstitial collagen, although a wide range of α-chain sizes is observed. There are, however, some notable differences; nematodes and some insects [e.g. Drosophila (De Biasi and Pilotto, 1976) but not the locust (Ashhurst and Bailey, 1980)] appear to lack striated fibrils.

The genes for one Drosphila (Monson et al., 1982) and two nematode (Kramer et al., 1982) collagens have recently been isolated. Sequencing has revealed discontinuities in their helical regions and so the higher order structures of these collagens may resemble vertebrate type IV.

I.2.iii. Biosynthesis of Collagen

Collagen α-chains are translated on ER-bound ribosomes as pre-pro-collagen peptides, the structures of which have been reviewed in detail by Miller and Gay (1982). The pre-pro leader sequence is proteolytically removed during traversal of the membrane by the peptide. Blobel and Dobberstein (1975) suggest that such processing events are coupled to translation, but Randall (1983) argues against this.

As the peptide enters the cisternae, a number of post-translational modifications occur (see below). Assembly of the collagen molecule occurs concomitantly with these modifications and probably involves alignment of the α-chains by the carboxy-propeptide.

A number of prolines are enzymatically hydroxylated by prolyl hydroxylase, which appears to recognise the β-turn conformation of nascent procollagen. The hydroxylation process results in a "straightening out" of this conformation into the linear triple helical form of native collagen (Chopra and Ananthanarayanan, 1982). It would seem that hydroxylation of prolines proceeds *pari passu* with helix formation and that as the helix forms, steric hindrance limits the number of prolines that can be modified.

Another enzyme, lysyl hydroxylase, is also active at this time; it converts some lysine residues to hydroxylysines. Some of these hydrosylysines are subsequently glycosylated, perhaps to prevent their oxidative deamination to aldehydes (Yamauchi *et al.*, 1982). This reaction of lysines and hydroxylysines, catalysed by lysyl oxidase

during the assembly of mature collagen molecules into fibrils generates reactive groups able to form cross-linking aldols and aldimines.

Mature, modified procollagen molecules are exocytosed *via* the Golgi. Specific extracellular proteases remove the propeptides, although there are conflicting data as to when this occurs. Davidson *et al.* (1975), and others since then, have shown that the amino-propeptide is removed before the carboxyl-propeptide, and that both are removed prior to, or concomitantly with, assembly into fibrils. However, using specific antibodies, Fleischmajer *et al.* (1983) have demonstrated the presence of amino-propeptides, but not carboxyl-propeptides, with a periodicity of approximately 60 nm, in newly formed embryonic types I and III fibrils. Small amounts of amino-propeptide were also detected in adult human skin type I (only in small fibrils) and type III (in 20-80 nm fibrils). These workers postulate that fibril formation involves the deposition of pN-collagen, (i.e. collagen with the amino-propeptide still attached) with the amino terminal protease perhaps exercising a role in regulating fibril growth.

The method by which fibrils self-assemble *in vivo* is not well understood.

I.3. COLLAGEN GENES AND THEIR EXPRESSION

The first, albeit indirect, analyses of collagen genes were undertaken using RNA populations physically enriched for procollagen sequences (Benveniste *et al*,. 1973; Frischauf *et al*., 1978). Isolation of such RNA was facilitated by the fact that although many cell types (even

those not generally associated with connective tissue, such as hepatocytes [Diegelmann et al., 1983; Saber et al., 1983]) are capable of synthesising collagen, some cell types, notably chick embryo calvaria (parietal and frontal bones), devote most of their protein synthesizing activity to the synthesis of type I procollagen (Boedtker et al., 1974).

Since these early experiments, molecular clones of the genes coding for a number of collagen types, from a number of species, have been constructued, although the genes and messengers for chicken type I, especially the pro α 2(I) chain, remain the best characterised.

I.3.i. Collagen mRNA

Collagen α-chains are translated from large (see Table I.2.) polyadenylated (Boedtker et al., 1974) mRNAs. The size heterogeneity observed, represents, for the human (Chu et al., 1982b; Myers et al., 1983) and chick (Aho et al., 1983) pro α 2(I) mRNA at least, multiple transcripts encoded from the same gene but varying in the length of their 3' untranslated regions. Similar size heterogeneity of 3' ends has been observed for the transcripts of other gene systems. Setzer et al (1982) have identified seven cytoplasmic, polyadenylated transcripts arising from the mouse dihydrofolate reductase gene, and Early et al (1980) have shown that an immunoglobulin μ gene can generate two messengers. Whilst it is clear that the variant antibody messengers have a functional role (they give rise to products which differ at their carboxy termini, thereby generating secreted or membrane bound forms [Rogers et al.,

TABLE I.2.

Size of mature collagen mRNAs, determined by northern blot analysis (Alwine *et al.*, 1977), coding for different ∝-chains and from different organisms.

(* represents "minor species.")

TRANSLATION PRODUCT (pre-pro∝-chain)	ORGANISM	SIZE IN BASES	REFERENCE
1 (I)	Chick	6400* 5600* 4900 7100* 5000	Rave et al. (1979), Adams et al. (1979).
2 (I)	Chick	5100 5700* 5200	Rave et al. (1979) Adams et al. (1979)
1 (II)	Chick	7000* 5300	Vuorio et al. (1982)
1 (III)	Chick	6000	Yamada et al. (1982a)
1 (I)	Human	7200 5900	Chu et al. (1982a)
2 (I)	Human	6200 5700 5500*	Chu et al. (1982a)
Cuticle or basement membrane type.	Drosophila	6400	Monson et al. (1982)
Cuticle type	C. elegans	1200	Kramer et al. (1982)

1980]), it is possible that both the dihydrofolate reductase and collagen polymorphism merely reflect a less than perfect fidelity of processing prior to polyadenylation. It is though that most 3' termini of polymerase II transcribed genes are generated by endonuclease cleavage, just 3' to the AAUAAA sequence, of a transcript which may have extended 1 kb or more further 3' [see Proudfoot, 1982].

Extensive sequence analysis of molecular clones made from chicken pro α 1(I) and pro α 2(I) mRNAs (Fuller and Boedtker, 1981; Tate et al., 1983) and human pro α1(I) (Chu et al., 1982a) and pro α 2(I) (Bernard et al., 1983) messengers has, in addition to providing protein sequence data (see Table I.3.), revealed some of the structural features of these RNA species (see Figure I.1.). They are, in most respects, typical eukaryotic messengers. However, the chick pro α 2(I) message, at least, has two AUG codons 5' to the one used for initiation of translation (which gives rise to the pro α-chain). Whether or not short peptides are ever translated from these upstream AUGs is not known, although Vogeli et al (1981) have postulated that they are unavailable for translation initiation due to their involvement in secondary structure.

Analysis of codon usage for both chicken and human α-chain domains indicated a strong third base preference for U and C in codons for glycine, proline and alanine, although a similar preference for codon usage was not seen for the propeptide domains.

I.3.ii. Collagen Gene Structure

The existence of distinct classes of collagen,

TABLE I.3.

Regions of collagen for which the amino acid sequence has been deduced from DNA sequencing analysis of cDNA clones.

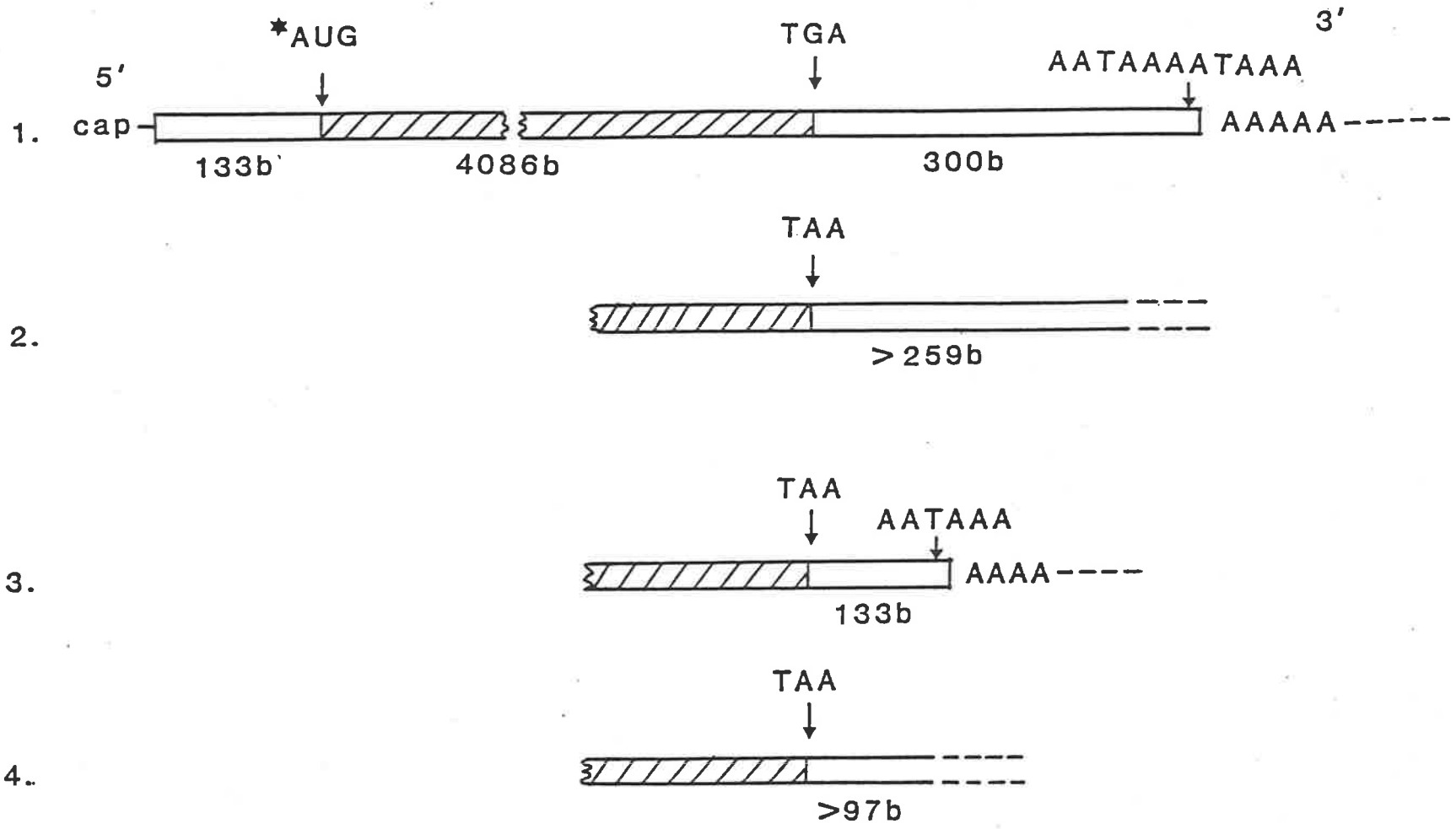
CHAIN	ORGANISM	SPAN OF DEDUCED SEQUENCE (NUMBER OF AMINO ACID RESIDUES)	REFERENCE
α 1(I)	Chick	Residue 813 - TAA codon (473 residues)	Fuller and Boedtker (1981)
α 2(I)	Chick	Residues 1 - 89	Tate *et al.* (1983)
		Residue 813 - TGA codon (360 residues)	Fuller and Boedtker (1981)
α 1(I)	Human	Residues 247 - 312 787 - 861	Chu *et al.* (1982a)
		Final 12 residues at the carboxyl terminal.	
α 2(I)	Human	Residue 533 - TAA codon (740 residues)	Bernard *et al.* (1983)

Figure I.1

Structure of collagen mRNAs, determined from sequence analysis of cDNA clones:

1. Chick α 2(1); Fuller and Boedtker, 1981; Tate *et al.*, 1983.
2. Human α 2(1); Bernard *et al.*, 1983.
3. Chick α 1(1); Fuller and Boedtker, 1981.
4. Human α 1(1); Chu *et al.*, 1982a.

* See text.


*AUG
TGA
3'
5'
AATAAAATAAA
1.
cap
AAAAA-----
133b
4086b
300b
TAA
2.
>259b
TAA
AATAAA
3.
AAAA----
133b
TAA
4.
>97b

composed of α-chains whose primary sequence is unique to that class, suggests that the collagens are encoded by a multigene family.

In vertebrates, this family must contain at least fifteen members to account for the known α-chains (Solomon, 1980). However, it is likely that this estimate of the family size will need to be revised to account for:

(a) The genes of variants of already characterised types, such as the variant bovine α 1(I) chain identified by Yamauchi et al. (1982), although whether or not this represents a new gene, as these workers suggest, or is merely an allelic variant, such as that found by Driesel et al. (1982), is open to question.

(b) Variants such as the human α 1(I)-like gene cloned by Weiss et al. (1982), shown by Chu et al. (1982a) not to be the α 1(I) gene predominantly expressed in normal human fibroblasts, and possibly a pseudogene.

(c) The genes coding for as yet uncharacterised variant collagen types.

Southern blot analysis indicates that the nematode genome also has a family of collagen genes, with fifty or more members (Kramer et al., 1982).

Very little is known about the gross organisation of the collagen gene family. The unique human α 2(I) gene (Dalgleish et al., 1982) has been unambiguously localised to the region 7pter → 7q22 (Solomon et al., 1983), suggesting that the mapping of human type I procollagen sequences to chromosome 17 by Church et al. (1980) might have been of the α 1(I) gene. Whether or not any of the collagen genes exist in a clustered array, as is common for other gene

families, such as the chicken histone genes (Harvey et al., 1981) is not known.

In contrast to the small amount known about the genomic organisation, highly detailed analyses of the structure of some collagen genes, especially the chick α 2(I) gene, have been made.

Vertebrate Collagen Genes

Vogeli et al. (1980), Ohkubo et al. (1980) and Wozney et al. (1981a) have isolated overlapping genomic clones containing the entire chick α 2(I) gene. It is large (38 kb), it contains many (at least forty-nine) introns and it has a highly ordered pattern of exons. Analyses of the 3' halves (i.e. encoding the carboxyl halves of the proteins) of the sheep (Boyd et al., 1980; Schafer et al., 1980) and human (Dalgleish et al., 1982) α 2(I) genes, the mouse α 1(I) gene (Monson and McCarthy, 1981), the chick α 1(III) gene (Yamada et al., 1983a) and an entire human α 1(I)-like gene (Weiss et al., 1982) have revealed that these features are typical of vertebrate collagen genes.

Whilst the chick α 2(I) gene contains the greatest number of introns so far identified in a single gene, some other eukaryotic genes also are highly interrupted [for example, the 21 kb Xenopus vitellogenin A1 gene contains 33 introns (Wahli et al., 1980)]. The pattern of this interruption in collagen genes (described below) is, however, somewhat unique and may reflect the evolutionary origins of these genes.

Initial R-loop analysis (Schafer et al., 1980; Wozney et al., 1981b) revealed that the exons coding for the

helical region of the protein were small (approximately 100 bp or less). The introns ranged in size from about 100 bp to about 3 kb. Detailed sequence analysis of the chick α 2(I) gene (Yamada et al., 1980; Dickson et al., 1981; Wozney et al., 1981a) revealed that, of the sixteen exons spanning the helical region sequenced, eight were 54 bp long, two were 108 bp (54 x 2), four were 99 bp (108 - 9) and two were 45 bp (54 - 9) long. Furthermore, ten adjacent exons spanning amino acid residues 175 to 411 were found to alternate in size between a large (99 bp or 108 bp) and a small (45 bp or 54 bp) exon.

Exons spanning the region coding for both the amino and carboxyl globular domains of collagen are, however, not simply based in size upon 54 bp ± multiples of 9 bp. Wozney et al (1981b) showed exons 1 to 4 (numbered from the 3' end) to be 444 bp, 243 bp, 189 bp and 249 bp respectively (although Dickson et al. (1981) found exons 3 and 4 to be 191 and 247 bp respectively) and Tate et al. (1983) found exons 49 to 45 (numbered from the 3' end) to be 203 bp, 11 bp, 18 bp, 36 bp and 84 bp respectively.

Comparison of the four exons encoding the carboxyl termini of chick α 2(I) and α 1(III) collagens (Yamada et al., 1983b) has revealed similarities both in exon size and nucleotide sequence. If one deducts the sequence encoding the 3' untranslated region, exons 1 and 2 of α 2(I) and α 1(III) are identical in length and exons 3 differ in length by only 1 (or 3) nucleotide(s). Furthermore, in the middle of exon 2, 48 out of 49 nucleotides are identical between α 2(I) and α 1(III) and 44 out of 48 are identical between α 1(I) and α 1(III). Similar degrees of homology have been observed in the same region in other chick and

human collagen genes (Tolstoshev and Solomon, 1983). Although this sequence encodes the unique carbohydrate attachment site of the carboxyl propeptide and so presumably requires a conserved amino acid sequence, it is not clear why a conserved nucleotide sequence is also required. It is possible that this sequence is the DNA or RNA binding site for an, as yet unidentified, regulatory or processing molecule.

Myers et al. (1982) have identified two blocks of middle repetitive sequence in both the human α 1(I) and α 2(I) genes, located in the 3' untranslated region and the large 3' intron. Although the sequences are not conserved between these two genes, their positions are conserved. These workers propose (although they provide no evidence to support this proposition) that these sequences are hotspots of recombination.

The promoter elements of the chick α 2(I) gene (Vogeli et al. 1981) are typical of those found 5' to eukaryotic genes transcribed by RNA polymerase II. the TATA box, 33 bp 5' to the adenine mapped as the CAP site (Merlino et al., 1981 and 1982) is identical in sequence and position to the consensus sequence (Breathnach and Chambon, 1981). The CAT box (GCCCATTGC) exhibits some differences from the consensus sequence (GG$\frac{T}{C}$CAATCT) , but its distance 5' from the CAP site (79 bp) is typical.

The sequence spanning 160 bp of the promoter 5' to the CAP site contains three regions of dyad symmetry. These potentially are able to give rise to three mutually exclusive stem and loop structures, none of which encompass the TATA box, but all of which involve, to varying degrees,

the CAT sequence. It is not known whether such structures are able to form *in vivo* (Courey and Wang, 1983), but if they do, it is conceivable that they may play a role in the regulation of expression of this gene.

Invertebrate Collagen Genes

In contrast to the vertebrates, the invertebrate collagen genes that have been examined, *viz.*, two nematode cuticle genes (Kramer *et al.*, 1982) and a *Drosophila* cuticle or basement membrane gene (Monson *et al.*, 1982), have a simple and compact structure.

Of the 1.5 kb, of the 9.2 kb *Drosophila* clone sequenced, only 81 bp are occupied by intron sequences. the remaining coding region is organised into two large, uninterrupted blocks. The total size of the gene is not known, nor are any data regarding promoter structure available.

The nematode genes, *col*-1 and *col*-2, span from the predicted CAP site to the AATAAA sequence, approximately 1295 bp and either 1009 or 1323 bp (i.e. there are two AATAAA sequences) respectively. *Col*-1 contains two introns (102 bp and 52 bp long); *col*-2 is interrupted by a single 47 bp intron. Sequences resembling TATA and CAT boxes are apparently present 60 bp and 100 bp 5' to the *col*-1 AUG respectively (although no data are available).

I.3.iii. EVOLUTION OF COLLAGEN GENES

Yamada *et al.* (1980) have suggested that the primordial collagen gene was a 54 bp long sequence coding for a functional domain (Gilbert, 1978) which, by genetic means, was duplicated into a tandem array, with each 54 bp

unit separated by a block of DNA whose transcript was able to be recognised by a splicing mechanism as an intron. They suggest that large collagen genes were generated by recombination within introns. Other functional regions of collagen, the telo- and pro-peptides were added by recombination between DNA flanking their gene domains and DNA flanking the primordial collagen gene. In addition, they suggest that the collagenous sequences in C1q and acetylcholinesterase may have been derived from this primitive 54 bp unit.

Comparison of the nucleotide sequences of eight mouse α 1(I) exons (Monson and McCarthy, 1981), five of which are 54 bp and three 108 bp, has revealed a considerable degree of homology. In four of the ten possible pairwise comparisons of the 54 bp exons, the homology was direct, suggesting that they may have arisen by a duplication event. In three cases, maximum homology was observed when the sequences were staggered by 10 bp. Homologous recombination between these pairs would generate 45 bp and 63 bp long sequences, the former of which have been identified in the chick α 2(I) gene. Other staggered homologies were observed, notably a 76% homology between a 108 bp exon and a 54 bp exon at a 46 bp stagger. The 108 bp exon may have arisen by homologous recombination between a 63 bp exon and a 54 bp exon staggered by 46 bp (see Figure 8 in Monson and McCarthy, 1981). It is not clear how the alternating pattern of large and small exons, seen in the chick α 2(I) gene, may have arisen.

Thus, it seems possible that the early collagen gene may have arisen by a combination of a duplication of a

54 bp unit, fusion of other domains (coding for non-helical regions) and homologous recombination between exons, the latter process being limited by the strict requirement to maintain a regular protein structure (9 bp encodes one Gly-X-Y triplet). Additional restraints on recombination or random mutation by the necessity to maintain specific nucleotide sequence in various regions (for example, the conserved region encoding the carboxyl propeptide carbohydrate attachment site) have likely also played a role.

Although evident in all vertebrate collagen genes examined to date, sequences related to a 54 bp structure have not been found in invertebrate collagen genes. This may reflect an independent origin for the invertebrate collagens. Alternatively, it may indicate that the vertebrate collagens have been subjected to evolutionary constraints different from the invertebrate collagens.

I.3.iv. EXPRESSION OF COLLAGEN GENES

Analysis of the steady state levels of type I procollagen messengers in RNA isolated from whole organs and cultured cells by translation (Rowe *et al.*, 1978; Moen *et al.*, 1979), northern blot (Adams *et al.*, 1979) and liquid hybridization (Parker and Fitschen, 1980) assays has revealed a strong correlation between rates of procollagen synthesis and the level of specific messenger present. The reduction in the rate of type I procollagen synthesis induced by either Rous sarcoma virus (RSV) (Sandemyer and Bornstein, 1979; Adams *et al.*, 1979) or SV40 (Parker *et al.*, 1982) transformation of cultured fibroblasts was found

also to be directly correlated with the loss of specific mRNAs. Furthermore, Avvedimento et al. (1981), using intron specific probes, showed a decrease in the level of the nuclear RNA precursors to α 2(I) collagen in RSV transformed fibroblasts.

These results suggest that the rate of transcription is the major determinant regulating the rate of type I procollagen synthesis. However, analysis of procollagen synthesis in both cultured foetal lung fibroblasts (Tolstoshev et al., 1981a) and foetal sheep skin (Toltoshev et al., 1981b) has shown that both the rate of intracellular procollagen degradation and the efficiency of utilisation of procollagen mRNA play an important role in regulating collagen production.

The efficiency of translation of collagen mRNA may involve a negative feedback step. Wiestner et al. (1979) have demonstrated that the type I procollagen amino-terminal extension peptide, p_N α 1(I) - Col-1, can specifically inhibit type I collagen biosynthesis in cultured fibroblasts. This inhibition occurs at the level of translation (Paglia et al., 1979 and 1981); the peptide acts to inhibit either polypeptide chain elongation or termination (Hörlein et al., 1981). It is an intriguing possibility that the p_N α 1(I)-Col-1 peptide may exert its effect by destabilising secondary structure (Vogeli et al., 1981) at the 5' end of the mRNA, thereby allowing one of the alternative AUG codons to be used for translation initiation. The result of this would be the production of either an hexapeptide or a tetrapeptide, rather than the normal collagen propeptide. Whether or not propeptide

mediated inhibition of translation occurs naturally *in vivo* is not known.

Although production of collagen protein is stable and rigidly controlled under conditions of constant environment (Breul *et al.*, 1980) a variety of external factors, including prathyroid hormone (Kream *et al.*, 1980) ascorbic acid (Rowe and Schwarz, 1983) and 5-bromo-2-deoxyuridine (Pawlowski *et al.*, 1981) cause substantial changes to collagen mRNA levels and hence collagen biosynthesis. Chondrocytes, which normally produce type II collagen, can be reversibly induced, by growing them in a monolayer culture, to switch to type I synthesis (Benya and Shaffer, 1982). Surprisingly, there is a lag of several days between the appearance of α 1(I) mRNA and α 2(I) mRNA during the switch to type I synthesis (Duchene *et al.*, 1982) indicating that the mechanism(s) which normally regulates co-ordinate expression of these genes at a pro α 1(I):pro α 2(I) ratio of 2:1 (Vuust, 1975; Vuust *et al.*, 1983) is not always rigidly controlled.

A number of approaches have been used to investigate collagen gene expression during early embryogenesis. Ultrastructural analysis (Trelstad *et al.*, 1967) has revealed that type IV collagen is present in the chick embryo as early as gastrulation, and immunofluorescent studies (von der Mark *et al.*, 1976) have identified type II collagen at stage 15 and type I collagen at stage 17 of the chick embryo. Merlino *et al.* (1983) detected both type I collagen mRNAs in 2-day chick embryos and found DNase I hypersensitivity sites upstream from the α 2(I) gene in 2-day embryo DNA (the appearance of DNase I hypersensitivity

sites correlates with the potential of that gene to be transcriptionally active [Elgin, 1981]). Expression of collagen genes in mice, on the other hand, appears to occur somewhat later; the lack of functional α 1(I) genes is not lethal until day 13 of embryonic life (Schneike et al., 1983).

In nematodes, collagen mRNA is undetectable by cytological hybridization until the 100 cell stage (Edwards and Wood, 1983).

I.4. GENETIC DISEASES OF COLLAGEN

There is a considerable amount of evidence (for example, see Eyre, 1981) to implicate aberrant collagen as being the underlying defect in a number of the rare, heritable disorders of connective tissue of both man and other mammals.

Traditionally, these diseases have been grouped according to the nature of the clinical symptoms that they produce. Whilst such a classification is useful for the purposes of treating patients, it can be misleading for the purpose of identifying the lesion. For example, investigation of collagen synthesis in cells from two patients exhibiting the same symptoms of Ehlers-Danlos syndrome IV (EDS IV) revealed that one failed to make any type III collagen (Gay et al., 1976) whereas the other synthesised a normal amount of type III but secreted only a small amount (Byers et al., 1981a). Clearly, the site of the lesion in these two patients is very different.

As the molecular nature of some of the collagenopathies has been revealed, it has become possible to group

these diseases into three general classes, depending on the site of the defect, viz.,

(1) Primary, involving the sequence, transcription or translation of collagen genes.

(2) Secondary, involving the post-translational modification, assembly or secretion of collagen peptides.

(3) Tertiary, involving interactions between collagen and non-collagen elements of connective tissue.

I.4.i. PRIMARY COLLAGENOPATHIES

To date, only a small number of abnormal collagens resulting from simple mutations (substitutions, deletions or insertions of one or more bases) have been described, although a number of cases exist where such mutations are implicated.

Insertions/Deletions

Perhaps the best characterised coding region mutation is one which has given rise to the symptoms of osteogenesis imperfecta, type II (OI II). Although there is evidence implicating either a decreased concentration or decreased translation efficiency of pro α 1(I) mRNA (Penttinen et al., 1975; Steinmann et al., 1979), Barsh and Byers (1981) have shown there to be two electrophoretically separable pro α 1(I) chains. One of the α 1(I) alleles in this patient has recently been shown to contain a 500 bp deletion (Chu et al., 1983).

Byers et al. (1981b) have analysed both isolated protein (synthesised in vivo) and translation products (in vitro) from a Marfan syndrome patient and have found an

increase in the molecular weight of half of the α 2(I) chains, corresponding to an insertion of approximately twenty amino acids. This insertion was located, by peptide mapping, in the helical region, towards the carboxyl terminal of the protein.

The consequence of such an altered α 2 chain is likely a reduction of cross-linking, resulting in the observed increased solubility in non-denaturing solvents of this patient's skin collagen (Siegel and Chang, 1978).

It is interesting to speculate that such an insertion might have arisen as a duplication, during recombination, of a 54 bp exon in one α 2(1) allele; sequence analysis of this allele should enable accurate identification of the nature of the lesion.

The absence of α 2(1) chains in collagen secreted by the fibroblasts of an OI I patient (Nicholls et al., 1979) appears also to involve a deletion. This deletion has been mapped to the 3' ends of both pro α 2(1) alleles (Pihlajamiena, cited in Sykes, 1983) and presumably results in peptides unable to be incorporated into the triple helix. In addition, it has been reported (Deak et al., 1982) that pro α 2(1) mRNA from this patient is inefficiently translated.

Point Mutations

Steinmann et al. (1980) have found that half of the α 2(1) chains of a patient with Ehlers-Danlos syndrome, type VII, retain their amino-terminal propeptides. Since the level of amino-terminal propeptidase was found to be normal, it seems likely that the defect resulted from a mutation at the region encoding the peptidase cleavage site

of one pro α 2(1) allele.

The presence of an unusual, disulphide bonded cyanogen bromide peptide from a patient with OI II (Steinmann, cited in Sykes, 1983) implicates a glycine to cysteine substitution in a pro α 1(1) allele.

Expression Mutations

Francis et al. (1981) have demonstrated an increased ratio of α 1(III):α 1(I) in skin biopsy from eighteen patients with OI I. It is postulated (Penttinen et al., 1975) that this results from a decreased rate of type I synthesis rather than either an increased rate of type III synthesis or an increased rate of type I degradation.

de Wet et al. (1982) have observed a decreased rate of α 2(I) in three patients with OI II and Barsh et al. (1982) have examined three OI I patients who synthesised only half the normal amount of pro α 1(I).

It is likely that reduced synthesis of collagen chains may result from the same types of lesions that give rise to the thalassaemias.

The most common thalassaemias are the β^+ (reduced levels of β-globin) and β° (total absence of β-globin) types.

The β^+ thalassaemias thus far studied at the molecular level appear to result from either aberrant processing of pre-mRNA (Spritz et al., 1981; Westaway and Williamson, 1981; Spence et al., 1982) or decreased stability of the mRNA (Nienhuis et al., 1977).

A number of different lesions appear to cause the β° diseases, including:-

(a) abnormal processing of pre-mRNA, as indicated by

the presence of nuclear β-globin RNA, but the absence of cytoplasmic β-globin sequences (Comi et al., 1977) caused, in one individual at least, by a mutation at the 5' splice junction of the large intron IVS2 (Baird et al., 1981).

(b) failure to translate apparently normal mRNA. Specific lesions include the presence of suppressable nonsense mutations in the coding region (Chang et al., 1979) and an apparently defective initiation codon in one individual (Old et al., 1978).

(c) deletions in non-coding regions of the RNA (Flavell et al., 1979). There is also evidence that deletions in the DNA can act in cis over long distances (Fritsch et al., 1979).

The less common α-thalassaemias generally appear to have arisen as the result of deletions of whole genes.

I.4.ii. SECONDARY COLLAGENOPATHIES

Of the inherited collagen diseases thus far investigated, those diseases resulting from the reduced, or absence of, activity of processing enzymes constitute the majority.

Enzymes whose aberrant activity has resulted in collagen disease include lysyl hydroxylase (Krane et al., 1972), lysyl oxidase (Byers et al., 1980), amino-terminal propeptidase (Lichtenstein et al., 1973), cystathionine synthetase (Kang and Trelstad, 1973), homogentisic acid oxidase (Murray et al., 1979) and an unidentified, intracellular, copper-binding protein (Goka et al., 1976).

It is likely that a range of lesions will be identified in these aberrant enzymes. Quinn and Krane

(1976) have shown that in some cases of EDS VI, the defect in lysyl hydroxylase is not at its catalytic active site but at its cofactor binding site.

I.4.iii. TERTIARY COLLAGENOPATHIES

Primary and secondary defects of collagen are manifested by their effects on connective tissue. In some cases, these lesions also affect the non-collagenous components. For example, lysyl oxidase is required for cross-link formation in elastin, and so deficiency of this enzyme will result in faulty elastin as well as faulty collagen (Di Ferrante et al., 1975).

Some primary defects of non-collagenous connective tissue components have been investigated. Spondyloepiphyseal dysplasia patients appear to have abnormal proteoglycan metabolism (Byers et al., 1978). Both dominant and recessive forms of epidermolysis bullosa have been shown to result from elevated levels of skin collagenase (Bauer et al., 1977). Other disorders in which aberrant collagenase have been implicated exist, but it seems likely that the involvement of this enzyme may be a secondary effect.

I.5. AIMS OF THE PROJECT

Analysis of both normal and aberrant collagen genes and their transcripts at the molecular level requires the use of gene probes of high purity and specificity. Recombinant DNA technology enables the construction of such probes.

At the time that the work described in this thesis was initiated, no human collagen sequences had been

converted to recombinant form. The aim of this project, therefore, was to clone human collagen gene sequences which could subsequently be used as probes for the analysis of both collagen gene expression and the nature of the lesions in the naturally occuring mutations. In addition, DNA sequence analysis would enable the primary structure of the human α-chains to be elucidated.

This thesis describes the molecular cloning and characterisation of sequences encoding human type I collagens.

CHAPTER II

MATERIALS AND METHODS

II.1. ABBREVIATIONS

Abbreviations were as described in "Instructions to Authors" (1978). In addition:

BCIG	5-bromo-4-chloro-3-indolyl-β-D-galactoside
bis	N,N'-methyl-bisacrylamide
BSA	bovine serum albumin
DMEM	Dulbecco's modified Eagle's medium
DMSO	dimethyl sulphoxide
DTT	dithiothreitol
FCS	foetal calf serum
HEPES	N-2-hydroxyethylpiperazine-N'-2-ethane-sulphonic acid
IPTG	isopropyl-β-D-thio-galactopyranoside
PEG	polyethylene glycol
SDS	sodium dodecyl sulphate

II.2. MATERIALS

II.2.i. General Reagents and Materials

Reagents used were of technical grade, or higher,

purity. Most chemicals and materials were obtained from a range of suppliers, although the source of supply was found to be critical for some including:-

Material	Source
Low Gelling Temperature Agarose	B.R.L. Inc., Gaithersburg, MD., U.S.A.
Nitrocellulose Filters, BA85	Schleicher and Schüll GmbH, Dassel, F.D.R.
Nuclease Free Sucrose	Schwarz/Mann, Orangeburg, N.Y., U.S.A.
Oligo-dT-cellulose, Type III	Collaborative Research, Waltham, MA., U.S.A.
Optical Grade Caesium Chloride	Harshaw Chem. Co., Solon, OH, U.S.A. **or** Metallgesellschaft AG, Frankfurt, F.D.R.
PEG 6000	BDH Ltd., Port Fairy, Australia.
Trypsin 1:250	Difco Laboratories, Detroit, MI., U.S.A.

The following reagents were gifts:

Material	Source
Chloramphenicol	Parke-Davis Pty. Ltd., Sydney, Australia.
^{32}P-dNTPs	R.H. Symons
Tetracycline	Commonwealth Serum Laboratories, Melbourne, Australia

II.2.ii. ENZYMES

Enzymes were obtained from the following sources:

Enzyme	Source
AMV RNA-dependent DNA-polymerase (reverse transcriptase)	Gift from J.W. Beard and the N.I.H. Cancer Program.
Calf intestinal phosphatase	Sigma Chem. Co., St. Louis, MO., U.S.A.
E. coli DNA-polymerase I	Boehringer Mannheim GmbH, F.D.R. BRESA, Adelaide, Australia.
E. coli DNA-polymerase I, Klenow Fragment	Boehringer Mannheim, BRESA.

E. coli DN'ase I	Sigma Chemical Company.
Exonuclease Bal-31	B.R.L. Inc.
Pancreatic ribonuclease	Sigma Chem. Co.
Polynucleotide kinase	Boehringer Mannheim.
Proteinase K	Boehringer Mannheim.
Restriction endonucleases	New England Biolabs Inc. Beverly, MA., U.S.A. Boehringer Mannheim.
S_1 nuclease	Boehringer Mannheim.
T_4 DNA ligase	Boehringer Mannheim, BRESA.
Terminal deoxynucleotidyl transferase	P-L Biochemicals Inc. Milwaukee, WIS., U.S.A.

II.2.iii. Biological Reagents

Bacterial Strains

LE392: *E. coli* F^-, *hsd* R 514 (r_k^-, m_k^-), *sup* E 44, *sup* F 58, *lac* Y 1, *gal* K 2, *gal* T 22, *met* B 1, *trp* R 55, λ^-. Gift from J.B. Egan.

JM101: *E. coli lac*, *pro*, *sup* E, *thi*, F^1 *trad* D 36, *pro*

AB, *lac* I^q, Z Δ M15. Gift from A.J. Robins.

MC1061: *E. coli ara* D *139*, Δ (*ara*, *leu*) 7697, Δ *lac* X74, *gal* U^-, *gal* K^-, *hsr*$^-$, *hsm*$^+$, *str* A. Gift from R.P. Harvey.

Human Genomic Library was constructed by Lawn, R.M., Fritsch, E.F., Parker, R.C., Blake, G. and T. Maniatis, and supplied by T. Maniatis.

Sheep Collagen Genomal Clone (SpC3) was kindly supplied by P. Tolstoshev.

II.3. METHODS

II.3.i. *Culture of Human Fibroblasts*

(Magee and Moore, 1984)

All procedures involving cultured cells were performed aseptically.

Media

Two different media were routinely used, *viz.*,

(a) RPMI 1640 containing 2 mM glutamine, 24 mM sodium bicarbonate and 50 μg (50 U)/ml gentamicin.

(b) DMEM containing 4 mM glutamine, 5.5. mM glucose, 1 mM sodium pyruvate, 44 mM sodium bicarbonate, 10 mM HEPES and 50 μg (50 U)/ml gentamicin.

Both media were prepared with organic free, reagent grade water and were filter sterilized.

Growth and harvesting

Human foreskin fibroblasts were obtained approximately eight generations after establishment of the primary culture. They were grown in either of the above media,

plus 10% FCS (non-inactivated), at 37°C either in sealed bottles or in vented flasks in a humid atmosphere containing 5% CO_2. Culture vessels ranged in size from 25 cm^2 to 400 cm^2 (roller bottles).

Cells generally were split two days after attaining dense confluency, although this time was sometimes extended to a week or more. The split ratio was 1:2.

Fibroblasts were harvested for splitting by first washing the monolayer twice for five minutes at 21°C with PBS (137 mM NaCl, 3 mM KCl, 1.5 mM KH_2PO_4, 8 mM Na_2HPO_4) and then incubating at 37°C with a sufficient volume of PBS, containing 0.1% trypsin, 0.6 mM Na_2EDTA, to cover the cells. When the monolayer was observed to begin to lift, the cells were detached by sharply striking the side of the culture flask and then resuspended by trituration in either growth medium (for splitting into fresh flasks) or PBS (for harvesting or storage). Those cells harvested into PBS were washed by gentle centrifugation (400 g).

Storage

Washed cells were resuspended in DMEM, minus HEPES, containing 20% FCS and 10% DMSO, at a concentration of 4 x 10^6 cells/ml, and heat sealed in 2 ml plastic ampoules. The ampoules were packed into cardboard tubes and left overnight on the top shelf of a -80°C cabinet. The next day the ampoules were placed in liquid nitrogen.

Cells were recovered by rapid thawing at 37°C, followed by dilution into growth medium and plating into a suitably sized culture vessel.

II.3.ii. Isolation of Human Genomic DNA

Isolation of nuclei

(a) from cultured fibroblasts

Cells were harvested, as described above, washed twice in ice-cold PBS and resuspended in 5 ml/10^7 cells of ice-cold 0.5% Triton DF-16 in TE (10 mM Tris-Cl pH 8.0, 1 mM EDTA) by vigorous vortexing. Nuclei were recovered by centrifugation and resuspended in a minimum volume of TE.

(b) from human placenta

(Marshall and Burgoyne, 1976)

Fresh placenta was finely chopped and approximately 10 g were quickly homogenised in 20 ml of 11.6% sucrose in homogenisation buffer (2 mM EDTA, 0.5 mM EGTA, 60 mM KCl, 15 mM β-mercaptoethanol, 15 mM NaCl, 0.15 mM spermine, 0.5 mM spermidine, 15 mM Tris-Cl pH 7.4).

The homogenate was filtered through muslin and the nuclei pelleted by centrifugation (16,000 g, 20 minutes, 4°C) through a 10 ml pad of 47% sucrose in homogenisation buffer. The supernatant was aspirated and the crude nuclear pellet resuspended in a minimum volume of TE.

Preparation of high molecular weight DNA

(Gross-Bellard et al., 1973)

Nuclei were added dropwise to 10 volumes of gently stirring 10 mM NaCl, 10 mM EDTA, 10 mM Tris-Cl pH 8.0, 0.5% SDS, containing 100 µg/ml proteinase K and incubated at 37°C for four hours. Liberated DNA was gently phenol extracted three times and low molecular weight contaminants removed by extensive dialysis against TE. RNA was removed by incubating at 37°C for four hours with 20 µg/ml pancreatic RN'ase (previously heated to 80°C for 20 minutes

to destroy DN'ase activity) followed by phenol extraction and extensive dialysis against TE.

DNA was stored at 4°C.

II.3.iii. Preparation of RNA

All procedures involving RNA were carried out at 4°c using sterile solutions and glassware.

Isolation of RNA from chick embryo calvaria

Method 1 (Seeburg et al., 1977)

Calvaria (parietal and frontal bones) were removed from fifty 16 day-old chick embryos and snap-frozen in liquid nitrogen. Frozen tissue was homogenised in warm 7 M guanidinium-HCl, 20 mM Tris-Cl pH 7.5, 1 mM EDTA, 1% sarkosyl NL-97 in a Dounce homogeniser in a final volume of 20 ml, and RNA pelleted by centrifugation through a 5.7 M CsCl pad as described (Seeburg et al., 1977).

The clear RNA pellets were resuspended in TE containing 5% sarkosyl and 5% phenol, adjusted to 200 mM NaCl and extracted with an equal volume of phenol/chloroform. RNA was precipitated from the aqueous phase by the addition of 2.5 volumes of ethanol and recovered by centrifugation.

RNA was stored at -80°C as an ethanol precipitate.

Method 2 (Chirgwin et al., 1979; J. Brooker, pers. comm.)

Calvaria from fifty 16 day-old embryos were homogenised in a Dounce homogeniser in 7 ml of 6 M guanidinium-HCl, 200 mM Na-acetate pH 5.2 and 1 mM β-mercaptoethanol and the RNA precipitated by the addition of $\frac{1}{4}$ volume of

1 M acetic acid, 1/10 volume of 2 M K-acetate and an equal volume of ethanol. The precipitate was recovered by centrifugation (16,000 g, 15 minutes, 0°C) and resuspended in half the original volume of 6 M guanidinium-HCl, 200 mM Na-acetate pH 5.2, 10 mM EDTA. RNA was re-precipitated with an equal volume of ethanol, recovered, resuspended and re-precipitated using the same procedure.

After recovery, the pellet was dissolved in 10 ml of 25 mM EDTA and extracted with an equal volume of phenol/chloroform. Two volumes of 4.5 M K-acetate pH 6.0 were added to the aqueous phase and the RNA left to precipitate overnight at -20°C. The precipitate was pelleted by centrifugation (16,000 g, 15 minutes, 0°C) and any contaminating DNA and low molecular weight RNA dissolved in 2 M LiCl. High molecular weight RNA was centrifuged out of the LiCl solution, dissolved in water, adjusted to 200 mM NaCl and ethanol precipitated.

RNA was stored at -80°C as an ethanol precipitate.

Isolation of RNA from cultured human fibroblasts

Fibroblasts were lysed in Triton/TE and the nuclei removed from the lysate as described above (Section II.3.ii.). The lysate was extracted with an equal volume of phenol/chloroform, the aqueous phase adjusted to 200 mM NaCl and the RNA precipitated by the addition of 2.5 volumes of ethanol.

RNA was stored as an ethanol precipitate at -80°C.

II.3.iv. Preparative Fractionation of RNA

Size fractionation

RNA was resuspended in 10 mM Tris-Cl pH 7.5, 1 mM

EDTA, 0.5% SDS, heated at 65°C for 5 minutes, snap chilled on ice and centrifuged (180,000 g, 16 hours, 4°C) on 10% - 40% linear sucrose in 10 mM Tris-Cl pH 7.5, 1 mM EDTA gradients. RNA was identified by its absorbance at 254 nm and the desired fractions collected, adjusted to 200 mM NaCl and ethanol precipitated.

Selection of poly(A)$^+$ RNA (Aviv and Leder, 1972)

0.2 g of oligo-dT-cellulose were suspended in several ml of water and poured into a small column. After flushing with 100 mM NaOH, the column was equilibrated with high-salt buffer (0.5 M NaCl, 20 mM Tris-Cl pH 7.5, 1 mM EDTA, 0.1% SDS).

RNA was dissolved in water, heated at 80°C for 3 minutes then snap-chilled on ice, adjusted to high-salt buffer conditions and applied to the column. Elution of unbound RNA with high-salt buffer was monitored by its absorbance at 254 nm. Unbound RNA was re-applied and re-eluted.

Bound [poly(A)$^+$] RNA was eluted with 10 mM Tris-Cl pH 7.5, 1 mM EDTA, 0.05% SDS and ethanol precipitated.

Chromatography was performed at 25°C.

II.3.v. Restriction Enzyme Digestion and Analysis of DNA

Restriction digestions

Restriction endonuclease digestion of DNA was performed using the conditions for each enzyme described by Davis et al. (1980). ATP (100 µM) was also included when the restricted DNA was to be ligated. A two-fold excess of enzyme generally was used and the reactions were run for an hour, although this time was increased to up to eight hours

for preparative digestions.

Reactions were terminated either by the addition of EDTA to 5 mM, followed by phenol/chloroform extraction and ethanol precipitation, or the addition of half a volume of urea load buffer (4 M urea, 50 mM EDTA, 0.1% bromophenol blue, 50% sucrose).

Agarose gel electrophoresis

Analytical

Agarose (0.7% - 2%) was dissolved in TEA (40 mM Tris-acetate, 20 mM Na-acetate, 1 mM EDTA, pH 8.2) and cast either in 14 cm x 14 cm x 0.3 cm horizontal slab-gel templates or on to 7.5 cm x 5 cm microscope slides, for horizontal gels.

Vertical gels were electrophoresed between tanks containing TEA at 65 mA for approximately three hours. Horizontal gels were run submerged in TEA at 150 mA for approximately 15 minutes.

DNA was visualised by staining with 10 µg/ml ethidium bromide for 5 minutes and examination under UV light.

Preparative

Low gelling temperature (LGT) agarose was dissolved in TEA and cast either into vertical templates or on to horizontal slides, as described above. Electrophoresis was carried out at 4°C.

DNA was detected by brief ethidium bromide staining and the desired bands excised from the gel with a scalpel. Two volumes of 200 mM NaCl, 10 mM Tris-Cl pH 8.0, 0.1 mM EDTA were added to the slice, and the agarose melted at 65°C for 15 minutes. An equal volume of buffer-saturated phenol at 37°C was added, the phases rapidly

mixed then immediately separated by centrifugation. The aqueous phase was re-extracted with phenol, then with ether and the DNA ethanol precipitated.

Typically, 60% of the DNA present in any band was recovered.

Polyacrylamide gel electrophoresis

Electrophoresis of DNA species of less than about 1 Kbp in length was carried out on vertical 14 cm x 14 cm x 0.5 mm gels containing 4% - 20% acrylamide/bis (30:1) polymerised in 90 mM Tris-borate, 2.5 mM EDTA, pH 8.3. Electrophoresis was performed at 250 V for approximately 90 minutes. DNA was visualised under UV light following ethidium bromide staining.

DNA fragments that had been fractionated preparatively were excised from the gel and the DNA eluted into two changes of 200 µl 10 mM Tris-Cl pH 8.0, 0.1 mM EDTA at 37°C for between 1 and 16 hours. The eluate was adjusted to 200 mM NaCl and the DNA ethanol precipitated.

Efficiency of recovery depended on the size of the DNA fragments and ranged from 50-99%.

Transfer of DNA to nitrocellulose and hybridization with a labelled probe.

Restricted DNA fractionated on agarose slab gels was transferred to nitrocellulose filter paper using the method of Southern (1975), as modified by Wahl et al. (1979).

Prehybridization, hybridization and washing conditions were essentially as descriobed by Wahl et al. (1979), except that formamide and salt concentrations were altered when very short probes were used (Kidd et al., 1983) and both dextran sulphate and glycine were omitted from the

hybridization mix.

Washed, dried, nitrocellulose filters were placed in contact with X-ray film and exposed at -80°C in the presence of one or two tungstate intensifying screens.

Dot-blot analysis of DNA (Kafatos et al., 1979)

DNA (up to 5 µg/dot) was denatured in 0.5 M NaOH, neutralised with HCl, an equal volume of 20 x SSC (3 M NaCl, 0.3 M Na-citrate) added and the sample spotted on to nitrocellulose filter paper damp with 20 x SSC. The filters were then processed as described above.

II.3.vi. Construction of a cDNA Library

Preparation of vector (Roychoudhury et al., 1976; A. Hobbs, pers. comm.)

12 µg of un-nicked pBR322 DNA were cleaved with Pst 1 and 1 µg removed for a pilot tailing reaction. 1 nmol of ^{3}H-dGTP was dried down and resuspended in a 50 µl volume containing 4 nmol dGTP, 200 mM Na-Cacodylate (twice recrystalized from ethanol: ether; V:V, 1:1) pH 6.9, 2.8 mM β-mercaptoethanol, 1 mM $CoCl_2$, 10 µg BSA, 1 µg cut vector and 6 U of terminal deoxynucleotidyl transferase and the addition of poly-dG nucleotide tails, at 30°C, was followed by assaying the conversion of ^{3}H-dGTP into a trichloro-acetic acid-insoluble form.

The time required to add 15 dG residues/3' end was calculated and the remaining 11 µg of linear pBR322 was tailed in the 50 µl reaction mix for that time. The reaction was stopped by the addition of EDTA to 5 mM, followed by phenol/chloroform extraction. Tailed vector was isolated from LGT agarose (Section II.3.v.).

Synthesis of the first strand

(Efstratiadis et al., 1976)

25 µg of poly $(A)^+$ RNA and 250 ng of oligonucleotide primer were co-precipitated with ethanol, resuspended in 10 µl of 200 mM KCl, 10 mM Tris-Cl pH 8.30, heated to 100°C for 5 minutes and incubated at 41°C for three hours. The mix was adjusted to 10 mM $MgCl_2$, 60 mM KCl, 10 mM Tris-Cl pH 8.30, 10 mM DTT, 0.5 mM each of dTTP, dCTP, dGTP and dATP, 2.5 µl placental ribonuclease inhibitor and 25 U of reverse transcriptase added and synthesis allowed to proceed for 45 minutes at 41°C.

Synthesis of the second strand

Method 1 (Land et al., 1981)

After removal of the RNA template, by alkaline hydrolysis, and neutralization, unincorporated nucleotides and oligonucleotide primer sequences were removed by centrifugation through a 200 µl Sephadex G-50 column. The fraction containing the cDNA was adjusted to the tailing conditions described above and approximately 20 dC residues added to the 3' end.

The tailing reaction was stopped by boiling the mix for 5 minutes and the conditions adjusted to those used for the synthesis of the first strand, in a volume of 200 µl. 500 ng of oligo-dG_8 primer and 25 U of reverse transcriptase were added and the second strand synthesised at 41°C for one hour.

Synthesis was stopped by phenol/chloroform extraction and the double stranded (ds) cDNA was fractionated on a Sephadex G-50 column.

Method 2 (Efstratiadis et al., 1976)

The RNA template was degraded by boiling for two minutes. The reaction volume was increased two-fold and the conditions adjusted to those used for synthesis of the first strand. 25 U of reverse transcriptase were added and the second strand synthesised at 37°C for six hours.

Synthesis was stopped by phenol/chloroform extraction and the ds cDNA fractionated on a Sephadex G-50 column and ethanol co-precipitated with 2 µg E. coli tRNA.

cDNA was resuspended in a 200 µl volume containing 200 mM NaCl, 2 mM $ZnSO_4$, 50 mM Na-acetate pH 4.6 and 2000 U S_1 nuclease and incubated at 37°C for 20 minutes. The reaction was adjusted to 5 mM EDTA and phenol/chloroform extracted.

Size selection, tailing and annealing of ds cDNA to tailed vector

Double stranded cDNA was fractionated on a 10% - 40% linear sucrose gradient (180,000 g, 16 hours, 4°C); 0.5 ml fractions were collected across the gradient and different size classes of ds cDNA were pooled and ethanol precipitated.

An average of 20 dC residues/3' end were added to the ds cDNA using the conditions described above. The tailing reaction was terminated by the addition of an equal volume of 20 mM EDTA.

Half the tailed ds cDNA was annealed to 150 ng dG-tailed vector in 0.2 M NaCl, 10 mM Tris-Cl pH 8.0 by heating for 10 minutes at 65°C, incubating for one hour at 45°C and finally allowing the solution to cool slowly to 4°C. The annealed DNA was stored at 4°C for at least 16 hours

prior to transformation.

Transformation of E. coli

E. coli strain MC1061 was grown overnight at 37°C in Luria broth (L-broth; 1% bacto-tryptone, 0.5% yeast extract, 0.17 M NaCl, pH 7.0) and then diluted $^1/_{50}$ into fresh L-broth and grown to an A_{600} of 0.6. The cells were chilled on ice for 30 minutes, pelleted by centrifugation (500 g, 5 minutes, 4°C) and washed in $^1/_2$ volume of ice-cold 0.1 M $MgCl_2$. The cells were resuspended in $^1/_{20}$ of the original volume of ice-cold, freshly prepared 0.1 M $CaCl_2$ and stored at 4°C for between 4 and 24 hours.

0.2 ml of these competent cells were added to 0.1 ml of the DNA (typically 5 ng - 50 ng) in 0.1 M Tris-Cl pH 7.5, and stood, with occasional mixing, on ice for 30 minutes. The cells were heated at 42°C for two minutes, kept on ice for a further 30 minutes and then allowed to warm to room temperature. 0.5 ml of L-broth was added and the transformed cells incubated at 37°C for 20-30 minutes.

The transformed cells were mixed with 3 ml of 0.7% L-agar (at 42°C) and plated on to 1.5% L-agar plates containing 15 μg/ml tetracycline. Plates were incubated overnight at 37°C. Transformation frequencies were always better than 10^6 per microgram of pBR322.

II.3.vii. Detection and Examination of Recombinant Plasmid Clones

Colony screening (Grunstein and Hogness, 1975)

Colonies from a transformation were transferred by toothpick to a master plate and to a sheet of nitrocellulose that had been boiled three times in distilled water and lain on to an L-agar plate containing

15 µg/ml tetracycline. The colonies were grown overnight on the nitrocellulose at 37°C, and the colonies lysed by transferring the nitrocellulose sequentially on to 3 MM paper saturated with 10% SDS for three minutes, 0.5 N NaOH for 7 minutes, 1 M Tris-Cl pH 7.4 for two minutes, 1 M Tris-Cl pH 7.4 for two minutes and 1.5 M NaCl, 0.5 M Tris-Cl pH 7.4 for 4 minutes. The nitrocellulose filter was baked at 80°C, under vacuum, for two hours. Hybridization and washing conditions were as described for Southern blot experiments. In some cases, after the initial hybridization and detection of colonies, annealed probe was removed from the filters by boiling for 10 minutes in two changes of distilled water. The hybridization procedure was then repeated using a different labelled hybridization probe.

Miniscreen examination of plasmid recombinants

(Birnboim and Doly, 1979)

1.5 ml cultures of each recombinant were grown overnight in L-broth containing 15 µg/ml tetracycline. The cells were pelleted by centrifugation for 30 seconds in an Eppendorf centrifuge, resuspended in 100 µl of 15% sucrose, 25 mM Tris-Cl pH 8.0, 10 mM EDTA, containing 4 mg/ml lysozyme, and incubated at room temperature for 5 minutes. 200 µl of freshly prepared, ice-cold 0.2 M NaOH, 1% SDS were added and the solution gently mixed and returned to ice for 10 minutes. 125 µl of ice-cold 3 M Na-acetate pH 4.6 were added and the solution incubated on ice for a further 15 minutes.

Insoluble material was removed by centrifugation (10 minutes, Eppendorf centrifuge, 4°C) and the supernatant

phenol/chloroform extracted. Plasmid DNA was recovered from the aqueous phase by ethanol precipitation, resuspended in water and $^{1}/_{5}$ analysed by restriction digestion and agarose gel electrophoresis. 1 µl of 10 mg/ml DN'ase-free pancreatic RN'ase was included in the restriction reaction.

II.3.viii. Large-Scale Preparation of Recombinant Plasmid DNA

500 ml cultures of recombinant cells were grown in L-broth to an A_{600} of 1.0 and then chloramphenicol added to a final concentration of 150 µg/ml. The cells were incubated for 8-16 hours to allow amplification of the plasmid DNA (Clewell, 1972). Cells were harvested by centrifugation (10,000 g, 5 minutes, 4°C) and plasmid DNA isolated by either Triton or alkali/SDS lysis.

Triton lysis method (Guerry et al., 1973)

Cells were resuspended in 15 ml of 15% sucrose, 50 mM EDTA pH 8.0, containing 12.5 mg of lysozyme and incubated on ice for 15 minutes. 15 ml of 0.1% Triton X-100, 62.5 mM EDTA, 50 mM Tris-Cl pH 8.0 were added, with gentle mixing until the solution was homogeneous and the solution centrifuged (45,000 g, 30 minutes, 4°C). The supernatant was carefully removed and treated with 20 µg/ml (final concentration) DN'ase-free, pancreatic RN'ase, for 30 minutes at 37°C and 50 µg/ml (final concentration) Proteinase K for 30 minutes at 37°C. The solution was extracted with an equal volume of phenol/chloroform and the aqueous phase dialysed extensively against 10 mM Tris-Cl pH 7.4, 1 mM EDTA. Following dialysis, the solution was adjusted to 0.2 M NaCl

and the DNA recovered by ethanol precipitation. Contaminating RNA was removed by fractionating the DNA on a Sephadex G-150 column eluted with 0.2 M NaCl, 10 mM Tris-Cl pH 7.5, 1 mM EDTA. Plasmid DNA was identified by its absorbance at 254 nm and ethanol precipitated.

Alkali/SDS lysis method

Plasmid DNA was liberated from the cells as described above for the miniscreen method, except that the volumes were increased 40-fold, and plasmid DNA was treated with 20 µg/ml DN'ase-free, pancreatic RN'ase prior to phenol/chloroform extraction.

The ethanol precipitate was resuspended in 1.6 ml of water, adjusted to 0.4 M NaCl, 6.5% PEG, and the DNA precipitated on ice for one hour. The precipitate was recovered by centrifugation (10 minutes, Eppendorf centrifuge, 4°C), washed with 70% ethanol and dissolved in water. Plasmid DNA was stored at either 4°C or -20°C.

Isolation of supercoiled DNA

Plasmid DNA (approximately 200 µg) was resuspended in 7.00 ml H_2O/tube and 7.00 g solid CsCl added. In the dark, 0.700 ml 10 mg/ml ethidium bromide were added and the mixture centrifuged at 210,000 g for 40 hours at 15°C. The lower band was identified by brief exposure to weak UV light and recovered. Ethidium bromide was removed by five extractions with isoamyl alcohol and the DNA precipitated by the addition of two volumes of water and six volumes of ethanol. DNA was recovered by centrifugation and washed three times with 70% ethanol.

II.3.ix. Preparation of **In Vitro** Labelled DNA

Random-primed reverse transcription

(Taylor et al., 1976)

Random oligonucleotides were prepared from calf-thymus DNA as described by Taylor et al. (1976). The synthesis of cDNA was carried out in a 20 µl reaction mix containing up to 2 µg of mRNA, 1 mM each of the deoxyribonucleotides dATP, dTTP, dGTP, about 0.1 mM α-^{32}P dCTP, 50 mM Tris-Cl pH 8.3, 10 mM $MgCl_2$, 10 mM β-mercaptoethanol and the oligonucleotides to a final concentration of 2 mg/ml. 10 U of reverse transcriptase were added and the solution incubated at 41°C for 60 minutes. The RNA template was removed by alkaline hydrolysis with 0.3 N NaOH for 15 minutes at 65°C, and the solution neutralised by the addition of HCl to 0.3 M and Tris-Cl pH 7.5 to 0.1 M. The mix was extracted with an equal volume of phenol/chloroform and the aqueous phase loaded on to a 0.4 cm x 10 cm Sephadex G-50 column and eluted with 10 mM Tris-Cl pH 7.6, 1 mM EDTA. 200 µl fractions were collected and the cDNA detected by Cerenkov counting.

Oligo-dT-primed reverse transcription

DNA complementary to poly(A)$^+$ RNA was synthesised as described above, except that the random oligonucleotide primers were replaced with 20 µg/ml final concentration of oligo-dT_{10}.

Nick-translation of double-stranded DNA

(Maniatis et al., 1975)

200 ng of DNA were labelled in a 25 µl reaction mix containing 50 mM Tris-Cl pH 7.8, 5 mM $MgCl_2$, 10 mM

β-mercaptoethanol, 50 µg/ml bovine serum albumin, 5 µM each of ^{32}P-dCTP and ^{32}P-dGTP and 25 µM each of unlabelled dATP and dTTP. The DNA was nicked by the addition of 20 pg of E. coli DN'ase I and the reaction was started by the addition of two units of E. coli DNA-polymerase I. The solution was incubated at 15°C for 90 minutes, phenol/chloroform extracted and the unincorporated nucleotides removed by chromatography on Sephadex G-50 as described above. If the labelled DNA was to be used as hybridization probe, the DNA strands were separated by boiling the solution for two minutes and then snap-cooling on ice.

5'-end-labelling using polynucleotide kinase

Prior to labelling, DNA (up to 5 µg) generally was dephosphorylated at 37°C for two hours in a 60 µl volume containing 100 mM Tris-Cl pH 8.0, 0.15% SDS and 0.12 U calf intestinal phosphatase (previously dialysed against 100 mM Tris-Cl pH 8.0, 1 mM $ZnCl_2$). After dephosphorylation, the enzyme was inactivated by heating at 65°C for 5 minutes, followed by phenol/chloroform extraction. DNA was recovered by ethanol precipitation.

DNA was end-labelled in a 10 µl volume containing 50 mM Tris-Cl pH 7.5, 10 mM $MgCl_2$, 5 mM DTT, 0.1 mM spermidine, approximately 50 pmol γ-^{32}P-ATP and 2 U of T_4 polynucleotide kinase at 37°C for 30 minutes. The solution was either extracted with an equal volume of phenol/chloroform and the DNA recovered by ethanol precipitation or analysed directly by gel electrophoresis.

End-fill labelling.

DNA fragments (up to 5 µg) with 5'-overhangs were labelled in a 20 µl volume containing 50 mM NaCl, 5 mM Tris-Cl pH 7.4, 15 mM $MgCl_2$, 5 mM β-mercaptoethanol, 0.1 mM of an α-^{32}P-dNTP complementary to at least one base of the 5'-overhang, 0.5 mM each of the remaining dNTPs and 2 U of *E. coli* DNA-polymerase I, Klenow Fragment (Klenow), at 37°C for 30 minutes. When it was required that the DNA fragments be repaired to blunt-ends, unlabelled dNTPs were added so that all four were at 0.5 mM, an additional 1 U of Klenow added, and the reaction continued for 15 minutes at 37°C. The reaction was terminated by phenol/chloroform extraction and the DNA recovered by ethanol precipitation.

II.3.x. Isolation of Clones from a Recombinant Genomic Library

Plating and screening (Benton and Davis, 1977)

0.25 ml of a suspension containing 7.75 x 10^4 pfu in 10 mM Tris-Cl pH 7.4, 10 mM $MgCl_2$ were gently mixed with 0.5 ml of a mid-log phase culture of *E. coli* LE392 in L-broth and incubated at 37°C for 10 minutes. 9 ml of 0.7% L-agar, containing 10 mM $MgCl_2$, at 42°C, were added and the mixture poured on to fresh, dry 15 cm 1.5% agar plates containing 1% bacto-tryptone, 0.5% yeast extract, 0.5% NaCl, 0.2% glucose, 10 mM Tris-Cl pH 7.5, 1 mM $MgCl_2$. Plates were incubated, inverted, at 37°C overnight then stored at 4°C to harden the agar.

An unwashed, 14 cm nitrocellulose disc was lain on to the plate, orientation marks made with a needle and, when uniformly wet, peeled off and placed on to filter paper

saturated with 0.5 M NaOH, 1.5 M NaCl for one minute and then sequentially on to two filter papers saturated with 0.5 M Tris-Cl pH 7.4, 1.5 M NaCl for two minutes each. A duplicate filter was lain on to the plate, the orientation marks aligned and the filter processed as described for the first filter.

Filters were air dried, baked at 80°C *in vacuo* for one hour then pre-hybridised, probed and washed as specified above for Southern blots (Section II.3.v.). Autoradiography was carried out for two days.

Growth of 'phage

10^5 pfu/15 cm plate were absorbed on to LE392 and plated as described above. Plates were incubated right-side up overnight at 37°C and then stored at 4°C. Plates were overlayed with 10 ml PSB (100 mM NaCl, 10 mM Tris-Cl pH 7.4, 10 mM $MgCl_2$) and the 'phage allowed to diffuse into this solution at 4°C for eight hours. Debris was removed by centrifugation (10,000 *g*, 5 minutes, 4°C) and the 'phage precipitated at 4°C for two hours by adjusting the solution to 875 mM NaCl, 6% PEG. The flocculated 'phage were collected by centrifugation (10,000 *g*, 10 minutes, 4°C) and resuspended in 14 ml PSB.

This suspension was layered on to discontinuous CsCl gradients containing 2 ml blocks of CsCl in PSB, with densities of ρ = 1.40 and ρ = 1.60 and centrifuged at 210,000 *g* for 90 minutes at 15°C. 'Phage particles were collected from the 1.40/1.60 interface and stored at 4°C.

DNA was isolated from 'phage stocks by phenol/chloroform extraction following the addition of two volumes of 10 mM Tris-Cl pH 7.4, 5 mM EDTA, and concentrated by

ethanol precipitation.

II.3.xi. Subcloning DNA Fragments into Plasmid Vectors

Vector DNA was linearised with a suitable restriction enzyme, dephosphorylated and purified from uncut vector by passaging through an LGT-agarose gel.

Restriction fragments to be subcloned were preparatively isolated from either sucrose gradients or LGT-agarose or polyacrylamide gels.

Ligation of insert into vector was done in a 20 µl volume containing 20 mM Tris-Cl pH 7.6, 10 mM $MgCl_2$, 10 mM DTT, 0.6 mM ATP, insert and vector DNA and 0.5 U T_4 DNA ligase at 4-15°C for 4-16 hours. Sufficient vector to give a two-fold molar excess over insert generally was used, although the concentrations of both insert and vector were sometimes optimised according to the mathematical treatment of Dugaiczyk et al. (1975).

Recombinant molecules were transformed into MC1061, selected and characterised as described above (Sections II.3.vi and vii).

II.3.xii. Gilbert and Maxam DNA Sequencing Procedures

(Maxam and Gilbert, 1980)

End-labelling DNA molecules

DNA molecules were labelled at either their 5'-ends (by kinasing) or their 3'-ends (by end-fill labelling) as described above (Section II.3.ix.).

To isolate DNA labelled at only one end, either secondary restriction cleavage was performed and the products electrophoresed on a 6% polyacrylamide gel, or the

DNA strands separated by heating at 90°C for 2 minutes in 40 µl of 30% DMSO, 1 mM EDTA pH 8.0, 0.05% xylene cyanol FF, 0.05% bromophenol blue and electrophoresed on a 5% polyacrylamide gel with a 50:1 acrylamide to bis ratio.

Base modification

End-labelled DNA was dissolved in 30 µl of water and divided into four aliquots G(5 µl), A + G (10 µl), T + C (10 µl) and C(5 µl). 1 µl of carrier DNA (*E. coli* DNA, 1 mg/ml) was added to each aliquot before commencing the modification reactions.

(a) Modification of guanine only

200 µl of Cacodylate buffer (50 mM Na-Cacodylate pH 8.0, 10 mM magnesium chloride, 0.1 mM EDTA) and 1 µl of dimethylsulphate were added to the aliquot and the reaction mixture incubated at either 21°C for 1-2 minutes (for restriction fragments) or 37°C for 15 minutes (for synthetic primers). The reaction was stopped by the addition of 50 µl of G stop mix (3 M Na-acetate, pH 6.0, 2.5 M β-mercaptoethanol, 1 mM EDTA, 0.1 mg/ml *E. coli* tRNA), 1 ml of ethanol added and the DNA precipitated at -80°C.

(b) Modification of adenine and guanine

25 µl of formic acid were added to the A + G aliquot and incubation was at 21°C for 1-2 minutes for restriction fragments and at 37°C for 12 minutes for synthetic primers. 250 µl of A + G stop mix (0.3 M Na-acetate pH 6.0, 0.1 mM EDTA, 25 µg/ml *E. coli* tRNA) and 1 ml of ethanol were added before precipitation at -80°C.

(c) Modification of cytosine and thymine

10 µl of water and 35 µl of hydrazine were added to

the C + T aliquot, with incubation at either 21°C for 1 ½ to 4 minutes for restriction fragments or 45°C for 18 minutes for synthetic primers. 250 µl of C + T stop mix (0.3 M NaCl, 0.1 mM EDTA pH 7.6, 25 µg/ml *E. coli* tRNA) and 1 ml of ethanol were added and the mix precipitated at -80°C.

(d) Modification of cytosine only

15 µl of 5 M NaCl and 35 µl of hydrazine were added to the C aliquot with incubation at 21°C for 2-5 minutes (for restriction fragments) or at 45°C for 18 minutes (for synthetic primers), followed by the addition of 250 µl of C stop mix (0.1 mM EDTA pH 7.6, 25 µg/ml *E. coli* tRNA) and 1 ml of ethanol at -80°C.

After centrifugation all samples were washed with 70% ethanol to remove residual reagents, reprecipitated by the addition of 300 µl of 0.3 M Na-acetate pH 6.0, and 1 ml of ethanol at -80°C for 15 minutes, then centrifuged, washed again with 70% ethanol and evaporated to dryness in a vacuum.

Base removal and strand scission

All samples were redissolved in 100 µl of freshly prepared 1 M piperidine, heated at 90°C for 30 minutes and then evaporated to dryness in a vacuum. The DNA was then redissolved in 25 µl of water and again evaporated to dryness before dissolving in formamide loading buffer (80% deionised formamide, 0.01% bromophenol blue, 0.01% xylene cyanol FF, 0.1 mM EDTA).

This method did not remove the piperidine efficiently and it was later decided to dissolve the DNA samples in 95 µl of water after the first evaporation step and then

ethanol precipitate the DNA by adjusting to 0.2 M NaCl and adding 250 µl of nuclease free ethanol at -80°C. The DNA pellet was then washed and dried in a vacuum before being dissolved in formamide loading buffer. This method removed the majority of the piperidine and gave cleaner DNA samples for loading.

Samples were heated at 90°C for two minutes then quickly chilled on ice before loading on to 40 cm x 40 cm 0.3 mm, 6% - 20% polyacrylamide gels containing 8.3 M urea. Gels were pre-warmed (by pre-electrophoresis at high current) and debris and urea removed from the wells prior to loading. Electrophoresis was performed at 30 mA. After electrophoresis, gels were fixed with 1.7 M acetic acid for five minutes, washed with 4 l of 20% ethanol and baked at 110°C for 30 minutes. Autoradiography was carried out at room termperature for between 30 minutes and four days.

II.3.xiii. Subcloning into M-13 'Phage Vectors

Preparation of M-13 replicative-form (Rf) DNA

(Winter, 1980)

To 3 ml of 0.7% L-agar, at 45°C, were added 20 µl BCIG (8 mg/ml in dimethylformamide), 20 µl IPTG (8 mg/ml in water), 0.2 ml JM101 (A_{600} = 0.6) and 0.1 ml of diluted M-13 'phage (approximately 200 pfu). This mixture was poured on to a 1.5% agar in 2 x YT (1% yeast extract, 1.6% bacto-tryptone, 0.5% NaCl, pH 7.0) plate and incubated at 37°C for 9 hours.

A blue plaque was selected, toothpicked into 1 ml of 2 x YT broth and grown with shaking for 6 hours. Meanwhile, a 10 ml culture of JM101 from a single colony on a

minimal glucose plate was grown to an A_{600} of 0.5, and added to 1 litre of 2 x YT. When the A_{600} of this culture reached 0.5, the 1 ml of 'phage solution was added and grown for 4 hours. Replicative form M-13 DNA was prepared by the alkali/SDS method described above (Secton II.3.viii.).

Ligation and transformation

M-13 (mp 83 or mp 93) vectors were prepared and ligations performed as described above (Section II.3.xi).

Competent cells were prepared by growing JM101 to an A_{600} of 0.6 in 2 x YT-broth, harvesting by centrifugation (500 g, 5 minutes, 4°C) and resuspending in freshly prepared 50 mM $CaCl_2$. Cells were used after storage at 4°C for at least 4 hours, but up to 9 days. One fifth of a ligation mix was added to 0.2 ml of competent JM101 and left on ice for 40 minutes. The cells were heat-shocked at 42°C for 2 minutes and then added to 3 ml of 0.7% agar containing 20 µl BCIG (8 mg/ml), 20 µl IPTG (8 mg/ml), and 0.2 ml of exponential JM101 (A_{600} ~ 0.05). The mixture was plated on 2 x YT-agar plates and grown for 9-12 hours at 37°C.

Preparation of templates for sequencing

Recombinant plaques were toothpicked into 1 ml of 2 x YT containing 2 µl of overnight JM101 and grown with shaking for 6 hours at 37°C. Cells were pelleted by centrifugation in an Eppendorf centrifuge for 5 minutes. To each supernatant were added 0.2 ml of 2.5 M NaCl, 20% PEG 6000 and, after leaving at room temperature for 15 minutes, the 'phage pellet was collected by centrifugation. After removal of all the supernatant, the pellet was resuspended

in 0.1 ml of 10 mM Tris-Cl pH 8.0, 0.1 mM EDTA and extracted with an equal volume of neutralised phenol. The aqueous phase was re-extracted with 0.5 ml of diethyl ether and ethanol precipitated. The 'phage DNA was collected by centrifugation, resuspended in 25 µl of 10 mM Tris-Cl pH 8.0, 0.1 mM EDTA and stored at -20°C.

Complementarity testing of single-stranded M-13 recombinants

To determine which strand, of a particular sub-cloned DNA fragment, was present in a single-stranded M-13 recombinant (ssM-13 clone), hybridization analysis was carried out using an arbitrarily selected, or previously sequenced, ssM-13 clone, as a reference.

1 µl of the ssM-13 clone to be tested was added to 1 µl of reference ssM-13 clone and incubated with 1 µl of 10 x Hin buffer (100 mM Tris-Cl pH 7.4, 100 mM $MgCl_2$, 500 mM NaCl) at 65°C for 1 hour.

2 µl of loading buffer were added and the sample was electrophoresed on a horizontal, 1% agarose gel, next to 2 µl of reference clone (plus 2 µl loading buffer), until the dye had moved the desired distance. The DNA was visualised after ethidium bromide staining. Single-stranded M-13 clones with inserts identical to the reference clone co-migrate with the reference, whereas clones with the complementary strand are retarded as they have hybridized to the reference, thereby doubling their molecular weight.

II.3.xiv. Di-deoxy Sequencing Procedures

(Sanger et al., 1977)

Hybridization

2.5 ng of universal primer (17-mer) were annealed to 1 µg of M-13 single-stranded template in a 10 µl volume containing 10 mM Tris-Cl pH 7.4, 10 mM $MgCl_2$ by incubating at 70°C for 10 minutes, 37°C for 10 minutes and 25°C for 10 minutes.

Polymerisation

4 µl of α-^{32}P-dGTP (approximately 16 µCi) were lyophilized, the hybridization mixture was added, vortexed to resuspend the label and then 1 µl of 10 mM DTT was added. 1.5 µl each of the appropriate zero mix (T° for ddTTP: 10 µM dTTP, 200 µM dCTP, 200 µM dATP, 5 mM Tris-Cl, pH 8.0, 0.1 mM EDTA; C° for ddCTP : 200 µM dTTP, 10 µM dCTP, 200 µM dATP, 5 mM Tris-Cl, pH 8.0, 0.1 mM EDTA; A° for ddATP : 200 µM dTTP, 200 µM dCTP, 10 µM dATP, 5 mM Tris-Cl, pH 8.0, 0.1 mM EDTA; G° for ddGTP :200 µM dTTP, 200 µM, dCTP, 200 µM dATP, 5 mM Tris-Cl, pH 8.0, 0.1 mM EDTA) and ddNTP solutions (0.3 mM ddTTP, 0.15 mM ddCTP, 0.5 mM ddATP, 0.35 mM ddGTP, each in water) were added together. 2 µl of the zero mix - ddNTP mixtures were added separately to four Eppendorf tubes ("reaction tubes").

1 µl of Klenow fragment (0.5 U) was added to the hybridization mixture - label - DTT solution. 2 µl of this were then added to each of the four reaction tubes and the solutions were mixed by centrifugation for 1 minute. After 10 minutes incubation at 37°C, 1 µl of dGTP chase (500 µM dGTP in 5 mM Tris-Cl pH 8.0, 0.1 mM EDTA) was added to each of the four tubes, mixed by 1 minute centrifugation and

incubated for a further 15 minutes at 37°C.

4 µl of formamide loading buffer (formamide, deionised with mixed bed resin, 0.1% bromo cresol purple, 0.1% xylene cyanol FF and EDTA to 20 mM) were added to stop the reactions and mixed by a short centrifugation.

Samples were boiled for 3 minutes and then analysed on a sequencing gel (Section II.3.xii).

II.3.xv. Containment Facilities

All manipulations involving recombinant DNA were carried out in accordance with the regulations and approval of the Australian Academy of Science Committee On Recombinant DNA and the University Council of the University of Adelaide.

CHAPTER III

ISOLATION OF A PUTATIVE COLLAGEN GENOMIC CLONE

III.1. INTRODUCTION

Prior to the advent of recombinant DNA technology, the best characterised collagen gene sequences were those encoding chick type I collagen, as their messengers could be isolated in an enriched form from tissues such as embryonic calvaria (see Chapter I). Not surprisingly, these messengers were the first to be committed to recombinant form (Lehrach *et al.*, 1978; Sobel *et al.*, 1978; Lehrach *et al.*, 1979; Yamamoto *et al.*, 1980) and characterised in detail. The genes for these messengers were isolated soon after (Ohkubo *et al.*, 1980; Wozney *et al.*, 1981a). At about the same time, genomic clones encoding part of the sheep α 2(1) gene were isolated (Boyd *et al.*, 1980). Again, the ability to prepare RNA enriched for pro-collagen sequences from embryonic tissue was exploited to construct homologous probes for the sheep collagen gene.

Collagen genes for *Drosophila* (Monson *et al.*, 1982) *C. elegans* (Kramer *et al.*, 1982) and mouse (Monson and McCarthy, 1981) were isolated from recombinant libraries using cross-species hybridization probes. This approach avoided the need for purification of homologous probes for the selection of the genes.

At the onset of the research described in this thesis, very little was known about human collagen genes. One of the factors contributing to this paucity of information was the relative difficulty in obtaining large amounts of RNA, enriched for pro-collagen sequences, which could be used as homologous hybridization probe.

In considering approaches to the isolation of human type I collagen genes, it was decided that, rather than

attempt to develop homologous probes, the use of cross-species probes to screen a recombinant gene bank would be more practicable. The successful isolation of *Drosophila*, *C. elegans* and mouse collagen genes indicates the feasibility of such an approach.

However, although the use of heterologous probes was thought to be the most judicious on technical grounds, it was envisaged that probes used across species barriers might not necessarily detect their homologues (i.e. a probe from one class of collagen in one species may not detect the same class of collagen in another species). In fact, the collagen genes isolated from *Drosophila* and *C. elegans* clone banks, using chick type I probes, were quite different from vertebrate type I genes (see Chapter I). These, however, may be extreme examples as both insects and nematodes are, evolutionarily, very widely diverged from birds; the use of a chick α 1(1) probe detected only α 1(1) sequences in a mouse gene bank (Monson and McCarthy, 1981). Furthermore, chick α 1(1) cDNA clones have been shown *not* to cross-hybridize with chick α 2(1) cDNA clones (Lehrach *et al.*, 1979), suggesting that the repeated nature of the triple helical region of all classes of collagen (*viz.*, glycine as every third residue and a high frequency of lysines and prolines at other positions) is not reflected as homology at the nucleic acid level for chick type I genes at least.

To minimise the detection of non-homologous collagen types, it was decided to use two independent cross-species probes to screen a human recombinant gene bank, *viz.*, cDNA prepared from chick embryo calvaria RNA enriched for

pro α 1(1) and pro α 2(1) sequences and part of the sheep pro α 2(1) gene. Any recombinants that were scored as positives with both probes were likely to contain pro α 2(1) sequences.

This chapter describes the construction of the two cross-species probes, a partial analysis of the suitability of these probes to detect human procollagen sequences and their use to screen a human recombinant gene bank.

III.2. RESULTS

III.2.i. Preparation of Chick Embryo Calvaria RNA

A range of techniques is available for the preparation of RNA from tissues. The method most widely used to isolate RNA from chick embryo calvaria has been that of Benveniste et al. (1973) in which calvaria are frozen in liquid nitrogen, ground to a powder in a pestle and mortar, dissolved in a Na-acetate/detergent solution and extracted with phenol/chloroform.

Although this method has the advantage of being simple, both the yield and quality of RNA isolated were found to be variable. It was felt that, although the level of ribonuclease was probably low in calvaria, the delay involved in dissolving the powdered tissue in buffer, prior to phenol extraction was giving ribonucleases, liberated by disruption of the cells, an opportunity to degrade the RNA. The inclusion of competitive ribonuclease inhibitors was considered, but these have not always proved adequate in preventing RNA degradation (Chirgwin et al., 1979). The active site inhibitor, diethyl pyrocarbonate, which is effective in the inhibition of ribonucleases (Harding

et al., 1977), also modifies single stranded nucleic acids (Ehrenberg et al., 1976) and so was not suitable.

Rather, it was decided to use the method described by Seeburg et al. (1977) in which powdered, frozen tissue was homogenized directly into the potent denaturant, guanidinium hydrochloride, and RNA isolated by centrifugation through a CsCl pad (Section II.3.iii.).

Although the quality (as judged by its sucrose gradient profile) of RNA prepared in this matter was good, the yield was found to vary according to the number of calvaria used. Approximately 60 µg of calvaria total RNA per chick could be prepared when up to 50 embryos were used, but the yield per embryo dropped when more tissue was used. This reduced yield may have reflected the difficulty in centrifugation of RNA through homogenates made viscous with deproteinated DNA.

An alternative method, modified from that of Chirgwin et al. (1979) by J. Brooker (pers. comm.) (Section II.3.iii.), in which tissue was frozen in liquid nitrogen, ground to a powder, homogenized in a guanidinium salt and RNA recovered by repeated precipitation with ethanol and K-acetate, was used for large scale preparations of calvaria RNA. Although Chirgwin et al. (1979) reported that freezing and grinding tissue prior to homogenization resulted in degradation of RNA, this was not observed. In fact, failure to freeze and grind the calvaria resulted in a much reduced yield of RNA, probably as a result of inadequate disruption of this very resiliant tissue solely by homogenization in a Dounce homogenizer. Approximately 80 µg of calvaria total RNA per embryo were isolated by

this method, even when calvaria from as many as 100 embryos were used.

Total RNA was fractionated by sucrose gradient centrifugation, the 20-40S fraction collected and poly(A)$^+$ material isolated by oligo-dT chromatography (Section II.3.iv.) as shown in Figure III.1. Approximately 1 µg of fractionated RNA was recovered from 1 mg of total RNA.

Fractionated RNA was used as a template for oligo-dT primed cDNA synthesis (Section II.3.ix.).

III.2.ii. Preparation of Probe from a Sheep Genomal Clone, SpC3

Boyd et al. (1980) have isolated two overlapping genomal clones encoding part of the pro α 2(1) gene, from a sheep genomic library constructed by Kretschmer et al. (1980). Both these clones have been made available by P. Tolstoshev.

A restriction map of the larger of the two, SpC3, indicating the EcoRI sites [modified from Boyd et al. (1980)] is shown in Figure III.2.i. Digestion of SpC3 with EcoRI generates three general size classes of fragments viz., the small fragments (1.3, 1.4, 1.5 and 1.8 kb), the intermediate fragments (2.3 and 5.7 kb) and the large, Charon 4a arms (10.9 and 19.8 kb). It was decided, for two reasons, to use the fragments of the small class collectively as templates from which to make probe.

Firstly, because it was likely that some regions of the sheep gene (especially introns, but also exon regions) might not be sufficiently homologous with the human gene to be good hybridization probes, it was decided to use as

Figure III.1.

Fractionation of chick embryo calvaria RNA.

RNA was prepared from the calvaria of fifty, 16 day-old chick embryos by the method of J. Brooker (pers. comm.) (Section II.3.iii.) and fractionated on 10-40% linear sucrose gradients.

(i) A_{254} profile of total cellular RNA across the sucrose gradient.

Fraction A was collected and the poly $(A)^+$ RNA isolated by two rounds of oligo-dT chromatography (Section II.3.iv.).

(ii) Elution profile of RNA from the oligo-dT column

(a) total RNA applied

(b) RNA eluted in fractions 1-10 reapplied

(c) elution buffer applied

Poly $(A)^+$ RNA was collected as indicated.

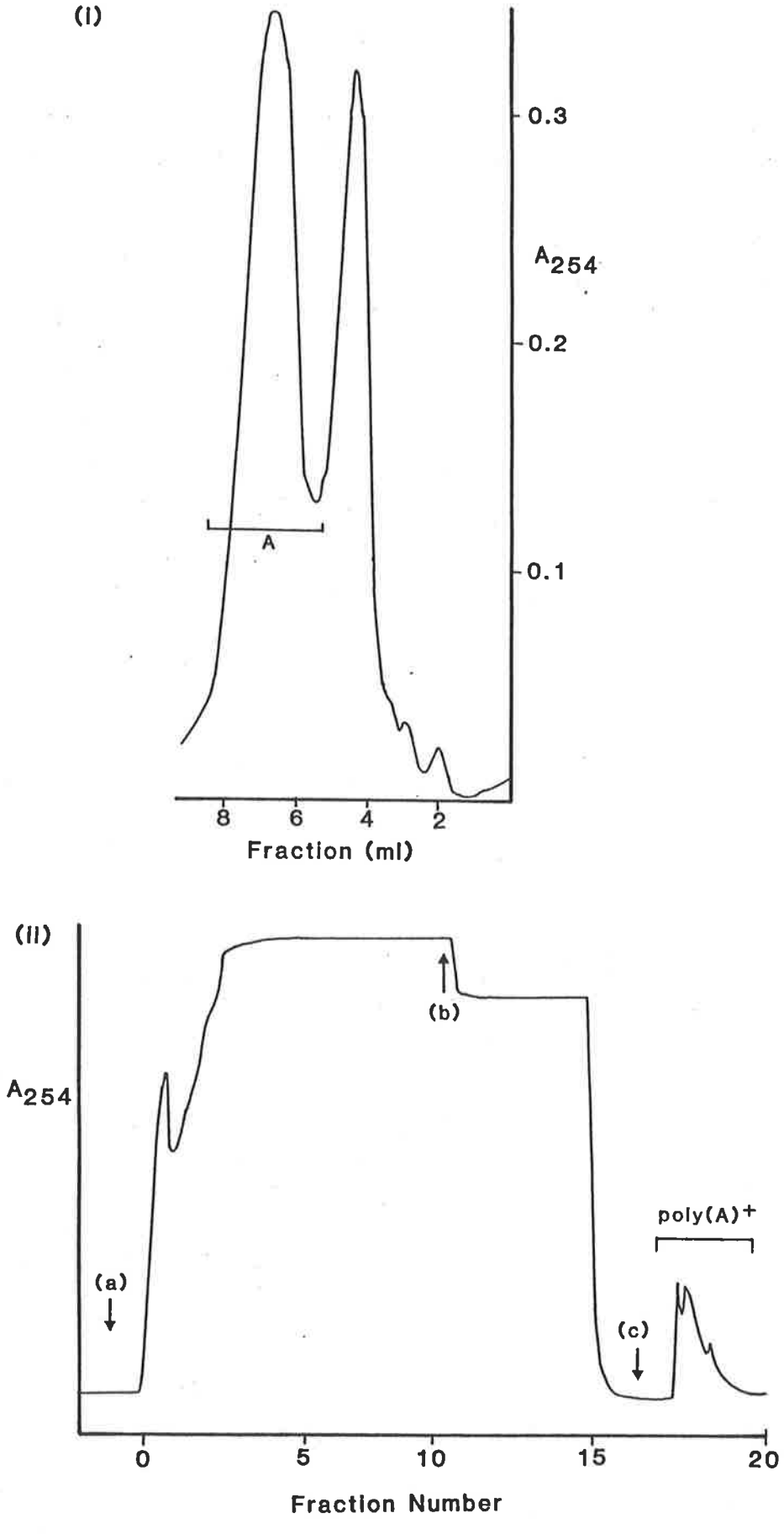

(I)
0.3
A_{254}
0.2
0.1
A
8
6
4
2
Fraction (ml)
(II)
A_{254}
(a)
(b)
(c)
poly(A)+
0
5
10
15
20
Fraction Number

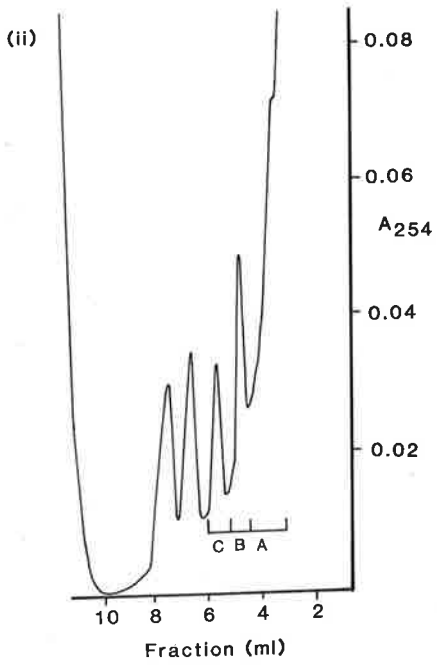
Figure III.2.

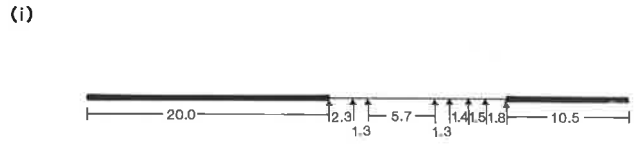
Fractionation of restriction fragments generated by EcoRI digestion of the sheep pro α 2(1) genomal clone, SpC3.

(i) Restriction map of SpC3 indicating the EcoRI sites [modified from Boyd et al. (1980)]. Solid lines denote the vector arms.

(ii) A_{254} profile of 50 µg of an EcoRI digest of SpC3 fractionated on a 10-40% linear sucrose gradient. Three fractions, A, B and C were collected.

(iii) An aliquot each of fractions A, B and C was analysed by agarose gel electrophoresis. Fraction A contained only the "small fragments" of SpC3 (Section III.2.ii.); these were used as templates for nick translation.


(ii)
0.08
0.06
A254
0.04
0.02
C B A
10 8 6 4 2
Fraction (ml)


(i)
20.0
12.3
1.3
5.7
1.3
1.4
1.5
1.8
10.5


(iii)
A B C

large a fragment of the sheep gene as was possible. The largest available restriction fragment containing only insert sequence is the 5.7 kb EcoRI fragment. However, this fragment is in the middle of the clone, shown by R-loop analysis to contain at least several introns, and so may not have been suitable. An alternative was to use, collectively, the small fragments which span 7.3 kb of the sheep genome over a region of 14 kb. The 1.8 kb fragment, at least, appears (by R-loop analysis) to contain no intron regions.

The second reason for using the small fragments was that they could be readily purified away from the Charon 4a arms. Because the probes were to be used to screen a recombinant λ library, it was essential that λ sequences be absent, otherwise all plaques would yield a positive response, regardless of the nature of the inserted sequence.

Fractionation of EcoRI digested SpC3 was performed on sucrose gradients and the purity of each fraction assessed by agarose gel electrophoresis, as shown in Figure III.2.

The small fragments were labelled by nick translation (Section II.3.ix.).

III.2.iii. Analysis of Probes

To assess the use of the two probes (i.e. nick translated SpC3 fragments and cDNA made to fractionated chick embryo calvaria RNA) for the detection of human procollagen sequences, two hybridization experiments were performed.

To verify that the cDNA did indeed contain sequences

complementary to procollagen mRNA, cDNA was hybridized to a Southern blot of EcoRI digested SpC3 (Figure III.3.). As is shown, chick cDNA detected sheep pro α 2(1) sequences, in particular those in the 1.5 kb EcoRI fragment. In addition, those fragments adjacent (in the sheep genome) to the 1.5 kb fragment (viz., the 1.4 and 1.8 kb EcoRI fragments) and the 2.3 kb EcoRI fragment were also detected, but less strongly.

The strong hybridization between the chick cDNA and the 1.5 kb sheep fragment may indicate a high degree of sequence conservation between homologous areas of the sheep and chick α 2(1) collagen genes. Alternatively, it may reflect, to some degree, a trivial reason, such as the paucity of introns in the 1.5 kb fragment. Whatever the cause of this strong cross-reaction is, it should not be assumed that the 1.5 kb fragment will cross-react strongly with human sequences, i.e. this result does not indicate that the 1.5 kb fragment would necessarily be a good probe for the human pro α 2(1) collagen gene.

It should also be noted that this hybridization result gives no quantitative information as to the abundance of sequences complementary to pro α 2(1) mRNA in the cDNA. However, it does indicate that they are sufficiently abundant for the cDNA to be an effective probe.

A second experiment was performed to verify that both probes were able to hybridize to human genomal sequences. Aliquots of human and sheep (gift from B. Powell) genomic DNA (along with SpC3 and λ DNA controls) were spotted on to filters (Section II.3.v.) and challenged with either probe. Figure III.4 shows the strong hybridization of both

Figure III.3.

Hybridization of chick embryo calvaria cDNA to SpC3.

1 µg of SpC3 DNA was digested with EcoRI and fractionated on a 1% agarose gel. Bands were detected by ethidium bromide staining. DNA was Southern blotted on to nitrocellulose and hybridized with radiolabelled cDNA made from fractionated chick embryo calvaria RNA (Section III.2.i.). The filter was washed in 2 x SSC/0.1% SDS at 65°C and autoradiographed overnight.

A: Ethidium bromide stained DNA.

B: Autoradiogram.

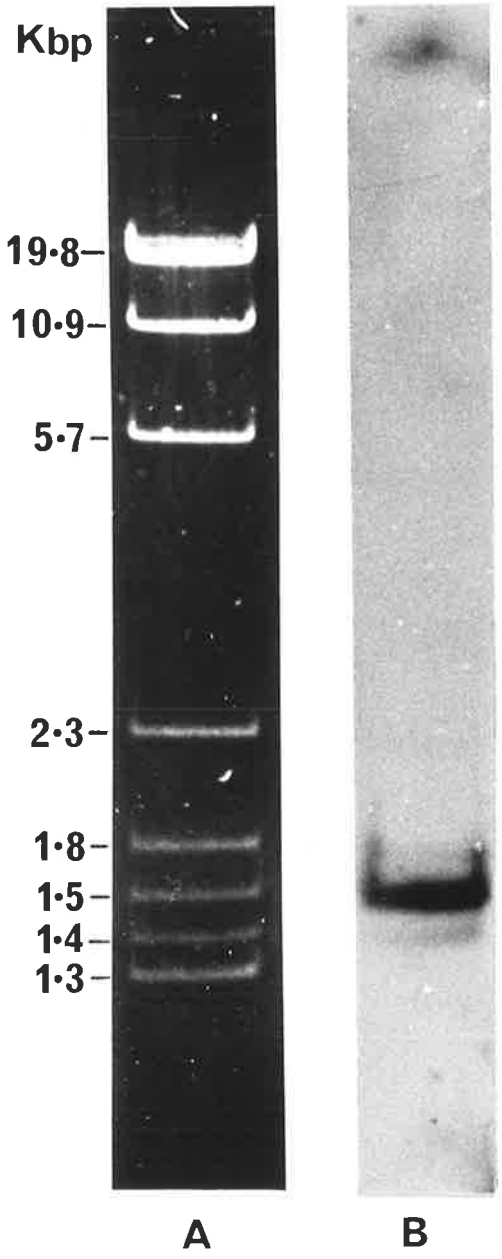

Kbp
19·8
10·9
5·7
2·3
1·8
1·5
1·4
1·3
A
B

Figure III.4.

Hybridization of library screening probes to human and sheep genomic DNA.

Samples were spotted on to nitrocellulose filters (Section II.3.v.):

A: 1 ng SpC3 DNA

B: 1 µg λ DNA

C: 10 µg human genomic DNA

D: 2 µg sheep genomic DNA (gift from B. Powell)

and hybridized with either:

(i) cDNA made to fractionated chick embryo calvaria RNA (Section III.2.i.) or,

(ii) nick translated SpC3 "small fragments" (Section III.2.ii.).

Filters were washed in 2 x SSC/0.1% SDS at 65°C and autoradiographed for 16 hours at -80°C.

i)

A B C D

ii)

probes to sheep and human DNA. λ sequences were not detected after this overnight exposure, although a longer exposure (not shown) revealed some hybridization between the nick translated probe and λ. Both probes hybridized to SpC3 DNA, although the signal was relatively weak when the cDNA probe was used.

Thus, both the chick cDNA and the nick translated sheep sequences should be suitable cross-species probes to detect human collagen genes.

III.2.iv. Selection of Recombinants

A human genomic library, constructed by Lawn et al. (1978) using the methods of Maniatis et al. (1978), was made available by T. Maniatis. Human genomic DNA had been partially digested with Hae III and Alu I and fragments with an average length of 15-20 kb inserted into the λ vector, Charon 4a (Blattner et al., 1977), by the addition of EcoRI linkers. At this time (and subsequently), the library had been used to examine a number of gene systems, notably the human globin genes (Efstratiadis et al., 1980); no major rearrangements attributable to the cloning process have been identified. The library had been amplified, by low density plating and subsequent recovery of 'phage into PSB, in this laboratory by S. Clarke.

The human haploid genome contains approximately 3×10^6 kb. Thus, 1.5×10^5 recombinants with 20 kb inserts should account for every gene. However, it is more appropriate to consider the library in statistical terms. Using the mathematical treatment of Clarke and Carbon (1976) it can be calculated that the probability of any sequence

being present in 7.75×10^5 recombinants with 20 kb inserts is > 99.5%.

This number of recombinants was plated at high density and screened in duplicate with either chick embryo calvaria cDNA or nick translated SpC3 fragments, using the methods of Benton and Davis (1977) (Section II.3.x.). A total of 6 positive signals was observed in duplicate at the primary screening. The recombinants giving rise to these responses were picked into PSB. Figure III.5. shows positive signals produced from plates containing 77,500 plaques.

In addition to detecting 6 plaques in duplicate, both probes each hybridized to additional recombinants. The nick translated probe detected, with equal intensity, approximately 50 additional recombinants. A longer period of autoradiography (not shown) failed to reveal any additional responses. Several of these recombinants were picked and stored in PSB at 4°C. cDNA detected, with a range of intensities, a larger number of plaques. A longer exposure (not shown) revealed in excess of 300 signals. It is likely that some of these recombinants would contain gene sequences encoding the pro α 1(1) gene. However, owing to the large number of these responses and the lack of a positive probe for human α 1(1) sequences, none of these plaques was picked.

In the primary screening, because plaques were in contact (confluent lysis of the bacterial lawn), isolates contain many recombinants in addition to the one producing the positive signal. It was therefore necessary to repeat screenings at lower plaque density to a point where a single plaque could be shown to give a positive response

Figure III.5.

Primary screening of a human recombinant library.

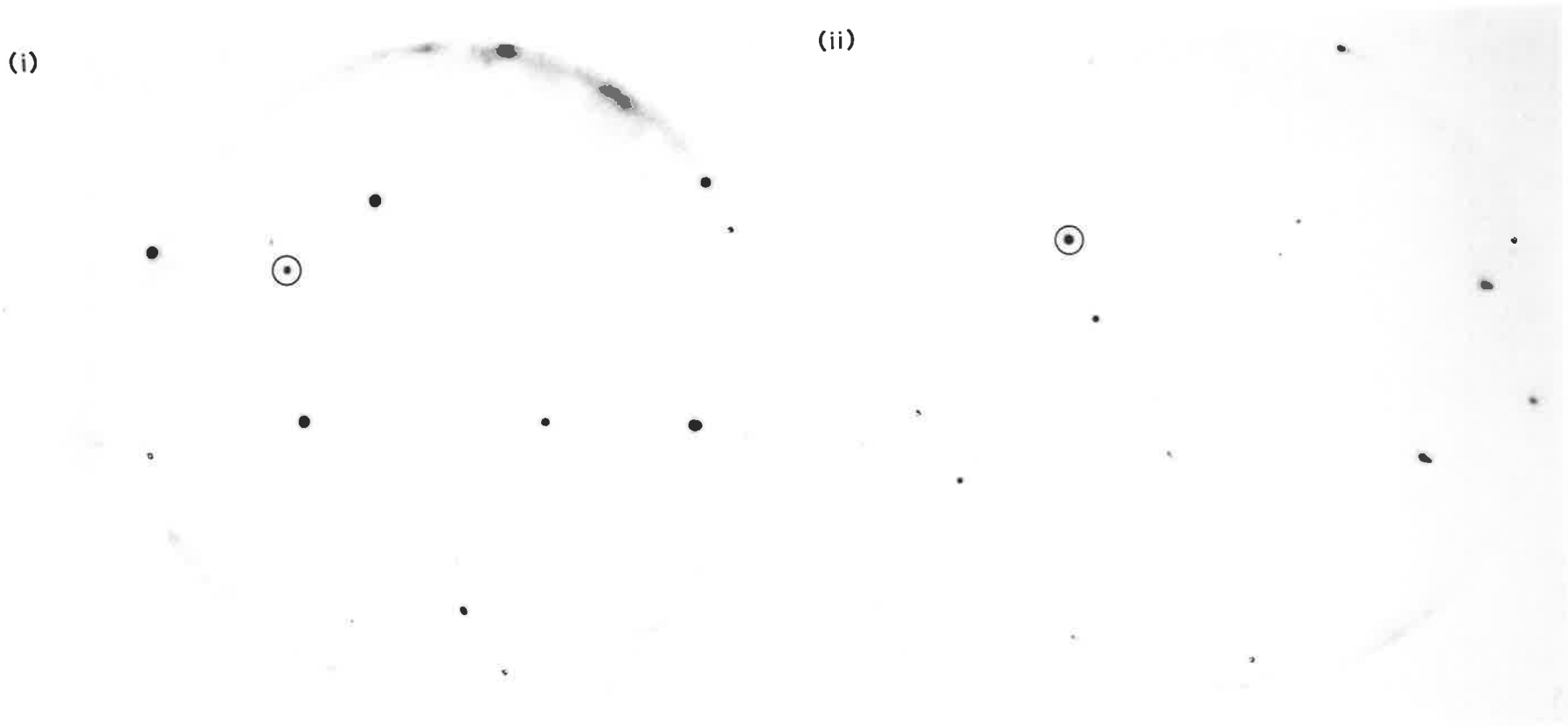
The human library was plated at a density of 7.75×10^4 pfu/plate on to ten, 15 cm plates and from these, Benton and Davis filters were prepared in duplicate (Section II.3.x.). Filters were probed with approximately 10^6 cpm per filter of either:-

(i) nick translated SpC3 "small fragments" (Section III.2.ii.) or,

(ii) cDNA made to fractionated chick embryo calvaria RNA (Section III.2.i.).

Filters were washed in 2 x SSC/0.1% SDS at 65°C and autoradiographed for two days at -80°c.

A total of six positive signals was observed in duplicate. The signal on this plate present in duplicate, is circled.


(i)
(ii)

and harvested to the exclusion of any contaminating plaques. This procedure required an extra two rounds of screening.

The six primary isolates were titred and replated to give distinguishable plaques. Filters were prepared and challenged with the nick translated probe. Positive signals were identified on all six second round plates. 'Phage giving rise to these responses were picked, replated a low density and rescreened with nick translated probe. Figure III.6. shows signals from the third round screening. Recombinants producing positive responses were picked for characterisation.

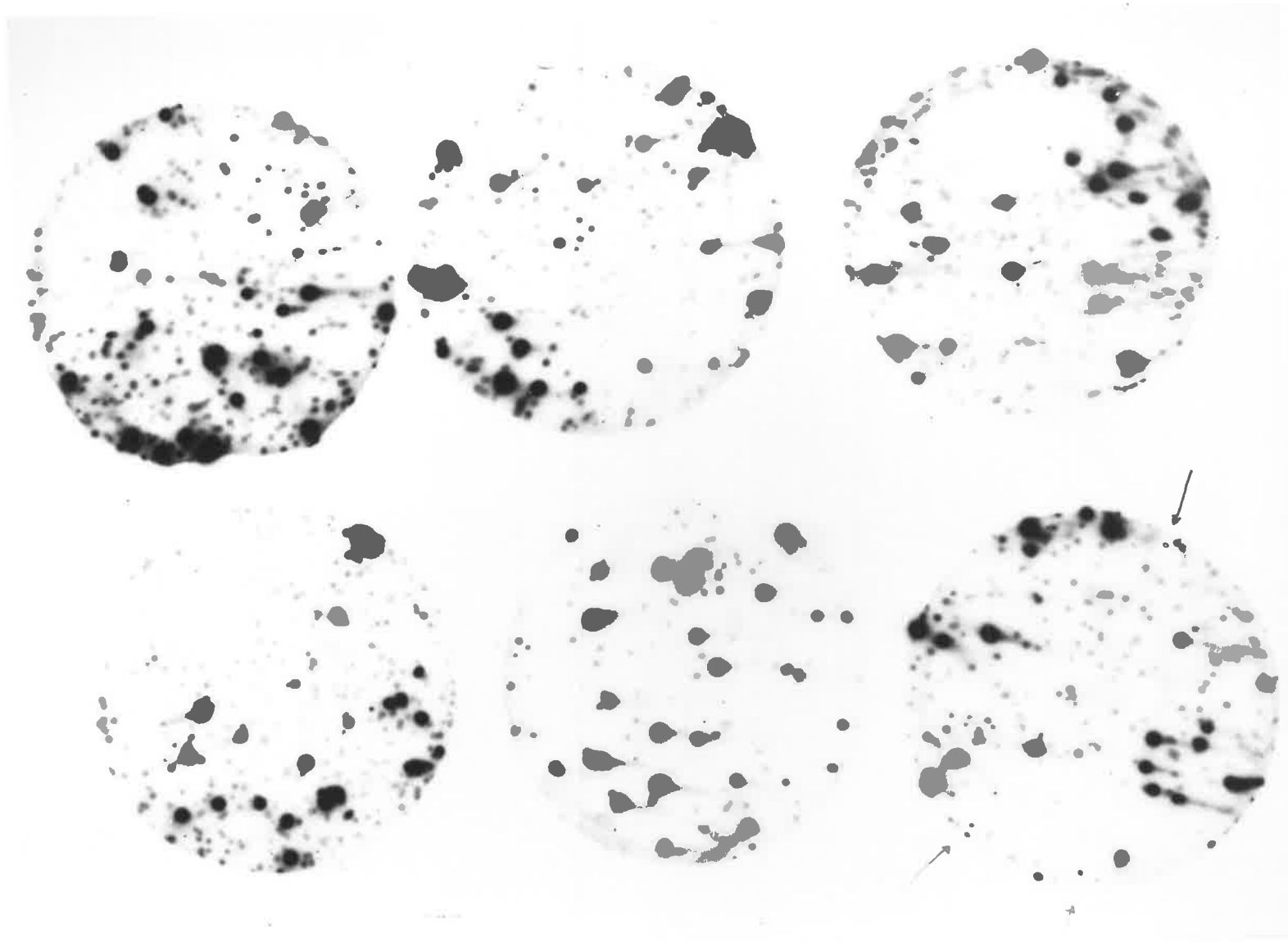
Figure III.6.

Third round screening of primary isolates from a human library.

Primary isolates from screening the human genomic library were picked into PSB, titred and replated at a density of 4000 pfu/plate. Benton and Davis filters were prepared (Section II.3.x.) and challenged with nick translated SpC3 probe (Section III.2.ix.) (not shown). Discrete plaques, corresponding to the six recombinants identified in the primary screening, were picked into PSB, titred and plated at a density of 100-500 pfu/plate. Filters were prepared and probed with 10^6 cpm/filter of nick translated SpC3 probe, washed in 2 x SSC/0.1% SDS at 65°C and autoradiographed at -80°C for 16 hours.

Strong, positive responses are evident on all six filters. The autoradiogram has been slightly overexposed to show the weak detection of other plaques as the consequence of a low level of vector arm DNA contaminating the probe.

Recombinants giving rise to positive responses were picked for analysis.



CHAPTER IV

CHARACTERISATION OF GENOMAL CLONES

IV.1. INTRODUCTION

Although the most definitive identification of collagen gene recombinants is provided by the DNA sequence of the genes themselves, a number of other techniques are available to allow an initial identification. Lehrach *et al.* (1979) showed that their putative (at the time) chick pro α 1(1) construct could hybridize to calvaria RNA to give S_1 nuclease resistant molecules and both Yamamoto *et al.* (1980) and Boyd *et al.* have used a hybrid-selected translation assay to identify recombinants. Monson and McCarthy (1981), who isolated the mouse pro α 1(1) gene using a cross-species probe, used the same probe to characterise the mouse gene prior to sequencing. It was decided to use this approach for the initial characterisation of the six recombinants isolated from the human gene bank (Chapter III).

This chapter decribes the characterisation of the putative human collagen genes by restriction analysis, hybridization studies and direct DNA sequencing.

IV.2. RESULTS

IV.2.i. Restriction Analysis of Recombinants

A plaque giving rise to a positive response at the third round screening (Figure III.6.) was picked from each of the six plates and 'phage particles amplified by plate culture, concentrated by PEG precipitation and DNA prepared (Section II.3.x.). A sample of each of the six DNA preparations was digested with EcoRI and analysed on a 1% agarose gel (not shown). The banding patterns of each recombinant appeared to be identical. To verify this, an

equivalent amount of each DNA was digested with EcoRI, the digests pooled and fractionated on an agarose gel. Such a comparison allows small differences in the size of apparently identical bands to be detected. The equivalent fragments from each recombinant co-migrated. It is therefore likely that each of the six positives represents a different isolate of the same recombinant. A large scale 'phage DNA preparation (Section II.3.x.) was performed on one of these isolates, called Pnc-8.

Pnc-8 DNA was digested with a range of restriction enzymes and the resultant fragments fractionated on agarose gels (Figure IV.1.). Hind III digested λ DNA and Hinf I digested pBR322 DNA were co-electrophoresed as molecular weight references. In addition, restriction fragments derived wholly from vector arm sequences were used as size markers. An artifactual, 7 kb band is present in the Bam H1 digests; this is generated by annealing of the cohesive ends of the 1.5 kb and 5.5 kb arm fragments, generating the cos site.

Maps, indicating the sizes and relative positions of each fragment are shown in Figure IV.2. and a complete map of the insert is shown, to scale, in Figure IV.3.

IV.2.ii. Hybridization Analysis of Pnc-8

Since Pnc-8 was selected from the recombinant bank by the criterion of its hybridization to both chick embryo calvaria cDNA and part of the sheep pro α 2(1) gene, both these probes should hybridize to restricted Pnc-8 DNA in a Southern blot experiment. However, because the cDNA was

Figure IV.1.

Restriction analysis of Pnc-8 DNA.

1 µg samples of Pnc-8 DNA were digested with a range of restriction enzymes, either singly or in combination, and the fragments resolved on 1% agarose gels (Section II.3.v.). Bands were observed under UV light following ethidium bromide staining.

M1 : λ DNA digested with Hind III.

E : Pnc-8 DNA digested with EcoRI.

H : Pnc-8 DNA digested with Hind III.

E/H: Pnc-8 DNA digested with EcoRI and Hind III.

B/K: Pnc-8 DNA digested with Bam H1 and Kpn I.

K : Pnc-8 DNA digested with Kpn I.

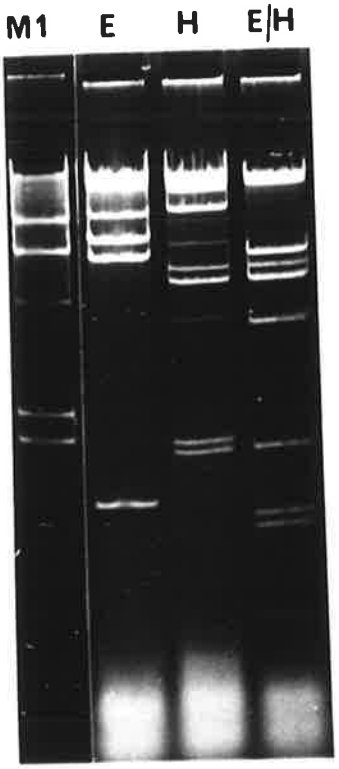
E/K: Pnc-8 DNA digested with EcoRI and Kpn I.

M2 : λ DNA digested with Hind III plus pBR322 DNA digested with Hinf I.

B : Pnc-8 DNA digested with Bam H1.

B/E: Pnc-8 DNA digested with Bam H1 and EcoRI.

Sizes of restriction fragments are summarised in Figure IV.2.


M1
E
H
E|H

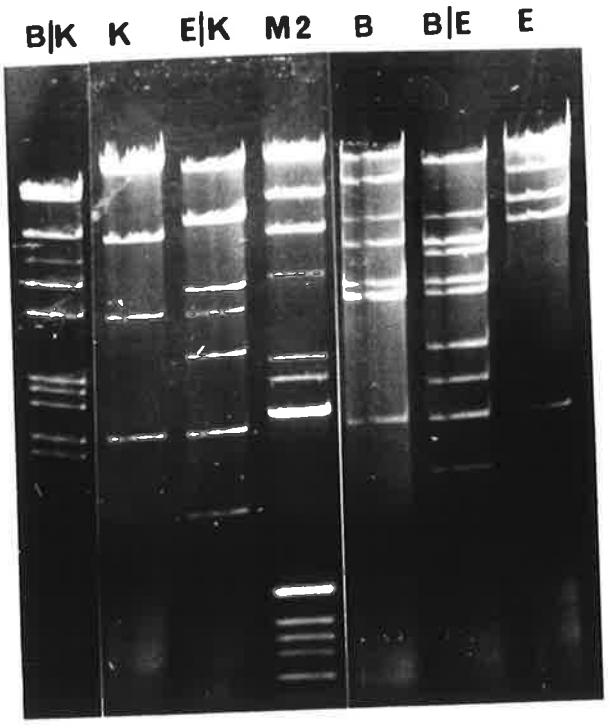

B|K
K
E|K
M2
B
B|E
E

Figure IV.2.

Restriction maps of Pnc-8 DNA.

The sizes of fragments obtained by restriction enzyme digestion of Pnc-8 DNA (Figure IV.1.) were determined using Hinf I digested pBR322 DNA and Hind III digested λ DNA as molecular weight markers. In addition, Hind III, Bam H1 and Kpn I fragments derived wholly from vector arm regions also served as size markers.

B: Bam H1 site.

E: EcoRI site.

H: Hind III site.

K: Kpn I site.

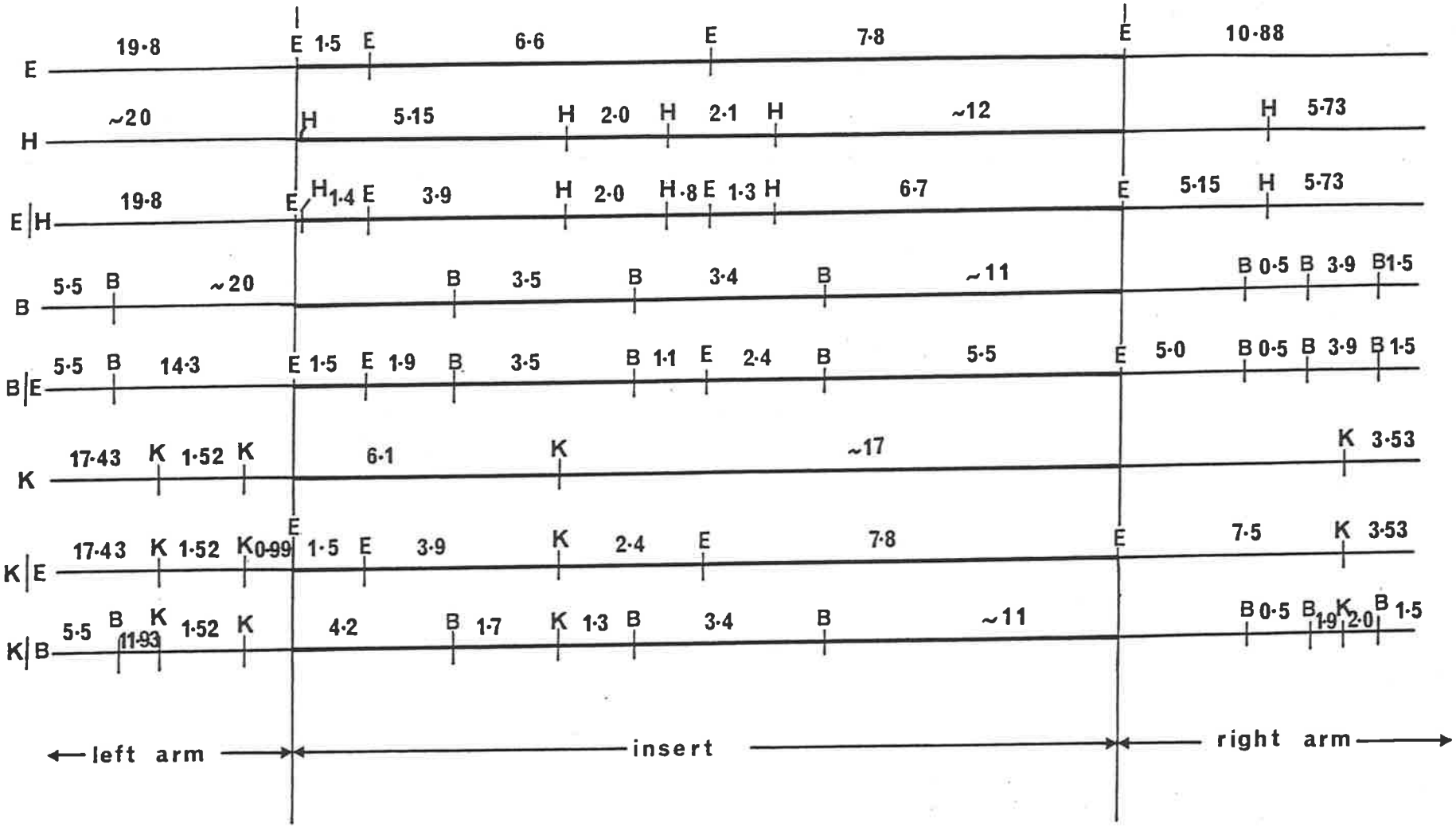

E
19·8
E 1·5 E
6·6
E
7·8
E
10·88
H
~20
H
5·15
H 2·0 H 2·1 H
~12
H 5·73
E H
19·8
E H 1·4 E
3·9
H 2·0 H ·8 E 1·3 H
6·7
E
5·15 H 5·73
B
5·5 B
~20
B 3·5 B 3·4 B
~11
B 0·5 B 3·9 B 1·5
B E
5·5 B 14·3
E 1·5 E 1·9 B 3·5 B 1·1 E 2·4 B
5·5
E 5·0 B 0·5 B 3·9 B 1·5
K
17·43 K 1·52 K
6·1
K
~17
K 3·53
K E
17·43 K 1·52 K 0·99 E 1·5 E 3·9 K 2·4 E
7·8
E 7·5 K 3·53
K B
5·5 B K 11·93 1·52 K
4·2
B 1·7 K 1·3 B 3·4 B
~11
B 0·5 B 1·9 K 2·0 B 1·5
left arm
insert
right arm

Figure IV.3.

Restriction map of Pnc-8 DNA.

Data shown in Figures IV.1. and IV.2. were used to construct a restriction map of the insert of Pnc-8. The insert is drawn to scale.

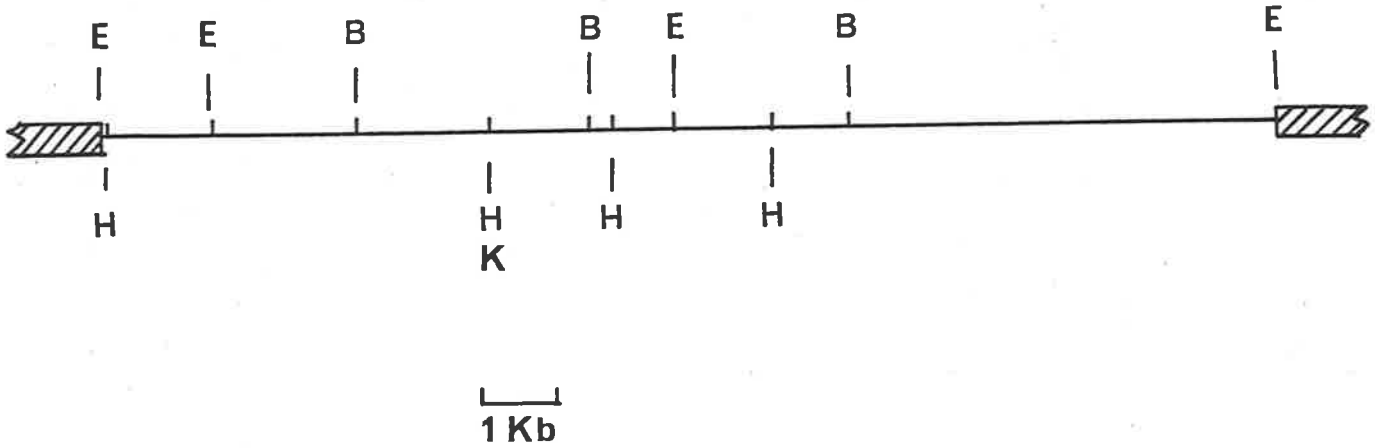
B: Bam H1 site.

E: EcoRI site.

H: Hind III site.

K: Kpn I site.

The shaded ares indicate the vector arms.


E
E
B
B
E
B
E
H
H
K
H
H
1 Kb

made from mature mRNA, it is complementary only to exon regions of the gene. Thus, probing a Southern blot of Pnc-8 DNA with cDNA should locate regions most readily identifiable by DNA sequencing.

Restriction analysis had shown that most of the insert could readily be digested into fragments of less than 2 kb (with the exception of those sequences adjacent to the right hand vector arm), which are of a convenient size for sequence analysis.

As a preliminary experiment, Pnc-8 DNA was digested with EcoRI and fractionated alongside EcoRI digested SpC3 DNA on an agarose gel. DNA was blotted on to a nitrocellulose filter and probed with chick embryo calvaria cDNA (Figure IV.4.). In addition to detecting sequences in SpC3 (shown previously; see Figure III.3.), this probe strongly hybridized to the 1.5 kb EcoRI fragment of Pnc-8. No other Pnc-8 sequences were detected. A longer period of autoradiography did weakly detect the 6.6. kb and 7.8 kb EcoRI fragments, but vector arms were also detected and so this response was probably artifactual.

Hybridization of cDNA to the 1.5 kb EcoRI fragment of Pnc-8 thus identified this molecule as being a good candidate for DNA sequence analysis.

IV.2.iii. Subcloning the 1.5 kb EcoRI Fragment of Pnc-8 into pBR325

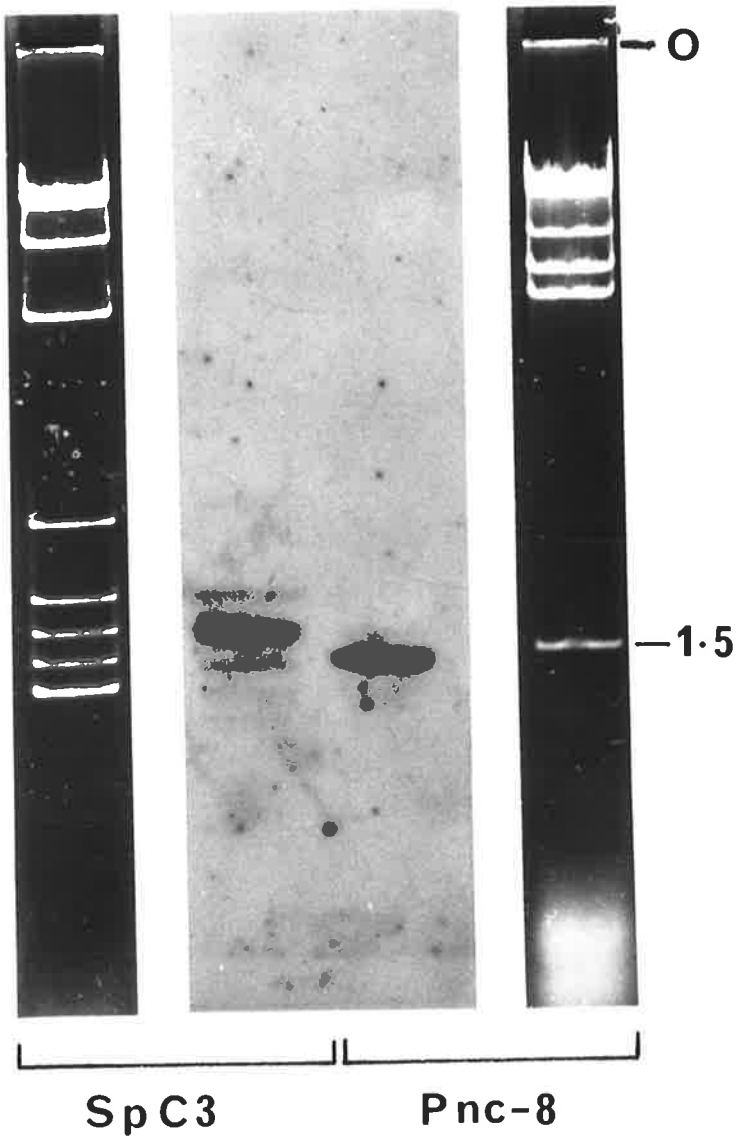
One requirement for direct DNA sequencing using the techniques of Maxam and Gilbert (1980) is the preparation, in pure form, of relatively large amounts of the molecule to be sequenced. Whilst it should be possible to isolate

Figure IV.4.

Hybridization of chick embryo calvaria cDNA to EcoRI digested Pnc-8 DNA.

1 µg each of SpC3 and Pnc-8 DNA were digested with EcoRI, the fragments resolved on a 1% agarose gel and the bands visualised under UV light following ethidium bromide staining (Section II.3.v.). Fragments were transferred to nitrocellulose and hybridized with 2 x 10^6 cpm cDNA made to fractionated chick embryo calvaria RNA (Section III.2.i.). The filter was washed in 2 x SSC/0.1% SDS and autoradiographed for two days at -80°C.

Bands on the autoradiogram have been aligned with the ethidium bromide stained fragments.


O
1·5
SpC3
Pnc-8

the 1.5 kb fragment from EcoRI digested Pnc-8, in sufficiently pure form, by sucrose gradient centrifugation, very large amounts of the clone would need to be digested, as the 1.5 kb fragment represents only approximately 3% of the total mass of the Pnc-8 genome.

To alleviate this problem of yield, it was decided to subclone the 1.5 kb fragment into a plasmid vector. pBR325 (Bolivar, 1978) was chosen, for although not as small as the commonly used vector, pBR322 (Bolivar et al., 1977), its unique EcoRI site is in the gene encoding resistance to chloramphenicol. Recombinants can thus be selected by their resistance to tetracycline and sensitivity to chloramphenicol. A 1.5 kb insert constitutes approximately 22% of the mass of a pBR325/1.5 kb-insert hybrid molecule.

Pnc-8 DNA was digested with EcoRI and the 1.5 kb fragment isolated from a sucrose gradient (Figure IV.5.i.). An aliquot was ligated into EcoRI linearised and dephosphorylated pBR325 and a portion was transformed into competent E. coli and transformants selected on plates containing tetracycline (Section II.3.xi.). Of the 100 colonies examined, 86 were found to be resistant to tetracycline but sensitive to chloramphenicol. Plasmid DNA was isolated from 10 of these colonies by the mini-screen procedure (Section II.3.vii.), digested with EcoRI and analysed on an agarose gel (Figure IV.5.ii.). At least 6 of the 10 recombinants contained the 1.5 kb insert. Insufficient DNA was present to clearly see if the remaining 4 contained the correct insert.

One of the 6 recombinants shown to contain the 1.5 kb insert, called p1.5E, was selected and a large scale

Figure IV.5.

Subcloning of the 1.5 kb EcoRI fragment of Pnc-8 into pBR325.

100 µg of Pnc-8 DNA were digested with EcoRI and the digest fractionated on a 10-40% linear sucrose gradient.

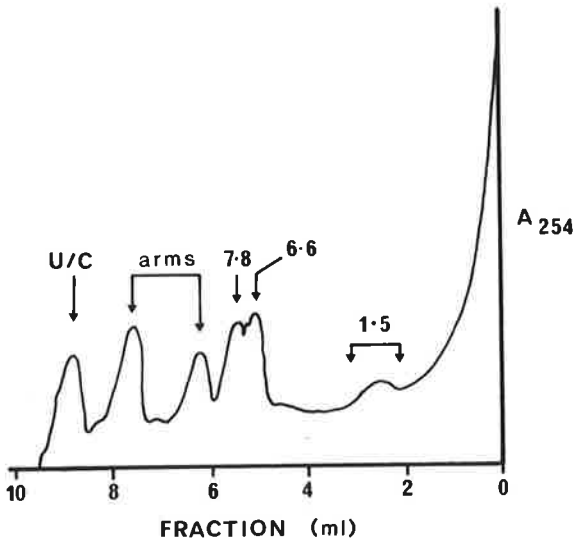
(i) A_{254} profile of Pnc-8, EcoRI restriction fragments across the sucrose gradient. The fraction containing the 1.5 kb molecule was collected and the DNA concentrated by ethanol precipitation.

100 ng of the 1.5 kb fragment were ligated with 350 ng of dephosphorylated, EcoRI cut pBR325 DNA and the ligation mix transformed into E. coli MC1061 (Section II.3.xi.). Ten colonies resistant to tetracycline, but sensitive to chloramphenicol, were analysed by the mini-screen procedure (Section II.3.vii.).

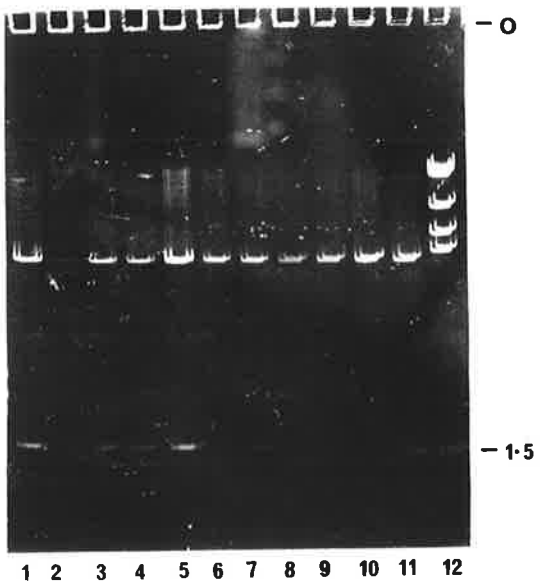
(ii) EcoRI digests of ten mini-screened subclones, fractionated on a 1% agarose gel (tracks 1-11, except for track 2 where a sample was not loaded) and stained with ethidium bromide. 500 ng EcoRI cut Pnc-8 DNA was run as a marker track (lane 12).

At least six of the ten recombinants have a 1.5 kb insert.

(i)


U/C
arms
7·8
6·6
1·5
A 254
10
8
6
4
2
0
FRACTION (ml)

(ii)


O
1·5
1 2 3 4 5 6 7 8 9 10 11 12

preparation of plasmid DNA undertaken (Section II.3.viii.).

IV.2.iv. Sequence Analysis of p1.5E Insert

Sequence analysis using the chemical degradation methods of Maxam and Gilbert (1980) can only be performed on DNA molecules which are of uniform length and which are labelled at one end only.

Molecules of uniform length are conveniently generated by digestion with restriction enzymes, as these cut at precisely defined sites. Because the gel systems used to fractionate the degraded DNA molecules can resolve a maximum of approximately 300 bases from the labelled end, it is convenient to generate restriction fragments, for sequencing, of approximately this size.

A number of techniques are available to end-label DNA. However, the anti-parallel nature of the strands in duplex DNA results in the labelling of both ends of the molecule. An additional step, either the dissociation of, and separation of the two DNA strands on an acrylamide gel, or restriction cleavage at an internal site, is therefore required to produce molecules labelled at one end only.

If a detailed restriction map of the insert to be sequenced is available, choice of restriction enzymes to construct a bank of fragments suitable for sequencing is straightforward. However, because a restriction map of p1.5E had not been determined, it was more convenient to identify suitable restriction sites in a series of pilot experiments.

It was decided initially to use the restriction enzyme Hae III to generate an array of smaller fragments from the

1.5 kb insert of p1.5E. Hae III recognises the sequence $^{5'}$GGCC$^{3'}$. This sequence is likely to occur moderately often in regions of collagen genes encoding the helical portion of the protein, as $^{5'}$GGC$^{3'}$ encodes glycine, and the next base, C, is the first residue of the proline codon.

To verify that Hae III digestion would convert the p1.5E insert into an array of suitably sized sub-fragments, a trial experiment was performed. Insert was resected from p1.5E DNA with EcoRI and isolated on a sucrose gradient. A sample was digested with Hae III and the bands displayed on a 5% polyacrylamide gel. Three fragments with sizes of approximately 530 bp, 520 bp and 430 bp were identified. These sizes were judged to be suitable for labelling and subsequent secondary cutting.

Digestion with Hae III generates molecules with blunt ends. These can be labelled at their 5'-termini by the polynucleotide kinase reaction, although blunt-ended molecules are not as efficiently labelled as are those with 5'-overhangs (such as generated by digestion with EcoRI).

Removal of 5'-phosphate groups prior to kinasing (Lillehaug et al., 1976) was found markedly to increase the efficiency of labelling. Dephosphorylation was performed using calf intestinal phosphatase. Some commercial batches of this enzyme were found to contain a low molecular weight contaminant which both became highly labelled and inhibited labelling of restriction fragments in the kinase reaction. Fractionation of dephosphorylated DNA on a polyacrylamide gel prior to kinasing avoided this problem.

1 µg of purified p1.5E insert was digested with Hae III. Terminal phosphate groups were removed (Section

II.3.ix.) and the fragments resolved on a 5% preparative polyacrylamide gel (Section II.3.v.) as shown in Figure IV.6. Bands were excised with a scalpel and the DNA eluted into TE. Using thin gels, it was not found necessary to crush the gel slice nor to use the long elution times described by Maxam and Gilbert (1980). DNA was concentrated by ethanol precipitation, and each fragment individually 5'-labelled by kinasing (Section II.3.ix.).

Examination of a portion of the kinased DNA on an acrylamide gel (not shown) revealed that both the 530 bp and 520 bp molecules were substantially contaminated with each other. Rather than attempt to purify these fragments, it was decided, in future experiments, to excise the two bands from the preparative acrylamide gel (prior to kinasing) as a doublet and to rely on resolution of the four labelled molecules following secondary cutting.

Pilot reactions were performed to identify suitable restriction sites for secondary cutting of the end-labelled DNA. 100 µM ATP was included in these digestions to protect the 5'-phosphates from phosphatases sometimes found to be present in commercially prepared restriction enzymes.

The unique Hind III site in the p1.5E insert was located conveniently within the 430 bp fragment, and Mbo II was found to cut uniquely in both the 530 bp and 520 bp fragments.

Preparatively kinased Hae III fragments were digested with the appropriate enzyme (i.e. either Hind III or Mbo II) and the resultant molecules fractionated on 6% polyacrylamide gels and identified by autoradiography (Figure IV.7.). Bands were excised from the gel, the DNA

Figure IV.6.

Preparative digestion of p1.5E insert DNA.

1 µg of p1.5E insert, fractionated away from vector sequences by sucrose gradient centrifugation, was digested with Hae III, dephosphorylated and the fragments resolved on a 5% polyacrylamide gel. Bands were visualised under UV following ethidium bromide staining and excised with a scalpel. DNA was eluted from each gel slice into 200 µl of TE, and concentrated by ethanol precipitation.

1: Hae III digested p1.5E insert.

2: Hinf I digested pBR322 markers.

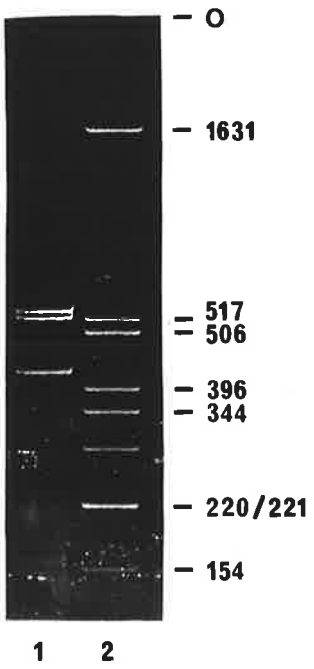

O
1631
517
506
396
344
220/221
154
1
2

Figure IV.7.

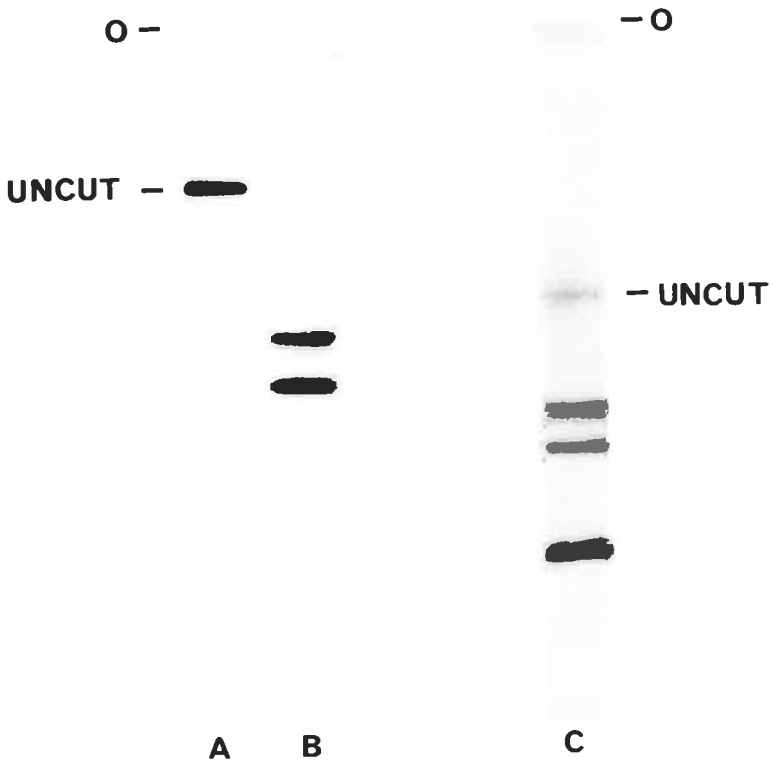
Secondary cutting of 5'-end labelled Hae III fragments derived from the p1.5E insert.

Dephosphorylated Hae III fragments, recovered from a preparative acrylamide gel (Figure IV.6.), were 5'-end labelled by kinasing (Section II.3.ix.) and digested with a suitable (determined by a pilot experiment) restriction enzyme. The resultant molecules, labelled at one end only, were resolved on 6% polyacrylamide gels. Bands were detected by autoradiography and excised with a scalpel. DNA was eluted from each slice into 200 µl of TE and recovered by ethanol precipitation.

A: Kinased, 430 bp Hae III fragment, uncut.

B: Kinased, 430 bp Hae III fragment, secondarily cut with Hind III.

C: Kinased, 530 bp and 520 bp Hae III fragments, secondarily cut with Mbo II.


O –
UNCUT –
– O
– UNCUT
A
B
C

eluted into TE and recovered by ethanol precipitation. The larger two Mbo II-generated bands (Figure IV.7.) were not always sufficiently well resolved for sequencing and so, after elution from the gel slice, the DNA was heated in 30% DMSO and refractionated on a strand separation gel (Maxam and Gilbert, 1980). This procedure was found to clearly resolve the two labelled molecules.

If, at this stage, a minimum of 10^5 dpm of ^{32}p-label was present in each end-labelled fragment, it was practical to proceed with the sequencing reactions.

Sequencing was carried out essentially as described by Maxam and Gilbert (1980) and the degraded DNA electrophoresed on high resolution, polyacrylamide-urea sequencing gels (Sanger and Coulson, 1978) (Section II.3.xii.). The acrylamide percentage of the gels and the duration of electrophoresis were varied to maximise resolution of different areas of the sequence. To read up to 200 bps of sequence, a single loading of sequencing reactions was run on a 20% gel (bps 5 → 60) and two staggered loadings on an 8% gel (bps 40 → > 200). After electrophoresis, gels were fixed with acetic acid, the urea washed away with 20% ethanol and the gel baked dry. Bands were detected by autoradiography. A typical sequencing ladder is shown in Figure IV.8.

The initial strategy of sequencing from the two internal Hae III sites and the EcoRI termini, towards Hind III and Mbo II sites, left some areas unsequenced. These areas included both the first few bases from each labelled end, which were found difficult to identify even on 20% gels, and the sequences approaching the sites of secondary

Figure IV.8.

DNA sequence determination.

End-labelled, Mbo II cut, Hae III fragments (Figure IV.7.) were degraded using the base-specific cleavage reactions of Maxam and Gilbert (1980) and electrophoresed on polyacrylamide DNA-sequencing gels (Section II.3.xii.). Shown here is the sequencing ladder of the smallest Mbo II - Hae III fragment on an 8% gel. The bands were detected by autoradiography.


G A T C

cutting, as sequence beyond about 250 bp was not reliably resolved. In addition, those regions that were sequenced, had only been sequenced in one direction. To be confident that data generated are accurate, it is necessary to sequence the DNA in both directions. Furthermore, although the sizes of the three Hae III fragments indicated that the entire 1.5 kb fragment had been accounted for, it was possible that one or more additional Hae III sites were present very close to those already identified. To avoid not detecting any small Hae III fragments, it was necessary to sequence through the Hae III sites. So, to completely sequence the p1.5E insert, it was necessary to utilise additional strategies.

One strategy was to use the same approach as that used for the Hae III fragments, but with fragments generated with a different restriction enzyme. The pilot experiments that had identified Hae III as being a suitable enzyme also indicated that Rsa I digestion would be suitable. Rsa I digestion of p1.5E insert generated four fragments with approximate sizes of 210 bp, a doublet at 370 bp and 517 bp (not shown). To identify the sequences at the termini of the p1.5E insert, it was decided to generate fragments spanning the insert-vector junctions, by digestion of the whole recombinant, and to sequence through these junctions. Figure IV.9 shows the fractionation, on an acrylamide gel, of 5 µg of p1.5E DNA digested with Sau96 I and labelled by end-fill labelling (Section II.3.ix.). The fragment containing the insert was isolated for subsequent secondary cutting and DNA-sequencing.

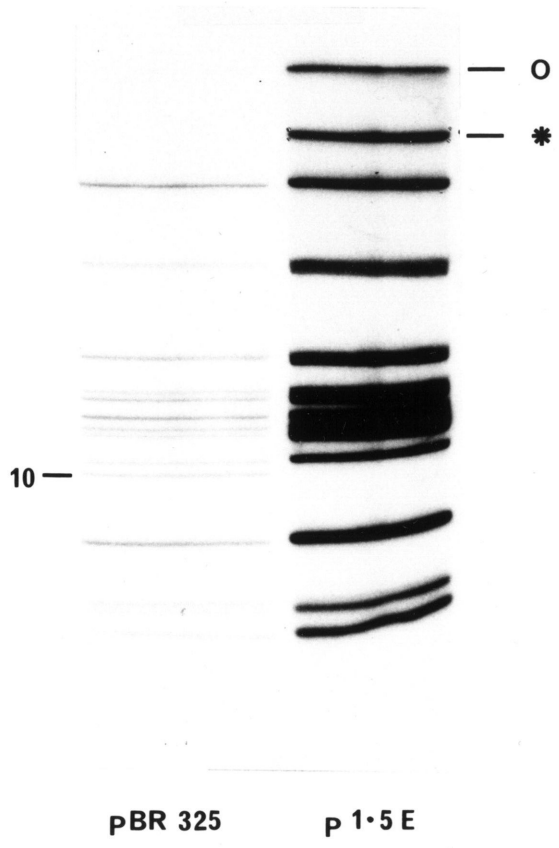
During the course of sequencing DNA molecules gener-

Figure IV.9.

End-fill labelling of Sau96 I digested p1.5E DNA.

5 µg of p1.5E DNA and 1 µg of pBR325 DNA were digested with Sau96 I (Section II.3.v.) and the 5'-overhangs repaired with Klenow in the presence of ^{32}P-dCTP and ^{32}PdGTP (Section II.3.ix.). Labelled fragments were electrophoresed on a 6% polyacrylamide gel and identified by autoradiography.

Band 10, containing the EcoRI site, in the pBR325 track is absent in the p1.5E track. Instead, a new band, indicated "*", containing the insert, is present. This band was excised and the DNA recovered by elution into TE and ethanol precipitation.


O
*
10
pBR 325
p 1·5 E

ated by these new strategies, M-13 cloning/di-deoxy sequencing techniques (Winter, 1980; Sanger et al., 1977) became available. Sequencing by these methods provides a number of advantages over the methods of Maxam and Gilbert (1980). Briefly, they are rapid, highly accurate and do not require the design of complex strategies. Central to the usefulness of these systems are the M-13 'phage vectors.

The M-13 genome exists in two forms during the life cycle of the 'phage viz., a double stranded replicative form (Rf) and a single stranded virion form. In the vector systems available, mp 83 and mp 93 (Messing and Vieira, 1982), a polylinker sequence (i.e. synthetic region of multiple, unique cloning sites) is present in the β-galactosidase gene. Vector is prepared from Rf DNA and sequences to be cloned are ligated into the appropriate site(s) in the polylinker; recombinants are selected by the resultant insertional inactivation of the β-galactosidase gene. Virions, extruded from infected cells, possess the insert DNA in single stranded form, the form required as template for chain termination sequencing (Sanger et al., 1977). Because the strand in the virion is always the same M-13 strand, a "universal" primer (Anderson et al., 1980) can be used to prime synthesis of the complementary strand. Sequences are generated by random insertion of dNTP-analogues (lacking 3'-hydroxyl groups) into a radiolabelled, complementary copy of the template, during its synthesis with Klenow.

A range of M-13 recombinants were constructed to generate templates for sequencing (Section II.3.xiii.).

The whole 1.5 kb insert was cloned into the EcoRI site of mp 83 and both Hae III and Rsa I fragments were cloned into the Sma I site of mp 83. In addition, a random library of fragments spanning the 1.5 kb insert was constructed. Insert was isolated from an LGT-agarose gel (Figure IV.10.i.) and ligated (Section II.3.xi.) at a high DNA concentration to generate concatamers. Concatamers were randomly sheared by sonication, the ends repaired with Klenow (Section II.3.ix.) and size fractions isolated from a polyacrylamide gel (Figure IV.10.ii.) and cloned into the Sma I site of mp 83.

In each case, the molecules cloned into the various M-13 vectors were able to insert in either orientation. Thus, from any given transformation, some 'phage contain one strand and the rest, the other. 'Phage containing opposite strands were identified by their ability to hybridize to each other in a complementarity assay (Section II.3.xiii.).

Sequencing from M-13 recombinant-'phage templates was performed as described in Section II.3.xiii. A typical sequencing ladder is shown in Figure IV.11.

The sequence of the p1.5E insert was determined by Gilbert-Maxam sequencing and confirmed in both directions by di-deoxy sequencing (Figure IV.12.).

IV.2.v. Examination of DNA Sequence

The primary reason for determining the DNA sequence of the 1.5 kb fragment was to enable the product it encoded, presumably a type I collagen α-chain, to be unambiguously identified. Although very little is known about the primary structure of the human type I collagen α-chains,

Figure IV.10.

Shotgun cloning into M-13 mp 83 of randomly generated fragments spanning the insert of p1.5E.

25 µg p1.5E DNA were digested with EcoRI and electrophoresed on a 1% LGT-agarose gel (Section II.3.v.) and the DNA visualised under UV light following ethidium bromide staining.

(i) Stained EcoRI cut p1.5E on a 1% LGT-agarose gel. The band was excised and the DNA recovered (Section II.3.v.).

2.5 µg of p1.5E insert DNA were ligated, in a 20 µl volume, overnight at 4°C (Section II.3.xi.). The ligation mix was diluted to 400 µl and subjected to eight, 30 second bursts of sonication at high power with a mini-probe. DNA was ethanol precipitated and the ends repaired with Klenow (Section II.3.ix.). DNA was fractionated on a 6% polyacrylamide gel and viewed under UV light following ethidium bromide staining.

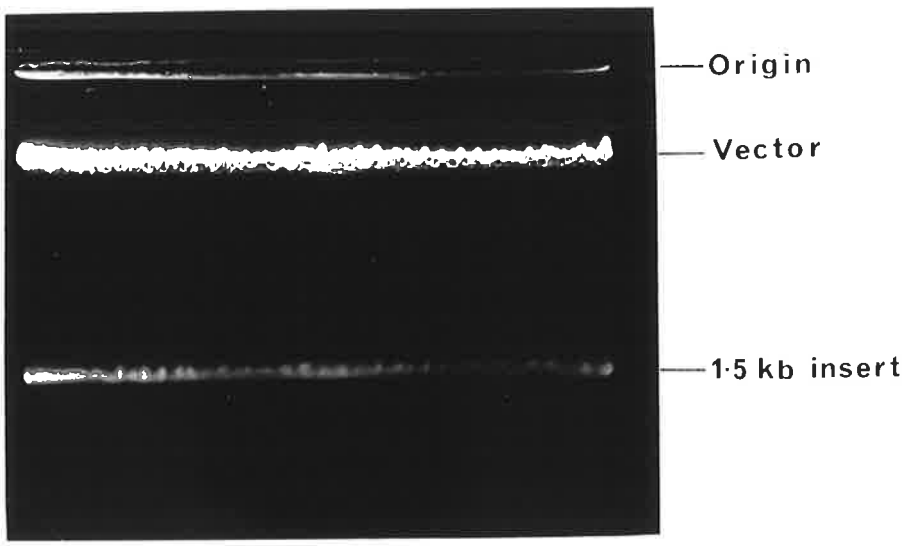
A: pBR322 digested with Hae III.

B: sonicated p1.5E insert DNA.

Three size classes viz., 1, 2 and 3 were selected, the regions excised from the gel and the DNA recovered.

DNA from each size class was cloned into the Sma I site of M-13 mp 83 (Section II.3.xiii.).

(i)



(ii)

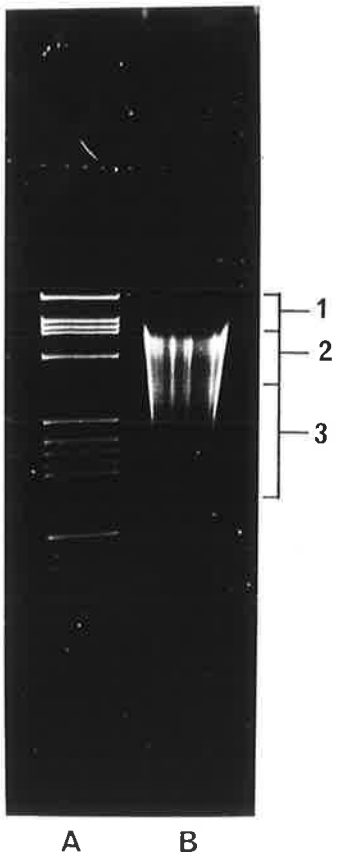


Figure IV.11.

DNA sequence determination.

DNA fragments generated by either restriction enzyme digestion or sonication, were subcloned into the M-13 vector, mp 83 and single stranded 'phage DNA isolated (Section II.3.xiii.). Single stranded DNA was used as a template for chain-termination sequencing (Sanger et al., 1977) (Section II.3.xiii.) and the synthetic products analysed on polyacrylamide DNA-sequencing gels (Sanger and Coulson, 1978) (Section II.3.xii.). Shown here is the sequencing ladder of the largest Hae III fragment of p1.5E insert, cloned into mp 83, running towards the EcoRI linker of Pnc-8, on a 6% gel. Bands were detected by autoradiography.

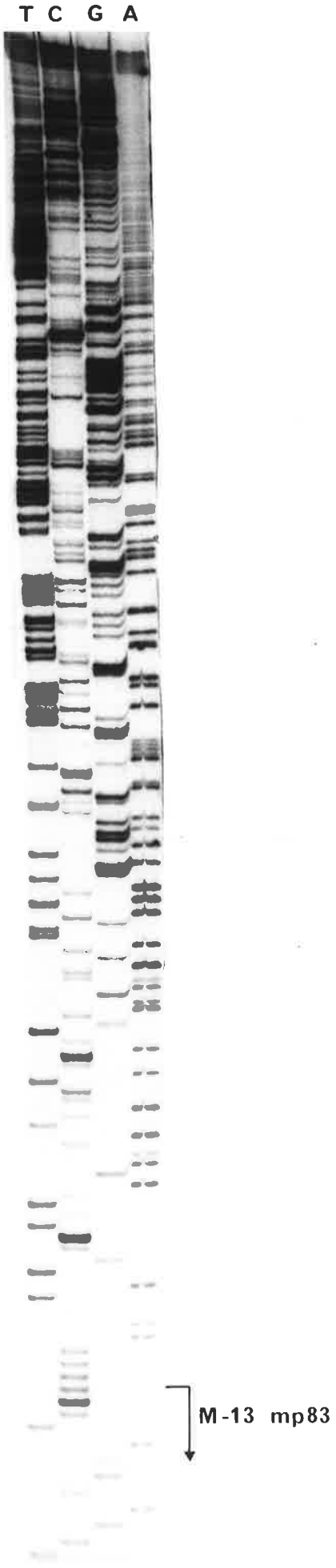

T C G A
M-13 mp83

Figure IV.12.

DNA sequence of the 1.5 kb EcoRI fragment of Pnc-8.

DNA sequence was determined by Gilbert-Maxam sequencing and confirmed in both directions by di-deoxy sequencing. Fragments from various sequencing runs were oriented by sequencing through restriction sites and confirmed by complementarity testing of M-13 subclones to the whole insert, cloned in M-13.

The sequence is shown, 5' → 3', from the EcoRI linker of Pnc-8.

```
       10         20         30         40         50         60
AATTCCATAA ATTACAAAAT GACAGAAAAA GAAAAATGCC AGAATTTCTA CCTCCTCATT

       70         80         90        100        110        120
CTCTTATTCT AAAGAAGAAG CATATGCAAA AAGGATTAAT TGAAACAGAT AACTTTTTTA

      130        140        150        160        170        180
GATGACCTTG CCTCAGTCTA GTAGGTCTTA TGTTCATCTA GGTAACTGAT ACTTCAAAGA

      190        200        210        220        230        240
CAAGTGAATT AAGTTTTCTT TAAAAGTACC CTTTTCCTAA GCTTGGATCT GAGTCCACTC

      250        260        270        280        290        300
TTCCTGAGAT CTTTTTTTTT TCTTTTTTTT TTTTTTTCAT GTTTGACTCT TAGTATCTGA

      310        320        330        340        350        360
GTCCTTCTCC ACTTAACTGG AATTTCATCC TATTTTCTGT AGTTTGAATA TAATGTAGAA

      370        380        390        400        410        420
GGAGTGACTT CCAAGGAAAT GGCTACCCAA CTTGCCTTGA TGCGCCTGCT GGGCCACTGA

      430        440        450        460        470        480
AGTCTCAAGA AATTTTAAGT TCTGGGCACA CATTACCATT TAAAAAACAT GCTAAAGAAA

      490        500        510        520        530        540
AAACATGCAT GGACATAGTA TACATTTCAG CATTTCAGAA GGCACTATCT GATAAGAATA

      550        560        570        580        590        600
CTCTGAATTT ATCATCTGAA TTTAAGTACG AAATTAATAC GTATAAATAA ATAAATATAA

      610        620        630        640        650        660
AATACATACA AAAAACTGAA TTTACTTATG AAATAAGTTA GTTCTCTCAT TCTACTGTTC

      670        680        690        700        710        720
TTTTAACCCC CTTTAGACCC CCTGGGGGCA GTCTAAGTTA GAACCCCCTC CATCCCACTT

      730        740        750        760        770        780
CCCAAAGCTA TTCCCATATT TCACTATTAC TTACGAGAGC AGCCATCTAC AAGAAGCAGT

      790        800        810        820        830        840
GTAAGTGAAC CTGCTGTTGC CCTCAGCAGA CAAGTTCAAG CATCATTAGA GCCCTGTAGA

      850        860        870        880        890        900
ATGACAGCCT TTTTCAGGTT GCCGAGTCTC CTCATCCATG TATGCAATGC TGTTCTTGCA

      910        920        930        940        950        960
GTGGTAGGTG ATGTTCTGAG AGGCATAGTT GGCCTACATT TAACTTTTAA TGCTTTTTCT

      970        980        990       1000       1010       1020
ACAATATGCT ATAAATATAA GAAAAATTAA AATTCACTAA CAGCAAGACT ACATACCCAC

     1030       1040       1050       1060       1070       1080
CCAGGTCCCA CTCCCAAAAG ACACACATAG AGGGGACATA CACACACATC CTAAAAATGA

     1090       1100       1110       1120       1130       1140
CTTTGTAAGG AGAATAAGGG TACACTTGGA ATGTGTGTTG AAATGTTGTT GGTTTTTTTG

     1150       1160       1170       1180       1190       1200
GTTTGTTTGT TTGTTTGTTT TTTGTTAGAC TGATAGGAGC CCCTCCCACT AAAGACACCC

     1210       1220       1230       1240       1250       1260
TTGATACTGT TATTTCAAGG ATGAACTTAT TTATCTGGGA CAGACATCTT CAGAATGACA

     1270       1280       1290       1300       1310       1320
CATGCCAAAC ATGGTTCTTA TTAAATCAAA GGTTCAGTAA TTATCAGATT CGAGAAATAG

     1330       1340       1350       1360       1370       1380
TGATGCTTTG TGTGATCTAT TTTCTTCTCT TTGAAACAGA AAAAGACAAA TGAATGGGGA

     1390       1400       1410       1420       1430       1440
AAGACAATCA TTGAATACAA AACAGAATAA GCCATCACGC CTGCCCTTCC TTGAGATTGC

     1450       1460       1470
ACCTTTGGAC ATCGGTGGTG CTGACCAGGA ATT
```

identification of the characteristic repetitive nature of the sequence in the helical region [viz., $(Gly\text{-}X\text{-}Y)_n$] would enable the gene to be classified as encoding some type of collagen and comparison with chick, rat or bovine α-chain sequences (Fietzek and Kühn, 1976) should enable the type to be determined.

The DNA sequence was translated into the six possible reading frames and searched for (Gly-X-Y) triplets. Although a number of glycine residues was found, none fitted the pattern $(Gly\text{-}X\text{-}Y)_n$ for $n > 2$. Whether or not those residues which did fit the pattern for 2 repeats (i.e. $n = 2$) were in fact exons encoding collagen, or just fortuitous pairings of glycine residues could not be determined. In four cases (Figure IV.13.), the "AG-GT" rule for splice junctions was satisfied, but conformity with the rest of the consensus sequence (Sharp, 1981) was not clearly evident. So, examination of the DNA sequence of the 1.5 kb insert had failed to positively identify sequences encoding a collagen α-chain, although four regions were located which may be very small exons encoding part of the helical domain of an α-chain. Insufficient data were available to classify which α-chain, if any, that these regions encoded.

Failure to positively identify helical regions in the amino acid sequences deduced from the DNA sequence of p1.5E does not mean that this recombinant does not encode part of a collagen peptide. In addition to containing intron sequences, the 1.5 kb sequence may encode either an amino or carboxyl terminal region of an α-chain. The position of the 1.5 kb fragment (the only region detected by the cDNA

Figure IV.13.

Potential helical-region encoding exons in p1.5E.

DNA sequence of the 1.5 kb EcoRI fragment of Pnc-8 (Figure IV.12.) was translated into the six possible reading frames and examined for regions encoding the structure $(Gly\text{-}X\text{-}Y)_n$. Several such regions were found. Potentially, four of these are exons as they obey the "AG-GT" rule for splice junctions (Sharp, 1981).

These potenial exon regions are shown (upper case) along with their immediate flanking sequences (lower case) and the protein sequences that they encode (italics). The first base of each putative exon is numbered from the 5'-end. Sequence 1 is oriented 5' → 3' from the EcoRI linker of Pnc-8 and sequences 2, 3 and 4 are in the opposite orientation.

1098

1. ag GGT ACA CTT GGA ATg t

Gly Thr Leu Gly Met

44

2. a gGA AGG GCA GGC gt

Gly Arg Ala Gly

757

3. a gTG GGA TGG AGG GGg t

Val Gly Trp Arg Gly

793

4. a gGG GGT CTA AAG GGG gt

Gly Gly Leu Lys Gly

probe) in Pnc-8 adjacent to a vector arm (Figure IV.3.) suggests that either this is one end of the gene or that the rest of Pnc-8 is at least predominantly (so as to preclude detectable hybridization with cDNA) intron region(s). In the absence of protein sequence data, the globular termini of α-chains cannot be identified by their deduced DNA sequence alone. However, an indication that protein coding sequences are present can be obtained by an analysis of the potential codon usage.

Codon usage of the p1.5E insert was compared to that of the average from a large number of eukaryotic protein coding genes (EMBL gene bank) using a computer programme written by Staden (1984). Four regions likely to encode protein sequence [i.e. log (P/(1-)) > 2] were identified (Figure IV.14.). One of these (number 1) lies within the longest open reading frame of p1.5E (39 amino acids). Although these data provide nothing more than a statistical correlation, they do suggest that protein coding regions are present in this sequence. It is interesting to note that none of the small, potential exons (Figure IV.13.) are detected as being typical protein coding regions, although the programme used may not be suitable for the detection of very short amino acid sequences.

If the 1.5 kb fragment does represent the end of a collagen gene it should be possible, by isolating clones spanning regions adjacent to the 1.5 kb fragment in the genome, to locate the entire gene.

IV.2.vi. Re-screening the Human Genomic Library

Since the 1.5 kb EcoRI fragment is immediately adjacent to the long vector arm (Figure IV.3.), it was

Figure IV.14.

Potential protein coding regions in the 1.5 kb EcoRI fragment of Pnc-8.

Codon usage of the 1.5 kb fragment was compared to a codon usage frequency matrix derived from the EMBL sequence bank, using the programme of Staden (1984). Four blocks of sequence were identified where, for every codon, the log [P/(1-P)] score was > 2 (Staden, 1984). These four derived amino acid sequences are shown, numbered at the first base (from the 5'-end) of their first codon. Sequences 1, 2 and 3 are oriented 5' → 3' from the EcoRI linker of Pnc-8 and sequence 4 is in the opposite orientation.

343

1. *Tyr Asn Val Glu Gly Val Thr Ser Lys*
   Glu Met Ala Thr Gln Leu Ala Leu Met
   Arg Leu Leu Gly His

1336

2. *Ser Ile Phe Phe Ser Leu Lys Gln Lys*
   Lys Thr Asn Glu

733

3. *Pro Tyr Phe Thr Ile Thr Tyr Glu Ser Ser*
   His Leu Gln Glu

421

4. *Pro Ser Met Cys Val Phe Trp Glu Trp Asp Leu*

decided to use this as a probe to isolate clones that overlap with Pnc-8, from the human genomic library.

The library was plated at a density of 7.2×10^4 pfu/plate on to ten, 15 cm plates and Benton and Davis filters prepared in duplicate (Chapter III; Section II.3.x.). Filters were probed with 8×10^5 cpm/filter of p1.5E insert (prepared by LGT-agarose gel electrophoresis), radiolabelled by nick translation (Section II.3.ix.), washed in 0.5 x SSC/0.1% SDS at 65°C and autoradiographed for 20 hours at -80°C. Thirty-six positive plaques were detected in duplicate (not shown).

Plaques giving rise to positive responses were picked and each purified to homogeneity by a further two rounds of screening. Five of the primary isolates proved to be false positives. 'Phage particles from the 31 positives remaining were, in batches of six, amplified by plate culture, concentrated by PEG precipitation and DNA prepared (Section II.3.x.). A sample of each of the DNA preparations was digested with EcoRI and analysed on 1% agarose gels. Figure IV.15. shows the patterns obtained from six of the samples. Four of these six clones gave EcoRI patterns identical to Pnc-8 and the other two gave patterns different from Pnc-8 but identical to each other. This new clone was called λ122. Restriction analysis of DNA from the 31 positives revealed that they were all isolates of either Pnc-8 or λ122.

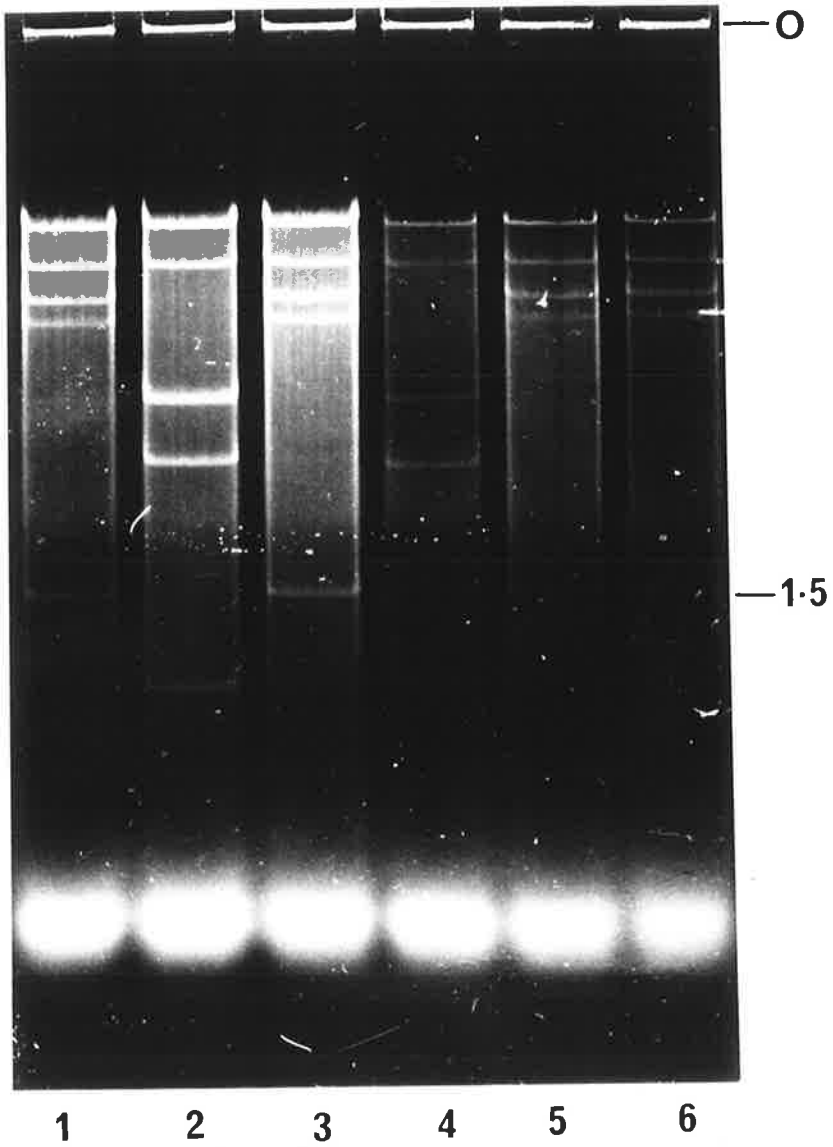
Rather than begin the characterisation of λ122 by constructing a detailed restriction map, it was decided to search for regions suitable for analysis by DNA sequencing. Regions that cross-reacted with cDNA would be suit-

Figure IV.15.

Restriction analysis of recombinants potentially overlapping Pnc-8.

'Phage DNA was prepared (Section II.3.x.) from each of the 31 recombinants isolated from the human gene bank as a consequence of their hybridization to the 1.5 kb EcoRI fragment of Pnc-8. Approximately 1 µg of DNA from each was digested with EcoRI and electrophoresed on a 1% agarose gel. DNA was viewed under UV light following ethidium bromide staining.

The restriction patterns of 6 of the 31 isolates are shown. Migration distance of a 1.5 kb fragment is indicated. Samples 1, 3, 5 and 6 have the same EcoRI pattern as Pnc-8. Samples 2 and 4 have the same pattern as each other and represent a new clone, called λ122.


O
1·5
1
2
3
4
5
6

able candidates for DNA sequencing.

1 µg of λ122 DNA was digested with EcoRI, fractionated on an agarose gel, blotted on to nitrocellulose (Section II.3.v.) and probed with 2 x 10^7 cpm of random primed cDNA prepared from fractionated chick embryo calvaria RNA (Section III.2.i.). The filter was washed at 2 x SSC/0.1% SDS at 65°C and autoradiographed at -80°C for three days. The probe failed to hybridize to any λ122 sequences (not shown). This experiment was repeated several times and although the cDNA could clearly hybridize to the 1.5 kb fragment of an EcoRI digest of Pnc-8, run on an adjacent gel track, it never specifically detected any λ122 sequences.

This result was surprising. λ122 was selected with the 1.5 kb fragment of Pnc-8, which strongly hybridizes with cDNA, and yet λ122 itself does not cross-react with cDNA. There are several possible explanations for this observation. It is possible that, although λ122 does overlap with Pnc-8, the region of overlap is short and fails to encompass the portion of the 1.5 kb fragment which hybridizes to the cDNA probe. Alternatively, λ122 might not overlap with Pnc-8 and may have been selected as the result of some cross-reaction (perhaps fortuitous) between an intron region of the 1.5 kb fragment and part of λ122.

No doubt, with further characterisation, it would be possible to determine both the relationship between λ122 and Pnc-8 and also to understand why λ122 failed to hybridize to the cDNA probe. However, the fact that cDNA does not detect λ122 sequences suggests that this isolate is unlikely to be useful, as potential coding regions

cannot be located.

At this stage, it was decided to re-appraise this approach for the isolation of human collagen gene sequences.

IV.3. DISCUSSION

In this chapter, the characterisation by restriction mapping, blot hybridization and direct DNA sequencing of a putative collagen genomal clone (called Pnc-8), isolated from a human gene bank, was described. Although hybridization results suggested that regions encoding collagen sequences were present, DNA sequence analysis failed to identify such regions. It seemed possible that Pnc-8 contained only a portion of a collagen gene and so attempts were made, by isolating possibly overlapping clones, to "chromosome crawl" into the rest of the gene. However, the clone isolated, called λ122, failed to hybridize to cDNA and was thus unlikely to contain more of the gene.

During the course of these experiments, several other workers reported the isolation of recombinants encoding parts of both the human α 1(1) and α 2(1) collagen chains (Myers et al., 1981; Chu et al., 1982a; Dalgleish et al., 1982). Significantly, the genomal clone, spanning the 3'-end of the pro α 2(1) gene, described by Dalgleish et al. (1982), called HpC1, was isolated from the Maniatis human library using the 1.5 kb EcoRI fragment from SpC3 as a probe; this approach was virtually identical to that described in Chapter III.

Comparison of HpC1 to Pnc-8 and λ122 reveals two substantial differences. Firstly, the restriction map of

HpC1 is entirely different from the maps of both Pnc-8 and λ122. The only potential similarity is that HpC1 and Pnc-8 both have a 1.5 kb EcoRI fragment, but in HpC1 this is internal (i.e. exists in the genome) whereas in Pnc-8, one of these EcoRI sites is a linker (i.e. does not exist in the genome). The second major difference between HpC1 and my isolates is the extent of mRNA coding region contained in each clone. HpC1 spans up to approximately 6.5 kb of mRNA coding region, whereas Pnc-8 and λ122 combined span less than 1.5 kb.

The fact that essentially the same probe has detected two entirely different recombinants (viz., Pnc-8 and HpC1) in what should be identical libraries, is perplexing. By a number of independent criteria (Dalgleish et al., 1982), HpC1 would appear to represent a substantial part of the human pro α 2(1) collagen gene. By one criterion at least (viz., hybridization to both cDNA and part of SpC3), Pnc-8 would also appear to contain part of a collagen gene. However, because of its dissimilarity to HpC1, Pnc-8 is unlikely to be the 3'-half of the human pro α 2(1) gene.

It is possible that Pnc-8 has arisen as the result of some perturbation of the library. At least two lines of evidence suggest that our aliquot of the library might not be truly representative of the human genome. Firstly, as discussed above, Pnc-8 does not resemble HpC1, yet these recombinants were isolated with essentally the same probe. Secondly, the number of isolates of both Pnc-8 and λ122 recovered from the library was very large. At the first screening (Chapter III), 6 isolates of Pnc-8 were detected from approximately 5 genome equivalents of 'phage

and at the re-screening (Section IV.2.vi.), 22 isolates were observed in less than 5 genome equivalents. The fact that Pnc-8 (and to a lesser extent λ122, although examination of one of the recombinants, detected by the nick translated probe but not the cDNA probe at the first library screening, revealed this to be λ122 [not shown]) was over-represented in the library implies that other sequences must be under-represented or absent. If some perturbation of the library had occurred, it is possible that this happened during amplification of our aliquot (performed prior to my first library screening).

The successful isolation of a recombinant encoding the human pro α 2(1) collagen gene by Dalgleish et al. (1982) demonstrated that the approach I had used should have yielded this gene. It seemed likely that my failure to isolate an identifiable collagen encoding clone resulted from our aliquot of the human library not being truly representative. So, to isolate collagen gene sequences it seemed necessary either to obtain or construct a representative library. Construction of a library is discussed in the next chapter.

CHAPTER V

CONSTRUCTION OF A HUMAN FIBROBLAST cDNA LIBRARY

V.1. INTRODUCTION

Chapter III described strategies aimed at identifying human type I procollagen sequences in a human genomal library. Six recombinants, thought likely to contain sequences encoding part of the pro α 2(1) gene, because of their hybridization to both fractionated chick embryo calvaria cDNA and to part of the sheep pro α 2(1) gene, were isolated.

Chapter IV described the characterisation of these recombinants. Although blot analysis using a cDNA probe identified regions possibly encoding a collagen gene, DNA sequence analysis failed to reveal identifiable collagen sequences. Attempts to isolate overlapping clones which could be identified as containing collagen gene sequence proved unsuccessful. These results suggested that our gene bank may not have been truly representative of the human genome.

At the same time, other workers constructed recombinants containing sequences encoding regions of human α 1(1) (Chu et al., 1982a) and α 2(1) (Myers et al., 1981; Dalgleish et al., 1982) collagen. Although unconditional access to these clones was not necessarily provided at the time, it was likely that they would be made available for use as probes in the future.

So, rather than attempt to construct recombinants which duplicated those already made, it was decided to isolate regions of the human type I collagen genes not present in those clones.

In each of the isolates of the human type I genes, recombinants spanned either the middle (i.e. encoding the

helical region) or the 3' end (i.e. encoding the carboxyl end of the protein) of the gene; none spanned the 5' ends of the genes. The only collagen gene for which the 5' end has been characterised is the chick pro α 2(1) gene (Vogeli et al., 1981; Tate et al., 1983). As well as providing protein sequence data (refractory to conventional protein sequencing techniques) several unusual structural features have been observed. These features include the presence of very small exons, one of which has overlapping donor splice sites, and the presence of two AUG codons 5' to the one normally used for translation initiation (Section I.3.i.). It is of considerable interest to see if these features are unique to the chick α 2(1) gene or are in fact typical of other (for example, human type I) collagen genes. In addition to providing fundamental information about their structure, recombinants spanning the 5'-ends of the human type I collagen genes would enable the construction of probes to assist in the analysis of this part of the gene and its transcript from cells manifesting primary collagenopathies (Section I.4.i.).

As part of their characterisation of the α 1(1) and α 2(1) mRNAs, Chu et al. (1982) and Bernard et al. (1983) respectively, have sequenced the most 5'-regions of their recombinant isolates. It is possible to utilise this sequence information to design synthetic oligonucleotides which can be used as primers for the synthesis of DNA complementary to the pro α 1(1) and pro α 2(1) mRNAs. Because DNA synthesis occurs in the direction 5' → 3', the cDNA will be synthesised towards the 5'-end of the mRNA. cDNA thus synthesised can either be converted to double

stranded cDNA and committed directly to recombinant form or, if made so as to be radiolabelled, used as probe to screen a recombinant gene bank.

It was decided to use the former approach viz., the direct cloning of double stranded cDNA whose first strand was synthesised using a synthetic oligonucleotide primer. This approach was favoured over the alternative, viz., the use of oligonucleotide primers to synthesise cDNA for use as library screening probes, for a number of reasons. Firstly, previous attempts to isolate collagen genes from the human genomal library (Chapters III and IV) had proved unsuccessful, possibly as the result of our aliquot of the library not being truly representative. Secondly, it is technically difficult to make cDNA labelled to a sufficiently high specific activity and yet still retain the hybridizational specificity conferred by the primer. If the primer alone is labelled (e.g. by kinasing), the specific activity of the resultant cDNA will be relatively low, although every labelled molecule of cDNA will have been primed from the synthetic oligomer. If, however, ^{32}P-dNTPs are included in the cDNA synthesis reaction to increase the specific activity, those cDNA molecules which originated by non-specific priming from RNA-derived, 3'-hydroxyl groups will now be labelled (Krieg et al., 1982). Unless the RNA template used was highly enriched for the desired sequences, these non-specific cDNAs will constitute a significant contaminant in the hybridization probe. Although the use of labelled primer alone (i.e. not extended against RNA) as a probe would appear to avoid these technical problems, the only sequence data available enable the synthesis of probes

complementary to the middle of the gene only; such probes would not detect 5' regions.

It was decided to use poly(A)$^+$ RNA from normal human fibroblasts as template for the synthesis of double stranded cDNA. Both the pro α 1(1) and pro α 2(1) cDNA clones previously isolated (Chu et al., 1982 and Myers et al., 1981, respectively) had been constructed from RNA from this source. Furthermore, although laborious, it was possible to grow large numbers of cells and thus isolate large amounts of RNA.

V.2. RESULTS

V.2.i. Synthesis and Characterisation of Oligonucleotide Primers

Since the DNA sequence has been determined for parts of both type I collagen mRNAs, the sequence of oligonucleotides to be used as primers for collagen cDNA synthesis can be unambiguously chosen. Although primers as short as eleven bases long have been used to construct and detect recombinants from low abundance messengers (Krieg et al., 1982), it is desirable to use primers larger than this so that hybridizations can be performed at higher stringency, thereby reducing the chance of non-specific priming events. Kidd et al. (1983) have shown that unique sequences can be identified in the human genome (albeit with some background) using a synthetic 19-mer and that by adjusting wash conditions, specific, single base mismatches could be detected. Thus, oligonucleotides of this length, or longer, should be of suitable length to prime specifically on human collagen mRNAs.

It was decided to use oligonucleotides with 21 residues as these should not only confer sufficient specificity for unique priming but are short enough to be straightforward to synthesise with high yields. Sequences were chosen which were both close to the most 5'-ends of the characterised recombinants and, as far as could be predicted from the known sequences, unique to those regions. To maximise the chance of these sequences being unique in the entire α-chain mRNAs, regions encoding infrequently used (in collagen α-chains) amino acids were chosen. Sequences of the pro α 1(1) and pro α 2(1) primers are shown in Figures V.1. and V.2. respectively.

Primers were synthesised from deoxynucleoside morphilinophosphoamadites (Beaucage and Caruthers, 1981; Dörper and Winnacker, 1983; McBride and Caruthers, 1983) by D. Skingle as part of an "in house" service provided in this Department. They were supplied as a crude mix (i.e. the entire synthesis reaction mix had been subjected to the detritylation reaction and subsequent steps to remove dimethoxytritanol) in aqueous solution.

To verify that the primers did exactly represent the chosen sequences, it was decided to subject them to DNA sequence analysis. Because the synthetic oligonucleotides were both single stranded and devoid of 5'-phosphate groups, they were ideal candidates for kinasing and sequencing by the chemical degradation methods of Maxam and Gilbert (1980).

Initial attempts to 5'-label the primers by kinasing proved disappointing. Not only was the incorporation of label on to the 5'-ends of the 21-mers low, but a back-

Figure V.1.

Sequence of the pro α 1(1) oligonucleotide primer.

A sequence (bold type), complementary to the DNA sequence (normal type), determined by Chu et al. (1982), encoding amino acids 252 → 258 of the human pro α 1(1) collagen α-chain (italics), was chosen for synthesis as a primer.

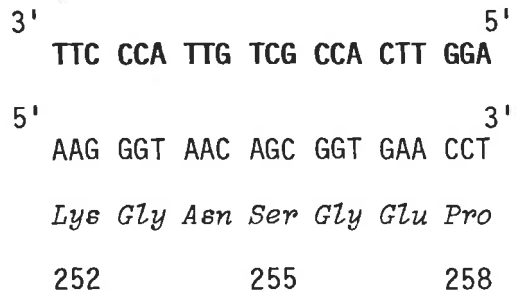
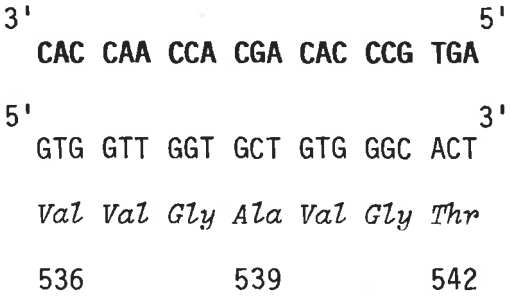

3' 5'
TTC CCA TTG TCG CCA CTT GGA
5' 3'
AAG GGT AAC AGC GGT GAA CCT
Lys Gly Asn Ser Gly Glu Pro
252 255 258

Figure V.2.

Sequence of the pro α 2(1) oligonucleotide primer.

A sequence (bold type), complementary to the DNA sequence (normal type), determined by Bernard et al. (1983), encoding amino acids 536 → 542 of the human pro α 2(1) collagen α-chain (italics), was chosen for synthesis as a primer.


3' 5'
CAC CAA CCA CGA CAC CCG TGA
5' 3'
GTG GTT GGT GCT GTG GGC ACT
Val Val Gly Ala Val Gly Thr
536 539 542

ground smear, detectable both by ethidium bromide staining and autoradiography, was evident when the kinased primers were analysed on a polyacrylamide gel (not shown). Surprisingly, a ladder of discrete fragments representing the failure sequences was not observed, even when the autoradiogram was overexposed (not shown).

Although the nature of the contaminant(s) which was both inhibiting the labelling of the oligomers and causing the background smear was not known, it was decided to try to remove it (them). The most appropriate approach seemed to be to purify the 21-mers prior to detritylation using trityl-specific reverse phase HPLC. A sample of the 2(1) synthesis reaction mix was fractionated by trityl-specific reverse phase HPLC (this procedure was performed by D. Skingle) and the 21-mer isolated (Figure V.3.i.) and detritylated. A 50 ng sample each of purified and unpurified α 2(1) 21-mer was kinased (Section II.3.ix.), fractionated on a denaturing, 20% polyacrylamide gel and autoradiographed (Figure V.3.ii.). The result was dramatic. The HPLC purification step not only eliminated the background smear but also enabled the oligonucleotide to become much more highly labelled. The α 1(1) 21-mer was purified in the same manner and all subsequent experiments were performed using primers purified by HPLC.

50 ng of each 21-mer were kinased and fractionated on a preparative, denaturing, 20% polyacrylamide gel. Bands were identified by autoradiography (Figure V.4.) and excised with a scalpel. DNA was eluted into 200 µl TE containing 4 µg E. coli tRNA and recovered by ethanol precipitation. Approximately 2 x 10^5 cpm/ng were

Figure V.3.

Purification of synthetic oligonucleotide primers by HPLC.

The α 2(1) 21-mer (Figure V.2.) was synthesised from deoxynucleoside morphilinophosphoamadites by D. Skingle, and the reaction mix, prior to detritylation, fractionated by trityl-specific reverse phase HPLC (performed by D. Skingle).

(i) HPLC elution profile

A: failure sequences

B: unknown contaminant

C: 21-mer

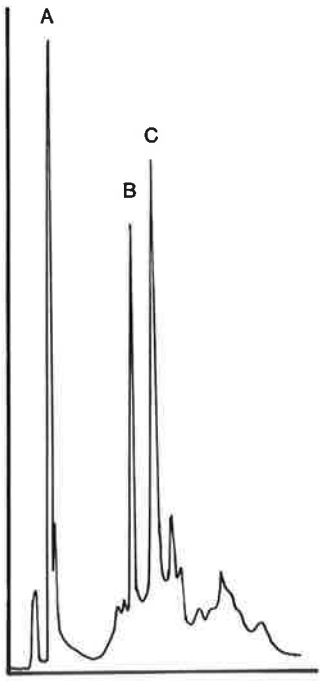
The fraction containing the 21-mer was collected and the oligomer detritylated. A 50 ng sample of this material, along with a 50 ng sample of α 2(1) 21-mer not fractionated by HPLC prior to detritylation, was 5'-end labelled using T_4 polynucleotide kinase (Section II.3.ix.) and electrophoresed on a denaturing, 20% polyacrylamide gel and autoradiographed at room temperature for ten seconds.

(ii) Autoradiogram of kinased α 2(1) 21-mers

1: unpurified 21-mer

2: 21-mer purified by trityl-specific reverse phase HPLC.

(i)



(ii)

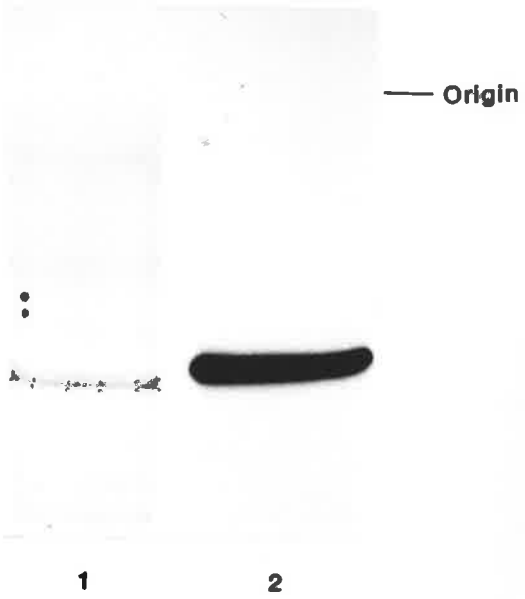


Figure V.4.

Preparative 5'-end labelling of α 1(1) and α 2(1) oligomers.

50 ng of each 21-mer were 5'-end labelled by kinasing (Section II.3.ix.) and fractionated on a 20% polyacrylamide gel containing 8.3 M urea. Bands were identified by autoradiography at room temperature for 10 seconds and excised with a scalpel. DNA was eluted into 200 µl TE containing 4 µg *E. coli* tRNA as carrier and concentrated by ethanol precipitation. Greater than 99% of the labelled molecules were recovered from the gel slice.


—Origin
α-1
α-2

recovered.

Each sample was resuspended in 30 µl H_2O and subjected to the base-specific cleavage reactions of Maxam and Gilbert (1980), modified as described in Section II.3.xii. The cleavage products were electrophoresed on a 20% DNA-sequencing gel (Figure V.5.) and bands identified by autoradiography.

Reading the sequencing ladders confirmed that the sequence of each primer was exactly as specified (Figures V.1. and V.2.), although the first two bases (from the 5'-end) were not able to be clearly resolved on the gel system used. However, since synthesis of the cDNA occurs from the 3'-hydroxyl group of the primer, even mis-match of both the two 5'-bases should not affect the ability of the primer to function, although its specificity may be slightly less.

The apparent one-base difference in size between the two primers (also evident in Figure V.4.) is an artifact. In each case only 21 bases are present in each sequencing ladder.

V.2.ii. Culture of Human Fibroblasts and Isolation of RNA

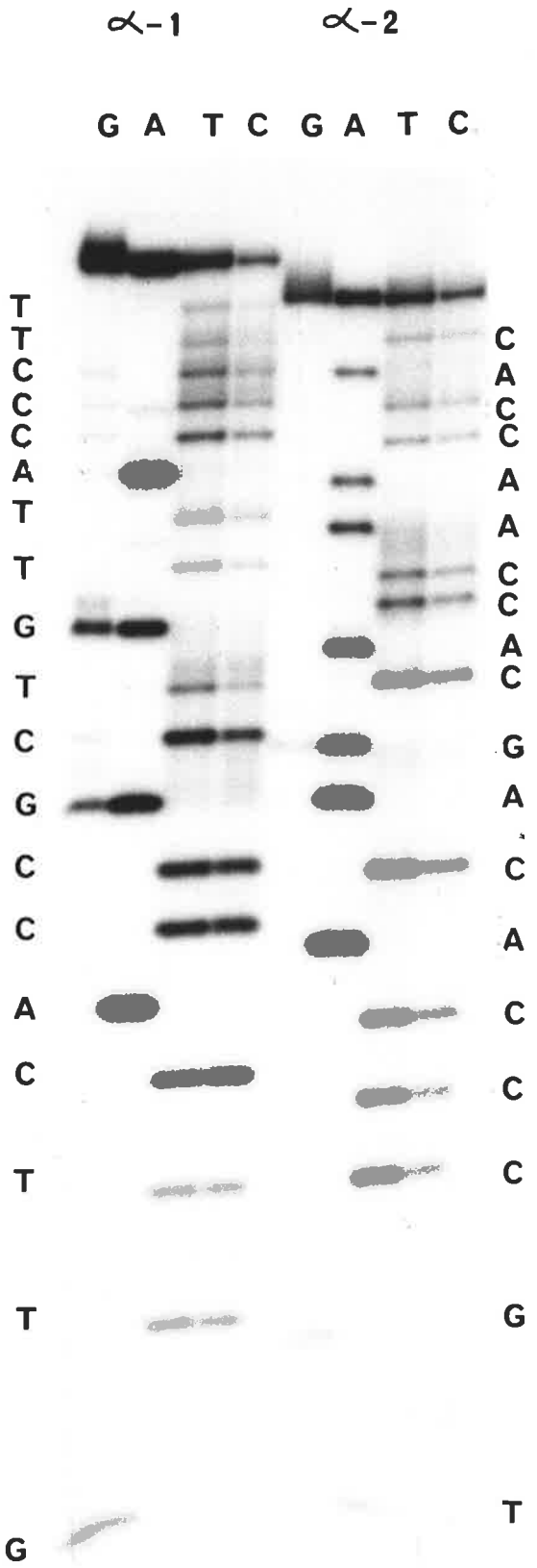
Human foreskin fibroblasts were obtained from R. Harris at approximately eight generations after establishment of the primary culture, and were grown as described in Section II.3.i. Several days prior to harvesting the cells for preparation of RNA, the culture medium was sometimes supplemented with 50 µg/ml ascorbic acid. This has been reported (Rowe and Schwarz, 1983) to cause an increase in the amount of type I pro-collagen mRNAs in cultured chick tendon cells.

Figure V.5.

DNA sequence determination of α 1(1) and α 2(1) oligomers.

End labelled 21-mers (Figure V.4.) were degraded using the base-specific cleavage reactions of Maxam and Gilbert (1980) and the products fractionated on a 20% polyacrylamide gel containing 8.3 M urea. Bands were identified by autoradiography at room temperature for 30 minutes.

The derived sequence of both 21-mers is written alongside the sequencing ladders.


α-1
α-2
G A T C G A T C
T
T
C
C
C
A
T
T
G
T
C
G
C
C
A
C
T
T
G
T
C
A
C
C
A
A
C
C
A
C
G
A
C
A
C
C
C
G
T

RNA was prepared by phenol exraction of cell lysates (Section II.3.iii.) and stored precipitated under 70% ethanol at -80°C. Typically, 150 µg cytoplasmic RNA/400 cm^2 roller bottle were prepared. The poly(A)$^+$ fraction was isolated by oligo-dT chromatography (Section II.3.iv.).

V.2.iii. Preparation of Vector

A number of vectors are available for the cloning of double stranded (ds) cDNA. Of these, perhaps the most widely and successfully used has been pBR322 (Bolivar et al., 1977). This vector is small, its complete DNA sequence is known (Sutcliffe, 1978) and it carries two antibiotic resistance genes. A number of unique restriction sites are present, enabling a number of different cloning strategies to be employed.

It was decided to clone the ds cDNA into the Pst 1 site of pBR322 by the addition of homopolymeric G-tails to the 3'-ends of the vector and C-tails to the 3'-ends of the ds cDNA, using terminal deoxynucleotidyl transferase (Villa-Komaroff et al., 1978). Not only is this method highly efficient, but recombinant colonies can be identified by their sensitivity to ampicillin and their resistance to tetracycline and the insert can be resected from the vector with Pst 1.

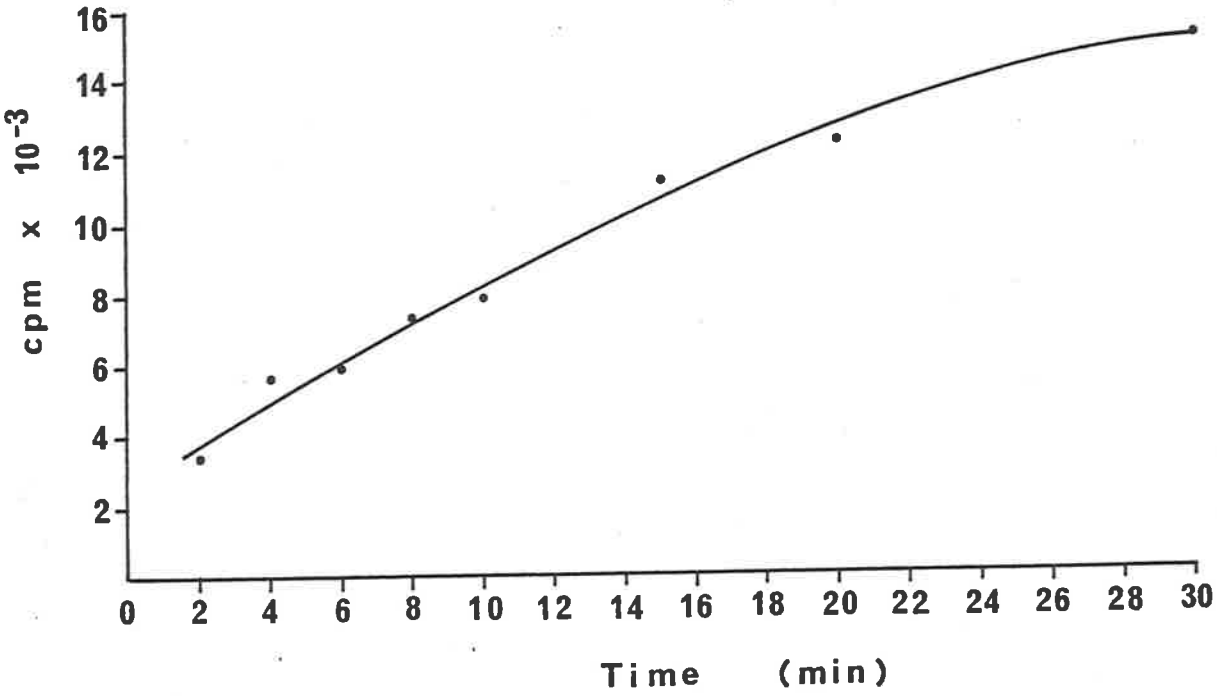
pBR322 DNA was prepared and the supercoiled form (i.e. un-nicked) isolated from a CsCl gradient (Section II.3.viii.). 12 µg of DNA were digested with Pst 1 and a 1 µg aliquot removed for a pilot tailing reaction (Section II.3.vi.). As shown in Figure V.6., incorporation of dG residues was essentially linear for up

Figure V.6.

Poly-dG-tailing of Pst 1 linearised pBR322.

1 µg of pBR322 DNA, linearised by digestion with Pst 1, was incubated with terminal deoxynucleotidyl transferase and ^{3}H-dGTP in a 50 µl volume (Section II.3.vi.). 5 µl aliquots were removed from the reaction at time intervals, and the amount of trichloro-acetic-acid-insoluble radioactivity determined.

From the plot, it was calculated that, for times up to 20 minutes, an average of approximately 6.6 guanine residues per minute were added to each 3'-end of the pBR322 molecules.


16
14
12
10
8
6
4
2
cpm x 10^{-3}
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Time (min)

to 20 minutes. It was calculated that an average of 6.6 residues were added per 3'-end per minute. Because a large excess of enzyme is used, the rate of tailing is independent of DNA concentration over a wide range.

15 dG residues/3'-end were added to the remaining 11 µg Pst 1 cut vector and the vector purified from LGT-agarose (Section II.3.vi.).

V.2.iv. Oligonucleotide Primed cDNA Synthesis

The distance from the priming sites of the 21-mers to the CAP sites of their respective mRNAs can be estimated by assuming that both the 5'-untranslated regions and the non-helical domain encoding regions of the human mRNAs are approximately the same size as equivalent regions in the chick mRNAs. This distance is approximately 1,400 bases for the α 1(1) mRNA and 2,200 bases for the α 2(1) mRNA. cDNA synthesised from the α 1(1) and α 2(1) mRNAs using the synthetic oligomers should, therefore, yield discrete products of these sizes. To investigate these cDNAs, a number of trial experiments were performed.

Both primers were preparatively 5'-end labelled by kinasing (Section II.3.ix.) and purified from a 20% polyacrylamide sequencing gel (Figure V.4.). DNA was recovered by ethanol precipitation without carrier tRNA. In a typical experiment, 2 ng each of labelled primers were annealed to 2 µg poly(A)$^+$ RNA for three hours at 41°C (Tm - ~ 29°C), the conditions adjusted for synthesis of cDNA and the reaction allowed to proceed for 45 minutes at 41°C (Section II.3.vi.). Synthetic products were analysed on 6% polyacrylamide DNA-sequencing gels (Section II.3.xii.). A

typical result is shown in Figure V.7. A short exposure of the autoradiogram revealed a smear of products, extending from < 20 bases up to approximately 1.5 kb, overlayed with a ladder of discrete bands. It is likely that these bands represent premature terminations of the synthesis reaction caused by the reverse transcriptase encountering secondary structure (Hagenbüchle et al., 1978) rather than reflecting breaks in the template. A longer exposure of the autoradiogram revealed that only a small number of the products extended beyond 1.5 kb. Generation of smears was found to be both template and reverse transcriptase dependent (not shown).

These results showed that, although cDNA may be extending from the α 1(1) primer to the CAP site of the pro α 1(1) mRNA, only a very small number (if any) of cDNA molecules extended from the α 2(1) primer to the CAP site of the pro α 2(1) mRNA. It was also clear that a significant proportion of the cDNA was small and so a size fractionation step would be required at a later stage of the cloning procedure. Attempts to extend synthesis of α 2(1) cDNA beyond 1.5 kb by varying both annealing and synthesis conditions were not successful (not shown).

Oligonucleotide primed cDNA synthesis was performed using 25 µg poly(A)$^+$ RNA as template exactly as described in Section II.3.vi. Two methods were used to synthesise the second strand.

V.2.v. Synthesis of the Second cDNA Strand

AMV-reverse transcriptase leaves, at the 3'-end of the newly synthesised cDNA, a short hairpin loop which is able

Figure V.7.

Oligonucleotide primed cDNA synthesis.

2 ng each of 5'-end labelled (Section II.3.ix.) α 1(1) and α 2(1) 21-mers (Figures V.1. and V.2.) were annealed to 2 μg poly(A)$^+$ RNA from human foreskin fibroblasts for three hours at 41°C and then cDNA synthesised using reverse transcriptase (Section II.3.vi.). Samples were phenol extracted, ethanol precipitated and fractionated on a 6% DNA-sequencing gel (Section II.3.xii.). Synthesis products were identified by autoradiography.

1: 5'-end labelled, Hinf 1 cut pBR322.

2: Oligomer primed cDNA, autoradiographed for 4 hours.

3: Track 2, autoradiographed for 24 hours.


O –
1
2
3

to act as a self-primer for the synthesis of the second strand (Efstratiadis et al., 1976). This loop, which covalently links the two strands of the double stranded cDNA, can be removed following second strand synthesis using the single-stranded-specific nuclease, S_1. It has, however, been reported (see Land et al., 1981) that using this method sometimes results in either loss or insertion of bases into the cDNA at the 5' (of the mRNA) end. To avoid the generation of such artifacts, an alternative strategy, involving the addition of a poly-dC tract to the 3'-end of the cDNA and the use of oligo-dG to prime synthesis of the second strand (Land et al., 1981), can be employed. It was decided to use both of these methods to synthesise the second strand.

After completion of synthesis of the first cDNA strand, the reaction volume was divided into two equal portions. In one sample, the RNA template was removed by boiling and the single stranded cDNA converted to a ds form using the self-priming method exactly as described in Section II.3.vi. The hairpin loop was removed by digestion with 2000 U S_1 nuclease (Section II.3.vi.) and the DNA phenol extracted and recovered by ethanol precipitation. The RNA template was degraded in the second aliquot by alkaline hydrolysis and, after neutralization, unincorporated nucleotides and oligonucleotide primer sequences were removed by centrifugation through a mini Sephadex G-50 column. Conditions were adjusted for tailing and an average of 20 dC residues were added per 3'-end (Secton V.2.iii.). Oligo-dG_8 was used to prime synthesis of the second strand (Section II.3.vi.).

Double stranded cDNA prepared by both methods was sized by sucrose gradient centrifugation and fractions collected as indicated in Figure V.8. The gradient profiles for each sample were identical and so corresponding fractions were pooled. DNA was recovered by ethanol precipitation.

V.2.vi. Tailing of ds cDNA, Annealing to Vector and Transformation

An average of 20 dC residues/3'-end were added to the ds cDNA from size class 2 (Figure V.8.) using the conditions determined whilst tailing the vector (Section V.2.iii.). Half the tailed ds cDNA was annealed to 150 ng dG-tailed vector (Section II.3.vi.) and aliquots of the annealed, circular molecules used to transform competent E. coli MC1061. A total of approximately 2,500 colonies were generated. Three hundred of these were examined for their resistance to ampicillin and tetracycline; all were amp.^{S} and tet.^{R} and thus represented true recombinants.

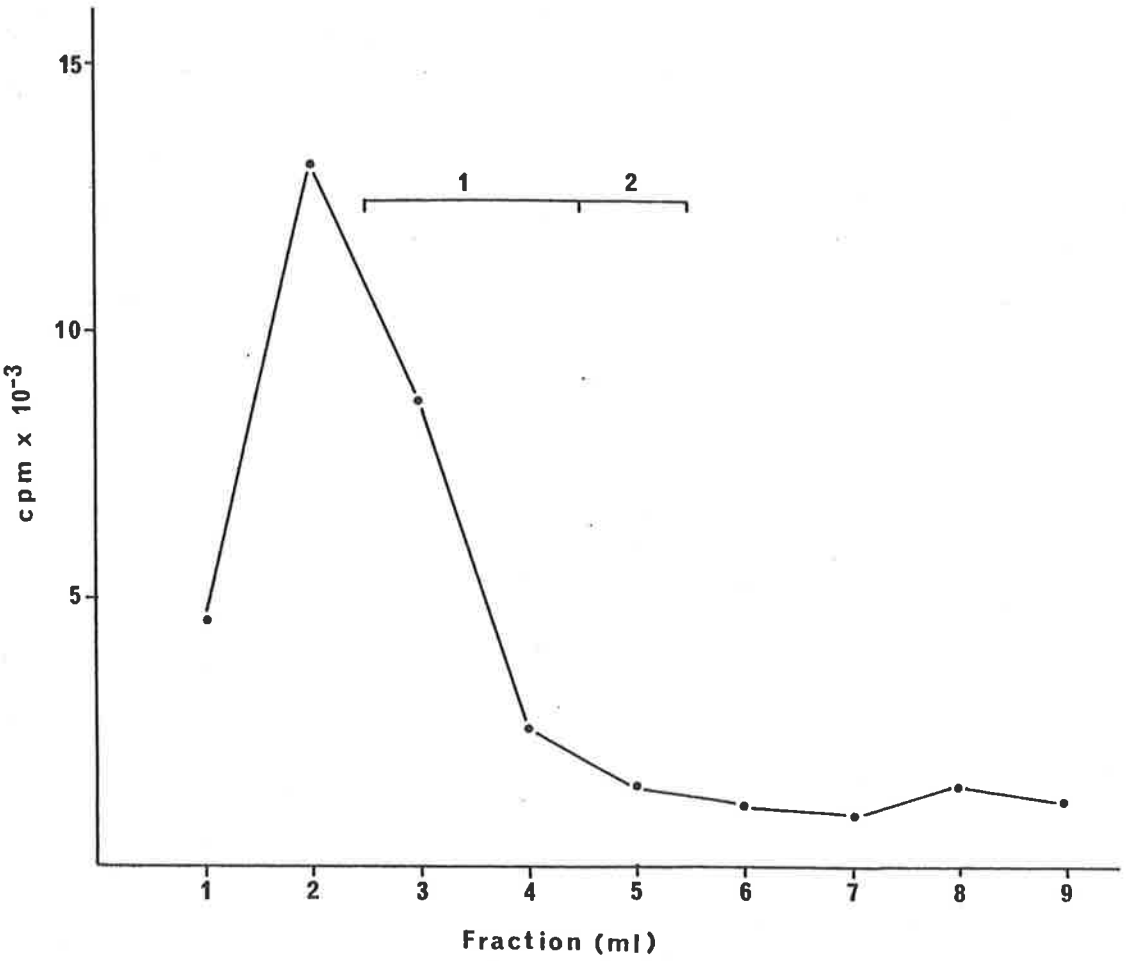
Detailed examination of these recombinants is discussed in the next chapter.

Figure V.8.

Size fractionation of ds cDNA.

Double stranded cDNA was loaded on to 10-40% linear sucrose gradients and centrifuged at 180,000 g for 16 hours at 4°C. Two size classes, 1 and 2 were selected and the DNA recovered by ethanol precipitation.

Only DNA from size class 2 was used in the preparation of recombinants.


15
10
5
cpm x 10⁻³
1
2
1
2
3
4
5
6
7
8
9
Fraction (ml)

CHAPTER VI

ISOLATION AND CHARACTERISATION OF RECOMBINANTS CONTAINING HUMAN COLLAGEN GENE SEQUENCES

VI.1. INTRODUCTION

The use of primers specific for pro α 1(I) and pro α 2(I) collagen sequences to prime the first strand cDNA synthesis in the construction of a library (Chapter V) should result in a very large proportion of the recombinants generated encoding collagen. Thus, only a small number of clones selected at random should need to be characterised before collagen encoding sequences could be identified. However, apart from this approach possibly being laborious, recombinants spanning regions not readily identifiable (such as those encoding the amino-propeptide, or the 5'-untranslated region of the mRNA) are likely to be discarded. It was therefore decided to identify recombinants suitable for further characterisation, by colony screening, using the primers as hybridization probe. Because the primers span helical region encoding sections of α 1(I) and α 2(I) collagen, regions of the recombinants adjacent to the primers will encode the characteristic (Gly-X-Y) pattern, readily identifiable by "translation" of their DNA sequence.

VI.2. RESULTS

VI.2.i. Colony Screening

Colonies containing recombinant molecules (Section V.2.vi.) were transferred to nitrocellulose filters by either of two methods. Those colonies which had been picked on to plates containing either ampicillin or tetracycline (Section V.2.vi.) were transferred by toothpick to circles of nitrocellulose on to which a grid pattern had been stamped. Each colony was transferred in duplicate.

The remaining colonies were lifted on to nitrocellulose filters using the technique developed for the transfer of 'phage (Benton and Davis, 1977). Bacteria were grown on the filters until the colonies were clearly visible and then lysed *in situ* using the method of Grunstein and Hogness (1975) as described in Section II.3.vii.

Clones were probed with a mixture of the α 1(I) and α 2(I) 21-mers, 5'-end labelled by kinasing (Section II.3.ix.), using the hybridization and washing conditions for primers described in Section II.3.v. The detection of four recombinants is shown in Figure VI.1.

A total of eighteen positive responses was observed.

VI.2.ii. Restriction Digestion and Hybridization Analysis

The colonies giving rise to these responses were picked and small-scale plasmid preparations performed on each one (Section II.3.vii.). As a preliminary step, it was decided to digest an aliquot of each plasmid DNA with EcoRI and to fractionate the products on an agarose gel rather than to try to resect the inserts with Pst 1 prior to analysis by electrophoresis. Although the G-C tailing procedure is designed to regenerate the Pst 1 site after annealing and ligation (*in vivo*), it has been reported that as few as 40% of DNA sequences inserted in this way are finally excisable by Pst 1 cleavage (Villa-Komaroff *et al.*, 1978). Furthermore, Pst 1 is more sensitive to the contaminants sometimes present in "mini-prep.'" DNA than is EcoRI. A small amount of EcoRI linearised pBR322 DNA was included in each track to serve as a size marker. An indication of the size of the insert is provided by

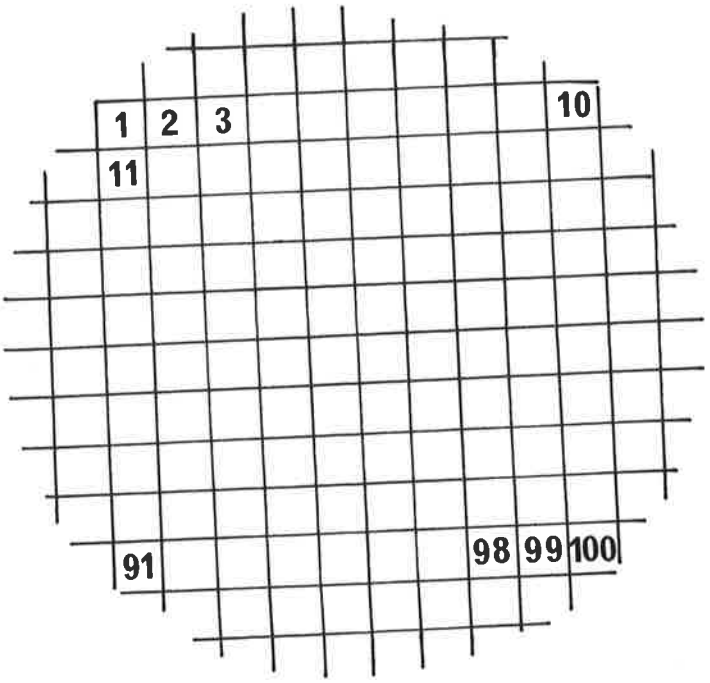
Figure VI.1.

Detection of sequences complementary to either the α 1(I) or the α 2(I) 21-mers amongst recombinants formed from human fibroblast RNA, using the colony screening procedure (Section II.3.vii.). Colonies containing recombinant plasmids were tooth-picked on to nitrocellulose filters, grown until clearly visible and then lysed in situ (Grunstein and Hogness, 1975). Filters were probed with ^{32}P-labelled α 1(I) and α 2(I) oligonucleotides, washed in 6 x SSC at 42°C and autoradiographed.

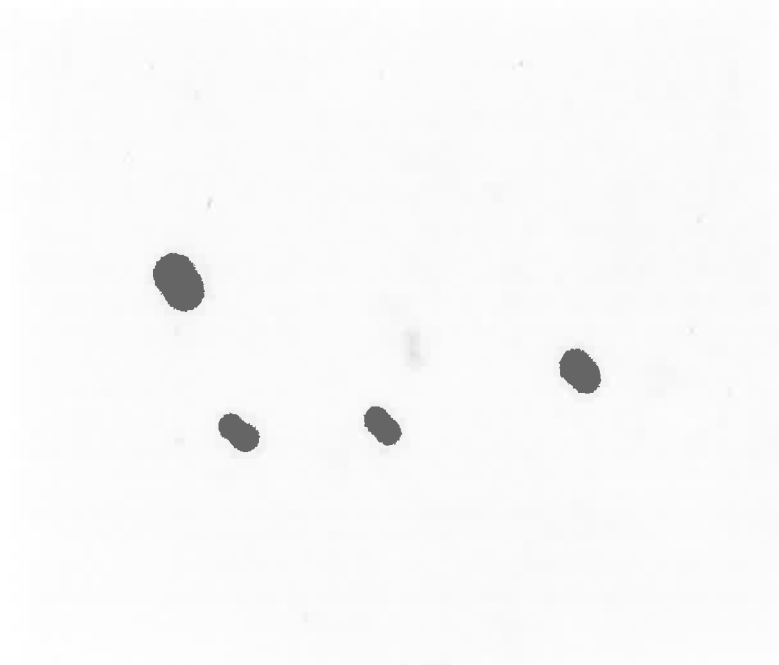
(i) The grid pattern on each nitrocellulose filter. One recombinant was grown, in duplicate, in each of the 100 squares of the grid.

(ii) Autoradiogram showing signals produced by hybridization of the probe to four recombinants. These four recombinants were picked from a master plate for further characterisation.

(i)



(ii)



observing the extent to which the recombinant plasmid migrates more slowly than pBR322. Figure VI.2.i. shows the pattern obtained for twelve of the recombinants. With the exception of clone number 12, a visible insert was evident in each track, although surprisingly (since the ds cDNA had been size fractionated) the inserts seemed to be quite small. The DNA was transferred to two nitrocellulose filters (Section II.3.v.) and probed with either kinased α 1(I) 21-mer or kinased α 2(I) 21-mer. The α 1(I) probe detected seven of the twelve recombinants and the α 2(I) probe, one recombinant (Figure VI.2.ii.). Failure to detect four of the recombinants was found to have resulted from selection of the wrong colonies from the master plates; the correct recombinants were subsequently located. Of the eighteen positives identified at the primary screening, seventeen cross-reacted with the α 1(I) probe and only one with the α 2(I) probe. One of the α 1(I) positives, called p1.57 (lane 3 in Figure VI.2.), and the α 2(I) positive, called p3.71, were chosen for further characterisation.

VI.2.iii. DNA Sequence Analysis

Large-scale plasmid DNA preparations were performed on p1.57 and p3.71 (Section II.3.viii.). A sample of each DNA was digested with Pst 1 and analysed on a polyacrylamide gel (not shown); both inserts were Pst 1 excisable. This result confirmed that the inserts were small; the α 1(I) insert was approximately 110 bp and the α 2(I) insert approximately 60 bp. In each case, the homopolymeric tails contribute approximately 20 bases/end to the size of the

Figure VI.2.

Mini-screen examination of plasmid recombinants.

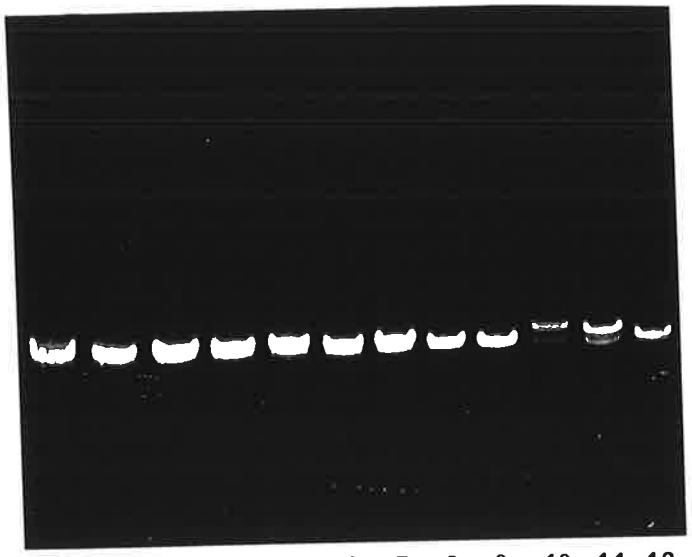
Colonies giving rise to positive responses by colony screening (Figure VI.1.) were picked from the master plates and plasmid DNA isolated by the "mini-prep.'" method (Section II.3.vii.). An aliquot of each was digested with EcoRI and electrophoresed on a 1% agarose gel. 50 ng of linear pBR322 DNA were included in each track as a size marker.

(i) EcoRI digested DNA from 12 of the positives, detected by ethidium bromide staining. Two bands (i.e. linear pBR322 and linear recombinant) were discernable in tracks 1-11 but were not evident in track 12.

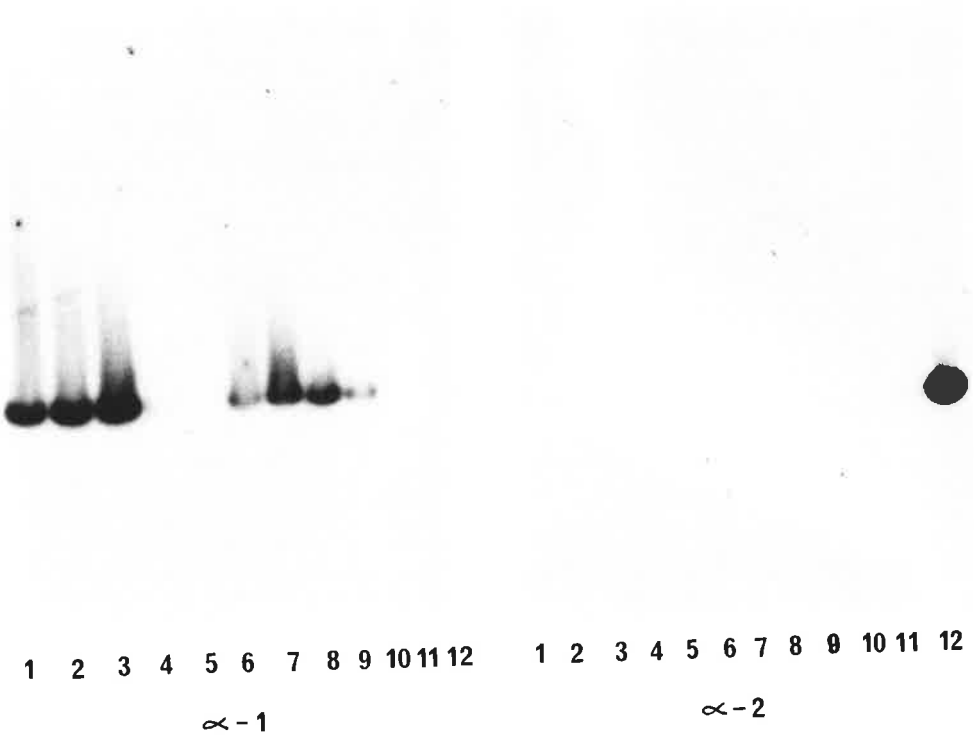
DNA was transferred to nitrocellulose filters (Section II.3.v.) and probed with either the α 1(I) 21-mer or the α 2(I) 21-mer, both radiolabelled by kinasing (Section II.3.ix.).

(ii) Autoradiograms showing detection of recombinant sequences by either the α-1 or α-2 probes.

(i)



(ii)



insert. Thus, the actual amount of potentially collagen-gene-derived sequence in each recombinant was ~ 70 bp for p1.57 and ~ 20 bp for p3.71. The latter size is consistent with this recombinant being derived from a molecule of α 2(I) 21-mer being dC-tailed, copied into ds form following oligo-dG_8 priming and subsequently retailed and cloned, and so this recombinant was discarded. It was decided to analyse the insert of p1.57 by DNA sequencing.

The insert was excised from p1.57 DNA and fractionated on a 5% polyacrylamide gel (Section II.3.v.). Insert DNA was recovered and ligated into Pst 1 digested M-13 mp83 vector (Section II.3.xiii.). Single stranded template was prepared from the resultant recombinant 'phage and subjected to di-deoxy sequencing procedures (Section II.3.xiv.). It was found that the sequence of the insert could not be determined using this strategy. The presence of the dC-dG tails flanking the insert severely disrupted the synthesis of the strand complementary to the template such that sequence could not be read (see Figure VI.3.). It was therefore necessary to adopt an alternative strategy to sequence this insert.

It was decided to use the α 1(I) synthetic oligomer as primer for DNA sequencing, as this should hybridize immediately adjacent to the tails at the 3'-end of the insert, on their 5'-side. An M-13 recombinant, with the insert in the orientation such that the α 1(I) 21-mer primer was complementary to it, was selected by a dot-blot assay using 5'-end labelled 21-mer as a probe (Section II.3.v.), and template prepared and sequenced (Figure VI.3.).

The sequence, shown in Figure VI.4., extends 41 bp

Figure VI.3.

DNA sequence determination of p1.57 insert.

5 µg of p1.57 DNA were digested with Pst 1, the insert purified from a 5% polyacrylamide gel (Section II.3.v.) and ligated into M-13 mp83 Pst 1 vector (Section II.3.xiii.). Recombinants were screened with 5'-end labelled α 1(I) 21-mer by a dot-blot assay (Section II.3.v.) and 'phage DNA prepared from one of the positives and used as template for di-deoxy sequencing using the α 1(I) 21-mer as a primer (Section II.3.xiv.). Products were resolved on a 6% DNA-sequencing gel and bands detected by autoradiography.

The sequence spans 41 bases of the insert from 3 dG residues adjacent to the primer to the poly-dc tract at the end of the insert.

Note the disruption to the sequencing ladder above the dc-tails.


T C G A

Figure VI.4.

DNA sequence determined from p1.57 (Figure VI.3.).

The sequence is shown 5' → 3'. Above are the deduced amino acid sequences in the three possible frames.

```
 GLY   LEU   LEU   SER   ILE   ASP   ARG   SER   GLU   GLY   ALA   ALA   LEU
   GLY   TYR   PHE   GLN   ***   ILE   ALA   ALA   ARG   GLU   LEU   LEU   CYS
     ALA   THR   PHE   ASN   ARG   SER   GLN   ARG   GLY   SER   CYS   SER   ALA
G G G C T A C T T T C A A T A G A T C G C A G C G A G G G A G C T G C T C T G C T
                10                  20                  30                  40
```

from the primer to the poly-dC tract at the end of the insert. Thus, the extent of the insert possibly derived from collagen gene sequences is 62 bp. The sequence was "translated" into an amino acid sequence in three reading frames (although, because the 3'-base of the primer was the third base of a codon, the 5'-base of the sequence should be the first base of the next codon) as shown in Figure VI.4. None of these reading frames (nor the three frames in the other direction) display the (Gly-X-Y) pattern characteristic of the helical region of collagen. Thus, p1.57 would appear not to have arisen from the specific priming of the α 1(I) 21-mer against human pro α 1(I) mRNA. This was an unexpected result and may indicate that the strategy used to construct the library had failed to generate predominantly collagen encoding clones. However, an alternative possibility was that p1.57, p3.71 and perhaps the other sixteen positive isolates were not representative of the library. The fact that these eighteen isolates all contained very small inserts even though the ds cDNA had been size fractionated was surprising. Analysis of eight clones picked at random from the library revealed that the sizes of their inserts ranged from approximately 300 bp to > 1 kb (not shown). Thus, in one sense at least, the eighteen positives were atypical of the library. It was therefore decided to examine clones that the α 1(I) and α 2(I) oligonucleotide probes had not detected.

A recombinant (called p03.3) with a typical insert size (approximately 580 bp) was selected at random and a large-scale plasmid preparation performed. Two strategies

were designed to sequence the insert of p03.3.

The first was to isolate insert, to remove the tails using the double-stranded-DNA-specific exonuclease, Bal-31, and to blunt-end ligate the truncated molecules into an M-13 vector. 10 µg of p03.3 were digested with Pst 1 and the digested DNA incubated with Bal-31 for a time (determined by a pilot experiment) sufficient to remove approximately 30 bp from each end of the DNA molecules. The ends were repaired to blunt-ends, the truncated insert molecules isolated from an LGT-agarose gel (Figure VI.5.i.) and cloned into the Sma 1 site of mp83. Recombinants in opposite directions were isolated by complementarity testing.

Although the first strategy enables the sequence of most of the insert to be determined, sequences adjacent to the G-C tails may be lost by Bal-31 digestion, and so an additional strategy was employed. 1.5 µg samples of whole p03.3 plasmid were digested with either Hae III, Fnu DII or Sau 3A and electrophoresed, alongside pBR322 digested with the same enzyme, on a 5% polyacrylamide gel (Figure VI.5.ii.). Fragments spanning the vector-insert junctions were isolated and cloned into the appropriate site of mp83.

Recombinants generated by both strategies were sequenced using the di-deoxy technique (Section II.3.xiv.). Initial results revealed a high G+C content in the sequence and so to avoid the possible mis-reading of sequences due to gel compressions, sequencing ladders were resolved on sequencing gels containing 25% formamide. Figure VI.6. shows a typical ladder on a formamide gel.

Figure VI.5.

Subcloning fragments spanning the insert of pO3.3 into M-13.

Strategy 1: 10 µg pO3.3 DNA were digested with Pst 1. A 1 µg aliquot was removed and the remaining 9 µg were digested with 1U Bal-31 in a 25 µl volume at room temperature for 10 seconds (conditions found, from a pilot experiment, to remove approximately 30 bp/end) and the reaction terminated by phenol extraction. DNA was recovered by ethanol precipitation and the ends repaired with Klenow (Section II.3.ix.). Fragments were electrophoresed on a 1% LGT-agarose gel (Section II.3.v.) and bands identified under UV by ethidium bromide staining:-

(i) - 1: 1 µg Pst 1 digested pO3.3 DNA.

2: 9 µg Pst 1, Bal-31 digested pO3.3 DNA.

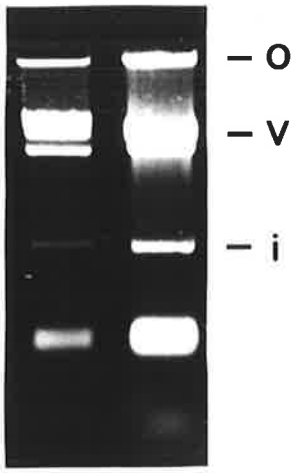
The small reduction in size of the insert (i) caused by the Bal-31 digestion is evident as a small increase in migration of this fragment in lane 2. This band was excised from the gel and the DNA recovered.

Strategy 2: 1.5 µg of either pO3.3 DNA or pBR322 DNA were digested with either Hae III, Fnu DII or Sau 3A and fractionated on a 6% polyacrylamide gel. Bands were detected by ethidium bromide staining.

(ii) Fnu DII digested pO3.3 DNA (O3.3) and pBR322 DNA (322). The 493 bp band in the 322 lane is absent in the O3.3 lane but is replaced by a band at ~ 670 bp. This band was excised from the gel and the DNA eluted into TE.

DNA fragments generated by both strategies were ligated into the appropriate M-13 vectors and transformed into JM101 (Section II.3.xiii.).

(i)



(ii)

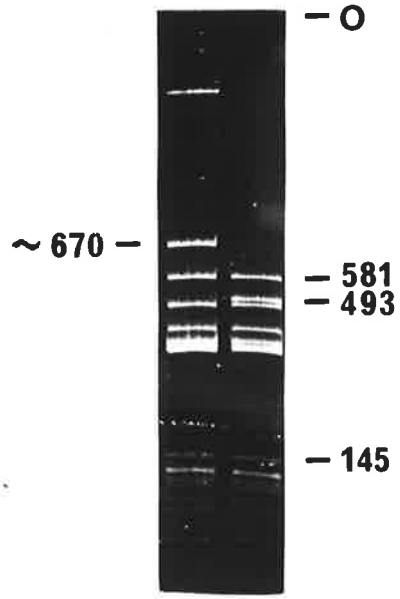
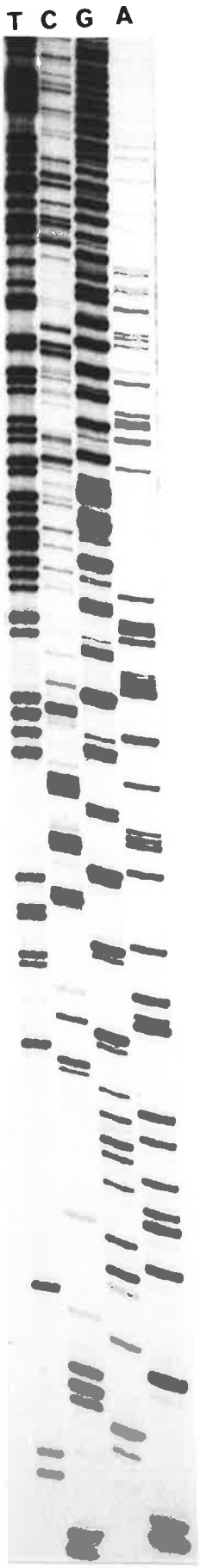


Figure VI.6.

DNA sequence determination of pO3.3 insert.

Sequences spanning the insert of pO3.3 were subcloned into mp83 using the strategies outlined in Figure VI.5. Sequences were generated using the di-deoxy method (Section II.3.xiv.) and sequencing ladders resolved on DNA-sequencing gels containing 25% formamide.

The sequence shown here begins at base 31 and extends 5' (Figure VI.7.).


T C G A

IV.2.iv. Examination of DNA Sequence

The sequence of 438 bp of the p03.3 insert is shown in Figure VI.7. This sequence was "translated" into the six possible reading frames. In one reading frame, shown in Figure VI.8., the sequence shows the pattern characteristic of the helical region of collagen, viz., $(Gly\text{-}X\text{-}Y)_n$ for 146 consecutive amino acids. Thus, this sequence encodes a collagen α-chain.

To determine which class of collagen p03.3 encodes, its deduced amino acid sequence was compared with the published amino acid sequences for a variety of classes from a variety of species. The best alignment was found to be for the α 2(I) chain (from rat and/or ox) between amino acids 379 and 524 of the helical region. Of the 87 rat and/or bovine amino acids known to span this region, perfect match was found with 81 residues from the deduced sequence. The homology between the human and rat and/or bovine sequences is summarised in Figure VI.9.

Thus, p03.3 clearly encodes the amino acid sequence spanning residues 379 to 524 of the human pro α 2(I) collagen chain. It was therefore appropriate to re-name this clone pHpC α2 (Human pro Collagen, α2). The deduced amino acid sequence of pHpC α2 provides the first primary sequence for this region of the human α 2(I) chain.

Whilst aligning the deduced amino acid sequence with the sequences of other collagen α-chains, a number of regions of sequence homology were observed between analogous regions of different chains. The most extensive of these is shown in Figure VI.10. Seventeen consecutive

Figure VI.7.

DNA sequence determined from p03.3.

The sequence, which spans 438 bp, is shown 5' → 3'.

```
        10          20          30          40          50          60
GGCCTCCCTG GCATCGACGG CAGGCCTGGC CCAATTGGCC CAGCTGGAGC AAGAGGAGAG

        70          80          90         100         110         120
CCTGGCAACA TTGGATTCCC TGGACCCAAA GGCCCCACTG GTGATCCTGG CAAAAACGGT

       130         140         150         160         170         180
GATAAAGGTC ATGCTGGTCT TGCTGGTGCT CGGGGTGCTC CAGGTCCTGA TGGAAACAAT

       190         200         210         220         230         240
GGTGCTCAGG GACCTCCTGG ACCACAGGGT GTTCAAGGTG GAAAAGGTGA ACAGGGTCCC

       250         260         270         280         290         300
GCTGGTCCTC CAGGCTTCCA GGGTCTGCCT GGCCCCTCAG GTCCCGCTGG TGAAGTTGGC

       310         320         330         340         350         360
AAACCAGGAG AAAGGGGTCT CCATGGTGAG TTTGGTCTCC CTGGTCCTGC TGGTCCAAGA

       370         380         390         400         410         420
GGGGAACGCG GTCCCCCAGG TGAGAGTGGT GCTGCCGGTC CTACTGGTCC TATTGGAAGC

       430
CGAGGTCCTT CTGGACCC
```

Figure VI.8.

Amino acid sequence deduced from the DNA sequence of pO3.3 (Figure VI.7.).

The DNA sequence of pO3.3 was "translated" into the six possible reading frames. In one frame, shown, the sequence showed the pattern characteristic of the helical region of a collagen α-chain, viz., $(Gly\text{-}X\text{-}Y)_n$.

```
 GLY   LEU   PRO   GLY   ILE   ASP   GLY   ARG   PRO   GLY   PRO   ILE   GLY   PRO   ALA   GLY   ALA   ARG   GLY   GLU
G G C C T C C C T G G C A T C G A C G G C A G G C C T G G C C C A A T T G G C C C A G C T G G A G C A A G A G G A G A G
                  10                  20                  30                  40                  50                  60

 PRO   GLY   ASN   ILE   GLY   PHE   PRO   GLY   PRO   LYS   GLY   PRO   THR   GLY   ASP   PRO   GLY   LYS   ASN   GLY
C C T G G C A A C A T T G G A T T C C C T G G A C C C A A A G G C C C C A C T G G T G A T C C T G G C A A A A A C G G T
                  70                  80                  90                 100                 110                 120

 ASP   LYS   GLY   HIS   ALA   GLY   LEU   ALA   GLY   ALA   ARG   GLY   ALA   PRO   GLY   PRO   ASP   GLY   ASN   ASN
G A T A A A G G T C A T G C T G G T C T T G C T G G T G C T C G G G G T G C T C C A G G T C C T G A T G G A A A C A A T
                  130                 140                 150                 160                 170                 180

 GLY   ALA   GLN   GLY   PRO   PRO   GLY   PRO   GLN   GLY   VAL   GLN   GLY   GLY   LYS   GLY   GLU   GLN   GLY   PRO
G G T G C T C A G G G A C C T C C T G G A C C A C A G G G T G T T C A A G G T G G A A A A G G T G A A C A G G G T C C C
                  190                 200                 210                 220                 230                 240

 ALA   GLY   PRO   PRO   GLY   PHE   GLN   GLY   LEU   PRO   GLY   PRO   SER   GLY   PRO   ALA   GLY   GLU   VAL   GLY
G C T G G T C C T C C A G G C T T C C A G G G T C T G C C T G G C C C C T C A G G T C C C G C T G G T G A A G T T G G C
                  250                 260                 270                 280                 290                 300

 LYS   PRO   GLY   GLU   ARG   GLY   LEU   HIS   GLY   GLU   PHE   GLY   LEU   PRO   GLY   PRO   ALA   GLY   PRO   ARG
A A A C C A G G A G A A A G G G G T C T C C A T G G T G A G T T T G G T C T C C C T G G T C C T G C T G G T C C A A G A
                  310                 320                 330                 340                 350                 360

 GLY   GLU   ARG   GLY   PRO   PRO   GLY   GLU   SER   GLY   ALA   ALA   GLY   PRO   THR   GLY   PRO   ILE   GLY   SER
G G G G A A C G C G G T C C C C C A G G T G A G A G T G G T G C T G C C G G T C C T A C T G G T C C T A T T G G A A G C
                  370                 380                 390                 400                 410                 420

 ARG   GLY   PRO   SER   GLY   PRO
C G A G G T C C T T C T G G A C C C
                  430
```

Figure VI.9.

Alignment of protein sequence deduced from the DNA sequence of pO3.3 with known α 2(I) collagen protein sequence.

The protein sequence, deduced from the DNA sequence of pO3.3 (Figure VI.8.), was aligned with bovine or rat α 2(I) collagen sequences (Fietzek and Kühn, 1976), spanning residues 379 to 524 of the helical region. The residues where the human sequence differs from the rat or bovine sequence are indicated; dashes represent those residues which are either common to both human and rat or ox, or are not known in these species (residues 406-425 and 451-489).

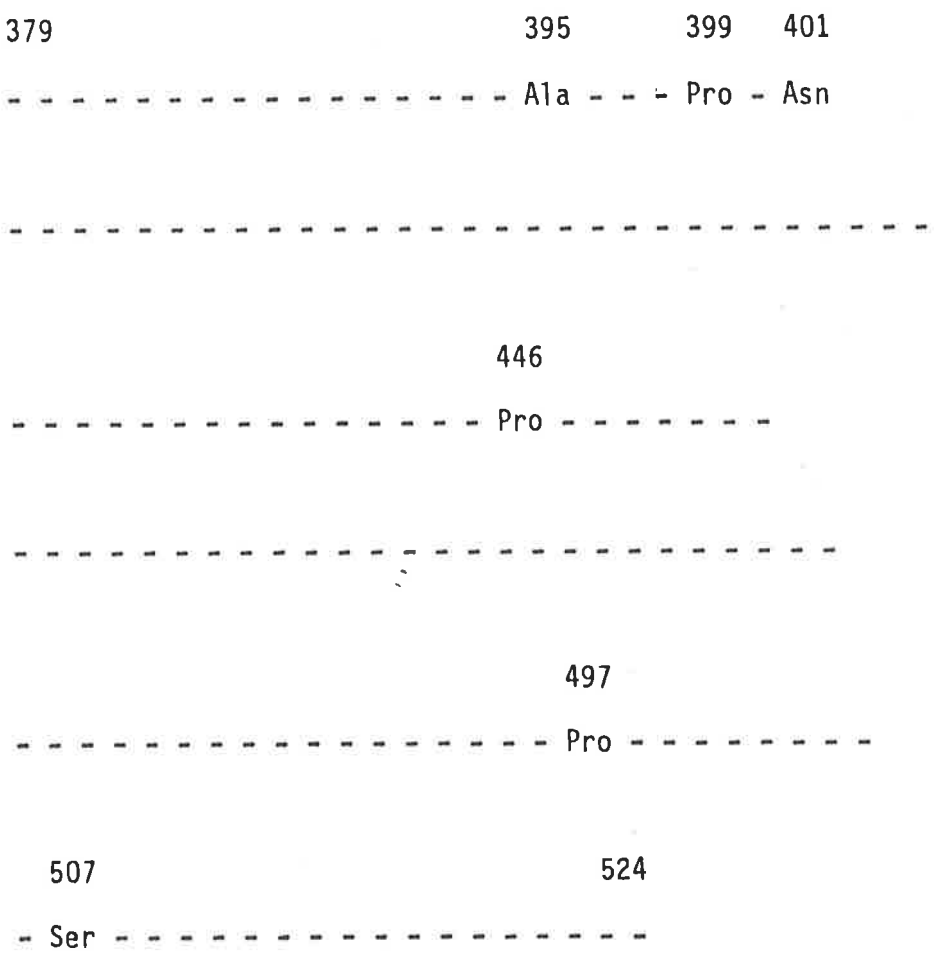

379
395
399
401
Ala
Pro
Asn
446
Pro
497
Pro
507
524
Ser

Figure VI.10.

Amino acid sequence homology between α 2(I), α 1(I) and α 1(III) peptides.

Comparison of the amino acid sequence of the chick, rat and bovine α 1(I), the bovine α 1(II) (Fietzek and Kühn, 1976) and human α 2(I) (deduced from the DNA sequence of pHpC α2) between residues 454 and 470. Residues identical to the α 2(I) sequence are indicated by dashes; amino acids which differ are named.

		454						460								470
human	α 2(I)	Gly	Glu	Gln	Gly	Pro	Ala	Gly	Pro	Pro	Gly	Phe	Gln	Gly	Leu	Pro	Gly	Pro
chick	α 1(I)	-	-	-	-	-	-	-	Ala	-	-	-	-	-	-	-	-	-
rat	α 1(I)	-	-	-	-	-	-	-	Ser	-	-	-	-	-	-	-	-	-
bovine	α 1(I)	-	-	-	-	-	-	-	Ser	-	-	-	-	-	-	-	-	-
bovine	α 1(II)	-	-	-	-	Ala	Hyp	-	-	Ser	-	-	-	-	-	-	-	-

residues of the human α 2(I) sequence can be aligned with the rat, chick and bovine α 1(I) sequence with only one residue change, and with the bovine α 1(II) chain with three changes. Shorter regions of perfect homology (i.e. no residue changes) were also observed, notably between residues 403 and 409 of the human α 2(I), chick, rat and bovine α 1(I), bovine α 1(II) and bovine α 1(III) chains. Regions outside these blocks showed, apart from having glycine as every third residue, no significant homology between the different chains.

VI.2.v. Discussion

Oligonucleotide primers complementary to sequences encoding regions of the human α 1(I) and α 2(I) collagen helix have been used to construct a cDNA library from human fibroblast mRNA. Although the size of the library was small, it is possible that many of its members encode collagen sequences as a recombinant (p03.3) selected at random was found to encode part of the human α 2(I) chain. Consequently, this clone was renamed pHpC α2.

Surprisingly, probing the library with the α 1(I) and α 2(I) primers detected only a small number of recombinants. In each case, these recombinants contained atypically small inserts. Sequence analysis of one of these revealed a sequence inconsistent with this recombinant having arisen from a specific priming event on a collagen mRNA. The reason why the recombinants that contain the primer sequence appear to be atypical is not known. It is unlikely to reflect some failure of the synthesis of the first strand of cDNA as alignment of the deduced amino acid

sequence of pHpC α2 with known α 2(I) sequences, positions its start at between residues 530 and 540, and since the primer extended between residues 536 and 542, it is likely that pHpC α2 arose from specific priming at this point.

pHpC α2 enables the first primary sequence of the human α-chain for the region it encodes to be deduced. In addition, it will be able to be used as a molecular probe to detect both human α 2(I) transcripts and the region of the human gene that it spans.

CHAPTER VII

FINAL DISCUSSION

The aim of the work described in this thesis has been to construct molecular probes for human type I collagen gene sequences. Initial attempts to isolate type I sequences from a human recombinant gene bank using cross-species hybridization probes proved unsuccessful, probably as a result of our aliquot of the library not being truly representative of the genome. An alternative strategy, involving the construction of a cDNA library from human fibroblast RNA, was undertaken. Synthetic oligonucleotide primers were employed to bias the library towards being enriched for the 5'-halves of type I collagen sequences. A recombinant, called pHpCα2, containing sequences encoding part of the human α 2(I) chain, was identified.

DNA sequence analysis of pHpCα2 enabled the primary structure of the human α 2(I) chain between residues 379 and 524 of the helical region to be deduced. These data expand upon those determined for both α 1(I) and α 2(I) chains from human and other species determined by other workers (see Section I.3.).

In addition to providing primary sequence data, pHpCα2 is able to be used as a molecular probe for the specific analysis of human pro α 2(I) collagen genes and their transcripts. Of particular interest is the analysis of potentially aberrant α 2(I) genes associated with primary collagenopathies (see Section I.4.i.). Determination of the exact site of a lesion in a type I collagen gene giving rise to an aberrant phenotype, and its correlation with the nature of that phenotype, should contribute significantly to an understanding of the complex interactions occurring in connective tissue.

Although most studies to date have only identified the nature of the lesion in a general sense [for example, Barsh *et al*. (1982) conclude from *in vivo* protein labelling experiments that one pro α 1(I) allele is non-functional in a case of O.I. I], recent approaches using gene probes have enabled some lesions to be precisely identified. The best documented of these is that of a 0.5 kb deletion in one pro α 1(I) allele, resulting in amino acids 325 to 410 being absent from the α 1(I) peptide (Chu *et al*., 1983). Trimers containing the shortened pro α 1(I) chain were rapidly degraded, resulting in a deficiency of type I collagen and the clinical symptoms of perinatal lethal O.I. Deletions at the 3'-end of the α 1(I)-like gene have also been identified as the causative lesion in several cases of perinatal lethal O.I. (Sykes, 1983).

The fact that in both these cases the genes encode aberrant collagens as the result of the presence of large deletions probably reflects the relative technical ease in which genes containing deletions can be identified. Experience with other gene systems, particularly the globin genes, indicates that a wide range of lesions are possible. Of particular interest, since collagen genes contain many introns, are lesions giving rise to aberrant splicing, a wide range of which have been identified in the globin system (reviewed by Mount and Steitz, 1983). However, identification of lesions other than large deletions using gene probes alone may not be technically straightforward, as the type I genes (and their primary transcripts) span approximately 40 kb. The most fruitful approach may be to localise the lesion using techniques such as high resolu-

tion peptide mapping and then to use gene probes to characterise the lesion in detail.

Ultimately, sequences spanning the entire type I mRNAs and genes will need to be constructed to fully characterise both normal and abnormal collagen genes and their expression. Although pHpCα2 was the only collagen-encoding recombinant examined in detail, it is likely that many more collagen-encoding clones are present in the cDNA library. Characterisation of additional isolates should enable identification of sequences spanning from the CAP sites to the priming sites (used during construction of the library), of both the α 1(I) and α 2(I) mRNAs.

However, even in the absence of complete type I sequences, some experiments can be performed on collagen genes. A preliminary experiment was performed in which 5 µg of human fibroblast, cytoplasmic RNA was glyoxylated, fractionated on an agarose gel, blotted on to nitrocellulose (Thomas, 1980) and probed with pHpCα2 insert (Figure VII.1.). A single 5.7 kb transcript was detected. This result is a little surprising, as Myers et al. (1983) have reported detecting four polyadenylated transcripts from the human α 2(I) gene, each differing in the length of their 3'-untranslated region. It is interesting to note that these workers examined total RNA whereas I have used only cytoplasmic RNA. Wickens and Gurdon (1983) have reported that partitioning of transcripts between the nucleus and cytoplasm occurs on the basis of their 3'-end. It is therefore possible that the 5.7 kb α 2(I) collagen transcript is the only one with a fully mature 3'-end. Further analysis should enable this hypothesis to be tested.

FIGURE VII.1.

Northern blot analysis of human α 2(I) transcripts.

5 µg of human fibroblast cytoplasmic RNA (Section II.3.iii.) were glyoxylated, fractionated on a 1% agarose gel and blotted on to nitrocellulose as described by Thomas (1980). The filter was probed with 10^6 cpm of pHpCα2 insert, labelled by nick translation (Section II.3.ix.), and washed at 0.2 x SSC/0.1% SDS at room temperature. Bands were detected by autoradiography.

Position of the 18S and 28S RNAs, determined by ethidium bromide staining an adjacent gel lane, is indicated.


O
28S
18S

In addition to enabling the size of transcripts to be determined, northern blot analysis is able to provide some information as to the abundance of a transcript. However, a more sensitive assay is to perform DNA-excess liquid hybridization followed by S_1 nuclease digestion and analysis of the amount of RNA-protected DNA. Radiolabelled, single stranded DNA synthesised from pHpCα2 sequences in M-13 templates would be a suitable probe to quantitate human α 2(I) transcripts. Alternatively, the synthetic 21-mers, 5'-end labelled, would also be suitable for this type of analysis.

pHpCα2 should also prove useful in the preparation of purified human pro α 2(I) mRNA. Single stranded pHpCα2 sequences bound to filters can be used to specifically isolate α 2(I) mRNA from a complex mixture of RNA, even if it is present in only small amounts. Specific mRNA can subsequently be eluted for further analysis (e.g. by translation).

Thus, both pHpCα2 and the oligonucleotide primers are useful, characterised tools for investigation of human type I collagen genes and their expression and it is intended that they should be freely available.

POSTSCRIPT

Although DNA sequence analysis of the 1.5 kb EcoRI fragment of Pnc-8 had failed to identify this fragment as encoding collagen sequences, the fact that it hybridized strongly to cDNA prepared from size fractionated, poly(A)+ RNA, isolated from chick embryo calvaria, suggested that it encoded some protein. With the advent of computer data bases containing large amounts of DNA sequence, it has become possible to compare sequences to virtually all the known DNA sequence and hence possibly identify the unknown sequence.

The sequence of the p1.5E insert was compared to the EMBL Nucleic Acid Data Base (Kindly performed by Dr. A. Rienser, C.S.I.R.O., Molecular Biology Division). Several stretches of homology were observed, but the most significant of these was a continuous stretch (if one base was removed from the Pnc-8 sequence) of 77 bases between the most 3'-end of Pnc-8 and part of Col H1.1 (Bernard et al., 1983) encoding amino acids (numbered from the carboxyl terminal) 24 to 49 of the human pro α 2(I) collagen chain. This homology is shown in Figure P. 1.

Significant homology between the remaining ~1400 bases of Pnc-8 and Col H1.1 were not observed. This observation, combined with the hybridization results using cDNA probes to both Pnc-8 and λ 122 suggests that Pnc-8 may be an aberrant clone, although to confirm this would require the demonstration of a difference of restriction pattern between Pnc-8 and uncloned human genomic DNA. An alternative and intriguing possibility is that Pnc-8 encodes an, as yet unidentified, protein which either fortuitiously, or for functional reasons, shares homology with the α 2(I) collagen chain. Further analysis would be required to address this possibility.

Figure P.1.

Comparison of the sequence of the 1.5 kb EcoRI fragment with those of a data base.

The sequence of the 1.5 kb EcoRI fragment of Pnc-8 was compared to the sequences in the EMBL Nucleic Acid Data Base. The most significant homology was between the seqence at the 3'-end of Pnc-8 and two blocks of ColH1.1 (Bernard et al., 1983). Removal of the G residue at position 1406 in Pnc-8 (indicated thus : ↑) enabled a continuous homology of 75 bp to be found.

```
                2109       2119       2129       2139       2149       2159
COLH1.1    > CAATCATTGA ATACAAAACA AATAAGCCAT CACGCCTGCC CTTCCTTGAT ATTTAAA|TA
               *    * *     ***    ********** ********** *********  ***  * *
PNC8.1     > AATCATTGAA TACAAAACAG AATAAGCCAT CACGCCTGCC CTTCCTTGAG ATTGCACCTT
             1386       1396       1406       1416       1426       1436
```

```
          2064       2074       2084       2094       2104       2114       2124       2134       2144
COLH1.1  > TTCTTGTAGA TGGCTGCTCT AAAAAGACAA ATGAATGGGG AAAGACAATC ATTGAATACA AAACAAATAA GCCATCACGC CTGCCC
           ** *  *    *          ********** ********** ********** ********** ***** *  *   *        *   **
PNC8.1   > TTTTCTTCTC TTTGAAACAG AAAAAGACAA ATGAATGGGG AAAGACAATC ATTGAATACA AAACAGAATA AGCCATCACG CCTGCC
          1340       1350       1360       1370       1380       1390       1400       1410       1420
                                                                              ▲
                                                                              |
```

SUMMARY

1. The work presented in this thesis describes the construction of gene probes specific for human type I procollagen sequences. Such probes enable the genes (and their transcripts) encoding both normal and aberrant collagens to be examined. Correlation of the nature of the defects giving rise to abnormal collagens with the changes in the overall properties of connective tissue may contribute to an understanding of the complex interactions in this tissue.

2. A recombinant human genomal library was screened with two, independent, cross-species collagen gene probes, viz., part of the sheep pro α 2(I) gene (derived from the genomal clone SpC3 - gift from P. Tolstoshev) and cDNA made from fractionated chick embryo calvaria RNA. A recombinant, called Pnc-8, was isolated and characterised by restriction mapping and hybridization analysis. A region contained on a 1.5 kb EcoRI fragment of Pnc-8 was identified as probably containing collagen gene sequences, by virtue of its strong hybridization to cDNA.

3. DNA sequence analysis of the 1.5 kb EcoRI fragment of Pnc-8 failed to reveal identifiable collagen gene sequences. It seemed likely that this inability to identify collagen sequences resulted from the fact that these sequences spanned one end of the gene (for which protein sequence data were not available) and so could not be recognised as encoding collagen. The 1.5 kb fragment was therefore used as a probe to re-screen the genomic library in an attempt to "chromosome crawl" into the rest of the putative collagen gene. Although a potentially

overlapping clone, λ122, was isolated, it failed to cross-react with cDNA and thus was unlikely to encode collagen sequences.

These results suggested that our aliquot of the human library was not truly representative of the human genome and so an alternative approach for the isolation of recombinants encoding human collagen gene sequences was investigated.

4. Oligonucleotide primers complementary to helix encoding regions of type I collagen genes were synthesised and used to prime cDNA synthesis from RNA isolated from human fibroblasts. Double stranded cDNA was used to make a library of plasmid recombinants. Characterisation of one of the recombinants identified a sequence encoding approximately 146 amino acids of the human pro α 2(I) gene.

5. The use of recombinants isolated from the cDNA library to analyse collagen genes and their transcripts is discussed.

REFERENCES

Adams, E. **Science**, 202, 591, 1978.

Adams, S.L., Alwine, J.C., de Crombrugghe, B. and I. Pastan. **J. Biol. Chem.**, 4935, 1979.

Aho, S., Tate, V. and H. Boedtker. **Nuc. Acids Res.**, 11, 5443, 1983.

Alwine, J.C., Kemp, D.J. and G.R. Stark. **Proc. Natl. Acad. Sci. USA**, 74, 5350, 1977.

Anderson, S., Gaït, M.J., Mayol, L. and I.G. Young. **Nuc. Acids Res.**, 8, 1731, 1980.

Ashhurst, D.E. and A.J. Bailey. **Eur. J. Biochem.**, 103, 75, 1980.

Aviv, H. and P. Leder. **Proc. Natl. Acad. Sci. USA**, 69, 1408, 1972.

Avvedimento, E., Yamada, Y., Lovelace, E., Vogeli, G., de Crombrugghe, B. and I. Pastan. **Nuc. Acids Res.**, 9, 1123, 1981.

Baird, M., Driscoll, C., Schreiner, H., Sciarratta, G.V., Sanscone, G., Niazi, G., Ramirez, F. and A. Bank. **Proc. Natl. Acad. Sci. USA**, 78, 2418, 1981.

Barsh, G.S. and P.H. Byers. **Proc. Natl. Acad. Sci. USA**, 78, 5142, 1981.

Barsh, G.S., David, K.E. and P.H. Byers. **Proc. Natl. Acad. Sci. USA**, 79, 3838, 1982.

Bauer, E.A., Gedde-Dahl, Jr., T. and A.Z. Eisen. **J. Invest. Derm.**, 68, 119, 1977.

Beaucage, S.L. and M.H. Caruthers. **Tetrahedron Letts.**, 22, 1859, 1981.

Benton, W.D. and R.W. Davis. **Science**, 196, 180, 1977.

Bentz, H., Morris, N.P., Murray, L.W., Sakai, L.Y., Hollister, D.W. and R.E. Burgeson. **Proc. Natl. Acad. Sci. USA**, 80, 3168, 1983.

Benveniste, K., Wilczek, J. and R. Stern. **Nature**, 246, 303, 1973.

Benya, P.D. and J.D. Shaffer. **Cell**, 30, 215, 1982.

Bernard, M.P., Myers, J.C., Chu, M-L., Ramirez, F., Eikenberry, E.F. and D.J. Prockop. **Biochem.**, 22, 1139, 1983.

Birnboim, H.C. and J. Doly. **Nuc. Acids Res.**, 7, 1513, 1979.

Blattner, F.R., Williams, B.G., Blechl, A.E., Denniston-Thompson, K., Faber, H.E., Furlong, L-A., Grunwald, D.J., Kiefer, D.O., Moore, D.D., Schumm, J.W., Sheldon, E.L. and O. Smithies. **Science** 196, 161, 1977.

Blobel, G. and B. Dobberstein. **J. Cell Biol.**, 67, 835, 1975.

Boedtker, H., Crkvenjakov, R.B., Last, J.A. and P. Doty. **Proc. Natl. Acad. Sci. USA**, 71, 4208, 1974.

Bolivar, F., Rodriguez, R.L., Greene, P.J., Betlach, M.C., Heymeker, H.L. and H.W. Boyer. **Gene**, 2, 95, 1977.

Bolivar, F. **Gene**, 4, 121, 1978.

Bornstein, P. and H. Sage., **Ann. Rev. Biochem.**, 49, 957, 1980.

Borstein, P. and W. Traub. **In** "The Proteins", Neurath, H. and R.L. Hill, eds., Vol. 4 411, Academic Press Inc., New York, 1979.

Boyd, C.D., Tolstoshev, P., Schafer, M.P., Trapnell, B.C., Coon, H.C., Kretschmer, P.J., Nienhuis, A.W. and R.G. Crystal. **J. Biol. Chem.** , 255, 3212, 1980.

Braconnot, H. **Annls. Chim. Phys.**, 13, 113, 1820. **Cited In**, "Treatise On Collagen", Ramachandran, G.N., ed., Vol. 1, Academic Press Inc., New York, 1968.

Breathnach, R. and P. Chambon. **Ann. Rev. Biochem.**, 50, 349, 1981.

Breul, S.D., Bradley, K.H., Hance, A.J., Schafer, M.P., Berg, R.A. and R.G. Crystal. **J. Biol. Chem.**, 255, 5250, 1980.

Brown, R.A., Shuttleworth, C.A. and J.B. Weiss. **Biochem. and Biophys. Res. Commun.**, 80, 866, 1978.

Byers, P.H., Holbrook, K.A., Hall, J.G., Bornstein, P. and J.W. Chandler. **Hum. Genet.**, 40, 157, 1978.

Byers, P.H., Siegel, R.C., Holbrook, K.A., Narayanan, A.S., Bornstein, P. and J.G. Hall. **New Eng. J. Med.**, 303, 61, 1980.

Byers, P.H., Holbrook, K.A., Barsh, G.S., Smith, L.T. and P. Bornstein. **Lab. Invest.**, 44, 336, 1981a.

Byers, P.H., Siegel, R.C., Paterson, K.E., Rowe, D.W., Holbrook, K.A., Smith, L.T., Chang, Y-H. and J.C.C. Fu. **Proc. Natl. Acad. Sci. USA**, 78, 7745, 1981b.

Chang, C.J., Temple, G.F., Trecartin, R.F. and Y.W. Kan. **Nature**, 281, 602, 1979.

Cheung, D.T., DiCesare, P., Benya, P.D., Libaw, E. and M.E. Nimni. **J. Biol. Chem.**, 258, 7774, 1983.

Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and W.J. Rutter. **Biochem.**, 24, 5294, 1979.

Chopra, R.K. and V.S. Ananthanarayanan. **Proc. Natl. Acad. Sci. USA**, 79, 7180, 1982.

Chu, M-L., Myers, J.C., Bernard, M.P., Ding, J-F. and F. Ramirez. **Nuc. Acids Res.**, 10, 5925, 1982a.

Chu, M-L., Myers, J.C. and F. Ramirez. **Fed. Proc.**, 41, 852, 1982b.

Chu, M-L., Williams, C.J., Pepe, G., Hirsch, J.L., Prockop. D.J. and F. Ramirez. **Nature**, 304, 78, 1983.

Chung, E. and E.J. Miller. **Science**, 183, 1200, 1974.

Clarke, L. and J. Carbon. **Cell**, 9, 91, 1976.

Clewell, D.B. **J. Bacteriol.**, 110, 667, 1972.

Comi, P., Giglioni, B., Barbarano, L., Ottolenghi, S., Williamson, R., Novakova, M. and G. Masera. **Eur. J. Biochem.**, 79, 617, 1977.

Courey, A.J. and J.C. Wang. **Cell**, 33, 817, 1983.

Dalgleish, R., Trapnell, B.C., Crystal, R.G. and P. Tolstoshev. **J. Biol. Chem.**, 257, 13816, 1982.

Davidson, J.M., McEneany, L.S.G. and P. Bornstein. **Biochem.**, 14, 5188, 1975.

Davis, R.W., Botstein, D. and J.R. Roth. "Advanced Bacterial Genetics", C.S.H. Press, New York, 1980.

Deak, S.B., Chu, M-L., Myers, J.C., Nicholls, A.C., Pope, F.M., Rowe, D. and D.J. Prockop. **Fed. Proc.**, 41, 3402, 1982.

De Biasi, S. and F. Pilotto. **J. Submicr. Cytol.**, 8, 337, 1976.

Dickson, L.A., Ninomiya, Y., Bernard, M.P., Pesciotta, D.M., Parsons, J., Green, G., Eikenberry, E.F., de Crombrugghe, B., Vogeli, G., Pastan, I., Fietzek, P.P. and B.R. Olsen. **J. Biol. Chem.**, 256, 8407, 1981.

Diegelmann, R.F., Guzelian, P.S., Gay, R. and S. Gay. **Science**, 219, 1343, 1983.

Di Ferrante, N., Leachman, R.D., Angelini, P., Donnelly, P.V., Francis, G. and A. Almazan. **Connect. Tissue Res.**, 3, 49, 1975.

Dörper, T. and E-L. Winnacker. **Nuc. Acids. Res.**, 11, 2575, 1983.

Driesel, A.J., Schumacher, A.M. and R.A. Flavell. **Hum. Genet.**, 62, 175, 1982.

Duchene, M., Sobel, M.E. and P.K. Müller. **Expt. Cell Res.**, 142, 317, 1982.

Dugaiczyk, A., Boyer, H.W. and H.M. Goodman. **J. Mol. Biol.**, 96, 171, 1975.

Early, P.. Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R. and L. Hood. **Cell**, 20, 313, 1980.

Edwards, M.K. and W.B. Wood. **Develop. Biol.**, 97, 375, 1983.

Efstratiadis, A., Kafatos, F.C., Maxam, A.M. and T. Maniatis. **Cell**, 7, 279, 1976.

Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., De Riel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and N.J. Proudfoot. **Cell**, 21, 653, 1980.

Ehrenberg, L., Fedorcsak, I. and F. Solymosy. **Proc. Nuc. Acid Res. Mol. Biol.**, 16, 189, 1976.

Elgin, S.C.R. **Cell**, 27, 413, 1981.

Eyre, D.R. **Clin. Orthopaed. Rel. Res.**, 159, 97, 1981.

Fessler, J.H. and L.I. Fessler. **Ann. Rev. Biochem.**, 47, 129, 1978.

Fietzek, P.P. and K. Kühn. **Int. Rev. Connect. Tiss. Res.**, 7, 1, 1976.

Fietzek, P.P., Rexrodt, F.W., Wendt, P., Stark, M. and K. Kühn. **Eur. J. Biochem.**, 30, 163, 1972.

Flavell, R.A., Bernards, R., Kooter, J.M., de Boer, E., Little, P.F.R., Annison, G. and R. Williamson. **Nuc. Acids Res.**, 6, 2749, 1979.

Fleischmajer, R., Olsen, B.R., Timpl, R., Perlish, J.S. and O. Lovelace. **Proc. Natl. Acad. Sci. USA**, 80, 3354, 1983.

Francis, M.J.O., Williams, K.J., Sykes, B.C. and R. Smith. **Clin. Sci.**, 60, 617, 1981.

Frischauf, A.M., Lehrach, H., Rosner, C. and H. Boedtker. **Biochem.**, 17, 3243, 1978.

Fritsch, E.F., Lawn, R.M. and T. Maniatis. **Nature**, 279, 589, 1979.

Fuller, F. and H. Boedtker. **Biochem.**, 20, 996, 1981.

Furthmayr, H., Wiedemann, H., Timpl, R., Odermatt, E. and J. Engel. **Biochem. J.**, 211, 303, 1983.

Furuto, D.K. and E.J. Miller. **J. Biol. Chem.**, 255, 290, 1980.

Furuto, D.K. and E.J. Miller. **Biochem.**, 20, 1635, 1981.

Gay, S., Martin, G.R., Müller, P.K., Timpl, R. and K. Kühn. **Proc. Natl. Acad. Sci. USA**, 73, 4037, 1976.

Gilbert, W. **Nature**, 271, 501, 1978.

Glimcher, M.J. and S.M. Krane. **The Organisation and Structure of Bone and the Mechanism of Calcification.** **In** "Treatise on Collagen", Ramachandran, G.M. and B.S. Gould, eds. Vol. 2, Pt. B., Academic Press Inc., New York, 1968.

Goka, T.J., Stevenson, R.E., Hofferan, P.M. and R.R. Howell. **Proc. Natl. Acad. Sci. USA**, 73, 604, 1976.

Goodman, M., Greenspon, S.A. and C.A. Krakower. **J. Immunol.**, 75, 96, 1955.

Gross, J. **Some Aspects of the Biology of the Extracellular Matrix.** **In** "Gene Families of Collagen and Other Proteins", Prockop, D.J. and P.C. Champe, eds. Elsevier, North-Holland, 1980.

Gross-Bellard, M., Oudet, P. and P. Chambon. **Eur. J. Biochem.**, 36, 32, 1973.

Grunstein, M.L. and D.S. Hogness. **Proc. Natl. Acad. Sci. USA**, 72, 3961, 1975.

Guerry, P., Le Blanc, D.J. and S. Falkow. **J. Bacteriol.**, 116, 1064, 1973.

Hagenbüchle, O., Santer, M., Steitz, J.A. and R.J. Mans. **Cell**, 13, 551, 1978.

Hall, C.E., Jakus, M.A. and F.O. Schmitt. **J. Am. Chem. Soc.**, 64, 1234, 1942.

Hall, Z.W. and R.B. Kelly. **Nature New Biol.**, 232, 62, 1971.

Harding, J.D., MacDonald, R.J., Przybyla, A.E., Chirgwin, J.M. Pictet, R.L. and W.J. Rutter. **J. Biol. Chem.**, 252, 7391, 1977.

Harvey, R.P., Krieg, P.A., Robins, A.J., Coles, L.S. and J.R.E. Wells. **Nature**, 294, 49, 1981.

Harwood, R. **Int. Rev. Connect. Tissue Res.**, 8, 159, 1979.

Hodge, A.J. and F.O. Schmitt. **Proc. Natl. Acad. Sci. USA**, 46, 186, 1960.

Hofmann, H., Fietzek, P.P. and K. Kühn. **J. Mol. Biol.**, 125, 137, 1978.

Hörlein, D., Fietzek, P.P. and K. Kühn. **FEBS Letts.**, 89, 279, 1978.

Hörlein, D., McPherson, J., Goh, S.H. and P. Bornstein. **Proc. Natl. Acad. Sci. USA**, 78, 6163, 1981.

"Instructions to Authors". **Biochem. J.**, 169, 2, 1978.

Jander, R., Rauterberg, J., Voss, B. and D.B. von Bassewitz. **Eur. J. Biochem.**, 114, 17, 1981.

Jander, R., Rauterberg, J. and R.W. Glanville. **Eur. J. Biochem.**, 133, 39, 1983.

Kafatos, F.C., Jones, W.C. and A. Efstratiadis. **Nuc. Acids Res.**, 7, 1541, 1979.

Kang, A.H. and R.L. Trelstrad. **J. Clin. Invest.**, 52, 2571, 1973.

Kidd, V.J., Wallace, R.B., Itakura, K. and S.L.C. Woo. **Nature**, 304, 230, 1983.

Kramer, J.M., Cox, G.N. and D. Hirsh. **Cell**, 30, 599, 1982.

Krane, S.J., Pinnell, S.R. and R.W. Erbe. **Proc. natl. Acad. Sci. USA**, 69, 2899, 1972.

Kream, B.E., Rowe, D.W., Gworek, S.C. and L.G. Raisz. **Proc. Natl. Acad. Sci. USA**, 77, 5654, 1980.

Kretschmer, P.J., Kaufman, R.E., Coon, H.C., Chen, M-J. Y., Geist, C.E. and A.W. Nienhuis **J. Biol. Chem.**, 255, 3204, 1980.

Krieg, P.A., Robins, A.J., Gait, M.J., Titmas, R.C. and J.R.E. Wells. **Nuc. Acids Res.**, 10, 1495, 1982.

Land, H., Grez, M., Hauser, H., Lindenmaier, W. and G. Schütz. **Nuc. Acids Res.**, 9, 2251, 1981.

Lawn, R.M., Fritsch, E.F., Parker, R.C., Blake, G. and T. Maniatis, **Cell**, 15, 1157, 1978.

Lehrach, H., Frischauf, A.M., Hanahan, D., Wozney, D., Fuller, F., Crkvenjakov, R., Boedtker, H. and P. Doty. **Proc. Natl. Acad. Sci. USA**, 75, 5417, 1978.

Lehrach, H., Frischauf, A.M., Hanahan, D., Wozney, J., Fuller, F. and H. Boedtker. **Biochem.**, 18, 3146, 1979.

Lichtenstein, J.R., Martin, G.R., Kohn, L., Byers, P.H. and V.A. McKusick. **Science**, 182, 298, 1973.

Lillehaug, J.R., Kleppe, R.K. and K. Kleppe. **Biochem.**, 15, 1858, 1976.

Madri, J.A., Foellmer, H.G. and H. Furthmayr. **Biochem.**, 22, 2797, 1983.

Magee, B. and B. Moore **in** "Mackie and McCartney : Medical Microbiology 13E, **Vol.** 2", eds. Collee, J.G., Duguid, J.P. and B.P. Marmion, Churchill Livingstone, Lond., 1984 **in press.**

Maniatis, T., Jeffrey, A. and D.G. Kleid. **Proc. Natl. Acad. Sci. USA**, 72, 1184, 1975.

Maniatis, T., Hardison, R.C., Lacy, E., Lauer, J. O'Connell, C., Quon, D., Sim, G.K. and A. Efstratiadis. **Cell**, 15, 687, 1978.

von der Mark, H., von der Mark, K. and S. Gay. **Develop. Biol.**, 48, 237, 1976.

Marshall, A.J. and L.A. Burgoyne. **Nuc. Acids Res.**, 3, 1101, 1976.

Maser, M.P. and R.V. Rice. **Biochim. Biophys. Acta**, 63, 255, 1962.

Maxam, A.M. and W. Gilbert. **Methods in Enzymology**, 65, 499, 1980.

McBride, L.J. and M.H. Caruthers. **Tetrahedron Letts.**, 24, 245, 1983.

Merlino, G.T., Vogeli, G., Yamamoto, T., de Crombrugghe, B. and I. Pastan. **J. Biol. Chem.**, 256, 11251, 1981.

Merlino, G.T., Tyagi, J.S., de Crombrugghe, B. and I. Pastan. **J. Biol. Chem.**, 257, 7254, 1982.

Merlino, G.T., McKeon, C., de Crombrugghe, B. and I. Pastan. **J. Biol. Chem.**, 258, 10041, 1983.

Messing, J. and J. Vieira. **Gene**, 19, 269, 1982.

Miller, E.J. and S. Gay. **Methods in Enzymology**, 82, 22, 1982.

Moen, R.C., Rowe, D.W. and R.D. Palmiter. **J. Biol. Chem.**, 254, 3526, 1979.

Monson, J.M. and B.J. McCarthy. **DNA**, 1, 59, 1981.

Monson, J.M., Natzle, J., Friedman, J. and B.J. McCarthy. **Proc. Natl. Acad. Sci. USA**, 79, 1761, 1982.

Mount, S. and J. Steitz. **Nature**, 303, 380, 1983.

Murray, J.C., Lindberg, K.A. and S.R. Pinnell. **J. Clin. Invest.**, 59, 1071, 1979.

Myers, J.C., Chu M-L., Faro, S.H., Clark, W.J., Prockop, D.J. and F. Ramirez. **Proc. Natl. Acad. Sci. USA**, 78, 3516, 1981.

Myers, J.C., Chu, M-L. and F. Ramirez. **J. Cell Biochem. (Supp).**, 6, 318, 1982.

Myers, J.C., Dickson, L.A., de Wet, W.J., Bernard, M.P., Chu, M-L., Liberto, M.D., Pepe, G., Sangiorgi, F.O. and F. Ramirez. **J. Biol. Chem.**, 258, 10128, 1983.

Nicholls, A.C., Pope, F.M. and H. Schloon. **Lancet**, 1, 1193, 1979.

Nienhuis, A.W., Turner, P. and E.J. Benz Jr. **Proc. Natl. Acad. Sci. USA**, 74, 3690, 1977.

Odermatt, E., Risteli, J., Van Delden, V. and R. Timpl. **Biochem. J.**, 211, 295, 1983.

Ohkubo, H., Vogeli, G., Mudryj, M., Avvedimento, V.E., Sullivan, M., Pastan, I. and B. De Crombrugghe. **Proc. Natl. Acad. Sci. USA**, 77, 7059, 1980.

Old, J.M., Proudfoot, N.J., Wood, W.G., Longley, J.I., Clegg, J.B. and D.J. Weatherall. **Cell**, 14, 289, 1978.

Paglia, L., Wilczek, J., de Leon, L.D., Martin, G.R., Hörlein, D. and P. Müller. **Biochem.**, 18, 5030, 1979.

Paglia, L.M., Wiestner, M., Duchene, M., Ouellette, L.A., Hörlein, D., Martin, G.R. and P.K. Müller. **Biochem.**, 20, 3523, 1981.

Parker, I. and W. Fitschen. **Nuc. Acids Res.**, 8, 2823, 1980.

Parker, M.I., Judge, K. and W. Gevers. **Nuc. Acids Res.**, 10, 5879, 1982.

Parry, D.A.D. and A.S. Craig. **Nature**, 282, 213, 1979.

Pawlowski, P.J., Brierley, G.T. and L.N. Lukens. **J. Biol. Chem.**, 256, 7695, 1981.

Penttinen, R.P., Lichtenstein, J.R., Martin, G.R. and V.A. McKusick. **Proc. Natl. Acad. Sci. USA**, 72, 586, 1975.

Proudfoot, N. **Nature**, 298, 516, 1982.

Quinn, R.S. and S.M. Krane. **J. Clin. Invest.**, 57, 83, 1976.

Randall, L.R. **Cell**, 33, 231, 1983.

Rauterberg, J., Fietzek, P. Rexrodt, F., Becker, U., Stark, M. and K. Kühn. **FEBS Letts.**, 21, 75, 1972.

Rave, N., Crkvenjakov, R. and H. Boedtker. **Nuc. Acids Res.**, 6, 3559, 1979.

Reese, C.A. and R. Mayne. **Biochem.**, 20, 5443, 1981.

Ried, K.B.M. **Biochem. J.**, 179, 367, 1979.

Ried, K.B.M., Lowe, D.M. and R.R. Porter. **Biochem. J.**, 130, 749, 1972.

Ried, K.B.M. and R.R. Porter. **Biochem. J.**, 155, 19, 1976.

Risteli, J., Bächinger, H.P., Engle, J., Furthmayr, H. and R. Timpl. **Eur. J. Biochem.**, 108, 239, 1980.

Rogers, J., Early, P., Carter, C., Calame, K., Bond, M., Hood, L. and R. Wall. **Cell**, 20, 303, 1980.

Rosenberry, T.L. and J.M. Richardson. **Biochem.**, 16, 3550, 1977.

Rowe, D.W., Moen, R.C., Davidson, J.M., Byers, P.H., Bornstein, P. and R.D. Palmiter. **Biochem.**, 17, 1581, 1978.

Rowe, L.B. and R.I. Schwarz. **Molec. Cell. Biol.**, 3, 241, 1983.

Roychoudhury, R., Jay, E. and R. Wu. **Nuc. Acids Res.**, 3, 863, 1976.

Saber, M.A., Zern, M.A. and D.A. Shafritz. **Proc. Natl. Acad. Sci. USA**, 80, 4017, 1983.

Sage, H., Pritzl, P. and P. Bornstein. **Biochem.**, 19, 5747, 1980.

Sandmeyer, S. and P. Bornstein. **J. Biol. Chem.**, 254, 4950, 1979.

Sanger, F. and A.R. Coulson. **FEBS Letts.**, 87, 107, 1978.

Sanger, F., Nicklen, S. and A.R. Coulson. **Proc. Natl. Acad. Sci. USA**, 74, 5463, 1977.

Schafer, M.P., Boyd, C.D., Tolstoshev, P. and R.G. Crystal. **Nuc. Acids Res.**, 8, 2241, 1980.

Schnieke, A., Harbers, K. and R. Jaenisch. **Nature**, 304, 315, 1983.

Schuppan, D., Timpl, R. and R.W. Glanville. **FEBS Letts.**, 115, 297, 1980.

Seeburg, P.H., Shine, J., Martial, J.A., Ullrich, A., Baxter, J.D. and H.M. Goodman. **Cell**, 12, 157, 1977.

Setzer, D.R., McGrogan, M. and R.T. Schimke. **J. Biol. Chem.**, 257, 5143, 1982.

Sharp, P.A. **Cell**, 23, 643, 1981.

Siegel, R.C. and Y-H. Chang. **Clin. Res.**, 26, 501A, 1978.

Sobel, M.E., Yamamoto, T., Adams, S.L., Di Lauro, R., Avvedimento, E.V., de Crombrugghe, B. and I. Pastan. **Proc. Natl. Acad. Sci. USA**, 75, 5846, 1978.

Solomon, E. **Nature**, 286, 656, 1980.

Solomon, E., Hiorns, L., Dalgleish, R., Tolstoshev, P., Crystal, R. and B. Sykes. **Cytogenet. Cell Genet.**, 35, 64, 1983.

Southern, E.M. **J. Mol. Biol.**, 98, 503, 1975.

Spence, S.E., Pergolizzi, R.G., Dovano-Peluso, M., Kosche, A., Dobkins, C.S. and A. Bank. **Nuc. Acids Res.**, 10, 1283, 1982.

Spritz, R.A., Jagadeeswaran, P., Choudary, P.V., Biro, P.A., Elder, J.T., De Riel, J.K., Marley, J.L., Gefter, M.L., Forget, B.G. and S.M. Weissman. **Proc. Natl. Acad. Sci. USA**, 78, 2455, 1981.

Staden, R. **Nuc. Acids Res.**, 12, 551, 1984.

Steinmann, B.U., Martin, G.R., Baum, B.I. and R.G. Crystal. **FEBS Letts.**, 101, 269, 1979.

Steinmann, B.U., Tuderman, L., Peltonen, L., Martin, G.R., McKusick, V.A. and D.J. Prockop. **J. Biol. Chem.**, 255, 8887, 1980.

Sutcliffe, J.G. **Nuc. Acids Res.**, 5, 2721, 1978.

Sykes, B. **Nature**, 305, 764, 1983.

Tanzer, M.L. **Science**, 180, 561, 1973.

Tate, V.E., Finer, M.H., Boedtker, H. and P. Doty. **Nuc. Acids Res.**, 11, 91, 1983.

Taylor, J.M., Illmensee, R. and J. Summers. **Biochim. Biophys. Acta**, 442, 324, 1976.

Thomas, P. **Proc. Natl. Acad. Sci. USA**, 77, 5201, 1980.

Timpl, R., Wiedemann, H., Van Delden, V., Furthmayr, H. and K. Kühn. **Eur. J. Biochem.**, 120, 203, 1981.

Tolstoshev, P., Haber, R., Trapnell, B.C. and R.G. Crystal. **J. Biol. Chem.**, 256, 9672, 1981a.

Tolstoshev, P., Berg, R.A., Rennard, S.I., Bradley, K.H., Trapnell, B.C. and R.G. Crystal. **J. Biol. Chem.**, 256, 3135, 1981b.

Tolstoshev, P. and E. Solomon. **Nature**, 300, 581, 1983.

Trelstad, R.L., Hay, E.D. and J-P. Revel. **Develop. Biol.**, 16, 78, 1967.

Trüeb, B., Odermatt, B.F., Sahu, A.P., Spiess, M., Rüttner, J.R. and K.H. Winterhalter. **Renal Physiol., Basel,** 3, 23, 1980.

Villa-Komaroff, L., Efstratiadis, A., Broome, S., Lomedico, P., Tizard, R., Naber, S.P., Chick, W.L. and W. Gilbert. **Proc. Natl. Acad. Sci. USA,** 75, 3727, 1978.

Vogeli, G., Ohkubo, H., Avvedimento, V.E., Sullivan, M., Yamada, Y., Mudryj, M., Pastan, I. and B. de Crombrugghe. **Cold Spring Harb. Symp. Quant. Biol.,** 45, Pt. 2, 777, 1980.

Vogeli, G., Ohkubo, H., Sobel, M.E., Yamada, Y., Pastan, I. and B. de Crombrugghe. **Proc. Natl. Acad. Sci. USA,** 78, 5334, 1981.

Vuorio, E., Sandell, L., Kravis, D., Sheffield, V.C., Vuorio, T., Dorfman, A. and W.B. Upholt. **Nuc. Acids Res.,** 10, 1175, 1982.

Vuust, J. **Eur. J. Biochem.,** 60, 41, 1975.

Vuust, J., Abildsten, D. and T. Lund. **Connect. Tissue Res.,** 11, 185, 1983,.

Wahl, G.M. Stern, M. and G.R. Stark. **Proc. Natl. Acad. Sci. USA,** 76, 3683, 1979.

Wahli, W., Dawid, I.B., Wyler, T., Weber, R. and G.U. Ryffel. **Cell,** 20, 107, 1980.

Weiss, E.H., Cheah, K.S.E., Grosveld, F.G., Dahl, H.H., Solomon, E. and R.A. Flavell. **Nuc. Acids Res.,** 10, 1981, 1982.

Westaway, D. and R. Williamson. **Nuc. Acids Res.,** 9, 1777, 1981.

de Wet, W., Njieha, F.K. and D.J. Prockop. **Fed. Proc.**, 41, 619, 1982.

Wickens, M.P. and J.B. Gurdon. **J. Mol. Biol.**, 163, 1, 1983.

Wiestner, M., Krieg, T., Hörlein, D., Glanville, R.W., Fietzek, P. and P.K. Müller. **J. Biol. Chem.**, 254, 7016, 1979.

Winter, G. EMBO Summer Course, 1980.

Wozney, J., Hanahan, D., Tate, V., Boedtker, H. and P. Doty. **Nature**, 294, 129, 1981a.

Wozney, J., Hanahan, D., Morimoto, R., Boedtker, H. and P. Doty. **Proc. Natl. Acad. Sci. USA**, 78, 712, 1981b.

Yamada, Y., Avvedimento, V.E., Mudryj, M., Ohkubo, H., Vogeli, G., Irani, M., Pastan, I. and B. de Crombrugghe. **Cell**, 22, 887, 1980.

Yamada, Y., Mudryj, M., Sullivan, M. and B. de Crombrugghe. **J. Biol. Chem.**, 258, 2758, 1983a.

Yamada, Y., Kühn, K. and B. de Crombrugghe. **Nuc. Acids. Res.**, 11, 2733, 1983b.

Yamamoto, T., Sobel, M.E., Adams, S.L., Avvedimento, V.E., Di Lauro, R., Pastan, I., de Crombrugghe, B., Showalter, A., Pesciotta, D., Fietzek, P. and B. Olsen. **J. Biol. Chem.**, 255, 2612, 1980.

Yamauchi, M., Noyes, C., Kuboki, Y. and G.L. Mechanic. **Proc. Natl. Acad. Sci. USA**, 79, 7684, 1982.